Australasian Association for Digital Humanities (aaDH)
Association for Computers and the Humanities (ACH)
Canadian Society for Digital Humanities / Société canadienne des humanités numériques (CSDH/SCHN)
centerNet
European Association for Digital Humanities (EADH)
Humanistica
Japanese Association for Digital Humanities (JADH)

Digital Humanities 2018
Puentes-Bridges

Book of Abstracts
Libro de resúmenes

Mexico City
26-29 June 2018

ALLIANCE OF DIGITAL HUMANITIES ORGANIZATIONS

EL COLEGIO DE MÉXICO

RedHD

PROGRAM COMMITTEE / COMITÉ PROGRAMA ACADÉMICO

Élika Ortega – Northeastern University (PC Co-chair)
Glen Worthey – Stanford University (PC Co-chair)
Sarah Kenderdine – aaDH
Chris Thomson – aaDH
Lisa Rhody – ACH
Alex Gil – ACH
Constance Crompton – CSDH/SCHN
Dan O'Donnell – CSDH/SCHN
Nancy Friedland – centerNet
Brian Rosenblum – centerNet
Bárbara Bordalejo – EADH
Elisabeth Burr – EADH
Björn-Olav Dozo – Humanistica
Emmanuel Chateau Dutier – Humanistica
Akihiro Kawase – JADH
Maki Miyake – JADH

LOCAL ORGANIZING COMMITTEE / COMITÉ LOCAL ORGANIZADOR

Isabel Galina – Universidad Nacional Autónoma de México (UNAM) (Co-chair)
Ernesto Priani – Universidad Nacional Autónoma de México (UNAM) (Co-chair)
Miriam Peña – Universidad Nacional Autónoma de México (UNAM)
Jonathan Girón Palau – Universidad Nacional Autónoma de México (UNAM)
Ernesto Miranda – Secretaria de Cultura
Micaela Chávez Villa – El Colegio de México (Colmex)
Alberto Santiago Martinez – El Colegio de México (Colmex)
Silvia Gutiérrez – El Colegio de México (Colmex)
Natalie Baur – El Colegio de México (Colmex)
León Ruiz – El Colegio de México (Colmex)

SPONSORS / PATROCINADORES

Agenda Digital de Cultura. Secretaría de Cultura
Consejo Nacional de Ciencia y Tecnología (Conacyt)
Gale, Cengage
Stanford University Press
Tecnológico de Monterrey. Escuela de Humanidades y Educación
The Association for Computers and the Humanities (ACH)
Universidad del Claustro de Sor Juana

We would like to thank the support of the Instituto de Investigaciones Sobre la Universidad y la Educación (IISUE) and the Instituto de Investigaciones Bibliográfica (IIB) of the Universidad Nacional Autónoma de México (UNAM). Also the generous funding from Conacyt, project number 293068 - Convocatoria 2018 del Programa de Apoyos para Actividades Científicas, Tecnológicas y de Innovación de la Dirección Adjunta de Desarrollo Científico.

La elaboración del libro de resúmenes fue posible gracias al apoyo del Instituto de Investigaciones Sobre la Universidad y la Educación (IISUE) y el Instituto de Investigaciones Bibliográfica (IIB) de la Universidad Nacional Autónoma de México. También fue posible gracias al financiamiento Conacyt proyecto número: 293068 - Convocatoria 2018 del Programa de Apoyos para Actividades Científicas, Tecnológicas y de Innovación de la Dirección Adjunta de Desarrollo Científico.

Digital Humanities 2018
Puentes-Bridges

Book of Abstracts
Libro de resúmenes

El Colegio de México

Universidad Nacional Autónoma de México

Red de Humanidades Digitales

26 - 29 June 2018
Mexico City

26 - 29 de junio 2018
Ciudad de México

Edited by / Editores
Jonathan Girón Palau
Isabel Galina Russell

DHConvalidator service
Aramís Concepción Durán
Christof Schöch

On-line abstracts / Resúmenes en línea
Reynaldo Crescencio

Design and typesetting / Diseño y maquetación
Yael Coronel Navarro
Juan Carlos Rosas Ramírez

Proof-reading / Revisión
Karla Guadalupe González Niño
Jessica América Gómez Flores

Online abstract available at: dh2018.adho.org/abstracts

Title: Digital Humanities 2018: Book of Abstracts / Libro de resúmenes.
Contributor (Corporate Author): Alliance of Digital Humanities Organizations.

Publisher: Red de Humanidades Digitales A. C.

Date of Publication: 2018

ISBN: 978-0-911221-62-6

Welcome to DH2018

Élika Ortega and Glen Worthey, Program Committee Co-chairs
Isabel Galina and Ernesto Priani, Local organizers, Co-chairs

As many old-timers and some newcomers know, this is the first time that the annual international Digital Humanities conference takes place in the Global South. This is a momentous achievement for an organization that has always strived to be truly global, diverse, and inclusive. The geographic movement of the conference has brought with it a renewed awareness of the differences among the numerous communities that constitute ADHO and the DH field at large. As we celebrate these differences, we have also made every effort for DH2018 to create meeting points, foster connections, and build bridges across the many Digital Humanities.

Making the conference bilingual, a tradition that we're following from DH2017, has been central to our work. Indeed, although English continues to be a powerful *lingua franca* in our field, about 20% of the presentations, posters, and panels this year are in another language. This development in the program is the result not only of the Program Committee's work; it was possible thanks to the 'backstage' volunteer labor of hundreds of reviewers who lent both their DH expertise, and their strong linguistic capacities. We also endeavored to make as much of the information and official communications of DH2018 bilingual, including its website, our email communications, the Convalidator tool, and this *Book of Abstracts*, to mention a few. There is still much left to do, and many interfaces are still available only in English, but we hope that our collective efforts will encourage all future ADHO conference organizers to continue in this tradition.

This year the conference includes twenty-two long paper sessions, twenty-two short paper sessions, thirty-three panel sessions, and sixteen workshops. Additionally, a two-part poster session will showcase the work of over 150 scholars. The topics and approaches represented span from linked data to digital ethnography; from classical antiquity to online activism; from pedagogy to theory; from indigenous languages to natural disasters. The broad scope of the program attests to the long-standing practices that first propelled the consolidation of the field of Digital Humanities, while making ample room for new approaches that increasingly bring us closer to the social, political, and natural challenges the world currently faces.

Our two DH2018 keynote speakers, Janet Chávez Santiago and Schuyler Esprit, bring our attention to the territories of the Central Valleys in Oaxaca in Mexico and the Caribbean island of Dominica. Impacted in distinct ways by colonial and neo-colonial powers, these sites are sources of *other* ways of seeing, weaving, and redesigning the world. They are also a locus sustaining the communities, academic and otherwise, that seek to utilize digital technologies for cultural, epistemological, and sometimes physical, survival.

Organizing DH2018 in Mexico City has been a challenge and a learning experience. Certain cultural assumptions have come to light simply by holding the conference in a different geographical location. We are sure that these experiences will be helpful as the conference continues to move to new and different locations. For us, Mexico's sociocultural diversity makes it an ideal location for converging digital humanists from distinct cultures, contexts, and socio-political realities. We believe that our steps towards bridging cultural, technological, political, and ideological borders will lead to the creation of a Digital Humanities community that is truly global, diverse, and inclusive.

Bienvenidos a DH2018

Élika Ortega y Glen Worthey, Co-presidentes del Comité Científico
Isabel Galina y Ernesto Priani, Co-presidentes del Comité Organizador Local

Como saben muchos veteranos y algunos novatos de DH, esta es la primera vez que la conferencia internacional Humanidades Digitales se lleva a cabo en el Sur Global. Se trata de un logro memorable para una organización que siempre se ha esforzado por ser verdaderamente global, diversa e incluyente. El cambio de ubicación de la conferencia ha aportado una conciencia renovada de las diferencias entre las diversas comunidades que forman ADHO y el campo de las HD, en general. Con el mismo entusiasmo con el que celebramos estas diferencias, nos hemos esforzado por crear puntos de encuentro en DH2018, establecer conexiones y construir puentes entre las muchas humanidades digitales.

Un aspecto central de nuestro trabajo ha sido preparar una conferencia bilingüe, una tradición que seguimos desde DH2017. Y si bien el inglés continúa siendo una importante *lingua franca* en nuestro campo, cerca de 20% de las presentaciones, pósters y paneles en el programa de este año están en otro idioma. Esta característica del programa no es el resultado solamente del trabajo del Comité Científico; fue posible gracias a la labor voluntaria "tras bambalinas" de cientos de dictaminadores que ofrecieron tanto su experticia en HD como sus habilidades lingüísticas. Asimismo, nos esforzamos para que gran parte de la información y las comunicaciones oficiales de DH2018 fueran bilingües, incluidos el sitio web, los correos electrónicos, la herramienta Convalidator, y este Libro de Resúmenes, por mencionar algunos. Aún falta mucho por hacer y muchas interfaces todavía se encuentran disponibles solamente en inglés, pero esperamos que el esfuerzo colectivo alentará a futuros organizadores de la conferencia de ADHO a continuar esta tradición.

Este año la conferencia incluye veintidós sesiones de presentaciones largas, veintidós sesiones de presentaciones breves, treinta y tres paneles y dieciséis talleres. También incluye una sesión doble de pósteres, que mostrará el trabajo de más de 150 académicos. Los tópicos y las aproximaciones presentados en el programa comprenden los datos conectados a la etnografía digital; de la antigüedad clásica al activismo en línea; desde la pedagogía a la teoría; de las lenguas indígenas a los desastres naturales. Este amplio rango de temas da cuenta de las prácticas que impulsaron la consolidación de las humanidades digitales y, al mismo tiempo, abre espacios para nuevas aproximaciones que, cada vez más, nos acercan a los desafíos sociales, políticos y naturales que el mundo encara actualmente.

Las dos ponentes magistrales para DH2018, Janet Chávez Santiago y Schuyler Esprit, nos transportan a los territorios de los Valles Centrales de Oaxaca, México y a la isla caribeña de Dominica. Impactados de formas distintas por las potencias coloniales y neocoloniales, estos sitios son la fuente de otras formas de ver, tejer y rediseñar el mundo. Son también los *loci* que sostienen comunidades, académicas y no académicas, que buscan utilizar las tecnologías digitales para la preservación cultural, epistemológica y, a veces, incluso la supervivencia física.

Organizar DH2018 en la Ciudad de México ha sido un reto y un aprendizaje. El simple hecho de que la conferencia se lleve a cabo en una región diferente ha sacado a la luz ciertas presuposiciones culturales y estamos seguros de que el aprendizaje se irá enriqueciendo en la medida en que la conferencia se realice en distintas ubicaciones. Consideramos que, por su diversidad sociocultural, México es un lugar ideal para la convergencia de humanistas digitales de culturas, contextos y realidades sociopolíticas particulares. Estamos convencidos de que, al encaminarnos hacia la creación de puentes entre fronteras culturales, tecnológicas, políticas e ideológicas nos acercaremos cada vez más a formar una comunidad de humanidades digitales verdaderamente global, diversa e incluyente.

Table of Contents

Plenary lectures

Panels

Long Papers

Short Papers

Posters

Preconference Workshops

Sig Endorsed

Plenary lectures

Weaving the Word

Janet Chávez Santiago
jazoula.10@gmail.com
Indigenous Languages Activist

The weft is a thread that is woven among the warp's yarns; these are our paper and pencil in the creation of a rug. Together, warp and weft are the bridge that unites the threads with our past and our present, and we weave the patterns of Mitla's friezes as a form of reading, or of interpreting, and of writing our ancestors, but also as a way to recount our dreams and our experiences. We weave in Zapotec. When we complete a rug, we share it with the world, and although the weave is in Zapotec, it can be interpreted in English, in Spanish, in Mixtec, or in Chatino.

Digital media can be seen as a warp on which the speakers of indigenous languages have an opportunity to weave their word and to share it within their own community and beyond. Although in our times digital media and social networks are a practical part of our daily lives and of our interactions with the world, we as speakers of indigenous languages must truly appropriate these spaces, to weave our word well, in order to liberate ourselves from the denial of the present.

Tramando la palabra

La trama es el hilo que se teje entre la urdimbre, son nuestro papel y lápiz para crear un lienzo. Juntos, trama y urdimbre, son el puente que unen los hilos con nuestro pasado y nuestro presente, tejemos las grecas de Mitla como una forma de leer o interpretar y escribir a nuestros ancestros, pero también para contar nuestros sueños y nuestras experiencias. Tramamos en zapoteco. Cuando terminamos un tapete lo compartimos con el mundo, y aunque el tejido está en zapoteco se puede interpretar en inglés, en español, en mixteco o en chatino.

Los medios digitales se pueden ver como una urdimbre en donde hablantes de lenguas indígenas tengan la oportunidad de tramar su palabra y compartirla dentro de su propia comunidad y más allá. Aunque hoy en día los medios digitales y las redes sociales son prácticamente parte de nuestra vida cotidiana y de nuestra interacción con el mundo, como hablantes de lenguas indígenas todavía nos hace falta apropiarnos realmente de estos espacios, tramar bien nuestra palabra para liberarnos de la negación del presente.

Digital Experimentation, Courageous Citizenship and Caribbean Futurism

Schuyler Esprit
schuyleresprit@gmail.com
Research Institute at Dominica State College

The violence and trauma of climate change have arrived. The Caribbean region is the unfortunate recipient of the impacts of climate change and, much like its inheritance of plantation slavery and colonialism, it is left with the infrastructural, social and cultural pillage of imperial and neocolonial imposition. My talk will consider whether and how the humanities, and digital humanities in particular, can produce the ideal intersection between planetary responsibility, community accountability and sustainable living.

In this talk I discuss Create Caribbean Research Institute's digital humanities praxis through the example of the environmental sustainability project, *Carisealand*. Through the exploration and discussion of theories, tools, methodologies and praxis of digital humanities applied to the project, I position Caribbean afrofuturism in the context of contemporary Caribbean digital environments and the lived experience of Caribbean people in the aftermath of climate change.

I apply discourses of afrofuturism to imagine an alternate Caribbean future represented in the redesign, digital imagination and representation of selected Caribbean communities. By offering models for rethinking, visualizing and rebuilding physical spaces, I hope to raise questions and offer insights about the power of digital humanities for social and environmental justice in the contemporary and future Caribbean. The goal is to also offer the model as a template for developing other mapping projects that can propose an alternate future for the Global South.

Experimentación Digital, Ciudadanía Valiente y Futurismo Caribeño

La violencia y el trauma del cambio climático ya comenzaron. La región del Caribe es la desafortunada receptora de los impactos del cambio climático y, al igual que con la herencia de esclavitud en las plantaciones y del colonialismo, sufre del saqueo infraestructural, social y cultural de la imposición imperial y neocolonial. Mi charla considerará si, y de qué forma, las humanidades, y las humanidades digitales en particular, pueden producir una intersección ideal entre la responsabilidad planetaria y comunitaria, y una vida sustentable.

Asimismo, en mi charla, discuto la práctica de las humanidades digitales en el Instituto de Investigación Create Caribbean utilizando como ejemplo el proyecto de sustentabilidad ambiental *Carisealand*. Por medio de la exploración y discusión de las teorías, herramientas, metodologías y prácticas de las humanidades digitales aplicadas en el proyecto, ubico el afrofuturismo caribeño en el contexto de los ambientes digitales contemporáneos del Caribe y la experiencia de los caribeños que viven con las repercusiones del cambio climático.

Finalmente, pongo en práctica los discursos del afrofuturismo para imaginar un futuro caribeño alternativo representado en el rediseño, la imaginación y representación digitales de ciertas comunidades caribeñas. Al ofrecer modelos para repensar, visualizar y reconstruir los espacios físicos, deseo despertar preguntas y ofrecer entendimiento acerca del poder que las humanidades digitales tienen para crear justicia social y ambiental en el Caribe contemporáneo y futuro. La meta es también ofrecer este modelo como una plantilla para desarrollar otros proyectos de mapeo que pueden proponer un futuro alternativo para el Sur Global.

Panels

Digital Humanities & Colonial Latin American Studies Roundtable

Hannah Alpert-Abrams
h.alpert-abrams@austin.utexas.edu
University of Texas at Austin, United States of America

Clayton McCarl
clayton.mccarl@unf.edu
University of North Florida, United States of America

Ernesto Priani
epriani@gmail.com
Universidad Nacional Autónoma de México, Mexico

Linda Rodriguez
lmr273@nyu.edu
New York University, United States of America

Diego Jimenez Baldillo
diego_jimenez@inah.gob.mx
Instituto Nacional de Antropología e Historia, Mexico

Patricia Murrieta-Flores
p.murrietaflores@chester.ac.uk
University of Chester, United Kingdom

Bruno Martins
bruno.g.martins@ist.utl.pt
University of Lisbon, Portugal

Ian Gregory
i.gregory@lancaster.ac.uk
Lancaster University, United Kingdom

Overview

Colonial Latin American studies is an interdisciplinary field that crosses methodological frontiers in order to expand our understanding of the colonial past. This involves the bridging of disciplinary divides, as scholars trained in archaeology, literature, art history, and linguistics come together to define, examine, and seek to understand the historical record, even as it remains elusive and heterogeneous. As in comparable fields based in other parts of the world, including Europe and North America, this interdisciplinary work has depended on the use of computational methods, digital platforms, and digital pedagogy. Yet in the case of colonial Latin American studies, the field has yet to directly address the unique impact of the digital humanities on colonial research. How do the particular cultural and material circumstances of Latin American studies inform the application of digital methods to colonial research? What are the responsibilities of scholars using digital platforms to represent colonial materials? And how should scholars of colonial Latin America respond to the political, cultural, and economic structures that shape transnational collaborations in the digital age?

This bilingual panel addresses these questions by uniting scholars at different career stages, across disciplinary and national boundaries, who are applying the methods of digital humanities to the field of colonial Latin American studies. The papers represented explore the construction of colonial corpora, the application of computational methodology, and the development of digital systems for encoding and display. Panelists will make 10-minute presentations of their work, providing points of departure for a more general discussion about how digital tools and methodologies can alter the way we interact with textual and visual objects within colonial Latin American studies, as well as how we might create sustainable corpora within our field and preserve them for the long term. We hope that this session can contribute to the construction of a DH community within the study of colonial Latin America, in order to create a space for experimentation and the exploration of theoretical and methodological concerns, and to give greater visibility to digital work currently underway. We believe this will have implications for the growth of the field and for our ability to value this work in the specific context of professionalization, tenure, and promotion.

Métodos digitales: repatriación o expatriación de documentos coloniales

Ernesto Priani

¿Son los métodos digitales un instrumento para "repatriar" documentos del patrimonio histórico colonial o, por el contrario, son un instrumento para una nueva expatriación de esos materiales? Los estudios coloniales en Latinoamérica requieren de estrategias multinacionales para desarrollar un mejor conocimiento de la cultura colonial por, al menos, dos diferentes razones: la dispersión de los materiales en repositorios de diversos países y la conformación geográfica de la colonia que no corresponde con los actuales estados nacionales. Los métodos digitales representan una oportunidad para llevar a cabo una estrategia que rebase fronteras (Oceanic Exchanges Project Team), pero su uso no está exento de problemas de orden cultural, epistémico y geopolítico (Fiormonte et al.). Una revisión rápida de los proyectos de estudios coloniales con herramientas digitales muestra un desbalance entre proyectos iniciados y desarrollados en Latinoamérica y los que se llevan a cabo sobre todo en Estados Unidos, así como desequilibrios entre países latinoamericanos. Este desnivel tiene que ver, por supuesto, con cuestiones de recursos y perspectivas culturales, pero concretamente con la distinta penetración de los métodos digitales en las academias de Latinoamérica y de Estados Unidos. Tal desequilibrio representa una distorsión que nos obliga a cuestionar el sentido que tiene el uso de métodos digitales. No únicamente está el problema de a quién pertenecen los materiales digitalizados, sino a qué horizonte

cultural responden sus formas de representación o de análisis, qué implicaciones tiene el uso de tal o cual tecnología, y cómo se reciben en los distintos países.

Building Early Colonial Corpora for Digital Scholarship

Hannah Alpert-Abrams

The application of digital humanities methodologies to early colonial texts from Latin America depends on the development of digital corpora that represent colonial discourse with reasonable accuracy. Difficulties arise, however, when we seek to describe such a corpus in the colonial case. Regional variation in the use of historical orthography, the unique conditions of colonial printing, and the widespread integration of Spanish and indigenous languages significantly impacted the shape of inscription during the colonial period. Processes of transcription, lemmatization, and analysis, however, require linguistic normalization. This process is made more difficult when we consider the technological limitations of tools for textual processing, which often originated for use on modern, monolingual, Anglophone texts. In this talk, I will address the challenges of developing a colonial corpus from these Anglophone tools, drawing on the Reading the First Books project as a case study. Reading the First Books was a two-year, multi-institutional, NEH-funded effort to develop tools for the automatic transcription of early modern printed books. The project, which concluded in December of 2017, was successful in expanding automatic transcription tools for use on multilingual, orthographically variant, early-modern printed books. It was not successful, however, in using that tool to automatically transcribe an early colonial corpus. In reflecting on these outcomes, this talk will identify key challenges in colonial corpus construction, and propose ways forward for the automatic transcription of early colonial texts.

Addressing the Challenges in the Semi-automated Identification, Extraction and Analysis of Information from Early Colonial Documents and the XVI Corpus Known as Relaciones Geográficas

Patricia Murrieta-Flores
Diego Jiménez-Badillo
Bruno Martins
Ian Gregory

With the advent of digitization of original and edited collections of historical documents, as well as the creation of novel methods such as Geographical Text Analysis and the use of techniques derived from Natural Language Processing (NLP), Machine Learning and Corpus Linguistics, opportunities have recently emerged to develop new approaches for the study of vast collections of early colonial sources. Within the project "Digging into Early Colonial Mexico: A Large-Scale Computational Analysis of Sixteenth-Century Historical Sources," funded by the Trans-Atlantic Platform for Social Sciences and the Humanities, we will be refining and developing computational methodologies to identify, extract, cross-reference, and analyse the sixteenth-century corpus of the Relaciones Geográficas of New Spain. Over the course of the next three years, we will be looking not only to advance the creation of computational techniques for the mining of information from these early colonial sources and to solve different historiographical questions related to the geographies contained within them, but also to confront a set of challenges that have rarely been addressed before. For example, although the field of Digital Humanities has seen progress in the use of NLP methods for the automated identification of proper names in historical documents, this research has been carried out substantially in the context of the English language, and rarely with documents in which two languages are combined. In the case of this project, we are dealing with documents written in a combination of early modern Spanish and other non-European languages such as Nahuatl. Another important challenge lies in the geoparsing of these documents, where we are confronting issues that range from spelling variations of place names in Spanish and Nahuatl, to the geographic disambiguation of these places. This paper will address, through this particular example, the challenges that any scholar attempting data mining and/or macro-analysis of colonial Latin American documents would face, delving into the ways we are dealing with them.

Theoretical Problems in the Semantic Markup of Colonial American Maritime Texts

Clayton McCarl

To date, little work has been done on semantic markup as an area of editorial theory, or as a theoretical domain of relevance to the field of colonial Latin American studies. In my current research, I address this situation in part by considering parameters for the markup of maritime texts. Such writings deal largely in references to external worlds of people, places and objects—named and unnamed, known and unknown, real and imagined—, and our understanding of such texts hinges on our ability to decipher their codification of complex, unfamiliar realities. In developing a markup scheme for exploring taxonomies of externality in colonial-era maritime texts, I have encountered several theoretical issues that I believe have consequences beyond my current project. In this presentation, I will consider specifically the conceptual ambiguities that such a markup scheme may expose; consider the interpretive danger that such an editorial approach might pose; and examine ways

in which, through such a process of markup, we might come to experience differently these textual objects.

Digital Aponte: Mitigating Archival Loss through Digital Methods

Linda Rodríguez

Archives are political projects. In the context of the colonial Americas, historian Kathryn Burns suggests "we make our archives and sources part of our research, looking at them as well as through them." To do so, she argues, expands our understanding of the historical relationships of power that condition their production. In this paper, I analyze how digital methods can help us look at, and through, documents that register loss. I focus on the Digital Aponte project that aims to make present a lost work of art. Jose Antonio Aponte (?-1812) was a free man of color, soldier, sculptor, and creator of a "book of paintings" in colonial Havana. Aponte's book has been lost, or destroyed, but his descriptions of the book's pages survive in the archival record, part of his testimony following his arrest for conspiring to plan slave rebellions. Digital Aponte presents this trial testimony, with plans to add explanatory annotations of Aponte's robust descriptions, along with contextual information. I explore how digital methods enable the project's objective to foreground the archival document as a generative text.

References

Burns, K. (2010). Into the Archive: *Writing and Power in Colonial Peru.* Durham: Duke University Press.

Fiormonte, D., Schmidt, D. Monella, P. and Sordi, P. (2015). "The Politics of Code. How Digital Representations and Languages Shape Culture." Extended Abstract, ICTs & Society Conference. http://infolet.it/files/2015/06/politics-of-code-fiormonte-et-al-def.pdf (accessed 1 May 2017).

Oceanic Exchanges Project Team (2017). *Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914.* DOI 10.17605/OSF.IO/WA94S (accessed 1 May 2017).

Bridging Cultures Through Mapping Practices: Space and Power in Asia and America

Cecile Armand
cecile.armand@gmail.com
Stanford University, United States of America

Christian Henriot
christian.r.henriot@gmail.com
Aix-Marseille University, France

Sora Kim
gorgeousora@gmail.com
Seoul National University, South Korea

Ian Caine
ian.caine@utsa.edu
University of Texas, United States of America

Jerry Gonzalez
jerry.gonzalez@utsa.edu
University of Texas, United States of America

Rebecca Walter
rjwalter@uw.edu
University of Washington, United States of America

This panel brings together four papers that span from modern Asia to contemporary Texas: two studies of land ownership, based on historical cadasters in modern Shanghai and imperial Korea, a spatial analysis of advertising development in modern Shanghai, and a survey of municipal annexation as a mechanism for suburban expansion in San Antonio, Texas. In this panel, we argue that spatial concepts and practices can serve as a bridge to connect distant topics, spaces and times.

While grounded in separate contexts, all papers address issues of space and power. More precisely, they reveal the intricate relationships between space, power and mapping practices. Two papers point out the significance of historical cadasters as a record of land property and a basis of land management, while the two others focus on municipal regulations toward suburban expansion in San Antonio, or advertising development in modern Shanghai. Everywhere, spatial policies and mapping practices appear crucial to assert municipal or imperial control. This panel further suggests that mapping practices can play as a bridge between distant cultures and territories, either by transplanting Japanese and European cadastral techniques in Korea and Shanghai, or through the municipal attempts to avoid fragmentation in San Antonio. Yet spatial policies also create new boundaries among local communities: Chinese/foreign residents in Shanghai, Korean/Japanese in imperial Korea, Latinos/Anglos and working-class/elites in San Antonio. While the study of land cadasters in Korea focuses on the political impulses underlying spatial policies, the surveys of land ownership and advertising development in modern Shanghai, or that of suburban expansion in San Antonio, also emphasize the importance of economic factors in shaping urban spaces (real estate market, transportation networks), either reinforcing or conflicting with municipal policies.

At the methodological level, the panel demonstrates the values and challenges of using digital tools to conduct spatial analyses and to bridge past and present landscapes. Each project relies on a wide range of mapping software and practices, from the systematic digitization of original maps (Shanghai and Korean cadastral

maps), to the uses of Geographical Information System (GIS) to build a geospatial database and bring together separate sets of data. GIS and spatial modeling even allow to reconstruct spatial layers that provide substitutes for missing data, as in the cases of Shanghai and Korean cadasters. Digital tools further enable the visualization of gaps and overlapping patterns, or tracing spatial changes across time. In the case of San Antonio, going a step further would lead to imagine a digital chronology of its suburban expansion, including flat mapping, a filmed spatial narrative, and an interactive timeline. Two projects eventually provide a digital interface open to sharing and cooperation (San Antonio, MADSpace). Although they do not provide ready-made arguments, digital and non-digital mapping tools open untrodden paths to interpret the past, and raise new research questions. Through their digital experience, the four projects bridge various disciplines and fields of expertise. They rely on interdisciplinary collaboration between historians, geographers, economists and sociologists, as well as sustained cooperation between researchers, engineers, designers and software developers.

One of the challenges the authors address is the combination and integration of heterogeneous materials, the use of modeling to process data extracted from textual sources, or to rely on directories (digitized/ocerized/extraction) to identify unregistered land owners, especially in the case of historical cadasters, etc. This approach goes far beyond a conventional use of primary sources in historical research. Moreover, the two studies on historical cadasters actually serve as a bridge between Korea, Japan and China in substantive and methodological terms. In the studies of the urban expansion of San Antonio and the development of advertising in Shanghai, spatial concepts (demographic expansion, socio-spatial divisions and segregation) and mapping tools (GIS, spatial analysis) serve to trace lines between an American and a Chinese city, and across time. All four papers contribute to a reflection about land control and management, about power and urban society, and about urban landscape and its transformation. We believe these are real and reasonable bridges between the four contributions.

As a panel, we find significant cross-fertilization between DH and geography, or even DH and social sciences. We believe that the "digital" affects the whole array of disciplines in the humanities and social sciences, tearing down walls and borders, and creating bridges and intersectional analyses. We contend that "DH questions" lay at the very heart of what we have proposed. For instance, archival documents offer insight on the localized political debates that shaped the terrain of our respective sites. Similarly, oral histories, public records, contemporaneous publications help us to analyze the changes in metropolitan spatial practices over time, as well as popular responses to such shifts. We also argue that "spatial humanities" are part and parcel of DH. In fact, spatial humanities represent a major part of DH worldwide. The questions we ask start from our terrains and our disciplines, but we work through methods, tools and notions that are deeply rooted in the digital practice of humanities.

Critical Theory + Empirical Practice: "The Archive" as Bridge

James William Baker
james.baker@sussex.ac.uk
University of Sussex, United Kingdom

Caroline Bassett
c.bassett@sussex.ac.uk
University of Sussex, United Kingdom

David Berry D.M.
berry@sussex.ac.uk
University of Sussex, United Kingdom

Sharon Webb
sharon.webb@sussex.ac.uk
University of Sussex, United Kingdom

Rebecca Wright
r.k.wright@sussex.ac.uk
University of York, United Kingdom

Digital humanities can be understood as a "trading zone" between different disciplinary traditions (McCarty, 2003). Critical theory and empirical practice may appear to operate at different extremes of research enterprise, and yet – as our panel seeks to demonstrate – the notion of "the archive" can function as a bridge between them, as a pathway or method between the trading zones. For us, the bridge has its strongest resonances in that academic endeavour where "the archive" is most revered, where it is used as a rite of passage, a marker of authority buried in footnotes: history writing. In this incarnation "the archive" often represents physical buildings with physical holdings.

And yet this version of "the archive" is as much imagined as it is real, a particular articulation or incantation of mid-nineteenth century state bureaucracy woven into the mystic of archival research, a place of dust, labour, boredom, and very occasional discovery (Steedman, 2006). Most archives do not conform to this incantation, not only because buildings with physical holdings that call themselves archives are not all remnants of mid-nineteenth century state bureaucracy, but also because many archives are not buildings and the holdings of many archives are not physical: instead they are lofts, shoeboxes, and server racks; web pages, word documents, and digital media. Here then, archives are much more and much less than buildings with physical holdings.

This latter bridge might seem less assured, "the ar-

chive" in this form might invoke gephyrophobia in some users, but – in the work of our panel – it has proven vital in traversing between critical theory and empirical practice (Berry 2017). Constituting contributions individuals from a range of traditions – critical theory, historical research, information science – our papers explore ways in which a critical-digital conception of the archive shines light on topics as diverse as the historical method, the responsibilities of researchers, the politics of technology and how the archive can help the empirical and the critical talk to each other.

Each short paper is presented by a faculty member of the Sussex Humanities Lab: a unique venture based at the University of Sussex, a digital humanities lab that takes an interdisciplinary and collaborative approach to digital research in the humanities, and includes a multi-disciplinary grouping of researchers in philosophy and information technology, history and archaeology, media and communications, music and performance technology, and sociology that is dedicated to developing and expanding research into how digital technologies are shaping our culture and society:

James Baker is Lecturer in Digital History and Archives. James cares about how people in the past interacted with things.

Caroline Bassett is Professor of Media and Communications and Director of the Sussex Humanities Lab. Her current work explores anti-computing.

David M. Berry is Professor of Digital Humanities. His new work examines the historical and philosophical genealogies of the notion of an „Idea of a University" and how they are relevant in a digital age.

Ben Jackson (panel chair) is a Research Fellow in Digital Humanities (Library). His interests include computer graphics, 3d modelling, and archival systems.

Sharon Webb is Lecturer in Digital Humanities. Her current research interests include community archives and identity, social network analysis (method and theory), and research data management.

Rebecca Wright is Research Fellow at the University of York and a Sussex Humanities Lab Associate. In 2017 she was a Research Fellow in Mass Observation at the Sussex Humanities Lab examining energy practices and digital methodologies within the Mass Observation archive.

Missing Dust: Born Digital Archives and the Historical Method

James Baker

The advent of the personal computer catalysed the second major break in production of Western manuscripts. These machines, interactions with which consolidated around WIMP-like Windows interfaces during the early- to mid- 1990s, rendered the manuscript anew. Hitherto physically and ontologically unique, the manuscript in the age of the personal computer increasingly did not exist as a physical object and was infinitely reproducible.

These 'born digital' archives have been accessioned, catalogued, and maintained by archivists for two decades. Personal papers have been archived using forensic approaches that capture documents (and the file and operating systems on which they are contained) as bitstreams and that interrogate documents for their forensic features: system metadata, deleted passages. Work by Kirschenbuam (2016), Reside (2011), and Reis (2017) has brought these born-digital archives into the purview of literary scholars and raised questions about analysis of contemporary literature. Little comparable work has focused on the historical method, on the implications of born digital archives for questions and problems common in History.

This paper describes three cases studies of empirical work that attempt bring the methodological challenges and opportunities created by born digital archives to the attention of historians: to bridge a gap between archival practice and historical research. First, a workshop organised in partnership with the Wellcome Library (London) at which a small group of contemporary historians – selected for their range of interests and expertise within the field – were invited to browse, interact with, and reflect on their encounters with born digital archives (e.g., born digital manuscript materials created by the geneticist Ian Dunham between 1997 and 2006). Here attention is paid to the applicability of existing methods, questions, and concerns (Sloyan et al, 2018). Second, a training event on forensic capture of data storage devices. This pedagogical activity used the BitCurator software suite to prompt historians into considering what a record, a series, and an archive are in the context of hard, floppy, and flash storage as repositories of archival materials. Third, archival research using the Mass Observation Archive at the University of Sussex: an anthropological initiative that has, since 1981, issued each year three Directives (a series of questions about a social, political, or everyday subject) to hundreds of UK-based volunteer writers. This work explored how people in Britain between 1991 and 2004 talked about writing and archiving on personal computers, their excitement and anxiety about these processes, and how their perception of self was refracted through their encounters with the machines they used to make documents. Here, attention focuses on the tensions between contemporary observations of behaviour and behaviour observed in the examination of born digital archives.

Together, these case studies address a series of problematics about historical work in the age of born digital archives: Do born digital manuscripts disrupt and undermine assumptions around historical practice? Does the manuscript remain a relevant source category when that manuscript is born digital? How can archival professions validate authority through infinitely reproducible documents that

leave no (or few) physical traces? What might replace dust in how historians feel and imagine the archive?

The Bridge: Accretion as the Principle of The Hybrid Archive

Caroline Bassett

William Gibson's 1994 science fiction novel *Virtual Light*, explores the end of cyberspace and the beginning of what was later termed the post-digital. At its heart is a bridge – a passage point, a habitation, and a player – which startles with its impossible geometry: "The integrity of its span was rigorous as the modern program itself, yet around this had grown another reality, intent upon its own agenda. This had occurred piecemeal, to no set plan, employing every imaginable technique and material. The result was something amorphous, startlingly organic."

Virtual Light was notable at the time because it pointed to the beginning of a transition from cyberpunk and the internet dream of disembodied virtuality to something more quotidian; the digital as the taken for granted, the fabric of the every day. But the bridge is also – at least in part – a heterotopia. It celebrates the opportunities arising, the pace between territories where many kinds of activity are possible, and where these activities make a difference. The bridge is a hybrid construction; through the central span the project of planning, order, and control endures – but what has been added, soldered, sutured on, has become integral. The result is something amorphous; a matter of rigorous structure and *ad hoc* accretion, an architecture comprehending organization and improvisation, mathematics and poetics. The bridge stands because something long-standing still stands, and the bridge is changed through additions that do not so much challenge this structure, but mutate it, and mutate with it.

The bridge might be understood as a microcosm of the archive today, exhibiting in its fictional structure the monstrously barnacled form this now takes. Many studies of archives in a digital age focus on either on the barnacles or the inner structure, on professional archives and archiving or on the actions and practices of community or *ad hoc* archivists. Taking its inspiration from Gibson's bridge, which also becomes an empirical object of study, this paper sets out to focus on what is generated between them. Specifically, this is explored through a consideration of archiving practices in science fiction – where the formal economy of the official archive, explored through a critical exploration of genre, is complemented by a study of the *ad hoc* collection practices of the informal reader economy. The intention is to use this to explore the hybrid archive as a new cultural form and in particular to conceptualize the distribution or organisation within it of expertise on the one hand, and authority and power on the power.

De-Archiving the Archive

David M. Berry

The traditional pre-digital structure of archives and practices of archivization were captured and stabilized through memory institutions such as museums, national libraries, universities and national archives, often funded by the state. These institutions provided an organizational form and institutional structure which made possible a political economy for archives as such and hence an economic stability to archives. Institutions provided a decision-making centre around the collection of archives, in essence an institutionalized archivization process that performed judgment in combination with curatorial functions. Indeed, the archive became defined as a preselected quantity of artifacts evaluated according to their worth for being preserved. The structure of traditional institutional arrangements around the archive was legitimated through a complex chain of practices and institutionalizations that authorized decisions to be taken about what of the present (and past) should be kept and what should be discarded. In contrast, in an age when digital technologies are delegated greater responsibility for a collection, computational rationalities are increasingly granted the task of archiving and re-presenting materials, through computational analytics and user data, the archive creates a second-order archive. Indeed, we are faced with new archival machines that demand a different social ontology but also a different way of exploring and interacting with archives. These new gateways to social memory are manifested in algorithms that instantiate a new archival imaginary – a new archival constellation that is constantly in motion, modulated and mediated. The digital creates a different kind of collection: digital archives are much more malleable and reconfigurable, and do not necessarily need to conform to traditional archives' organizational structures or systems. This new possibility of "infinite archives" create their own specific problems, particularly in born-digital and digitized collections, such as huge quantities of articles, texts and "Big Data" suddenly made available combined with the ability to generate comprehensive and exhaustive archives rather than curated ones. Computation therefore threatens to *de-archive the archive*, disintermediating the memory institutions and undermining the curatorial functions associated with archives. Many of the concerns of humanists have reflected an uncertainty about what the loss (or change) of archives might mean – although of course this could also reflect a loss of paper-ish culture – especially where medial changes imply epistemic change. In changing the structure of archives, and the memory institutions that curate and store them, computation renders them anew through a grammatization process which discretizes and re-orders. This process can be as simple as the infinitely re-orderable process of creating a database. It is also amenable to spatial planning and algorithmic analysis

that presents the opportunity for a logic of objectification. This is the recasting of the material world into the shapes dictated by computational analysis or computational processes. Through the principles of instrumentality, partially embedded in computational systems, but also in the neoliberal order that legitimates through principles of performativity, efficiency and a political economy of value, forces action on the archive to conform and interoperate. It is here, crucially, that critical theory can contribute to cultural critique of computational forms of archival logics.

Community Archives, Preservation and Practice

Sharon Webb

The University of Sussex, home of the Sussex Humanities Lab (SHL), sits just outside the seaside town of Brighton. South of London, Brighton boasts a rich, varied and complex LGBTQ+ history. It is a place of celebration for all things queer, as well as a place for vocal and energetic activist movements. In addition to its queer identity, Brighton is also hub of digital innovation, and annually hosts the Brighton Digital Festival (indeed a number of SHL members actively participate in this event). It is within this context, Brighton as a cultural and innovation hub, that this paper will discuss the fourth paradigm of archival theory, as both inherently "digital" and community driven, using Brighton as a case study. It will consider the development and creation of community archives, specifically LGBTQ+, both as a challenge to archival practice and theory, and as an opportunity.

The fourth archival turn or 'paradigm' (Cook, 2013) can be seen as both a response to official archival practices and policies that have failed in the past to represent, comprehensively, the narrative and history of minority groups in society. It can also be seen as an affect and influence of the Internet and digital methods which create opportunities for communities to create and manage their own representation in the digital, public, record. Cook states,

> …community is the key concept…of the fourth archival paradigm now comng into view, a democratizing of archives suitable for the social ethos, communication patterns, and community requirements of the digital age. (Cook 2013:116)

In effect, community pressures and the opportunities afforded by digital environments are pushing the boundaries of previous definitions of an "archive". Indeed, as we know, the Digital Humanities community have had a significant influence on these developments. Archives, that is digital ones, create a bridge between the formal structures that the humanities have traditionally accessed sources, knowledge, and reason. A digital archive is a place where we manifest discourse, memory, and importantly, create and reinforce community – communities of scholars, communities of users and specific communities self-identified by common interests, values, etc. (i.e. LGBTQ+ communities).

Brighton, as a case study, provides important examples of how communities generated and reinforce identity through archival practices. Projects like BrightonOurStory (now defunct physical archive), Queer in Brighton (Oral histories, LGBTQ History Club), Into the Outside (Photographic exhibitions), Brighton Transformed (Oral Histories) create memory and meaning through work that captures and records a specific community memory.

This presentation will consider tensions between these community driven endeavours and their capacity to support projects in the long-term, especially with regards to digital preservation. It will use the loss of the BrightonOurStory Archive (1989-2013) as a reminder of our responsibilities as researchers to these archival projects, and to think further about 'community requirements [in] the digital age'.

Media Imprints within the Digital Interface: Typewriting Mass Observation Online

Rebecca Wright

This paper examines how the digital politicizes medium within the historical archive. Through a critical analysis of *Mass Observation Online* (MOO) (the online portal of the Mass Observation Archive, University of Sussex) the paper will assess how Optical Character Recognition (OCR) has established a new hierarchy within a key archive of British social memory—centred around the typewriter. In doing so, the paper will address the historical contingency of media and the historiographical issues at stake when media structures digital archives.

Mass Observation (MO) was founded in 1937 to conduct a form of reverse anthropology observing the ordinary people of Great Britain. Democratic in mission (if not always in practice) the organisation developed a national panel of over 700 Observers to record the intricacies of everyday life. The digital interface of MOO, however, is undermining the democratic promise of MO by elevating type (which consists 30% of materials produced by the national panel) over the larger collection of materials written by hand. This has occurred because typewritten documents remain the only ones to have been OCRd for digital text. Due to the impact of the availability and quality of OCR text on research results (Hitchcock 2013), a new economy of representation has developed based not on what Observers wrote, but what they wrote in.

The elevation of one medium is not without consequences for our understanding of MO materials. After all, as media archaeologists such as Friedrich Kittler and

Marshall McLuhan taught us media is never neutral but embedded in the politics of identity, form, and representation. Typewriting during the inter-war period was connected to wider historical forces, including changes in white-collar work, gender roles and new cultures of representation. Important historiographical issues, therefore, are at stake in foregrounding medium for how we understand the nature of the MO project: from the constitution of the national panel, to the self-identity of Observers, the form of written materials, and the structure of life-writing. The digital has thus forced us to confront how medium transformed which, what, and how Observers wrote.

These issues will only be exacerbated when the New Mass Observation Project (MOP), restarted in 1981, is digitised. The historical contingency of media will be translated into the context of the early information age that embraced handwriting, typewriting, word processing and the PC—each with their own challenges for OCR software—re-working rankings within the archive.

This paper will thus use the example of MOO to examine how the digital is forcing historians to pay more attention to the site of production of our historical documents to consider how medium shapes our source materials and in turn the material interfaces of the digital archive. Drawing on critical frameworks from media archaeology it will ask how the media of source materials and digital interfaces is merging in new ways to re-work the politics of the archive and social memory.

References

Berry, D. M., and Fagerjord A. (2017). *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge: Polity Press, 2017. http://politybooks.com/bookdetail/?isbn=9780745697659.

Cook, T. (2013). "Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms." *Archival Science* 13, no. 2–3: 95–120. https://doi.org/10.1007/s10502-012-9180-7.

Gibson, W. (1994). *Virtual Light*. Spectra Books.

Hitchcock, T. (2013). "Confronting the Digital: Or How Academic History Writing Lost the Plot." *Cultural and Social History* 10, no. 1: 9–23. https://doi.org/10.2752/147800413X13515292098070.

Kirschenbaum, M. G. (2016). *Track Changes: A Literary History of Word Processing*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

McCarty, W. (2014). "Getting There from Here. Remembering the Future of Digital Humanities Roberto Busa Award Lecture 2013." *Literary and Linguistic Computing*: fqu022. https://doi.org/10.1093/llc/fqu022.

Reis, T. (2017). "The Rationale of the Born Digital Dossier Genetique: Digital Forensics and the Writing Process. With an Exemplary Discussion of Born Digital Draft Traces from the Thomas Kling Archive". *Digital Scholarship in the Humanities*: fqx049.

Reside, D. (2011). "'LAST MODIFIED JANUARY 1996': THE DIGITAL HISTORY OF RENT." *Theatre Survey* 52, no. 02: 335–340. https://doi.org/10.1017/S0040557411000421.

Sloyan, V., Demissie S., Eveleigh A., and Baker J. (2018) "Overview of a Born-Digital Archives Access Workshop Held at Wellcome Collection." Wellcome Trust. https://doi.org/10.6084/m9.figshare.6087194.v1.

Steedman, C. (2006). *Dust*. Manchester: Manchester University Press.

Networks of Communication and Collaboration in Latin America

Nora Christine Benedict
nbenedict@princeton.edu
Princeton University, United States of America

Cecily Raynor
cecily.raynor@mcgill.ca
McGill University, Canada

Roberto Cruz Arzabal
rcruz.arzabal@gmail.com
Universidad Nacional Autónoma de México / Universidad Iberoamericana, México

Rhian Lewis
rhian.lewis@mail.mcgill.ca
McGill University, Candada

Norberto Gomez Jr.
norbertogomezjr@gmail.com
Montgomery College, United States of America

Carolina Gaínza
carolina.gainza@udp.cl
Universidad Diego Portales, Chile

Panel Overview:

Drawing on this year's conference theme of "bridges/puentes," this panel examines the ways in which networks emerge among individuals working and operating in Latin America and beyond during the twentieth and twenty-first centuries. We use digital tools to explore how artists and intellectuals connected and collaborated across countries in the early part of the twentieth century. We assess the linguistic and cultural dimensions of web readership and its communities. We investigate alternative digital distribution methods for contemporary Mexican poetry in the twenty-first century. We analyze how the visibility of (digital) narratives surrounding sexual violence in Latin America creates a unique space for necessary dialogues. We look to the particular expressions of disappearance, mortality and even spirituality in Latin American Post Internet culture. And we study how collaborative practices in digital literary creation alter the various ways in which we produce and consume texts.

We, thus, consider not only how networks are formed within Latin America, but also the ways in which these links and connections extend to other regions of the world. The networks we analyze range from the literary and social, to the economic and political. How, for instance, are contemporary print forms the product of their settings, their individual publics, and their social networks? Are artistic networks conceived and maintained differently prior to the digital age? How might contemporary hashtag projects in the region expand the notion of the trans-Hispanic web? How do certain social media platforms alter our conception of self, nation, and world through their unique development of networks? How do cultural and artistic narratives eliminate social hierarchies and reveal networks of social justice? Viewed together as a collective whole (or a network of their own, perhaps), these projects explore what it means to be connected across geographies, cultures and time.

Global Networks of Cultural Production

Victoria Ocampo, her world-renowned journal Sur, and her publishing house of the same name, all loom large over Latin American cultural production in the twentieth century. While much has been written about this Argentine socialite and her impressive literary enterprises, a great deal of work still remains to be done with regard to the extent of her global reach. In an effort to address these issues, I am engineering a digital project, "Global Networks of Cultural Production," that details a complex web of connected intellectuals, both inside and outside of Latin America, through their correspondence, translations, prologues, and edited editions. In this presentation I will describe the central cruxes of my digital project as well as provide an initial demonstration of the database I am creating. The first layer, "The *Sur* Enterprise," presents users with the option to navigate among three modules: People, *Sur*, and Editorial Sur. Within each module, users can interact with data pertaining to Ocampo's networks. For instance, in the "People" module, users can explore the occupation(s), birthplace, death place, and sex for each person that is linked to Ocampo's literary network (and pinpoint overlaps among individuals), while the "*Sur*" module allows users to interact with contributions to the literary journal *Sur* (grouped by genre, author, and issue). The second layer, "Visual Essays," provides a series of network analysis visualizations that demonstrate the spatial and temporal impact of Ocampo's efforts on the Latin American, European, and Asian populaces. Critical essays that narrate the significance of the queried data and its visual iteration accompany all of these visualizations. Each of these layers is fueled by a relational database that holds up the established links with an archive of metadata gleaned from a variety of documents, including correspondence, contracts, and even physically published books and magazines. All of these dimensions work together to digitally model Victoria Ocampo's work in creating networks, literary circles, and literary canons.

The Digital Readership Networks of the Trans-Hispanic Web

Despite the position of Spanish as the fifth most prominent language in overall web content, scholars are only beginning to explore the nuances of the trans-Hispanic web (2015). Drawing on case studies from Spain, Argentina, Chile, Peru and Mexico, my research assesses the web as a linguistic and cultural territory that can be mapped using digital tools and methods. "The Digital Geographies of the Hispanic World" is the first comprehensive geographical study of the web as an arena for reading and engaging with literary content. This is a project with two intersecting goals: 1) to map the readership of web-based content related to Latin American literature through a series of Spanish-language websites, identifying the networks they establish; 2) to determine if digital literary production conforms to a broader post-national aesthetic observed in print literature. Indeed, in the twenty-first century, digital content comes to life as it intersects with web analytics, aiding scholars in grasping the cultural and linguistic configurations that emerge around web content. Given the new possibilities of data analysis of this rich content, scholars are beginning to realize that how readers engage with web-based literary content often has more to do with language communities than the IP addresses or national contexts from which literary content arises. This presentation will explore some of the most recent data collected on web readership and network analysis of a series of leading literary websites from the Hispanic world.

Post-Print Culture and Publishing Networks in Contemporary Mexican Poetry

In the last fifteen years, independent publishing houses have been the central space that has defined the most relevant literary themes and forms of contemporary Mexican poetry. These publishers have changed the inertia of the literary field through strategies that produce an aggregate value to their books, based on symbolic frameworks and alternative distribution practices. These unique methods of book circulation enable experimental poetics to find an auspicious space in independent publishing houses. The publishers in question use the academic prestige of experimental poetics, while speculative poetics nourish intellectual distinction when published by independent firms.

Nevertheless, despite the efforts to distribute books, independent publishers do not have the economic re-

sources for national or international shipping. As a response to those problems, in recent years web platforms and collaboration networks have appeared, allowing the free circulation of books in PDF format. The existence of both types of distribution poses questions regarding the social forms of circulation for contemporary Mexican poetry, particularly in terms of how literary forms establish a dialogue or refuse to deal with those alternative practices of distribution and distinction. To answer these questions, I propose "post-print" as a concept that can be broad enough to explain the relationship between print publishing and digital distribution, as well as the use of consent and collaboration in the reproduction of experimental literary forms.

Nuestro Primer Acoso: Digital Networks and Collective Action against Sexual Violence

In the spring of 2016, new digital activist networks emerged to address gendered and sexual violence in Latin America Of the hashtags generated by these movements, few gained the public recognition of #MiPrimerAcoso (or "My First Harassment" or "My First Abuse"), a hashtag that encouraged individuals to tweettheir first experiences of sexual violence.When evaluating #MiPrimerAcoso's popularity, it is necessary to contextualize the concretemetrics of #MiPrimerAcoso within the intangible, affective dimensions that characterize the streams of discourse that grew out of the hashtag: the networks of #MiPrimerAcoso formed on the basis of shared experiences and shared public feelings.This analysis seeks to surpass traditional metrics of Twitterengagement and delve deeper into the kinds of connections that users form with each other within the intangible streams of discourse generated by the hashtag. For example, what makes retweeting a news story about #MiPrimerAcoso different from retweeting another user's story of sexual violence?The quantitative dimensions of#MiPrimerAcoso's digital proliferation – the prevalence of retweets, for example, orthe use of other hashtags to link formerly disparate currents of digital conversation – are explored alongside a critical analysis of the discursive conditions generated under the hashtag's narrative premise.In examining this dialogue, this project illustrates the networks of affect that stitched recollections of trauma into a political outcry.

Critical Networks: Latin American Death, Remembrance, and Recovery in the Post-Internet

The faceless of Latin America, the *desaparecidos* (disappeared), historically total in the tens of thousands within multiple nations, with extreme numbers in Chile and Argentina, due primarily to military dictatorships. Thus, a history of disappearance and loss have become embedded into the national psyche of many in Latin America, leading to the advent of "truth commissions" during the "memory boom" of the 1970s and 1980s. Today, Latin America, one of the fastest growing Internet populations in the world, now finds itself rapidly joining a globalized electronic culture. Arguably, the network leads to monoculture, and a commoditization of the individual, a result of the electronic Culture Industry. Therefore, today, Latin America may find itself disappearing digitally.

With digital remembrance and disillusionment in mind, this paper investigates the particular expressions of disappearance, mortality and even spirituality in Latin American Post Internet culture through the work of Brazilian artist Eva Rocha, primarily, as well as Teresa Margolles of Mexico, both of whose work is devoted to the outcast, the dead, and the forgotten. Through an analysis of these critical works, one may find a foil to the new electronic colonialism of the global digital networ

Collaborative Practices in Digital Literary Creation

In Latin America, digital literature is a relatively new phenomenon. In the analysis of LatinAmerican digital texts, I have considered both their material composition as well as aspects of authorship and reception practices. Materiality here refers to the technologies that have been used by the author in the production of the digital text. Depending on the technology used in digital narratives, we find texts that range from simple productions—like hypertext based productions—to more complex texts that include music, images, moving text, and also make use of many different software. Thus, the effects produced in readers can be aesthetically varied, and are determined by the technologies used to create the literary works in question. Although it is true that the specificity of the medium is a main component in the study of digital literature, the sole attention to the material elements of the texts is not enough to grasp some features that are unique to these productions, especially in a region where the introduction and uses of new technologies are strongly related to politics.

In this presentation,I examine how both the production and the reception of literature have been affected by digitaltechnology, with special emphasis on issues related to Latin American digital literature. I willanalyze Jaime Alejandro Rodriguez's *Narratopedia*, Doménico Chiappe's *La Huella de Cosmos*, andLeonardo Valencia's and Eugenio Tiselli's *El Libro Flotante* in order to highlight collective practices ofcreation involved in digital productions. Through a discussion of these issues, I offer an overview of ongoing changes wrought by digital technology in contemporary Latin American cultural production.

Digital Decolonizations: Remediating the Popol Wuj

Allison Margaret Bigelow
amb8fk@virginia.edu
University of Virginia, United States of America

Pamela Espinosa de los Monteros
espinosadelosmonteros.1@osu.edu
Ohio State University, United States of America

Will Hansen
hansenw@newberry.org
Newberry Library, United States of America

Rafael Alvarado r
ca2t@virginia.edu
University of Virginia, United States of America

Catherine Addington
ca2bb@virginia.edu
University of Virginia, United States of America

Karina Baptista
kab7hg@virginia.edu
University of Virginia, United States of America

The Maya K'iche' book of creation, known today as the *Popol Wuj* (Council Book), challenges some of the foundational categories of literary, historical, and anthropological studies: the stories reflect a decidedly Mayan way of understanding the world and one's place in it (Lepe Lira 2016, Florescano 2002), but the history of the book cannot be disentangled from Spanish colonial power (Quiroa 2011, 2017) or contemporary national ideologies (González 2014). The *Popol Wuj* invites and resists analysis precisely because it defies a series of binary oppositions that underlie received frames of interpretation – indigenous and Spanish, print and image, spoken and written, sacred and secular, literary and theoretical.

Researchers on this panel – graduate students, teaching and research faculty, and librarians – have experimented with different methods of digitizing the *Popol Wuj*. In 2007, a collaborative team from Marshall University, the Newberry Library, and the Ohio State University Library created the first digital facsimile of the *Popol Wuj* so that Mayan people could electronically access the oldest surviving written version of their *tzijs*. In 2017, Multepal, which began in a graduate seminar at the University of Virginia and continues in development with colleagues in Guatemala and the US, aims to create a thematic resource collection that reveals the story's multiple layers of meaning and remediation.

Each paper on this panel analyzes some of the diverse intellectual, ethical, and technical challenges and opportunities that we face in trying to represent the graphic, oral, and narrative complexities of the *Popol Wuj*. Papers #1-#3 (Espinosa de los Monteros, Hansen, Bigelow and Alvarado) examine issues of access and artifact preservation in digitization, using TEI to mark up non-Western texts, and vexing ethical questions of how DH scholars can tell a story that is not ours. Papers #4 and #5 (Addington and Baptista) present original research on literary genre and colonial encounter that emerged from the digitization projects. Taken together, these five papers suggest how digitization efforts enable new possibilities for humanistic inquiry and dissertation projects in indigenous and Latin American studies. Because these projects are ongoing, we hope that our panel will create a space for brainstorming ideas with colleagues from other fields, universities, and countries. Our panelists are bilingual English/Spanish speakers, and we will present examples in both languages (see sample thematic entries that point to each other en castellano e inglés).

Following Gallon's (2016) model of a "technology of recovery," we aim to "bring forth the full humanity of marginalized peoples through the use of digital platforms and tools" in order to address issues of diversity (McPherson 2012) and the question of whether "information can be unfettered" (Earhart 2012). This is an especially complex issue in indigenous studies, where debates about data sovereignty are informed by a history of state-sponsored appropriations of Native knowledges (Gaertner 2017). The intellectual, multilingual, transnational community of scholars at DH 2018 would be an ideal platform to tackle these questions and identify new *puentes* in digitizing cultural legacies and acquiring a decolonized consciousness.

References

Earhart, Amy E. (2012). "'Can Information Be Unfettered?' Race and the New Digital Humanities Canon." In Gold and Klein, eds. *Debates in the Digital Humanities*. Minnesota, pp. 309-318.

Florescano, Enrique. (2002). "Los paradigmas mesoamericanos que unificaron la reconstrucción del pasado: el mito de la creación del cosmos; la fundación del reino maravilloso (Tollán), y Quetzalcóatl, el creador de estados y dinastías." *Historia mexicana* 52(2): 309-359.

Gaertner, David. (2017). "Why We Need to Talk About Indigenous Literature in the Digital Humanities." *Novel Alliances: Allied Perspectives on Art, Literature, and New Media*, 20 paragraphs. January 26, 2017. Accessed 20 December 2017. Available at: https://novelalliances.com/2017/01/26/indigenous-literature-and-the-digital-humanities.

Gallon, Kim. (2016). "Making a Case for the Black Digital Humanities." In Gold and Klein, eds. *Debates in the Digital Humanities*. 2nd ed. Minnesota, 15 paragraphs. http://dhdebates.gc.cuny.edu/debates/text/55.

González, Ann. (2014). "The *Popol Vuh* for Children: Explicit and Implicit Ideological Agendas." *Children's Literature Association Quarterly* 39(2): 216-233.

Lepe Lira, Luz María, ed. (2016). *Oralidad y escritura: Experiencias desde la literatura indígena*. Michoacán.
McPherson, Tara. (2012). "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation." In Gold and Klein, eds. *Debates in the Digital Humanities*. Minnesota, pp. 139-160.
Quiroa, Néstor. (2017). "Friar Francisco Ximénez and the Popol Vuh: From Religious Treatise to a Digital Sacred Book." *Ethnohistory* 64(2): 241-270.
---. (2011). "The Popol Vuh and the Dominican Religious Extirpation in Highland Guatemala: Prologues and Annotations of Fr. Francisco Ximénez." *The Americas*, 67(4): 467-494.

Mid-Range Reading: Manifesto Edition

Grant Wythoff
grant.wythoff@gmail.com
Pennsylvania State University, United States of America

Alison Booth
ab6j@virginia.edu
University of Virginia, United States of America

Sarah Allison
sallison@loyno.edu
Loyola University New Orleans, United States of America

Daniel Shore
daniel.shore@georgetown.edu
Georgetown University, United States of America

Overview

This panel intervenes in debates about interpretative methods that are often lumped under "reading," and often measured by metaphors of scale, from close to distant. Mining data in vast corpora promises to transform literary history, and all scholars in the humanities rely upon online materials and tools. Yet many humanists stand aloof from DH because of its presumed hyperbolic claims, its apparent blurring of the detailed artifact (the domain of humanities), and to some, its collusion, post-critique, with neo-liberal globalization. Four panelists, collaborating for the first time, have encountered provocative concepts in each other's work that moderate such stark oppositions between the humanist and the computational. The panelists' previous studies have demonstrated the "payoff" or mutual instruction of DH and other recognized standards of scholarship. At the same time, in meticulous capture of language, style, form, and cultural production, the panelists highlight the limits that some champions of algorithms might want to leap in a single bound. Technological approaches to literary studies require highly curated corpora and modulation, often excision, of noisy results. Each paper addresses the loss inherent in categories and models, and the gain in tracing discarded, fuzzy, or inaccessible data. While our fields span centuries of Anglophone culture, our work advocates diversity, women's history, and the DH community's values of open access and collaborative technological innovation.

Our papers address disruptions as well as continuities in observational scale as the tools and materials shift. Each panelist speaks from experience with a different dataset and her or his innovative approach to interpretation, touching on both language and technology. The first two speakers propose forms of mid-range reading to describe imaginative and interpretive leaps that scholars make between individual documents/texts and broader social forces; the second two address the reductions and abstractions that are necessary to the research project, themes common to all papers. As an archeologist of technologies, Wythoff rediscovers the concept of the gadget as an instance of human-inanimate interaction mirrored in DH. Booth expands on her response in *PMLA* to Franco Moretti's *Distant Reading*, highlighting typologies as well as specific textual features in biographical nonfiction that enforce communal narratives. Allison, co-author on Stanford Lit Lab pamphlets associated with distant reading, proposes reductive reading, or explicit acknowledgment of necessary simplification, even of such ambitious problems as the nature of fictionality, which has been differently framed in studies by Piper, Underwood, and Eliot. While concepts of scale pervade claims for methods, Shore offers the approaches of construction grammar and corpus linguistics for particular insights into abstractions and categorizations. Shore, like Allison, calls on us to acknowledge the motivated reductions that are necessary to the research process. Our talks reflect on the history of technology and biographical representation, the forms of fiction and nonfiction, and the preconditions of selection and labeling of data—enduring issues in the humanities that become more telling with the expanding digital capacity to "read" at large and at speed.

Tacit computing and method in the humanities

Grant Wythoff

Humanistic research has always involved imaginative and interpretive leaps from the person to „the social," from the text to „the historical." Think for instance of the Annales school and its emphasis on the history of collective mentalities, or how Foucault described „discourse" by reverse-engineering historical ways of constituting knowledge. Today however, with the availability of big data, many of these forms of humanistic interpretation have become second nature. The search for broad cultural formations is implicit in the earliest steps we take in a research project, from keyword searches to frequency analyses. To what degree are certain

kinds of historical argumentation baked into these mundane, day-to-day research activities, and what other kinds of cultural formations might we be overlooking?

In my current book project, *Gadgetry: A History of Techniques,* I reconstruct the history of a discourse on technology. The book focuses on the many kinds of objects that were described as „gadgets" across the twentieth century, from dashboard gauges to atomic bombs, can-openers to smartphones. While "gadget" can be a placeholder for any kind of object, even imaginary ones, I argue that its evolving application to particular tools and techniques reveals important lessons about our relationship to technology.

In this book, I explore the user's imagination of how their gadgets work. For example, a single iPhone contains over half the elements of the periodic table, extracted from almost every continent on the planet and compressed into a thin slab that allows the user to dip her toes into a river of collective affect generated by the social network of everyone she's ever met. This is a fantastically science-fictional experience that is now part of our everyday lives. But the emergence of new digital cultures, political movements, and forms of intimacy are all predicated on the unique habits each user adopts in order to understand these complex gadgets.

For this book, I text mine archives of novels, magazines, and newspapers in order to explore the distinctly vernacular philosophies––the media theories from below––that emerge from users and their everyday practices. Using databases like the Corpus of Historical American English, Historical American Newspapers, and the Media History Digital Library, I proceed by collecting as many instances of the word „gadget" as possible and plugging them into categories of my own making based on how the term is applied: is the gadget handmade or mass produced, seen as important or a trinket, does the word refer to the entirety of the tool or a component within it, and so on. Because I have hand-coded this „dataset" and designated myself the categories into which I sort each instance of the word, the portrait that emerges of a discourse on technology could be described as entirely of my own making, as opposed to algorithmically-generated. But what really is the distance between these two categories of interpretation? In this talk, I will compare my digital methods to other methods throughout the history of the humanities that have attempted to paint a portrait of collective feeling.

Mid-Range reading: typologies, events, and discourse in a network of women's biographies

Alison Booth

Although many investigate fictionality, scholars have attended much less to nonfiction and biography than to imaginative forms such as novels or film. Digital humanities (DH) expand the scale of literary history while building on existing maps of period, genre, and notable authors, with finding aids shaped by previous scholarship. Thus Andrew Piper's impressive textual analysis, "Fictionality," neglects life narrative. Collective Biographies of Women (CBW) accesses a corpus of 1270 English-language biographical collections published across centuries, in a feminist historical study of a "hidden collection" of nonfiction. CBW developed before Google Books glimmered on the horizon; we worked with WorldCat and analogue materials to rediscover such publications as *Noted Negro Women* (1893). Reversing the usual DH phases, I published the book before collaborating on an online resource. What could we learn about the trends in gender ideology already constructed by biographers and publishers, publication data, and contents? Biography is a model (i.e. reduction) of a life within networks of typologies based on social difference. Distant reading is not best adapted to ramifications within curated corpora, where there is no mystery of author or genre. We capture the distinctive form and rhetoric of biography (and changing meaning of words such as "noble") in relation to such scenarios as inter-class contact or recognition of genius. Sentiment analysis or word vectors developed for large corpora of novels or newspapers would miss the mark. The actual dynamics of gender representation, for example, can hardly be captured as a grammatical binary or by rates of male or female agents per 300 words, while nationality is a shifting attribute across geopolitical and individual transformations.

This paper extends Booth's "Mid-range reading: not a manifesto" and builds on the findings from CBW's method of mid-range reading as well as from the typologies and networks of women in the CBW database. CBW researchers are tagging discourse in biographies, such as first-person plural and plural proper names, and quantifying the distribution of types of events across versions of the same person or occupational types. Both scales of reading and typologies press upon ethics as well as epistemology: how to classify the individual text, or the character/person. Attention must be paid, yet cognition and knowledge depend on generalizations. CBW has focused on sets of books that document the ways women's lives have been typologically interpreted. Our "sample corpora" range from all the books that include a short life of the saintly Victorian nurse, Sister Dora, and the distinct set of books that feature the famous adventuress, Lola Montez; other networks cluster around Queen Cleopatra, Frances Trollope, African Americans, women in medicine, Latinas, presenters (publishers, biographers), and others among the 8500 persons. A method we call mid-range reading uses the Biographical Elements and Structure Schema (BESS), a stand-aside XML schema (not TEI editing within the text file) that links element types (of stage of life, events, discourse, persona description, topos) to numbered paragraphs. BESS analyses, then, measure ra-

tes and distributions of element types across versions of lives sorted typologically by the contents of interrelated books. In 2018 we will obtain TEI files of remaining texts, with non-consumptive use of the copyright materials, through the HathiTrust Research Center. Becoming in this sense an archive as well as a testing ground for narrative theory of biography and network analysis across centuries of representation of women, CBW can demonstrate the comparative rewards of large-scale textual analysis and mid-range reading, and add to the understanding of biographical representation in many forms.

Harnessing Pegasus: On Setting Reasonable Limits

Sarah Allison

This paper takes up a theoretical question in digital humanities practice: how we understand the borders or boundaries of projects. "Reductive reading" is my term for critical methods that call attention to how they subordinate, or reduce, textual complexity. I argue that the explicit way with which DH research acknowledges this act of simplification creates an ethos of critical frankness. As Stephen Ramsay argues, code must "assert its utter lack of neutrality with candor, so that the demonstrably non-neutral act of interpretation can occur." "Harnessing Pegasus" focuses on the poignant question of setting limits. How do researchers establish the right distance from the texts under consideration, or reduce the scope of their inquiry? Here, I consider how researchers set limits in three projects that aim to understand what we might take to be the constitutive feature of the novel: fictionality.

It is axiomatic that the most irritating questions after a talk--but often also the best--are those that deal with a project's limits. Researchers announce what they have done, and the members of the audience say, Ah, but why didn't you do something else? This practice can help establish that one has taken a reasonable approach to a legitimate question in the field or open up future possibilities for research. It can also bring home the importance of narrowing one's approach in order to answer a specific question, as in We *didn't* do something else. We did the thing we did. In sharing work publically, researchers are called to account for the boundaries they have set--or, as it is often framed, that they have been forced to set.

It is the latter attitude that interests me here, the moment when the scalar ambitions of distant reading meet pragmatic reality and intellectual justification. Mid-range reading leaves space to account for both. In this paper, I will consider three approaches to fictionality in literary history: by Andrew Piper in *Cultural Analytics*, by Ted Underwood, Michael L. Black, Loretta Auvil, and Boris Capitanu in their work on genre in the HathiTrust Digital Library, and by Simon Eliot in his bibliographic work on trends in publishing, 1800-1914. In considering the way each project treats its limitations, I seek to create connections--bridges--across them. How do their definitions of fictionality intersect with Catherine Gallagher's theoretical treatment of the topic, and what that that tell us about nonfictionality? In each of these three studies, non-fiction is represented by a discrete collection of texts. How does limiting the generic canon change the way we understand fictionality?

Other than Scale

Daniel Shore

This paper explores the limits of the concept of scale in digital inquiry. Quantitative scholars in particular have naturally chosen scale as what sets their approach apart from other established methods. They speak of the computer as a "macroscope" that permits "macroanalysis." Scholars counted things before computers, but computers let them count and compute lots of things. Contrasting themselves with close readers, distant readers propose, with the help of machines, to step back from the page to see more and see bigger. Claims of scalar difference are often quite quantitatively precise. Instead of offering a reading of a single novel, distant readers study the titles of 7,000 British novels from 1740-1850, or ask how not to read a million books, or search through the 60,237 full texts in EEBO TCP I and II. For nearly all quantitative analyses of texts, the authors could tell the reader exactly how many words they count in how many documents, in light of sophisticated metrics and models.

Talk of scale in the digital humanities has not been simply ill advised. In spite of quantitative precision, we don't really know what we talk about when we talk about scale. Individual texts are much bigger than are usually acknowledged. Even when bag-of-words approaches are forthright about discarding word order and syntax, they rarely itemize what they are discarding. What has been characterized as an increase in scale can be more accurately described as the sacrifice of one sort of information for another. The point is not to oppose reductionism, but to be fully aware of what is being reduced.

Scalar conceptualization of digital tools and methods has tended to crowd out other, non-scalar distinctions. Some, like experimental design, theories of evidence, and falsifiability (an account of what it would mean to be wrong) should be more prominent in the conversation. I'll focus on concepts - abstraction, categorization, hierarchy - that are central to meaning and linguistic creativity across languages. Here I turn to the insights of construction grammar and corpus linguistics to suggest further possibilities for investigation. The bigram *thought leader* is two words, but it is also a single compound noun, the meaning of which can't be fully predicted from the meaning of its parts. How big is it? An abstract construction like *Once upon a time… [] and*

they lived happily ever after may be only ten words, and yet as big as the fairy tale that fills its blank. How long is it? The relevant distinctions in these examples are not scalar in any simple sense, and the methods for understanding them cannot be captured by distance or proximity. I start with linguistic examples at the level of the utterance, propose a few ways forward for qualitative and quantitative inquiry, and close by suggesting how the non-scalar distinctions at work in construction grammar might be relevant for specifically literary questions such as genre and narrative form.

References

Allison, Sarah. *Reductive Reading: A Syntax of Victorian Moralizing*. Baltimore: Johns Hopkins University Press, forthcoming 2018.

Allison, Sarah. "Other People's Data: Humanities Edition," *Cultural Analytics*, Dec. 8, 2016. http://culturalanalytics.org/2016/12/other-peoples-data-humanities-edition/

Bode, Katherine. "The Equivalence of 'Close' And 'Distant' Reading; Or, toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78, no. 1 (March 1, 2017): 77–106, https://doi.org/10.1215/00267929-3699787.

Booth, Alison. *How to Make It as a Woman: Collective Biographical History from Victoria to the Present*. Chicago: University of Chicago Press, 2004.

Booth, Alison. "Mid-Range Reading: Not a Manifesto." *PMLA* 132: 3 (May 2017): 620-27.

Burguiere, Andre. *The Annales School: An Intellectual History*. Trans. Jane Marie Todd. Ithaca, NY: Cornell University Press, 2009.

Eliot, Simon. "Some Trends in Book Publishing, 1800-1914" in John O. Jordan and Robert L. Pattern (eds.), *Literature in the Marketplace*. Cambridge: Cambridge University Press, 2003.

Eliot, Simon, and Jonathan Rose, eds. *A Companion to the History of the Book*. Malden, MA: Wiley-Blackwell, 2009.

Gallagher, Catherine. *Nobody's Story: The Vanishing Acts of Women Writers in the Marketplace, 1670-1820*. Berkeley, U. of California P, 1994.

Gallagher, Catherine. "The Rise of Fictionality." *The Novel*. Ed. Franco Moretti, Vol. 1. Princeton: Princeton UP, 2006. 336-63.

Goldberg, Adele E. *Constructions at Work: The Nature of Generalization in Language*. New York: Oxford UP, 2006.

Goldberg, Adele E. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: U of Chicago P, 1995.

Hancher, Michael. "Re: Search and Close Reading," in *Debates in the Digital Humanities 2016*. University of Minnesota Press, 2016. 118–38. http://conservancy.umn.edu/handle/11299/181603.

Langacker, Ronald W. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford UP, 2008.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. "Idioms," *Language* 70 (1994): 491–538.

Piper, Andrew. "Fictionality." *Journal of Cultural Analytics*, December 20, 2016. https://doi.org/10.22148/16.011.

Robertson, Stephen, and Lincoln Mullen. "Digital History & Argument White Paper – Roy Rosenzweig Center for History and New Media." November 13, 2017. https://rrchnm.org/argument-white-paper/.

Shore, Daniel. *Cyberformalism: Histories of Linguistic Forms in the Digital Archive*. Baltimore: Johns Hopkins UP, forthcoming 2018.

Shore, Daniel. "Shakespeare's Constructicon," *Shakespeare Quarterly* 66.2 (2015): 113-136.

Smith, Barbara Herrnstein. "What Was Close Reading? A Century of Method in Literary Studies," *Minnesota Review* 87 (2016): 57–75.

Underwood, Ted. "Distant Reading and the Blurry Edges of Genre. " *The Stone and the Shell*. 22 Oct. 2014.

Underwood, Ted. "Understanding Genre in a Collection of a Million Volumes, Interim Report." Figshare. https://dx.doi.org/10.6084/m9.figshare.1281251

Wythoff, Grant. *Gadgetry: A History of Techniques*, in progress.

Wythoff, Grant. *The Perversity of Things: Hugo Gernsback on Media, Tinkering, and Scientifiction*. University of Minnesota Press, 2016.

Precarious Labor in the Digital Humanities

Christina Boyles
christina.boyles@trincoll.edu
Trinity College, United States of America

Carrie Johnston
johnstc@wfu.edu
Wake Forest University, United States of America

James McGrath
james_mcgrath@brown.edu
Brown University, United States of America

Paige Morgan
paige.c.morgan@gmail.com
University of Miami, United States of America

Miriam Posner
miriam.posner@gmail.com
UCLA, United States of America

Chelcie Rowell
chelcie.rowell@bc.edu
Boston College, United States of America

While crucial progress has been made toward the equitable distribution and acknowledgement of DH labor, important

work remains to be done to address and amend the overarching precarious nature of labor in the digital humanities. Many DH practitioners find themselves in situations that require both specialized and general knowledge, as well as a vast—and outlandish—technological skill set. In a field that relies heavily on grants and temporary positions, many of the people who occupy DH positions find themselves juggling the impossible task of keeping up with advances in technology, advising stakeholders from divisions across campus, researching, writing, and teaching, often without the security of long-term employment.

Ranging from mid- to early-career, the speakers comprising this panel have navigated myriad DH positions—staff, tenure-track faculty, contingent faculty, postdoc, graduate student, administrator, programmer, and librarian—and can speak to the expectations and pitfalls of digital humanities labor. Panelists include a tenure-track faculty member in Library & Information Science, a Digital Scholarship and Subject Specialist Librarian, a Digital Scholarship Coordinator in a university's IT department, a Digital Humanities Research Specialist in a university library, a Digital Humanities Librarian, and a Postdoctoral Fellow in Digital Public Humanities. This panel is largely an outgrowth of each panelist's efforts at her home institution and within her academic organizations to make DH labor more visible, to rethink standards of evaluation for digital scholarship, and to generate relationships that address the ethical dimensions of collaborative labor in the digital humanities.

Each of the six panelists will speak for ten minutes, allowing for thirty minutes of discussion among panelists and audience members. Taken together, the panelists' talking points indicate an emerging pattern in DH labor: the expectation of the novice practitioner or early-career scholar to act as an expert. The panelists agree that reorienting some of our collective focus away from cultivating digital technologies and projects and onto mentoring digital practitioners is a step in the direction of mitigating these unrealistic expectations and, in the long-run, generating more sustainable methods and practices around DH labor.

Of digital humanities research methods, Tara McPherson warns, "Our screens are cover stories, disguising deeply divided forms of both machine and human labor. We focus exclusively on them increasingly to our peril." The central goal of this panel is to initiate conversations about these "deeply divided forms of human labor" in the digital humanities, often neglected in favor of creating more DH projects. These divisions take many forms: the lack of ethnic, cultural, and economic diversity among DH practitioners; the contingent nature of DH positions; the exploitation of digital laborers within and beyond classrooms; and the challenging or outright dismissal of the value of digital humanities scholarship by tenure and promotion committees. In keeping with the conference theme, Bridges/Puentes, we aim to bridge these divides in DH through a better understanding of the precarious nature of labor of DH—its causes and manifestations—which in turn will generate better practices in creating DH positions and mentoring DH practitioners.

Miracle Workers

Miriam Posner

Alex Gil has identified the "miracle worker" as a particular kind of digital laborer, one who is expected to cover a range of roles, responsibilities and projects with a minimal amount of resources, support, and compensation. Miracle workers are expected to be competent scholars, accessible tech support, patient project managers, and more. When Alex created an "Open Directory of Miracle Workers" in digital humanities, almost 150 people added their names to this list, suggesting that this particular kind of labor is unfortunately endemic to academic environs. The risk of such a model is that it limits the ability of the "miracle worker" to implement significant and lasting change on campus. As Miriam Posner notes, "When we choose not to invest in our own infrastructure, we choose not to articulate a different possible version of the world." In other words, due to a lack of human and fiscal support, the "miracle worker" often is unable to challenge traditional modes of scholarship that may be ineffective or even harmful. This occurs both with infrastructure—like understaffing, limited resources, and/or tools that reinscribe Eurocentric biases—and with day-to-day operations—including consulting with faculty and staff and making presentations to administrators. It may be difficult, or even impossible, to find time to transform the digital humanities environment of an institution by identifying appropriate resources, developing and supporting innovative projects, or encouraging the use of new, more ethical, digital tools and contexts if a "miracle worker" is expected to serve a wide range of campus needs. And while there has been compelling digital scholarship that has modeled itself on more traditional forms of knowledge production like the scholarly monograph or journal article, collaborative, public-facing, and iterative digital scholarship proves challenging in environments that encourage "miracle worker syndrome" because they tend to privilege the monograph at the exclusion of digital work.

Examples of this abound throughout academia. One example is the tenure-track faculty member required to print hard copies of born-digital scholarship far afield from the monograph, whose portfolio may be read by a department and/or an administration with no clear guidelines for how to refer or promote the employees they hired to "do" digital work. Another is the graduate student encouraged to situate herself within digital humanities by completing unpaid labor in addition to a traditional dissertation, or by taking on part-time positions. Yet another is the DH librarian/coordinator whose success is measured in terms of grants received and projects completed, rather

than the quality of her digital labor, relationship-building, or program management.

While several humanities departments and professional organizations have taken steps to develop guidelines for professional evaluation of digital labor, these recommendations are not always implemented, or they may not serve the varied forms of academic labor beyond the tenure-track model. As such, they do little to alleviate the plight of the "miracle worker" or her precarity. Doing so would require transformative change that upends the hierarchical value-model of academia by both acknowledging and valuing the work of those precarious positions.

Flipped Mentorship

Carrie Johnston

This paper will address the lack of conversation on and effort around mentoring "miracle workers" in DH positions once they are hired. DH labor is a relatively new addition to the higher-education workforce, and often positions such as "Assistant Professor of Digital Humanities," "Postdoctoral Fellow in Digital Scholarship," and "Digital Scholarship Coordinator" have little in the way of clear expectations, guidelines for contract renewal, or institutional knowledge about the position.

The outsized expectations placed on DH positions create a mentorship paradox, in which the new hire—often an early-career scholar—is responsible for supporting multiple digital scholarship endeavors and advising more advanced scholars than herself. The mentor framework is therefore flipped, as the newcomer must take the lead in shepherding digital projects from early stages to finished products and providing proof of the staying power of DH on her campus.

This presentation will call attention to this flipped mentorship framework, highlighting the ways it requires many DH laborers to act as administrators, despite the fact that they are not trained for or paid to do such high-level work. They are often expected to start and run a digital humanities program on campus, advocate for themselves and the resources necessary for their program in meetings with high-level administrators and representatives from large granting agencies. Even if these positions are funded by permanent salary lines, these jobs are still tenuous due to the reality that there is insufficient support built in to to help the faculty or staff member navigate the idiosyncrasies of a new institution and to negotiate the tricky landscape of higher education.

In addition to bringing attention to the flipped mentorship framework, this presentation will offer potential ways to amend and to avoid multiplying this problem. In particular, it will outline conditions that must be attended to and in place before creating new DH positions and DH programs, and prior to hiring faculty and staff to fill DH positions. Better anticipating and charting specific technological, staffing, and academic requirements—from sufficient server space, to student-facing support for classroom projects, to new standards of evaluation for digital scholarship—will create the necessary conditions to build generative, sustainable DH labor practices on campus.

Public/Digital Humanities

James McGrath

Many digital humanities projects have argued that their work is designed or intended for general audiences or specific publics beyond communities residing within the institutional structures of higher ed: teachers and students in K–12 classrooms, users of particular social media networks, groups who embrace particular identities or geographic affiliations (among others). But the labor involved in attending to the needs for community outreach, interface design, user experience, accessibility, and other factors essential to making the metaphorical bridges materialize between these projects and their desired audiences is often precarious, underpaid, or even missing completely from the planning and implementation stages of these projects. Wendy Hsu argues that "we should think of public humanities work as a process, not a product" and that "we should do more to include the public at earlier phases of our work" ('Lessons on Public Humanities from the Civic Sphere'). If we agree with this sentiment (and similar ones raised by Sharon Leon, Steve Lubar, and other practitioners of public and digital humanities), and if we share in the desire of grant-funding organizations like the National Endowment for the Humanities to create "Digital Projects for the Public," then how might these investments in publics also require new forms of labor and collaboration?

Building on experiences doing digital public humanities work in both an English department (as a graduate student) and in a public humanities center (as a postdoctoral fellow), this presentation argues that the rhetorical aims of digital projects that seek to collaborate, serve, educate, or otherwise inspire particular publics must inevitably transform the material realities of how these projects are created, staffed, designed, and circulated. In North American academic contexts, these kinds of projects are generally supported by major grants or staffed by contingent labor like graduate students or postdoctoral fellows in humanities departments. While masters programs in public humanities (such as the one at Brown University) or public history (such as those at American University or Northeastern University) have created or collaborated on projects that seek to collaborate or engage with digital contexts, and certificate programs in Digital Public Humanities (such as the one offered by George Mason University) demonstrate the generative potential of pairing digital humanities and public humanities aims and methodologies, these initiatives are restricted

in many ways by their short timelines, by a lack of practitioners or available sustainable resources for public or digital projects, or by the privileging of more traditional academic forms of cultural production like journal articles, conference papers, thesis papers, coursework.

This presentation highlights the implications of introducing more varied forms of practitioners and labor into humanities departments, drawing on experiences collaborating with individuals (oral historians, public librarians, community archivists, filmmakers) and creating alternate forms of knowledge (crowdsourced archives, digital art installations, augmented reality tours). In order to reach particular publics as potential collaborators, audiences, participants, and even creators and instigators of digital projects, we must reimagine the forms and networks that our production of knowledge have traditionally inhabited.

Sustainability in the Digital Humanities

Christina Boyles

Issues surrounding long-term planning and sustainability have long haunted digital humanities initiatives. For example, in 2009 *Digital Humanities Quarterly* published a special cluster on what it means to be "Done" with work in our field. Strangely, the word "labor" only appears once in the entire corpus of writing completed for that cluster. In the time since its publication, digital humanities practitioners have become more cognizant of the benefits of creating work with an ending in mind: many grants now require that projects document plans for long-term preservation and sustainability, and institutions with digital scholarship hubs residing in libraries frequently design and support projects that store materials in digital repositories (when available). But the impact of these planning procedures does not always result in projects and procedures that are conscious of ethical forms of labor. Planning for long-term preservation may be seen by faculty members or graduate students more as outsourced labor for librarians than collaborative work; documentation protocols may fail to account for future audiences whose labor proves essential to a project's afterlife; contingent labor, by nature of its contingency, may depart a project without attending to its long-term needs (due to graduation, or the loss of available funding, or institutional re-structuring, among other factors). But an inattention to impact on dimensions of labor seems to persist; like the 2009 *DHQ* overview on endings and afterlives, a 2014 white paper by ITHAKA on "Sustaining the Digital Humanities" only mentions the word "labor" once in its entire overview of issues related to sustainability.

This presentation argues that collaborative methodologies and models of shared authority inevitably benefit planning for long-term preservation and sustainability. That being said, these modes of collaboration require renewed attention to where and how forms of labor and modes are described and valued on digital humanities projects. Julia Flanders has described the ways terms like "efficiency" and "productively" are selectively and voluntarily deployed and circulated in descriptions of academic forms of labor, and she observes the ways faculty labor is generally privileged at higher institutional, financial, and social levels despite the fact that "the vast preponderance of actual *work* involved in creating humanities scholarship and resources is not done by faculty." In some institutional contexts like the ones Flanders describes, resources for sustainability and preservation may be allocated, but the lived professional realities of the laborers invested in these methodologies may be undervalued, held separate, or under-utilized by faculty members and students interested in digital humanities scholarship. And an increase in the visibility of available resources for digital research does not address potential concerns about the value of library and archival expertise on the topics of preservation and sustainability. How can we incorporate these kinds of labor and the practitioners who value this work earlier and more authoritatively in digital project development? Where should ethical imperatives towards sustainability similarly influence ethical forms of collaboration between seemingly-disparate forms of academic labor?

Hybrid Positions, or 2-for-1 DH Labor

Chelcie Rowell

For some time now academic libraries have embraced a liaison model in which subject librarians support the research, teaching, and learning of specific academic departments. Many libraries have also responded to the changing needs of their campuses by hiring functional specialists in areas such as copyright, GIS, and data management. Still others combine subject specialist and functional specialists into one hybrid position, such as "DH and English Librarian," often filled by a person trained as a librarian with an MLIS degree who may or may not have an additional advanced degree in a discipline. The precarity of these positions stems not from their lack of permanence, but rather from the multiple demands placed on the person in this role. Not only must she navigate the changing field of librarianship in the relatively new position of "DH Librarian," but she must also continue to fulfill the more traditional expectations of a liaison librarian, such as building collections, conducting information literacy class sessions, performing outreach, and answering reference questions. These hybrid positions are therefore not hybrid at all, but rather a combination of the job responsibilities of two or more positions.

Library administrators often market the people in these hybrid roles as collaborative equals in faculty-directed

digital scholarship projects, and indeed, they could be. Nevertheless, this is an outsized expectation for many reasons, the first being that the DH Librarian does not have the luxury of focusing on one research project at a time, but must juggle a variety of projects brought to her from a range of faculty with different methodological approaches and specializations. And while training in librarianship cultivates knowledge indispensable to scholarly digital projects (metadata creation, data curation, research methods, archival principles), librarians enter these collaborations on unequal ground due to long-standing, rigid hierarchies in academia that subordinate librarians' broad expertise in project development to faculty researchers' disciplinary knowledge.

Librarians' labor is also at risk of being rendered invisible due to the quantitative models for tracking and evaluating library workers' performance. Recording the number of interactions with faculty, for instance, is an insufficient way to capture the variety of interactions (both in-person and virtual) that are required of a collaborative, long-term scholarly project. Just as the systems for evaluating tenure and promotion cases for faculty are often incompatible with digital humanities scholarship, so are the systems for evaluating librarians' performance.

Therefore, if library administrators invest resources in new positions such as "DH Librarian" (or add this title to existing job titles such as "English Librarian"), they must also invest resources in reimagining how to capture evidence of successful efforts to build campus DH community and capacity, how to empower the individuals who occupy those roles to effectively manage their portfolio of projects and responsibilities, and how to revise the evaluation methods for librarians called upon to collaborate on long-term digital scholarship projects.

While this paper highlights problems stemming from hybrid functional-subject specialist positions in academic libraries, it resonates with other papers on this panel, as many of these challenges are shared by librarians and junior faculty who are appointed as assistant professors or digital scholarship coordinators. For both of these kinds of "miracle workers," outsized expectations add up to precarious labor.

Delivering on the Deliverables

Paige Morgan

Many digital humanists in centers or libraries—interdisciplinary positions that cater to multiple departments—are expected to demonstrate the products of their digital labor to high-ranking administrators and stakeholders on a consistent basis. As such, they often are on tight and over-extended timelines to produce high-quality digital scholarship that will prove their value to the institution and demonstrate evidence of the success the institution's investment in digital humanities. To do so, many digital scholars are implicitly encouraged to cut corners in order to meet the unrealistic demands of the organization, or to compensate for faculty members' lack of experience scoping projects. When due attention is given to the project development process, the DH specialist in charge may be discredited or regarded as an unhelpful collaborator because she does not achieve project milestones in a timely manner or she has "failed" to deliver a completed product that met faculty members' expectations, whether or not they were realistic. Although stakeholder enthusiasm for digital humanities may be considerable, institutions are still learning how much and what sort of work is necessary to bring a project to completion effectively, sustainably, and without considerable exploitation. Even supportive and practical administrators may find that they vastly underestimated the work involved—but cannot provide more needed support without more evidence of success in the form of finished projects.

Such behavior places an overemphasis on the product rather than the process—resulting in the elision of the work around relationships developed, skills acquired, tools tested, and/or infrastructure created to produce a project. While prototypes and completed projects are certainly impressive, they cannot be sustained without these other, equally important, skills. Both building a project and developing the skills, policies, and partnerships needed to sustain energy and activity around it take significant time and training. Few DH specialists receive this guidance during their coursework and are therefore expected to acquire these traits upon hire. A further complication is that many long-running DH projects may involve navigating complex and fractious partnerships between departments and faculty members—and specialists may be in the difficult position of trying to quell squabbles and keep participants happy without having the authority to set boundaries that would be beneficial to the library, center, or department where they are housed. While the specialist may work "miracles," if the miracles come at the cost of burnout, then stakeholders' understanding of the labor necessary to achieve success and plan effectively for future endeavors is badly warped.

Many entities—whether libraries, centers, or departments—hope to become leaders in DH within their local campus communities and beyond. But what is required, not just to "make DH happen," but to make a particular entity a leader? This paper will explore the pressures that such goals place on DH specialists, as well as how pressure to deliver success shapes practices around collaboration within library/center project teams; and will offer suggestions for rethinking institutional strategy that could lead to better shared expertise and less precarity in the risk of specialist burnout.

References

Gil, A. (2016). 74 and counting. Open Directory of Miracle Workers. Tweet.

Gil, A. "Open Directory of Miracle Workers." Google Sheet.

Hsu, W. (2016). Lesson on Public Humanities from the Civic Sphere. In Gold and Klein (eds), *Debates in Digital Humanities*. Minneapolis: University of Minnesota Press, 2016.

Flanders, Julia. (2012). Time, Labor, and 'Alternate Careers' in Digital Humanities Knowledge Work. In Gold (ed), *Debates in Digital Humanities*. Minneapolis: University of Minnesota Press.

Kirschenbaum, M. (ed) (2002). *Digital Humanities Quarterly* 3(2).

Maron, N. and Pickle, S. (2014). Sustaining the Digital Humanities: Host Institution Support Beyond the Start-up Phase. Ithaka S+R.

McPherson, T. (2012). Why Are the Digital Humanities so White? or Thinking the Histories of Race and Computation. In Gold (ed), *Debates in Digital Humanities*. Minneapolis: University of Minnesota Press.

Posner, M. (2016). Money and time. Blog post. 14 March 2016.

Posner, M. (2012). The digital humanities postdoc. Blog post. 7 May 2012.

Experimental Humanities

Maria Sachiko Cecire
mcecire@bard.edu
Bard College, United States of America

Dennis Yi Tenen
dt2406@columbia.edu
Columbia University, United States of America

Wai Chee Dimock
wai.chee.dimock@yale.edu
Yale University, United States of America

Nicholas Bauch
nbbauch@ou.edu
University of Oklahoma, United States of America

Kimon Keramidas
kimon.keramidas@nyu.edu
New York University, United States of America

Freya Harrison
f.harrison@warwick.ac.uk
University of Warwick, United Kingdom / University of Pennsylvania, United States of America

Erin Connelly
erincon@upenn.edu
University of Pennsylvania, United States of America

Panel Abstract

Recent years have seen humanities scholars from different fields and types of institutions begin to call for an "experimental humanities," typically as a replacement for or an extension of digital humanities frameworks. Most recently, Wai Chee Dimock suggests drawing upon scientific understandings of "experimental" in the methodologies and aims of literary study – an approach that "test[s] the extent of isomorphism among different fields of knowledge" and that she argues can drive more original, collaborative, resilient, and publicly engaged humanities scholarship. While the term "experimental" was most closely associated with the rise of scientific methodologies and cultural power in the first half of the twentieth century, its adoption by artists over the past century means that the word now evokes the play and emotional investment of the artist's studio as much as it does the precision and rationalism of the scientist's lab.

This panel draws on the work and experiences of a range of scholars who conceive of their research and pedagogy in experimental terms. Looking beyond the digital, these academics work at the boundary between the university and the world, engaging in issues of social justice and advocacy, bringing traditionally disparate fields and approaches to bear on one another, or using empirical methods---such as map-making, microbiology techniques, or field work---in the study of history, literature, and culture. The papers in this panel explore how "experimental humanities" can be a useful paradigm for extending or reframing work in DH, addressing experimentation both in the sense of methods associated with scientific inquiry and in the sense of a radical, process-based practice, the outcome of which remains highly subjective, speculative, and unknown.

The papers in this panel represent work from diverse institutional settings, from the small liberal arts college to the big state school to the Ivy league. They theorize the experimental humanities as they appear in specific projects, undergraduate and graduate curricular initiatives, lab settings, and in partnerships with non-academic actors. Taken together, our panelists represent a new experimental turn in DH and humanities work more broadly, which is being institutionalized in named centers, groups, and labs across the United States and Europe. Each of the six presentations will take 10-12 minutes, grouped to frame the field (Dimock; Tenen), delve into experimental research methods (Harrison and Connelly; Bauch), and discuss the institutionalization of Experimental Humanities (Cecire and Keramidas), leaving 20-30 minutes for discussion and Q&A at the end.

DH and EH: A Symbiosis

Wai Chee Dimock

Taking a close look at two interlocking entities at the University of Pennsylvania -- the Price Lab for Digital Hu-

manities and the Penn Program for Environmental Humanities -- this paper argues for data and computation as a key partner in reshaping the humanities for the 21st century: as a science-informed, experiment-driven, and practice-rich discipline.

The Experimental Turn

Dennis Yi Tenen

My goal in this talk will be to situate a variety of "experimental" approaches to the study of literature in culture within a wider experimental turn, steering the academy toward critical practice, especially in fields long-dominated by speculative thought.

The experimental turn represents a generation's dissatisfaction with armchair philosophizing. Recall the burning armchair, the symbol of the experimental philosophy movement. Joshua Knobe and Shaun Nichols, some of the early proponents of the movement, explain that "many of the deepest questions of philosophy can only be properly addressed by immersing oneself in the messy, contingent, highly variable truths about how human beings really are." The emergence of spaces where research in the humanities is done exemplifies a similar trend. In naming the locations of their practice "laboratories," "studios," and "workshops," humanists reach for new metaphors of labor. These metaphors aim to reorganize the relationship between body, space, artifact, knowledge, and inscription.

As another example from the field of early modern history consider the preface to a recent volume on Ways of Making and Knowing, edited by Pamela Smith, Amy Meyers, and Harold Cook. They write that the "history of science is not a history of concepts, or at least not that alone, but a history of the making and using of objects to understand the world." Smith translates that insight in the laboratory, where, together with her students, she bakes bread and smelts iron to recreate long-lost artisanal techniques. For those who experiment, "book knowledge," "artifactual knowledge," and "knowledge at hand" connect in practice.

Somewhere between a lab experiment and experimental art, I join experimentalists like Smith to imagine a space for process-based scholarship, "to be judged not on its success or failure, but simply as an act the outcome of which is unknown."

Versioning and Visioning Geographic Language with Digital Design

Nicholas Bauch

The Experimental Geography Studio <http://geographystudio.org> is a research collective that focuses on a specific question: what happens when techniques from the creative arts are used to advance theoretical discourse in human geography? Despite the increasing ubiquity of digital/web mapping platforms in academic research, journalism, and popular usage, human geography ironically enough remains an overwhelmingly textual-theoretical enterprise. Categories like space, place, landscape, region, and territory are all fine-tuned in geographical discourse with rich and evocative metaphors, including "verticality," "hybridity," "oceanic," "phase," "bending," and "topology," among many others (see references below). As part of the formal minting of the Geo-Humanities subfield by the American Association of Geographers in 2011, there is a growing institutionalization and recognition of the art-geography nexus, involving media practices ranging from video, to sculpture, photography, performance, and not least digital/web design. What happens when the metaphors used to understand spatial mechanisms become re-versioned into graphic forms? In this paper I present an ongoing project in the Studio—Versioning Geographic Language—that aims to crystalize some of these metaphors in visualization. By creating a catalog of possibilities, the purpose is to clarify the textual slippage that inevitably occurs as metaphors accrue layers of intention by different authors. Far from obscure, these guiding concepts have great power in framing how students, scholars, and writers imagine the inner workings of—as Henri Lefebvre might have put it—the ways in which different spaces are produced. Pausing and seeing analytical concepts like "verticality," etc. allows us to sharpen them and return them back into empirical research about topics such as citizenship, urban environments, biotechnologies, and statecraft. The 'digital,' in this case, is not the same tool as it often has been in Digital Humanities, i.e. one that takes as its starting point a big data paradigm. Rather, it draws cues from graphic design and visual art to forge patterns and symbolizations.

Datamining Medieval Medical Texts for Antimicrobial Drug Discovery

Freya Harrison

Most antimicrobials are derived from natural products. Thus, ethnopharmacology (the study of traditional pharmacopeias) can help us find new antimicrobials to fight the rising scourge of multi-drug resistant infections. However, the standard approach of purifying individual compounds from natural materia medica rarely produces clinically-useful products. Historical medicinal remedies, in contrast, often involve complex preparations of several ingredients. Medieval European manuscripts describe numerous intricate remedies for infections: their efficacy may rely on creating a "cocktail" of natural products,

each of which has little or no antimicrobial potential when used alone. Modern science rarely investigates the activities of entire remedies, or explores how combinations of natural products work together. Further, while the folk medicines of other continents have long been mined for potential drugs, pre-modern European medicine has been dismissed as superstition or placebo. The remedies in these understudied manuscripts could be the products of rational drug design on the part of the physicians who created them, and if this is the case they could contain long-forgotten treatments for infectious diseases.

Our team (the Ancientbiotics consortium) investigated a 10th-century remedy for eye infections. This quadripartite cure is highly bactericidal against Staphylococcus aureus and other important pathogens. Crucially, its efficacy depends on combining ingredients exactly as specified by the text. Statistical analysis of the original manuscript revealed that this remedy's ingredients were combined in other remedies more often than would be expected if the author simply selected ingredients randomly. Tests of other common ingredients in medieval infection treatments also showed that combining ingredients can generate preparations with unexpectedly strong antibacterial activity. This has led us on a quest to turn collections of medieval remedies into databases amenable to statistical analysis to find the patterns and processes that underlay their construction. We will present our team's key findings so far, and place them in the context of an interdisciplinary, quantitative analysis of how medieval doctors chose, combined and used the materia medica available to them.

Inclusive Practices in Experimental Humanities

Maria Sachiko Cecire

DH prides itself on its attempts to break down institutional hierarchies and embrace the "fun" of scholarly inquiry and discovery. And yet the power dynamics of insiders and outsiders remains – including the long-running debate about whether or not one needs to be able to code in order to "do" DH (Gold, 2012). This paper argues that an Experimental Humanities model invites partners and participants with a wide range of backgrounds and kinds of expertise. By encompassing both the scientific implications of the word "experiment" and the creativity and affective commitment of the artist's studio, Experimental Humanities reorients the landscape of what may be prioritized as meaningful input and what constitutes technical skill in a humanities project. To theorize this approach I build on the work of medievalists Carolyn Dinshaw and Richard Utz, who argue for the importance of overcoming the barriers of professionalization that distinguish the "serious scholar" from the amateur in doing honest, self-aware humanities research. Dinshaw argues for recognizing the "radical love" of the amateur in scholarly production – in the case of medievalists, this manifests as the LARPers, Renaissance Faire attendees, and Tolkien enthusiasts whose activities academics tend to thrust away from themselves as unrelated to their work. Meanwhile, Utz uncovers medievalists' own histories of amateur passions, and suggests that repressing these origins results in less rigorous and useful scholarship.

I will describe two ways in which "amateurs" to the DH world are integral to the Experimental Humanities (http://eh.bard.edu) curricular initiative and center that has been running at Bard College since 2012. The first has to do with recruiting academics who study or use experimental technologies even if they do not see themselves as digital humanists, allowing us to build more robust, diverse teams. The second is our partnerships with people from outside of the academy who help us develop civically engaged experiments – both digital and non-digital – in the humanities. After a brief overview of Bard EH I will discuss a few specific ventures, such as our Digital History Lab which works alongside public servants, historical societies, and town libraries to create local projects that preserve and promote public history in the Hudson Valley. I suggest that drawing in, rather than distancing ourselves from, the passions and experiences of "amateurs" can help enrich our knowledge of the world and produce more interesting, useful experiments in the humanities.

Experiment as Experience, Practice as Pedagogy: Another Way of Rethinking Humanities in a Digital Age

Kimon Keramidas

In playing their part in assuaging the perceived crisis in the humanities, the experimental humanities do not center any one methodology or champion a particular augmentation of traditional humanistic approaches. Rather they engender a sense of adventure, encouraging new forms, promoting a creative mindset, and provoking scholars to apply the tenets of rigorous scholarship to a broader range of intellectual outputs. The experimental humanities encourage this work fully aware that digital technologies have a had a profound impact on our culture, but recognize their role as a defamiliarizing force and important part of our cultural context rather than as a motivator for epistemological change and forcible redefinition. This presentation will discuss three classes taught at NYU's Center for Experimental Humanities that combined traditional historical and theoretical humanities work with experiential learning and immersion in practical projects.

The first course, Telling the Sogdian Story, integrated students in a large-scale digital exhibition being organized by the Smithsonian's Freer|Sackler Asian Art Galleries. Along with traditional lectures from scholars specializing in the Sogdians, a medieval mercantile people from Central Asia, students participated in project planning discussions

with Freer|Sackler staff, and consultations and critiques with professional web designers and developers. Students were involved not only in researching and writing about objects and themes for the exhibition, but studied interface design and worked in teams to prototype different interactive design options for the final exhibition.

The second course, Queering the Web, put students at the center of a redesign of the public history site OutHistory.org. Combining queer history and theory, performance theory, interface design, and design history, the course considered whether web design and computer science carry implicit heteronormative practices that inherently impinge on the ability to properly represent queer history. Students were asked to write essays on the history of LGBTQ people in the United State, critique the current state of the site, and propose design modifications that not only updated the sites look and usability, but queered users' interaction with the materials.

The third course, Making Room for Youth, considered the Hardcore Punk movement of the late 1970s-1980s as a model of community-based cultural activism driven by DIY-practices and unique deployments of analog media. Students studied Hardcore both to understand the movement and to reflect on how digital technologies have changed our experience of culture, the capacity of cultural products to act as a force for initiating social change. For their culminating project, students organized an event that integrated analog and digital techniques to create a historical impression of Hardcore. This event highlighted how digital media have changed the music scene, and asked visitors to reflect on the role of cultural activism in a political era that shares many similarities with current conditions.

These three course, all co-taught, all involving project-based, process-oriented learning, and all using digital methodologies and technologies while actively interrogating them, represent how an experimental approach to the humanities can provide new perspectives and new experiences for humanities students.

Links:

Telling the Sogdian Story: http://kkeramidas.nyufasedtech.com/sites/telling-sogdian-story/

Queering the Web: http://kkeramidas.nyufasedtech.com/sites/queering-the-web/

Making Room for Youth: http://kimon.hosting.nyu.edu/sites/making-room-for-youth/

References

Allen, J. (2016). *Topologies of Power: Beyond territory and networks.* New York: Routledge.

Elden, S. (2013). "Secure the Volume: Vertical geopolitics and the depth of power." *Political Geography* 34:35-51.

Connelly, E. (forthcoming 2018). "Treating Infection in the Lylye of Medicynes." In C. Lee, E. Connelly; S. Künzel (eds.), *proceedings from the conference Disease, Disability and Medicine in Medieval Europe*, Studies in Early Medicine. Archaeopress.

Connelly, E. (2017). "My Written Books of Surgery in the Englishe Tonge: The London Company of Barber-Surgeons and the Lylye of Medicynes." *Manuscript Studies, A Journal of the Schoenberg Institute for Manuscript Studies* 2.2.

Dimock, W. C. (2017). "Experimental Humanities." *PMLA* 132.2: 241-9.

Dinshaw, C. (2012). *How Soon Is Now?: Medieval Texts, Amateur Readers, and the Queerness of Time*. Durham: Duke University Press.

Harrison, F. et al. (2015). "A 1,000-year-old antimicrobial remedy with antistaphylococcal activity." mBio 6:e01129-01115.

Harrison, F. & Connelly, E. (2018). "Could Medieval Medicine Help the Fight Against Antimicrobial Resistance?" In C. Jones, C. Kostick & K. Oschema (eds.), *Making the Medieval Relevant*. Berlin: De Gruyter.

Gold, M. K. (2012). "The Digital Humanities Moment," in M. K. Gold (ed.), *Debates in DH*. Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/1

Jones, M. (2009). "Phase Space: Geography, relational thinking, and beyond." *Progress in Human Geography* 33 (4):487-506.

Knobe, J., and S. Nichols (2008). *Experimental Philosophy*. Oxford; New York: Oxford University Press.

Smith, P. H., A. R. W. Meyers, and H. J. Cook, eds. (2014). *Ways of Making and Knowing: The Material Culture of Empirical Knowledge*. The Bard Graduate Center Cultural Histories of the Material World. Ann Arbor: University of Michigan Press.

Steinberg, P., and K. Peters (2015). "Wet Ontologies, Fluid Spaces: Giving depth to volume through oceanic thinking." *Environment and Planning D: Society and Space* 33:247-264.

Utz, R. J. (2017). *Medievalism: A Manifesto*. Past Imperfect. Kalamazoo: ARC Humanities Press.

Whatmore, S. (2002). *Hybrid Geographies*. London: Sage.

Reimagining the Humanities Lab

Tanya Clement
tclement@utexas.edu
University of Texas at Austin, United States of America

Lori Emerson
lori.emerson@gmail.com
University of Colorado at Bouldre, United States of America

Elizabeth Losh
lizlosh@gmail.com
William & Mary University, United States of America

Thomas Padilla
tgpadillajr@gmail.com
University of Nevada, Las Vegas, United States of America

A recent article has propped up the strawman that is the scientism of Digital Humanities (Brennan, 2017). This is not the first time that DH has been called out in terms of its supposed lack of success as a science (Allington, Brouillette, Golumbia, 2016). To be sure, scientific achievements are wonderfully measurable; they can be measured in terms of a year of progress – or ten years or the last 100 years – are marked by notions of societal impact. What has science done for us lately? We ask, as we point to discoveries and solutions. Likewise, a reputable humanities scholar asks, what has DH accomplished with all the money it has been bestowed? DH hasn't "reveal[ed] the secrets of complex social and cultural processes" and perhaps it is has not even made one of its central cases well since the author finds activities such as the "digitization, classification, description and metadata, organization, and navigation" of our cultural artifacts to be uncritical, apolitical work; rather, it is "a list, which leaves out that contradictory and negating quality of what is normally called "thinking" (Brennan, 2017). Yet, while these critics posit DH as a neoliberal pursuit to change higher education and thus the liberal arts into an education-for-hire endeavor, those who are the loudest critics of such a DH tend to measure DH's "success" according to a pay-for-hire scale. We paid for it, they seem to be saying, What did we get?

In marked contrast, DH scholars have long maintained "that scientific method and metaphor is, for the most part, incompatible with the terms of humanistic endeavor" (Ramsay, 2011; Drucker 2012; Binder 2016; Witmore 2016). To be sure, what have the *Humanities* "accomplished" in the last decade, digital or no? Have the Humanities revealed The Secrets? As humanists, we tend to work towards fissures and fractional tectonic shifts that resut in longer, slower, more nuanced and, in many cases, immeasurable impacts. For many of us involved in DH scholarship, singular accomplishments and "success" measured in terms of "done and out" or "problem solved" are not typically our goals.

This panel interrogates the polarities that remain present in the perceived differences between the proposed scientism of the "digital" on the one hand and the "humanities" on the other by discussing a current trend in DH towards establishing "digital humanities labs." Often, this oxymoronic title points to spaces of seemingly unscientific goings-on, of small doings, little happenings, and turtle-paced epistemological shiftings at the level of, and articulated through, infrastructure development. In the DH lab, "infrastructure" is understood as a "socio-technical phenomenon that enforces constraints on human experience at the same scale, complexity, and general cultural impact as the idea of 'culture' itself (Liu, 2016). This role of the DH lab has its roots in feminist inquiry by imbuing DH with science only in the sense of the "science question", which considers the politics underlying epistemologies of "purportedly value-neutral claims and practices" [Harding 1986, 23] and resonates with the work (the research, theory, and practices) being done to build information infrastructure in the humanities today. Ultimately, this panel situates "lab work" in DH as a site for humanistic rather than scientific work, as a site for interrogating what it means to be a "lab"; for generativity, legibility, and creativity; for exploring what is engendered by ad hoc arrangements, small scale problems, and low tech tools; and for considering how the co-construction of knowledge with stakeholders and community members can introduce participants to the theoretical, practical, and political implications of considering collections as data in the humanities.

Framing Potential

Thomas Padilla

Collections as data represents a mode of engagement that aims to surface the data driven potential of digitized and born digital library, museum, and archival collections. In the United States the Institute for Museum and Library Services supported *Always Already Computational: Collections as Data* project and the Library of Congress' *Collections as Data* events represent significant development initiatives that extend the aim of contemporary efforts like the Hathitrust Research Center into a more diversified set of institutional contexts. In Europe, related projects like DARIAH run apace. Increasingly, the collections as data concept is bolstered by attempts to play out the implications in theory and practice. Praxis in action spans emergent research library initiatives, national library initiatives, museum initiatives, cultural heritage position formation, and Information Science and Digital Humanities curricula development. It stands to reason that labs, including but not limited to those operating in the Humanities, can benefit from partnering on collections as data efforts. Collections as data provide malleable grounds for enhancing cross campus, cross institutional, and cross community partnerships that aim to support research, pedagogy, and civic engagement in a contemporary knowledge environment that shifts ever toward de facto digital knowledge creation. Possibilities in this space can be effectively pursued via resolution through three conceptual collections as data frames: **(1)** generativity - a question of meaning making capacity **(2)** legibility - a question of ability to convey provenance and possibility and **(3)** creativity - a question of the extent to which the effort provides the means or a context for empowered experimentation.

Surfacing the data driven potential of collections works toward the purpose of cultivating contemporary agency in a digital environment. Collections as data work within and

beyond labs constitutes a social imperative. With predominantly smooth GUI and application driven commodification of digital environments it becomes more and more difficult to push past the surface to gain purchase on the subjective forces that shape the data that constitute the digital objects that are trafficked throughout them. For many within and outside of academia it is not readily apparent that a Word document is not just a document, a website is not just a projection on a screen, an image is not merely a surrogate, and a tweet is much more than 280 characters. Meanwhile the facility and concordant power to control these composite environments with their composite objects resides in the hands of those that take a data first, representation second mentality - namely corporations, governments, law enforcement agencies, and researchers that exhibit ethically questionable engagement with digital traces of life. The collections as data imperative entails cultivation of the means to help all members of society, across all classes and backgrounds, working within the academy and outside of it to engage critically with digital traces of human activity in the fullest manner possible, native to the complexity of their form, and critically attuned to the possibilities and perils that come with their use. In what follows collections as data frames will be used to evaluate salient data driven research, professional practice, and lab oriented efforts in order to support speculative development of what a collections as data oriented lab could be.

The Unbearable Open-endedness of 'Lab: A Variantology

Lori Emerson

The second panelist will first discuss findings that she and her co-authors have accumulated in the course of writing *THE LAB BOOK: Situated Practices in Media Studies* (forthcoming from the University of Minnesota Press). The project investigates the history as well as the contemporary landscape of humanities-based media labs - including, of course, labs that openly identify as being engaged with the digital humanities - in terms of situated practices. Part of the book's documentation of the explosion of labs or lab-like entities around the world over the last decade or so includes a body of over sixty interviews with lab directors and denizens. As the third panelist will discuss, the interviews not only reveal profound variability in terms of these labs' driving philosophy, funding structures, infrastructures, administration, and outputs; but they also clearly demonstrate how many of hese labs do not explicitly either embody or refute scientificity so much as they pursue 21st century humanities objectives (which could include anything from research into processes of subjectivation, agency and materiality in computational culture to the production of narratives, performances, games, and/or music) in a mode that openly both acknowledges and carefully situates research process as well as research products, the role of collaboration, and the influence of physical and virtual infrastructure. While, outside of higher education, "lab" can now refer to anything from a line of men's grooming products to a department store display or even a company dedicated to pyschometric tracking, across the arts and humanities "lab" now has the potential to capture a remarkable array of methodically delineated and self-consciously documented entities for experimentation and collaboration.

Panelist two also views *THE LAB BOOK* as an opportunity to position the Media Archaeology Lab (MAL) in the contemporary landscape of these aforementioned humanities/DH/media labs. Since 2009, when panelist two founded the MAL, the lab has become known as one that undoes many assumptions about what labs should be or do; unlike labs that are structured hierarchically and driven by a single person with a single vision, the MAL takes many shapes: it is an archive for original works of early digital art/literature along with their original platforms; it is an apparatus through which we come to understand a complex history of media and the consequences of that history; it is a site for artistic interventions, experiments, and projects; it is a flexible, fluid space for students and faculty from a range of disciplines to undertake practice-based research; it is a means by which graduate students come for hands-on training in fields ranging from digital humanities, literary studies, media studies and curatorial studies to community outreach and education. In other words, the MAL is an intervention in "labness" insofar as it is a place where, depending on your approach, you will find opportunities for research and teaching in myriad configurations as well as a host of other, less clearly defined activities made possible by a collection that is both object and tool.

False Equivalencies: Addressing Scientific Positivism from a Feminist Digital Humanities Perspective

Elizabeth Losh

The Equality Lab unites diverse cohorts of scholars, students, and community members working broadly on inequality research in the mid-Atlantic region of the United States. It provides space for digital humanities teams working on community-based archival projects in which descendents and fictive kin may be important stakeholders. For example, its current partners include the Lemon Project, which uncovers and reconciles histories of slavery and segregation in higher education, and the LGBTIQ Research Project, which works with gay, lesbian, and transgender community groups in Virginia. The Equality Lab also provides technical expertise and material support for a broad range of individual student projects by humanities Ph.D. candidates that represent a variety of

disciplinary affiliations and theoretical interests in American studies, such as ethnic studies, disability studies, and environmental humanities.

As a digital humanities initiative, the Equality Lab foregrounds the ways that "equality" as a concept may suggest mathematical and scientific equivalence and thus it partners with the campus center for geospatial analysis and the university's data sciences initiative to assist in producing DH projects that foster what has been called "counter-data action" or "statactivism." (For example, its regional partners have worked on data projects involving redlining in housing policy and police shootings of unarmed persons of color.) Yet the Equality Lab also tests, tinkers, and examines scientistic assumptions about measurability, rationality, and surveillance in digital humanities work by approaching equality as a process rather than a product. The Equality Lab also explores the possibility that other formulations (equity, inclusion, etc.) might be more valuable as descriptors for social justice work in the digital humanities than equality itself.

Although making decisions about specific digital scholarship authoring platforms -- such as Omeka, which the Equality Lab uses to facilitate information exchanges with similar projects -- may be important, the Equality Lab avoids an exclusively tool-centric approach and validates the importance of what Christine Borgman has called "little data" as well as "big data." In creating a multi-functional digital humanities lab space, the Equality Lab adopts perspectives from feminist science and technology studies and its important work on lab culture, such as Adele Clarke and Joan Fujimora's *The Right Tools for the Job*, which argues that "tools," "jobs," and "rightness" are all situational and may reflect the specific contexts of ad hoc arrangements, doable problems, and disciplining tools. This feminist STS approach also validates craftwork and tacit knowledge practices as integral to digital humanities work, just as they are central to the labor of more traditional types of laboratories. The Equality Lab often builds on small-scale and relatively low-tech digital humanities interventions, such as a Wikistorming project with FemTechNet or programming bootcamps in Python or Processing with campus computer scientists.

The Equality Lab emphasizes the importance of working with communities as sites of the co-construction of knowledge to build trust, acknowledge expectations of reciprocity, and give appropriate credit for contributions, as exemplified in its recent three-day symposium on Race, Memory, and the Digital Humanities that featured digital humanities innovators like Jessica Marie Johnson, Gabrielle Foreman, and Marisa Parham.

Data as products of human intentions and world views

Tanya Clement

While collecting institutions at UT Austin such as the Harry Ransom Center (HRC), the LLILAS Benson Latin American Studies and Collections Library, and the Perry Castañeda Library (PCL) all have large repositories of images, audio-visual materials, and text around which UT researchers can conduct scholarly research, basic research around theories and practices that consider how we should prepare, provision, and support the use of these collections as data remains largely uninitiated. At UT Austin, we are conceiving of DH labs as providing new points of access to "collections as data" while also serving as invaluable opportunities for basic research on the very nature of humanities objects and the systems that circulate and represent them to us, now, in the past, and in the future. DH Labs can provide secure and scalable access to digitized and born-digital UT collections as data and support programming but DH labs can also introduce participants to the theoretical and practical implications of considering collections as data.

There are three particular "proof of concept" areas where we are focusing our intervention.

Ethics and Post-custodial Archives: In working with the Guatemalan National Police Archives, LLILAS Benson has made post-custodial archival development and digital scholarship strategic goals its institutional mission. In the post-custodial realm, the physical collection stays with the creator but digital collections can be accessed from other custodians, such as the Benson. The theoretical and political ramifications of this very sensitive work is heightened in the context of human rights documentation initiatives. To further these efforts, LLILAS Benson has been able to secure a Mellon grant to expand its digital holdings and international partnerships, and to establish a dedicated endowment to support digital scholarship initiatives, including a Digital Scholarship in the Americas Speaker Series, workshops, internships, and fellowships for UT and international faculty, graduate, and undergraduate students.

Sound Studies and Audio Preservation and Access: The HRC is home to world-class collections of images and audiovisual materials. In particular, recordings in the collection belonged to some of the 20th and 21st century's most notable writers, artists, and performers. As of January 2017, there are 14,682 audio recordings cataloged in the HRC's Sound Recordings Collection database; of these, 3,226 have been digitized and are available streaming onsite in the Reading and Viewing Room. These are often unique and rare non-commercial recordings often made for private use. A DH Lab with sound analysis technology will provide scholars with an opportunity to reflect on issues of infrastructure development--including data modeling and management; systems for security, integrity, and privacy; and the use of big data and machine learning algorithms—in the context of literary audio scholarship.

Design Thinking and Object Cataloguing: The Alexander Architectural Archives, The Benson Latin American Collection, the Fine Arts Library, the Blanton Museum of

Art, and the HRC all seek to modernize and coordinate their cataloguing processes in the visual arts and to implement shared technologies in order to improve efficiencies, leverage economies of scale, and promote the discovery and use of collections across campus. Structured Design Thinking workshops at the DH Lab at the Perry-Castañeda Library (PCL) will focus on cataloguing methodologies that incorporate multiple internal and external stakeholders; on evaluating available technology platforms; and considering current best practices in cataloguing, universal design, and user experience. A test corpus would include collections data for at least 500 objects-- an encyclopedic collection of approximately 2,000 works with particular depth in Western European art from the fourteenth through twentieth centuries and modern and contemporary art of the Americas from the Blanton and the Gernsheim Collection of approximately 35,000 images amassed by photo-historians Helmut and Alison Gernsheim between 1945 and 1963 currently held at the HRC. Both collections have little-to-no object-level cataloguing, making them ideal subjects for the project for building theory in design thinking and human computer interaction using collections as data.

Data are products of human intentions and world views. Empowering humanities scholars and students to better understand and to act with data entails building an ecosystem that provides opportunities for recognizing, interpreting, and acting upon the affordances of the cultural heritage artifacts that humanists have always studied refigured as data. "Labs" can provide opportunities for basic research in the organization, preservation, curation, analysis, representation, interpretation, and communication of data as well as in the digital tools and platforms, the publishing and open access models, and the social and technological systems these activities afford and inhibit.

References

Allington, D., Brouillette, S., Golumbia, D. (2016). Neoliberal Tools (and Archives): A Political History of Digital Humanities. *LA Review of Books*.

Brennan, T. (2017). The Digital-Humanities Bust. *The Chronicle of Higher Education* https://www.chronicle.com/article/The-Digital-Humanities-Bust/241424 (accessed 4 May 2018).

Binder, J. M. (2016). Alien Reading: Text Mining, Language Standardization, and the Humanities. In *Debates in the Digital Humanities*, edited by Matthew K. Gold and Lauren Klein. Minneapolis: University Of Minnesota Press.

Borgman, C. L. (2016). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press.

Bruno, I., Didier, E., and Vitale, T. (2014). *Statactivism: Forms of Action between Disclosure and Affirmation*. SSRN Scholarly Paper, ID 2466882, Social Science Research Network, *papers.ssrn.com*, https://papers.ssrn.com/abstract=246688 (accessed 4 May 2018).

Clarke, A. and Fujimora, J. (1992) *The Right Tools for the Job*. Princeton: Princeton University Press.

Currie, M., Paris, B.S., Pasquetto, I., and Pierre, J. (2016). The Conundrum of Police Officer-Involved Homicides: Counter-Data in Los Angeles County." *Big Data & Society* 3(2) http://journals.sagepub.com/doi/pdf/10.1177/2053951716663566 (accessed 4 May 2018).

Drucker, J. (2012). Humanistic Theory and Digital Scholarship. In *Debates in the Digital Humanities*, edited by Matthew K. Gold. Minneapolis: University Of Minnesota Press.

Harding, S. G. (1986). *The Science Question in Feminism*. Ithaca: Cornell University Press.

Liu, A. "Drafts for Against the Cultural Singularity" Alan Liu. 2 May 2016.

Witmore, Michael (2016). "Latour, the Digital Humanities, and the Divided Kingdom of Knowledge." *New Literary History* 47 (2): 353–75.

Padilla, T. (2016) *On a Collections as Data Imperative*. Library of Congress. http://digitalpreservation.gov/meetings/dcs16.html (accessed 4 May 2018).

Legado de las/los latinas/os en los Estados Unidos: Proyectos de DH con archivos del Recovery

Isis Campos
ircampos@central.uh.edu
Recovering the U.S. Hispanic Literary Heritage Program; University of Houston, United States of America

Annette Zapata
amzapata@uh.edu
Recovering the U.S. Hispanic Literary Heritage Program; University of Houston, United States of America

Maira E. Álvarez
mealvarez@uh.edu
Recovering the U.S. Hispanic Literary Heritage Program; University of Houston, United States of America

Sylvia A. Fernández
sferna109@gmail.com
Recovering the U.S. Hispanic Literary Heritage Program; University of Houston, United States of America

Introducción

A través de esta ponencia se propone hacer una presentación de varios proyectos digitales que se están llevando a cabo en el programa de Recuperación del Legado Escrito Hispano de los Estados Unidos en la Universidad

de Houston. El programa de Recuperación fue fundado en 1991, se dedica a rescatar, preservar y difundir la cultura escrita de las latinas y los latinos en los Estados Unidos desde tiempos coloniales hasta 1960. Este es uno de los centros más importantes a nivel internacional para la investigación de la documentación histórica, literaria y lingüística de la comunidad latina en Estados Unidos. Las herramientas y prácticas de las humanidades digitales permiten que estos materiales tengan visibilidad y salgan del estado de marginalidad al cual han sido relegados por los espacios literarios e históricos tradicionales por pertenecer a una comunidad minoritaria dentro de EUA.

La lista de autoridades: El legado escrito en la prensa latina de los Estados Unidos

El programa de Recuperación del Legado Escrito Hispano de los Estados Unidosha recuperado más de 1400 periódicos publicados por latinas/os en los Estados Unidos en un período que comprende desde 1808 a 1960. Dentro de este archivo se encuentra el legado escrito de autores y autoras que, a través de diversos géneros, documentan sus experiencias personales, políticas, económicas, religiosas, etc., como inmigrantes, exiliados o nativos de esta tierra. Estos colaboradores han provenido de diferentes países del mundo hispanohablante y representan la diversidad demográfica de los Estados Unidos. Sus publicaciones se han escrito en varios encuentran en una multiplicidad de idiomas: el inglés, el español, el francés, el ladino, entre otros y en forma bilingüe.

La lista de autoridades de Recuperación consta de un repertorio de alrededor de 5000 autores compilados a través de la indización de periódicos llevada a cabo desde 1992. Esta presentación abordará el proceso de recuperación y catalogación de estos autores. Así mismo, a través de perspectivas decoloniales y postcoloniales abordaremos la importancia de este tipo de trabajo de recuperación para evidenciar representaciones diversas y múltiples de las comunidades y culturas de las latinas y los latinos en los Estados Unidos.

Esta presentación aborda el proceso de creación y preparación de metadata con el propósito de facilitar, a través de una plataforma visual, el acceso a la información sobre los/as autores/as incluyendo sus vínculos transnacionales. El traslado de esta base de datos a una plataforma digital tiene como fin ampliar las limitadas nociones que se tienen del legado escrito latino en los Estados Unidos y a su vez aportar y contribuir con datos a la constante reescritura de la historia estadounidense y de otras naciones. Al mismo tiempo contribuye a contestar las preguntas quién habla, de quién se habla, por quién se habla, qué idiomas se emplean y qué conjeturas delimitan su producción, distribución y consumo (Roopika Rissam 2017).

Esta lista de autoridades facilita la contextualización de estas preguntas y a la vez cuestiona la construcción eurocéntrica de la historia literaria de los Estados Unidos. Este proyecto es un trabajo colaborativo con el potencial de servir como fuente de recursos para futuras lecturas.

Documentos personales del puertorriqueño Delis Negrón: Del archivo físico a un archivo digital

El archivo digital de Delis Negrón es un proyecto que captura la vida y el legado del puertorriqueño Delis Negrón, nacido en 1901, a los 16 años emigró a Nueva York y en 1917 se integró al ejército estadounidense. Durante su residencia en los Estados Unidos y México a mediados del siglo XX, quien fuera director, editor, poeta, escritor y activista. El archivo físico del poeta fue donado al programa de Recuperación del Legado Escrito Hispano de los Estados Unidos (Recuperación) por su familia. El archivo personal contiene fotografías familiares, poemarios, recortes de periódicos, manuscritos, correspondencia y notas escritas a mano.

La colección de Negrón representa uno de los tantos esfuerzos liderados por el programa de Recuperación de rescatar documentos y visibilizar historias de minorías no representadas en los archivos. Con el objetivo de trasladar el archivo físico de Negrón a una plataforma digital se creará no sólo el acceso a una historia particular de inmigración, sino también un contexto más amplio. Es decir, recompone la historia intelectual de las/los latinas/os en los Estados Unidos y sus contribuciones políticas, culturales y sociales al país. De ese modo, a través de la colección digital de Negrón, el exponer sus récords, su trayectoria profesional y su participación en la comunidad a través de los documentos disponibles se abren espacios alternativos de inclusión, documenta el rol que han cumplido las latinas y los latinos y su participación en diversas áreas como en el periodismo, la literatura y el activismo político, entre otras. Como sostiene Rodrigo Lazo, "They are the texts of the past that have not been written into the official spaces of archivization" (2010). Este esfuerzo significa que tanto el trabajo de preservación de la colección de Negrón, como el traslado a una plataforma digital, rompen con las formas tradicionales y crea espacios archivísticos alternativos que facilitan el acceso al material, con el fin de incorporar nuevas interpretaciones e identidades.

Durante la presentación se hablará del proceso de preservación de este tipo de archivos y las prácticas de humanidades digitales que permiten no sólo acceso a estas colecciones, sino que también invitan a reflexionar sobre el papel del archivo en general y a cuestionar las nociones de historia y quiénes forman parte de ella (Roopika Risam, 2014). Por ende, los temas a notar incluirán el proceso de recuperación del archivo personal de Negrón, su importancia en el contexto de estudios de los latinos en los Estados Unidos y cuáles fueron los factores que contribuyeron a la creación de la exhibición en línea. También se explicarán los protocolos seguidos en

la construcción de esta exhibición, sus funciones y cómo estos sirven de modelo y proporcionan el contexto para futuras exhibiciones basadas en colecciones del programa de Recuperación.

De la recuperación del archivo periodístico a la creación de Borderlands Archives Cartography

La reciente, constante y agresiva retórica política sobre la frontera entre México y Estados Unidos como una amenaza produce un contradiscurso representado por las comunidades fronterizas. Este contradiscurso representa la frontera como un espacio donde co-existen diversas culturas bajo el control transnacional de hegemonías políticas, económicas y sociales así como un espacio donde las regiones se influyen entre sí, pero mantienen sus propias identidades. De esta forma, Borderlands Archives Cartography (BAC) surge como una resistencia al discurso oficial para dar visibilidad a periódicos del siglo XIX y XX que reflejan historias de múltiples voces.

El corpus del proyecto se inició con un grupo de periódicos de 1808 a 1930 pertenecientes a la base de datos del archivo periodístico del programa de Recuperación del Legado Escrito Hispano de los Estados Unidos (Recuperación). Estos periódicos documentan las voces que existen en la frontera entre México y Estados Unidos por medio de historias personales, locales y nacionales, culturas y legados literarios. Los periódicos documentan las interacciones que dieron lugar a nuevas identidades como resultado de la pérdida de territorio, inmigraciones, exilio, desterritorialización y la vida transfronteriza. Por otra parte, como menciona Nicolás Kanellos, los periódicos ayudaban a los individuos y residentes a proteger sus derechos mediante la lucha contra la segregación y la discriminación, especialmente después de la cesión de una parte del territorio mexicano a los Estados Unidos en 1848 (Kanellos y Martell, 2000).

La necesidad de producir otras historias alternas a la oficial sobre todo que represente la frontera desde sus propias comunidades conlleva a que los archivos de periódicos sean visualizados en una plataforma digital. El proyecto utiliza un mapa digital para visualizar la ubicación geográfica de los periódicos haciendo uso de CARTO, un software de Sistema de Información Geográfica (SIG), y a su vez una página en línea, que sirva como repositorio del mapa, recursos visuales e información relacionada. Estas prácticas de visualización rompen con las formas tradicionales del archivo, transformándolo de un "static repository [to] an active site of knowledge production... [That re-]interpret, and even shapes knowledge from the ground [up]" (Cotera, 2015). De tal forma, estas dinámicas permiten, en este caso, otras nociones de la frontera, metodologías, análisis de data y uso del archivo con el objetivo de promover modos de investigación interdisciplinaria.

Esta presentación abordará el proceso de creación de BAC y su objetivo de localizar, digitalizar y facilitar acceso a archivos periodísticos de ambos lados de la frontera antes y después del establecimiento de la actual línea divisoria. Además, la visualización del contenido de este archivo digital desafía la perspectiva colonialista e imperialista de lo que es la frontera entre México y Estados Unidos. En definitiva, más que un proyecto, BAC se transforma en un compromiso personal con las comunidades fronterizas y su historia.

References

Cotera, M. (2015). Invisibility is an unnatural disaster: feminist archival praxis after the digital turn. *The South Atlantic Quarterly*, 114 (4): 781-801.

Kanellos, N. and Martell, H. (2000). *Hispanic Periodicals in the United States, Origins to 1960: A Brief History and Comprehensive Bibliography*. Houston: Arte Público Press.

Lazo, R. (2010). Migrant Archives: new routes in and out of American Studies. In Pinn, A.B., Levander, C.F., and Emerson, M.O. (eds), *Teaching and Studying the Americas: Cultural Influences from Colonialism to the Present*. Palgrave Macmillan, pp. 199-218.

Rissam, R. (2014). Professionalizing via Digital Humanities. In *New England American Studies Association Spring Colloquium, Professional Realities Inside and Outside the Academy*, 3 May 2014. https://www.slideshare.net/roopsi1/professionalization (accessed 25 April 2018).

Rissam, R. (2017). Breaking and building: the case of postcolonial digital humanities. In Singh, J.G. y Kim, D.D. (eds), *The Postcolonial World*. Routledge, pp. 345-362.

Social Justice, Data Curation, and Latin American & Caribbean Studies

Lorena Gauthereau
lgauthereau@uh.edu
University of Houston, United States of America

Hannah Alpert-Abrams
halperta@gmail.com
University of Texas at Austin, United States of America

Alex Galarza
agalarza@haverford.edu
Haverford College, United States of America

Mario H. Ramirez
mario.hugo.ramirez@gmail.com
Indiana University, United States of America

Crystal Andrea Felima
felima@ufl.edu
University of Florida, United States of America

Overview

The digitization of cultural heritage artifacts and historical documents offers new opportunities to preserve vulnerable records, undo archival silences, center marginalized voices, and enable the pursuit of justice for oppressed communities. In the case of records from Latin American, Caribbean, and Latinx communities, partnerships with academic institutions in the United States can help connect communities with the resources necessary to undertake large-scale digitization projects. These relationships, however, are not without questions. How can we maintain equitable partnerships in the face of the uneven distribution of resources? How can we prioritize community needs when they do not coincide with the goals of academic institutions? How do we support and promote multilingual digitization projects by way of largely Anglophone institutions and digitization workflows? And how do we ensure that our projects support social justice without creating new risks for vulnerable communities?

This panel addresses these questions by bringing together students and early career scholars whose work responds to the mandate, from institutions in the United States, to diversify the practice of data curation by building digital collections of Latin American, Caribbean, and Latinx materials. This mandate has been made explicit by the recent development of the Council on Library and Information Resources (CLIR) postdoctoral fellowship in data curation and Latin American and Caribbean studies, but it extends across U.S. institutions. Indeed, in many cases the projects represented here have been operating largely outside the framework of the digital humanities for many years. In reflecting on the history and future of these projects, we will explore how the digital humanities can serve as a useful framework for promoting these digitization projects, and where its limitations lie.

At the heart of this panel is the question of social justice. How do we ensure that these digitization projects, and the work that they enable, are always oriented towards justice? The projects represented here consider multiple frameworks for justice. In some cases, transnational digital preservation can be a kind of social justice for extremely vulnerable communities, particularly when it is accompanied by public institutional support. Other projects foreground a post-custodial approach toward archiving that shifts the role of archivists from custodians of records in a centralized repository, to that of managers of records that are distributed at the organizations where the records are created and used. This approach is considered as both a methodology and a means of building ethical partnerships with projects that aim at social justice. As we consider questions of access, description, and representation, new questions about justice emerge. How can we open the historical record without creating new risks for vulnerable community members? How can we support the recovery of historical memory without retraumatizing users? And how can we use diverse collections to decolonize the digital humanities? In answering these questions, we recall, in the words of Wendy Duff et al., that "social justice is always a process and can never be fully achieved" (324-325). It is as an engagement with this process that we address the practical concerns of funding, digitization workflows, storage, and platforms, remaining attentive to the downstream consequences of these practices for future research in digital humanities and Latin American, Caribbean, and Latinx studies.

Eternal sunrise: digital life cycles and long-term preservation for social justice archives

Hannah Alpert-Abrams

When the University of Texas at Austin partnered with the Guatemalan Police Archive (*Archivo Histórico de la Policía Nacional,* or AHPN) in 2011 to develop a digital archive, the plan was to provide secure, stable, long-term access to the collection. The AHPN contains approximately eighty million pages of records documenting the actions of the Guatemalan state police from 1882-1986, including detailed documentation of the period of the armed conflict (1960-1996). While the archive has significant value for historical research, the primary goal of the AHPN as an institution has been to facilitate the pursuit of justice through documentary evidence by providing support to court cases and by helping families to uncover the fates of their loved ones. For this reason, the AHPN is vulnerable both to physical destruction and political repression. A transnational digitization project promised to support the mission of the AHPN by protecting the documents from political or natural damage, and by creating paths to access that do not require the (sometimes dangerous) act of visiting the archive itself.

In this paper, written more than six years after the launch of the AHPN website, I reflect on how the social justice mission of the AHPN digitization project has been served or inhibited by the digital project lifecycle at the University of Texas. As at many institutions, digitization projects at UT frequently begin with urgent digitization, often timed according to an external grant cycle. This is then followed by preservation and maintenance; eventually, as needs transform, the project is sunsetted and archived. But the realities of institutional and political conditions do not always support this timeline. In the case of the AHPN, for example, political conditions in Guatemala in 2011 led the archivists to launch the website long before the collection had been fully digitized. One consequence of this urgency is that the burden of labor was shifted beyond the conclusion of the funded production period and well into the maintenance stage. As the collection has grown (doubling in the last six years alone), this has created an increasingly untenable situation for the university, resulting in significant inhibitions both to preservation and access.

This paper uses the case of the AHPN to consider how political contingencies can impact long-term maintenance and scope creep in the case of social justice archives. Given that political urgency is a normal condition for social justice work, I will argue that social justice digitization requires a shift in the way we value digitization work, including the way we design digitization timelines, distribute institutional resources, and value professional labor.

(Digital) Methodology of the Oppressed: Approaching US Latina/o Digital Humanities through Decolonialism and Affect

Lorena Gauthereau

Archives and Digital Humanities (DH) projects that highlight minority voices have the ability to disrupt the mainstream perception of history and literary canon through unacknowledged histories. All too often, large-scale DH projects and archives have reinforced Western epistemology and ontology. In response to this, scholars such as Alex Gil, Roopika Risam, micha cárdenas, Kara Keeling, and Syed Mustafa Ali have notably approached DH from a standpoint of thinking from the margins in order to encourage DH scholars to create and adopt methodologies that engage decolonial theory. Such methodologies consider the ways in which categories such as languages, borders, maps, Library of Congress subject headings, gender binaries, sexuality, Western religion, abled-bodies, and coding frame knowledge and knowledge-production through a colonial lens. While national archives help to structure knowledge through an authoritative national narrative, DH approached through a decolonial methodology seeks to address the gaps not only in digital scholarship, but also in the official (or "mainstream") archive. Fields such as US Latina/o studies, Black cultural studies, African American studies, Indigenous studies, multi-ethnic studies, border and borderland studies, transnational studies, hemispheric studies, Third World feminism, queer studies, and immigration literature have already begun the work of decentering traditional notions of nation, citizenship, sexuality, gender, language, and history.

In this paper, I will focus on the emerging US Latina/o Digital Humanities initiative at the Recovering the US Hispanic Literary Heritage project (aka "Recovery") in order to tease out what is truly at stake and how we can move toward decolonizing the Digital Humanities. Specifically, I rely on Women of Color (WOC) theory such as Chela Sandoval's *Methodology of the Oppressed* (2000) to tease out the implications of "oppositional consciousness" and Affect theory on developing a methodology for the Digital Humanities. Decoloniality lends itself to methodology in the way that it seeks to question history and hegemonic structures. One of the ways that decolonial theory approaches such a dimension is through what Emma Pérez (1999) calls the "decolonial imaginary." It is through the decolonial imaginary that we can push back against colonial legacies that structure our lives. Coloniality insists on the preference of Western ontologies and epistemologies and attempts to erase all non-Western forms of existing and knowing. It delegitimizes non-standard and non-Western languages and tries to put people and histories into strict categories. My goal, then, is to highlight the structural (colonial) problems encountered in US Latina/o DH and to point to the decolonial methodology that is evolving and coming out of these projects in response to such problems.

Revisiting the Archivo Mesoamericano: Digitization and the Revolutionary Histories of Central America and Mexico

Mario H. Ramirez

Culled from the archives of the Museo de la Palabra y la Imagen (MUPI) in El Salvador, the Institute of the History of Nicaragua and Central America (IHNCA) and the Center for Research and Advanced Studies in Social Anthropology in Mexico, the annotated videos that constitute the Archivo Mesoamericano at Indiana University, Bloomington document the rich activist history of the region, and the fervor and desire for profound social and political change that accompanied the period between the 1980s and early 1990s. Capturing well known political activists, indigenous traditions, local demonstrations and insightful reporting, the videos provide a vibrant tableau of a collective striving for equity, autonomy and agency that would have a significant impact on the present and futures of the nations involved. A post-custodial project, whose early digitization was achieved in concert with the organizations in question, the Archivo Mesoamericano is now on the precipice of a new endeavor to make the collection more dynamic and available through the migration of current holdings onto a new format, and the rekindling of relationships with stakeholders as a means of increasing holdings, collaborations and cross-institutional projects. Demonstrating a renewed commitment on the part of the university, the Digital Collections Services Department and the Center for Latin American and Caribbean Studies (CLACS) to the pedagogical and historical import of the collection, this project exemplifies a joint interest in highlighting the seminal importance of social justice efforts in the region to contemporary understandings of Mexican and Central American societies.

This presentation will provide an overview of the Archivo Mesoamericano, its content and particularities, but moreover focus on the process that brought the collection to the university and has framed its history there. Primarily collected and donated by Professor Jeffrey Gould (a noted scholar on the region), and supported at its incipience by special collections staff, the collection has

struggled to achieve the necessary research and pedagogical relevance at the university and beyond despite the rich opportunities that exist vis-à-vis CLACS, the numerous faculty members on campus that focus on Central America and Mexico, and the continued interest on the period documented in the broader research community. Ostensibly isolated in a special collections environment primarily dedicated to local Indiana history, the re-digitization of the Archivo Mesoamericano video portends a more central place at the university and its dissemination through teaching, exhibitions and research. Furthermore, with newly sparked connections between the Digital Collections Services Department and the Institute for Digital Arts and Humanities at the university, greater opportunities for the dynamic purposing of the collection in the research environment through digital humanities projects are created. But moreover, the renewal of relationships with previous institutional partners and inauguration of new collaborations with kindred organizations and repositories in Central America and Mexico has the potential to create rich opportunities for the sharing of resources and knowledge regarding the preservation and stewardship of audio-visual content, and to create new standards and venues for the featuring of the region's deep contributions to the histories of activism and social justice.

Digitizing a Human Rights Archive in Guatemala: Data Curation, Access, and Social Justice

Alex Galarza

In 1984, the friends and family of Guatemalan activists who were 'disappeared' by state security forces formed the Grupo de Apoyo Mutuo (GAM). Members searched for loved ones during a period of Cold War violence in which Guatemala's military and police routinely murdered anyone they considered subversive and waged a genocidal campaign against indigenous Guatemalans in the countryside. The Guatemalan Truth Commission's 1999 report documented over 200,000 deaths during internal conflict (1960-1996) and attributed 93% of human rights violations to state security and related paramilitary forces. The GAM has spent the past three decades collecting textual, visual, and audio-materials related to ongoing human rights trials and historical memory.

My talk describes the GAM Digital Archive Project, an ongoing collaboration between the GAM and the Digital Scholarship (DS) team at Haverford College to create a digital archive. The project follows a post-custodial model in which materials will remain in Guatemala with the Haverford Libraries providing support in the digitizing, hosting, and creating online resources to share the collection. I describe how we have built a sustainable partnership between the GAM and Haverford DS team by focusing on collaboratively developing an ethical digitization and descriptive workflow. I describe the risks and stakes inherent in digitizing sensitive materials and envisioning a public platform designed to shift historical memory about state violence and impact the legal efforts to seek justice for human rights violations. I also focus on the ways the project has incorporated students into the work of building the digital archive and conducting original research on archival materials as a way of ensuring a sustainable partnership and modeling digital scholarship that can emerge from these documents.

Teaching beyond Haitian Exceptionalism: Digital Decolonization and Social Justice Pedagogy in Caribbean Studies

Crystal Andrea Felima

The University of Florida offers a Haitian Studies program for students to further their interests in language and culture in Haiti. In one course, "Haitian Culture and Society," students can explore and learn about alternative narratives of Haiti. A singular narrative of Haiti assumes all Haitians are poverty-stricken and passive to the structures and powers that influence people's life course. Therefore, it is essential to engage counter-narratives that illustrate Haitians as active agents in social change and development. The course aims to impart knowledge and participate in critical discussions on Haiti's history, culture, and society while examining the complexity of the country's political instability and economic under-development. Also, the class requires students to review topics of the State, neoliberalism, development, gender, class, culture, religion, disasters, and public health. Structured as a digital humanities course, students use digital tools to create a final project that centers on socio-cultural life, human agency, and self-determination in Haiti. As result, questions of exclusion and inclusion, along with silencing and advocacy, via media technologies become critical inquiries that guide the course.

In this paper, I will share reflections on the challenges and opportunities presented by teaching this Haitian Studies course as a digital decolonizing project. As a scholar and teacher of Caribbean Studies and Anthropology, my pedagogy incorporates a social justice framework to teach beyond Haitian exceptionalism. Defined as the perception or condition that something is exceptional or unique, exceptionalism continues to shape, produce, and reify singular narratives and specific interpretations of Haitian people and culture. Furthermore, the construction of Haitian exceptionalism has reproduced the idea that suffering has a totalizing effect on Haitian lives. Therefore, I will discuss my teaching methodology in how I encouraged students to go beyond their current understanding of what they may know of Haiti. Through digital

scholarship and a social justice framework, I highlight how students structured their digital projects to showcase their understanding of Haiti and how that knowledge can be applied to other populations around the world. I argue that learning, conducting research, and presenting findings in digital humanities, data curation, and e-scholarship offers critical engagement to decolonization and social engagement.

References

Duff, W., A. Flinn, K. Suurtammn, and D. Wallace. (2013). "Social Justice Impact of Archives: A Preliminary Investigation." *Archival Science* 13 (4): 317-48.

Pérez, E. (1999). *The Decolonial Imaginary: Writing Chicanas into History*. (Theories of Representation and Difference). Bloomington: Indiana University Press.

Sandoval, C. (2000). *Methodology of the Oppressed*. Minneapolis, MN: Univ Of Minnesota Press.

Digital Humanities in Middle and High School: Case Studies and Pedagogical Approaches

Alexander Gil
colibri.alex@gmail.com
Columbia University Libraries, United States of America

Roopika Risam
rrisam@salemstate.edu
Salem State University, United States of America

Stan Golanka
sgolanka@trevor.org
Trevor Day School, United States of America

Nina Rosenblatt
nrosenblatt@trevor.org
Trevor Day School, United States of America

David Thomas
dthomas@trevor.org
Trevor Day School, United States of America

Matt Applegate
mapplega@gmail.com
Molloy College, United States of America

James Cohen
jcohen@molloy.edu
Molloy College, United States of America

Eric Rettberg
eric.j.rettberg@gmail.com
Illinois Mathematics and Science Academy, United States of America

Schuyler Esprit
schuyleresprit@gmail.com
Create Caribbean, Inc., Dominica

Overview

While scholarship on pedagogy in digital humanities has been growing, its focus has largely been on graduate and, to a lesser extent, undergraduate education. Yet, digital humanities pedagogy—namely its value for cultivating 21st century literacies tied to the production of knowledge and the ability to interpret digital media and computation—is as valuable, this panel argues, for middle- and high- school students as it is in higher education. Given that we are pursuing what Matthew Kirschenbaum describes as forms of "scholarship and pedagogy that are bound up with infrastructure in ways that are deeper and more explicit than we are generally accustomed to" (60), this panel examines the work of instructors who are beginning to plant the seeds of these new "customs" early on in humanities and social science training.

Using digital humanities pedagogy in the middle- and high-school classroom, panelists argue, can redress gaps in these literacies. It enables students, as Mark Sample suggests, "[to think] through their engagement with seemingly incongruous materials, developing a critical thinking practice about process and product" (405). In this way, the approaches to digital humanities pedagogy in middle and high schools articulated by panelists are not an attempt to teach students particular technical skills, applications, or platforms. Rather, this pedagogical approach enables students to envision a relationship between themselves and knowledge production.

The approaches to digital humanities voiced in this panel are rooted in digital humanities pedagogies in higher education, particularly project-based approaches to humanities knowledge that foster collaboration. As Tanya Clement has argued:

> Like pedagogy intended to teach students to read more critically, project-based learning in digital humanities demonstrates that when students learn how to study digital media, they are learning how to study knowledge production as it is represented in symbolic constructs that circulate within information systems that are themselves a form of knowledge production. (366)

In the case studies and pedagogical approaches discussed by panelists, the project form complements more traditional forms of knowledge production and evaluation in the classroom. As Brett D. Hirsch argues, this "introduces a new mode of work that emphasizes collectivity and collaboration in the pursuit and creation of new knowledge" (16). While these new modes can be linked to participatory forms of culture, made possible by low barriers for civic engagement and creative expression online (Jenkins et al. 9), panelists make the case for greater

attention to pedagogies that offer instruction to middle- and high-school students in collaborative production.

However, as panelists argue, middle- and high-school pedagogies for digital humanities require attention to the unique needs of students in curricula, the developmental trajectories of the students, and the socio-economic dimensions of these students' lives. In light of these concerns, what are the biggest challenges to doing digital humanities in middle and high schools? Which methods are most valuable and practically achievable? And how can we effectively prepare teachers to incorporate digital humanities into their teaching practices? In this panel we bring together an international team of researchers and faculty already engaged in answering these questions and implementing curricula in schools of education, digital humanities centers, and high schools. Our goal is both to present models and facilitate discussion with broader digital humanities communities about pedagogical infrastructures, methods, long-term goals, and the exciting possibility of cultivating digital humanities pipelines through intervention in middle and high schools.

Panel moderator: Alex Gil, Columbia University Libraries

Designing Digital Humanities Pedagogy Infrastructures for Teachers

Roopika Risam

While digital humanities pedagogy has increasingly received attention from practitioners who want to teach their own students more effectively, how do we prepare *teachers* for the challenging task of engaging with digital humanities in their own classrooms? This talk offers an answer to this question by examining the digital humanities pedagogy infrastructure for middle- and high-school teachers designed at Salem State University. I first discuss findings from a study undertaken with teachers in Massachusetts to identify their attitudes towards digital humanities. The results indicate lack of knowledge about digital humanities but significant interest in incorporating computational approaches to humanities into teaching. Teachers also raised concerns including the time needed to learn technologies and teach them to students, cost of software and hardware, uneven access to computers or the internet in classrooms and for students at home, fear of implementing unsuccessful lessons, and a lack of professional development opportunities for digital humanities.

This talk then considers the interdisciplinary graduate certificate in digital studies that Salem State University designed in response to the study. The program provides professional development while addressing teachers' perceived obstacles to including digital humanities in their teaching. I discuss the relationship between study results and program design, focusing on development of core courses, selection of elective courses, differentiation of course delivery methods, integration into existing master's programs, and creation of a directed study for curriculum design. To illustrate the impact of the program, I describe my work advising a team of teachers and administrators in the graduate certificate program who were planning technology needs for a new school building under construction and designing technology-infused curricula in English and History. While core and elective courses gave the teachers and administrators a solid background in digital humanities, a group directed study assisted the team with developing a scaffolded curriculum across middle-school humanities courses, designing classroom technology, and creating a professional learning community to provide in-school pedagogical support for teachers.

Finally, this talk discusses a follow-up study with graduates of the certificate programs that assessed program outcomes. These outcomes include assignments implemented by teachers in their classrooms, exemplar student work, and a marked difference in attitudes and perceptions of teachers who completed the certificate in comparison to those who participated in the initial study. Based on the outcomes and the success of the graduate certificate program, Salem State has begun integrating digital humanities pedagogy directly into its teacher training programs. Consequently, this talk argues, this digital humanities pedagogical infrastructure for teachers serves as an effective model for addressing the barriers to incorporating digital humanities into middle- and high-school curricula for teachers who are already in the classroom and those preparing for teaching careers in the humanities.

Digital Inquiry: The History of Youth

Nina Rosenblatt
David Thomas
Stan Golanka

On September 12th, 2017 Trevor Day School, an Independent School on the Upper East Side of New York City, launched two sections of an advanced history course entitled *Digital Inquiry: The History of Youth*. This course was the culmination of seven years of curriculum development work that began with a November 16th, 2010 article in the *New York Times* about Humanities 2.0 and the Stanford Republic of Letters Project. After an initial round of research we came to understand that digital projects had a role to play in our High School History curriculum. This realization coincided with our adoption of inquiry-based learning pedagogies. In a fundamental way, we argue, the techniques and disciplines involved in digital humanities allow high school students to conduct their own independent research in digital archives and become producers of history in their own right.

In order to motivate students to collaborate and learn unfamiliar working methods, we developed our course around a subject that would engage all students. We wanted a subject that would not require a textbook, was accessible to juniors and seniors in high school, and would lend itself to seminar style classes. In addition, we wanted to be able to supplement the subject matter with texts illuminating the nature of historical narrative, archives, and the use of digital techniques in academic research such as the paper by Lauren Klein, "The Image of Absence: Archival Silence, Data Visualization, and James Hemings" in *American Literature* Volume 85, Number 4, December 2013. The resulting course delved into the history of youth, looking at how being young is experienced and imagined differently in different times and places, and what we can learn about a society from its expectations for and attitude towards its youth, while teaching them production and analysis techniques for them to create new representations of that history.

The final consideration was to craft a series of lesson plans to embed a digital humanities knowledge-production laboratory in the class. The course lab was divided into three modules: Digital editions and markup (an introduction to the fundamentals of plain text and markup), digital collections/exhibits (an introduction to the fundamentals of metadata and databases), and cultural analytics (an introduction to the fundamentals of algorithmic thinking and data mining). Through these modules the students were immersed in the process of selection, digitization, mark-up, the creation of a database/archive, data extraction and cleanup and data analysis, all driven by the imperative to create and interpret history. Technologies taught included, but were not limited to, command line, git, GitHub, plain text editors, Markdown, YAML, Jekyll, Omeka, Python and Voyant Tools. These technologies were directly tied to the variety of ways in which historians collect "data" including using literary, psychological, sociological, statistical, and visual sources, working towards creating our own historical knowledge using the digital tools for collecting, visualizing, mapping, and analyzing the information.

In this panel we will present the results of our two course prototypes, lessons learned, future improvements, and argue for a generalizable model of instruction for high schools in the United States based on our experiences.

Digital Literary Studies in the High School Environment

Eric Rettberg

What are the challenges of adapting a course in Digital Humanities and Digital Culture from the pedagogical environment of the university to that of the high school classroom? What new challenges arise from asking minors to produce digital and public scholarship, and how can digitally inflected scholars and teachers foster innovative humanities work in school environments bound to pre-existing curricula? In this talk, I use my experience adapting a class in Digital Literary Studies to the high-school level to share unexpected challenges and opportunities and to suggest digital work as a strategy for promoting the humanities to administrators, peers, and students in STEM-oriented high school environments.

In early 2016, I left higher education to teach in the English department of the Illinois Mathematics and Science Academy, a state-funded boarding school for students talented in math and science. Given the immediate appeal of classes combining humanities with computing for STEM-focused students, I naively expected that I might be able to simply bring a college elective for English majors to my high school students. The actual challenges of doing so, however have been instructive: administrators have been less familiar with the existence of the methods of the Digital Humanities, digital assignments have had to be reframed to accommodate shared practice among teachers in my department, my school's technology environment has needed to be customized to accommodate the software installations that I took for granted before, oversight from administrators, colleagues, and parents has been more intensive, and without the support staff available to me at my higher education institutions, I've had to think creatively around constraints. By demonstrating small-scale digital humanities work in core classes, designing a week-long intersession class on a similar topic, and sharing my knowledge of University-level digital humanities, though, I've been able to design a class that has colleagues and students excited.

Heeding Ryan Cordell's call to embed digital humanities instruction in larger narratives beyond "recent scholarly revolution," I treat digital humanities praxis as one of three major components of change in literary production and study in the digital era. In addition to digital humanities projects centered on historical texts of students' choice, students read and discuss fictive works that represent cultures of technology in the digital era and computationally enabled works of electronic literature. Throughout the class, students sample digital humanities practice in lab sessions and build small-scale web resources and undertake digital-humanities experiments in group projects. By exploring electronic texts, they begin to more fully recognizes the affordances of digital technologies, and by reading print texts that represent digital culture, they think about their own roles as consumers of and creators of digital tools and cultures. While my school's student population and focus are especially suited to the STEAM focus that a class like this one offers, my experiences suggest that students at a wide variety of high schools would be engaged by these materials and skills.

Impact od Digital Humanities on High School History and Heritage Teaching and Learning in the Caribbean

Schuyler K Esprit

The experience of Create Caribbean Research Institute, the first Digital Humanities center in the English Speaking Caribbean tells an interesting story of how digital humanities can covertly and explicitly reshape the curriculum in history and literature of the Caribbean without necessarily requiring a massive paradigm shift of the national and regional curriculum requirements.

In Dominica (where Create Caribbean operates) and the Eastern Caribbean – among other islands – the secondary education curriculum responds to the mandates of the Caribbean Examinations Council (CXC) who sets the CSEC and CAPE syllabi for high school and post-secondary certification in the region. These examinations frame the education curriculum for the five to six years of high school in many islands and many educators in this system find themselves bound to deliver content in limiting and limited methods in order to ensure that students simply meet requirements to excel at subject exams at Caribbean History and English B (Literature), which has a heavy focus on Caribbean Literature.

However, students leave with an abstract and formalized understand of Caribbean history and culture, without a nuanced understanding of its relevance to their own lived experiences and the implications for their future. Create Caribbean uses digital humanities projects to reframe the conversation and disrupt traditional methods for learning. One of these projects uniquely highlights the potential for technology to change the face of education in Dominica and to get students more invested in Dominica's history and culture. This project, made by students for students, can be found at www.dominicahistory.org. The college student change-makers of Create Caribbean's internship program build digital humanities projects with a primary and secondary student audience in mind. The example of dominicahistory.org highlights one way that a collaboration with a national organization has allowed for a broader consideration of the methods of heritage and culture education for students while actually providing solid academic source material for their formal study requirements.

This presentation will discuss the origin, process and impacts of the Dominica History and Imagined Homeland digital projects of Create Caribbean as examples of disruptive secondary education. The presentation will also address the ways in which the projects have attracted the attention of high school teachers and transformed their interests in using technology to revise classroom experiences when they face limitations in adjusting other curricular frameworks.

Precarity and Practicality: DH, New Media, & Secondary Education
Matt Appegate

James Cohen

In 2015, faculty at Molloy College in Long Island worked with faculty and administrators to found the Baldwin High School New Media Academy, a co-organized effort to bring the study of New Media and Digital Humanities to underserved high school populations in Baldwin, New York. Working collectively, faculty at both institutions have established a curriculum and internship path at Baldwin High School that exposes students to methods of DH praxis and principles of New Media in a variety of means and environments (high school, college, in-person, online).

Our curriculum is based on five modules and two college-credit bearing courses. Our modules include Critical Making, Digital Storytelling, Multimodal Composition, Online Expression, and Social Media. Our college-credit bearing courses are Introduction to New Media and College Composition (the course is taught entirely on the methods of multimodal composition). Each module is integrated into existing high school courses, i.e., Social Studies, English, Wood Shop, etc., where students take college-credit bearing courses in their junior and senior years. Ultimately, the "academy" concept introduces students to DH methods and New Media in a gradated process--students choose their academy prior to entering their freshman year of high school and are enrolled in courses that employ our modules.

Our curriculum is based on principles of social good; it emphasizes both civic engagement and social justice, and provides sample assignments with grading rubrics for each module (Ratto). The civic-minded focus of our curriculum was developed in consultation with Baldwin High School, and fleshed out over 18 months of training. Our curriculum attempts to account for the precarious position women and people of color already inhabited in online spaces and demonstrate how DH methods and New Media principles can be mobilized to empower students via digital tools and languages.

The focus of this paper is to report on our work with underserved high school populations and relay the challenges of bringing this kind of material to a secondary education setting. We focus on the practicalities of bringing DH methods and New Media principles to high school (i.e., funding, time, expertise, bureaucracy), as well as the necessary training that takes places between high school and college faculty (PT days, on campus conferences, and student events). Finally, we discuss the opportunities that working with underserved high school populations provides both politically and pedagogically. In this context, DH operates on a minimal scale, but addresses communal needs.

Bios

Matt Applegate is an Assistant Professor of English and Digital Humanities at Molloy College. His work focuses on critical theory, digital humanities, digital literacy, and screen studies. His work as appeared in *Amodern*, *Theory & Event*, *Cultural Politics*, *Cultural Critique*, *Telos*, and more.

James Cohen is the director, co-founder and assistant professor of the New Media program at Molloy College in New York. Jamie is the author of *Producing New and Digital Media: Your Guide to Savvy Use of the Web* (Routledge 2015) and his published and presented research focuses on memes, YouTubers, populism, VR/AR/MR, and digital media literacy. He is a fellow of the Salzburg Academy on Media and Global Change and the Academy of Television Arts and Sciences.

Schuyler K Esprit is the Director of Create Caribbean Research Institute at Dominica State College, the first Digital Humanities center in the Caribbean. Dr. Esprit holds a PhD in English literature from University of Maryland – College Park. She is a scholar of Caribbean literature and cultural studies, and postcolonial theory. She is now completing her book entitled *West Indian Readers: A Social History* and its digital companion, both of which are historical explorations of reading culture in the Caribbean. She is currently Dean of Academic Affairs at Dominica State College.

Stan Golanka is Director of Academic Technology at Trevor Day School. He teaches computer programming and co-teaches Advanced History: Digital Inquiry. He holds a MA in Computing in Education from Teachers College at Columbia University.

Eric Rettberg teaches English at the Illinois Mathematics and Science Academy. He remains an active scholar of modernism, experimental poetry, sound studies, and the digital humanities. His work has appeared in *Comparative Literature Studies* and *Jacket 2*.

Roopika Risam is an assistant professor of English and English education at Salem State University. Her research considers the intersections of postcolonial cultures, African diaspora studies, and digital humanities. She is the author of *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Northwestern UP 2018) and her work has recently appeared in *Debates in the Digital Humanities 2016*, *Digital Scholarship in the Humanities*, and *South Asian Review.*

Nina Rosenblatt teaches US History, Art History, and Advanced History: Digital Inquiry at Trevor Day School. She holds a PhD in Art History from Columbia University.

David Thomas is Chair of the History Department at Trevor Day School, he teaches European History, Advanced European History, The History of China, and Advanced History: Digital Inquiry.

References

Clement, Tanya. "Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind," in *Digital Humanities Pedagogy: Practices, Principles, and Politics*, ed. Brett D. Hirsch (Cambridge: Open Book Publishers, 2012), 365-88.

Cordell, Ryan. "How Not to Teach Digital Humanities." *Ryancordell.org*, 1 Feb. 2015, http://ryancordell.org/teaching/how-not-to-teach-digital-humanities/.

Hirsch, Brett D. "</Parentheses>: Digital Humanities and the Place of Pedagogy." *Digital Humanities Pedagogy: Practices, Principles, and Politics,* ed. Brett D. Hirsch (Cambridge: Open Book Publishers, 2012), 3-30.

Jenkins, Henry and Ravi Purushotma, Margaret Weigel, Katie Clinton, Alice J. Robison. *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century* (Cambridge, MA: MIT Press, 2009).

Kirschenbaum, Matthew. "What Is Digital Humanities and What's It Doing in English Departments?" *ADE Bulletin* 150 (2010): 55-61.

Ratto, Matt. "OPEN DESIGN NOW." *Open Design Now*, Netherlands Institute for Design and Fashion and Waag Society, opendesignnow.org/index.html%3Fp=434.html.

Sample, Mark. "What's Wrong with Writing Essays?" *Debates in the Digital Humanities*, ed. Matthew K. Gold (Minneapolis: University of Minnesota Press, 2012), 404-5.

Remediating Machistán: Bridging Espacios Queer in Culturas Digitales, or Puentes over Troubled Waters

Carina Emilia Guzman
cartogeosapiens@gmail.com
University of Toronto, Canada

T.L. Cowan
tl.cowan@utoronto.ca
University of Toronto, Canada

Jasmine Rault
jas.rault@utoronto.ca
University of Toronto, Canada

Itzayana Gutierrez
itzayana.gutierrez@gmail.com
McGill University, Canada

Introduction

This panel consists of two papers in Spanish and two in English. They span a varied collection of subjects which foreground the remediation that bridging entails; bridging

spaces and architectures, bridging epistemologies and cultures, bridging the digital-analog divide, bridging into the imaginary and beyond. Remediation, these papers will address directly and indirectly, is a trouble: it implies the interpretation and translation of knowledge and creation. Who remediates and under what terms? How can we think of remediation as a bridge that invites us establish series of ethical considerations when approaching the sources and objects we will inevitably remediate?

In the first Spanish-language paper, author Carina Guzmán gestures towards a framework capable of bridging methods for the GIS-based study of a map/codex purported to have been generated within the highest realms of Aztec imperial administration: the Aztec Tribute Scroll, and its interpretations in 20th century historiography, on the one hand, and on the other, auto-ethnographic research in queer world-making in the 21st century Mexico City nightscape. Some of the bridging concepts this paper discusses are derived from Cultural and Radical Geography, such as space as a cultural product and the (economic) politics of territorialisation. Further bridging concepts explored come from the realm of Information and Archival theory, these include questions how space and memory function as archives, and how/what digital tools can contribute to developing an epistemology for their study.

In the second paper, in English, Jasmine Rault traces the conceptual and ideological bridge between twentieth century modern architectures and digital social design. It considers what queer architectures -- from Eileen Gray's domestic designs to Mexico City's *vecindades* -- can teach us about reimagining and redesigning the modern colonial logics of digital spaces, cultures and socialities. How might we remediate the decolonial, alter-modern potentialities of queer architecture for ethical digital social design? What futurities of decolonial digital cultures, ethics and spaces might be made possible by designing after these queer architectural histories?

T.L. Cowan proposes, in third paper, also in English, that studying genealogies of discourses of hygiene (cleanliness, health, safety) offers us an important lesson when we find ourselves in the situation in which "hygiene" is remediated to the digital realm. Bridging long-standing critiques of "hygiene" with contemporary digital culture, this paper presents a queer, decolonial reading of hygiene and contamination, re-thinking the ways that we recycle/repurpose/remediate/resuscitate these discourse and values in our digitally-mediated lives and offering practices of shared risk and mutual care as alternatives to the hygenic model.

The final paper, in Spanish, by Itzayana Gutiérrez, works with the remediation of a printed comic magazine into four online and for free download videogames. The comparison will be articulated considering the bridges between two media, the analog comic magazine and the videogame; and the bridges between two genres, the orientalist comic adventure and the war video game. It is important to mention that there are narrative as well as visual vocabulary continuities in this dual system, and both the comic and the videogame play with hyper-hygienic homoerotic scenarios to sanitize, purify and militarize interracial affects and national imaginaries.

El espacio y el archivo como puente: una propuesta epistemológica

Carina Emilia Guzmán

En esta ponencia abordaré de dos temas aparentemente dispares: un estudio sobre la Matrícula de tributos que realicé a través de un sistema de información geográfico (el software libre QGIS), y un proyecto de investigación basado en mi experiencia personal como organizadora de fiestas lésbicas en la Ciudad de México desde un espacio mágico: Machistán. Presento estos dos estudios académicos en planos paralelos y trazaré entre ellos un puente con dos bases. La primera conformada por en una serie de conceptos subyacentes compartidos: que el espacio es una creación cultural, que los lugares son producto de procesos políticos, y que –por ende- aquellos metros cuadrados de superficie terrestre que canónicamente solemos llamar "palimpsesto", están sujetos a procesos políticos y económicos de territorialización con sus respectivas disputas e identificaciones. La segunda base consiste en establecer la necesidad metodológica de valorar y procurar una epistemología planteada por el objeto de estudio no tradicional en sí; para tratar estos casos me acercaré a la epistemología de la Información (misma que subyace al concepto de Humanidades digitales y compartida con las demás ponencias de este panel) que considera que algunos objetos, como lo podrían ser un códice mesoamericano o una fiesta lésbica, han sido infravalorados por el quehacer académico al no existir un esfuerzo por conceptualizarlos y por lo tanto de archivarlos e investigarlos en sus propios términos.

La Matrícula de tributos es un texto de tradición mesoamericana que (se asume) reúne información sobre el tributo que se pagaba a la Triple Alianza al momento de la Conquista. Hacia 1949 Barlow localizó los poblados tributarios enlistados en el códice, realizando así la primera proyección cartográfica de lo que él llamó el imperio colhua-mexica. En 1996, Berdan et al. elaboraron una contrapropuesta cartográfica que pretendió dar cuenta de una diferenciación política entre territorios tributarios y estratégicos del imperio ahora llamado azteca. Aunque distintos, ambos mapas comparten la misma exigencia: ilustrar el aspecto de un imperio. En esta ponencia explico mis intentos por tender un puente hacia el conocimiento contenido en la Matrícula de tributos al elaborar un mapa (con la herramienta digital QGIS) cuya pretensión es el análisis de la información geográfica contenida en el códice sin que necesariamente rinda cuentas sobre

una dinámica fiscal, ni que represente el aspecto cartográfico de un territorio imperial.

En otro plano, Machistán es el universo imaginado de las "machas": así nos llamábamos entre nosotras dentro de una pequeña colectiva de lesbianas y mujeres queer a la que pertenecí hacia el año 2005. Nuestro proyecto político era la organización de fiestas que fueran una alternativa a los bares y antros gays y su hostilidad hacia las mujeres: la misión era crear el deleite lúdico colectivo que sólo la noche ofrece *al margen de la vida nocturna comercial* de la Ciudad de México. En esta ponencia explicaré que Machistán es una fuente en sí: un archivo o sistema de información. A través de este espacio mágico se puede entender la creación de conocimiento; la producción cultural; el trabajo (digital y análogo) y los afectos que se llevan a cabo sobre azoteas y bajo lonas para que decenas de queers se reúnan con el fin compartir un espacio/instante.

Así pues, propondré comprender la dimensión epistemológica de la información de estos objetos de estudio, y así entenderlos como archivos; de diseñar una metodología y quizás un sistema de información capaz de manejar el conocimiento y la creación presentes en estas fuentes, y que permita entenderlos e investigarlos críticamente

Decolonizing Architectures of Digital Modernity: Queer Futurity from Sapphic Modernist Design to Digital Vecindades

Jasmine Rault

What can early twentieth century modernist architecture and design teach us about contemporary decolonizing, feminist, queer and anti-racist practices and protocols for networked digital architecture? European modernist architecture was driven by ideals of internationally standardized open communication – the open floor plan, strip windows and glass walls, unimpeded visual access to interior and exterior space, eliminating walls and structural as well as decorative or symbolic obstructions to complete open access. There are some striking ideological continuities between these modernist architectural ideals (and aesthetics) and contemporary Euro-Atlantic values of unbridled digitization, designing global information networks for unobstructed open access – and these continuities need to be understood within a context of modern-colonial regimes of gender, sexuality and race. In my collaboration with T.L. Cowan to build the Cabaret Commons, have found ourselves taken, like so many researchers recently, by the fever to (digitally) archive precarious, precious, minoritized, invisibilized, intimate, forgotten trans- feminist and queer knowledges, scenes, resistance cultures and alternative futures. Bound by their beauty we are also, however, bound by the institutional and platform logics that we hope these archives can transform, and how we can be accountable to the "the people whose belongings have become [our] 'collections'" (Nowviski 2016). Although generally well- intentioned, conventionalised Western epistemological research practices reflect longstanding and ongoing acquisitional, abductive, possessive, extractive practices that prioritize the needs and values of the research community over the needs and values of the *researched* community.

If, as I argue, contemporary digital spaces, cultures and research practices have been designed by modern colonial discourses, politics and aesthetics of early twentieth century European architecture we might find generative decolonizing digital models, ethics and aesthetics in the queer architectures that have been designed against, or sometimes simply askew to, the modern. My presentation will take up two main types of queer architecture: the early twentieth century Sapphic modernist designs of women working with and deliberately against European modernist architectural ideals (Rault 2011); and the Mexican architectural innovation of the *vecindad*. In Spanish, the word *vecindad* can mean neighbourhood, but in Mexico City it typically describes a style of architecture: "The *vecindad* is usually known as a row of one room dwellings surrounding an open space or patio" (Alejandro M. Rebolledo 1998, 1). *Vecindades* were the main form of working-class architecture in Mexico City from around the sixteenth century until the mid-twentieth century, incorporate pre-colonial and Spanish colonial design and were more often built *by* than for their inhabitants. By the 1940s, they were cast as examples of unhygienic, degenerate, backward dwelling design that (largely European) modernist architecture should replace, and provide a compelling alternative to the gender, sexual, racial and class ideology of openness and transparency structuring architectural and digital modernity. What futurities of decolonial digital cultures, ethics and spaces might be made possible by designing after these queer architectural histories?

Dirty Queers, a love letter: Vulnerability, Care, Contamination, Migration, and Discourses of Digital Hygiene

T.L. Cowan

In this paper I propose that our current moment of technocultural embodiment offers us the opportunity to observe the long history of a particular configuration of power: namely, discourses of digital hygiene that fail to account for queer, feminist, decolonial and anti- racist critiques of 'hygiene' as a mode of social control. I argue that "digital hygiene" is promoted as a monitoring practice that is at once meant to keep you *and* your computer clean *and* safe—self-protection against both malware viruses and human trolls. Importantly, digital hygiene is also framed as responsible digital practices that keep

others safe from infection *by* you; if you have a dirty digital existence, you put yourself *and others* at risk. Rather than worrying about whether we are clean or dirty, I propose we worry about how our understandings of the conditions of cleanliness/purity and dirtiness/contamination have been shaped and how they continue to determine our digital existences.

Here, I trace the ways that hygiene protocols have been enacted as oppressive regimes through colonial, racist, classist, misogynist and homophobic surveillance and regulation against immigrant populations; women and other feminized populations; and Indigenous and Black youth and other negatively racialized populations. In our social imaginaries informed by the logics of digital hygiene, if we fail to practice all of the enforced/'recommended' forms of self-protection/inoculation/prophylaxis and we get 'infected' by viruses or attacked/infiltrated by identity-thieving or rape-and-death- threatening trolls, we are held at least partially at fault, due to our risky behaviour. As an alternative to discourses of digital hygiene, I focus here on queer theories and practices of mutual care, shared risk and activist methods like those practices by the Center for Solutions to Online Violence (http://femtechnet.org/csov/), which offers users ways to practice "safer" digital practices, while also--following Wendy Hui Kyong Chun and others-- taking an activist approach in which we "fight for a space in which one can be vulnerable and not attacked" (2016, 158). At once a history of hygiene, and a love letter to the "dirty" queer, migrant and other marginalized practices of shared risk and mutual care, this paper considers how a shift away from hygiene and towards other imaginaries may transform our digital lives.

Kalimán: el puente del lenguaje gráfico entre los comics y los videojuegos

Itzayana Gutiérrez

Durante el último tercio de 2016, distintos medios noticiosos dieron cobertura al emocionante lanzamiento de una serie de 4 videojuego basado en el comic mexicano más exitoso de todos los tiempos: *Kalimán el hombre increíble*.

Esta no es la primera vez que Kalimán cruza el puente entre un medio y otro. Sus aventuras comenzaron como radionovelas en la década de los 60 pero explotaron el mercado cultural como historias gráficas impresas en el mundo de las revistas de comic baratas. Su circulación impresa fue amplia y sostenida desde mediados de 1960 hasta principios de 1990 en México pero alcanzó importantes números de distribución en Sudamérica. Este éxito intentó ser recapitalizado en películas, novelas, readaptaciones gráficas digitales y ahora en esta serie de videojuegos.

Mi presentación en este panel de humanidades digitales hará un seguimiento de las actualizaciones y agregados del lenguaje gráfico de Kalimán a través de su traducción y rediseño para videojuegos. Desde mi punto de vista, Kalimán es una aventura veladamente homoerótica, codificadamente racista y abiertamente orientalista. Y este tono establecido en el mundo gráfico sólo se intensifica con un giro más belicista aún.

Kalimán eviste turbante blanco y mallas blancas que evocan modas del subcontinente Indio aunque sus frecuentes apariciones estelares con el torso desnudo hacen que su traje sea lo de menos (1). Su cuerpo musculoso y bello se completa con ojos azules y detalles faciales que acentúan ciertas convenciones de belleza masculina asociadas con conductas militares y de ocupación colonial. Kalimán tiene además de estas características utilizadas ampliamente en la gramática de otros superhéroes, poderes mentales y espirituales incomparables que le otorgan autoridad moral para el ejercicio libre de su violencia. Como en los lenguajes gráficos de vanguardia utilizados ampliamente por los fascismos y la propaganda de guerra, su cuerpo se beneficia de una autoridad seductora y vibraciones homoeróticas.

En este universo gráfico de convenciones raciales, orientalistas, estrategias misóginas y homoeróticas, mi presentación intentará aclarar las innovaciones y actualizaciones del lanzamiento de los videojuegos. Según el anuncio de su desarrollador Eduardo Guerrero, el producto tiene 4 ofertas.

Las 3 primeras están desarrolladas para dispositivos móviles y en un esquema de descarga gratuita. La primera, según dice Guerrero "se dedica a trabajar toda la parte mental" y recuerda esquemas lógicos de Candy Crush y otros sistemas de acumulación de puntos a través de niveles progresivos de dificultad. El segundo se basa en las historietas de Kalimán, en lo que parece ser un esquema de 8 bits similar a Zelda. El tercero usa el legendario modelo de enfrentamientos de Street Fighter. El cuarto videojuego, más ambicioso y complejo, por el que hay que pagar, está desarrollado para consolas y modelos inversivos de navegación en donde se presenta a "Kalimán evolucionado", quien se engancha en tareas de batalla en ambientes del Medio Este utilizando una variedad de tecnologías de observación, desplazamiento y ataque militares ("Kalimán ahora en videojuego" Entrevista con el realizador Eduardo Guerrero para NOTIMEX, https://youtu.be/ZaF31NUqmUg).

References

Banta, M. (2003). *Barbaric Intercourse: Caricature and the Culture of Conduct 1841-1936*. Chicago and London: University of Chicago Press.

Barlow, R. (1949). T*he Extent of the Empire of the Culhua-Mexica*. Berkeley: University of California Press.

Bartra, A. Piel de papel. Los pepines en la educación sentimental del mexicano. En *Revista Latinoamericana de Estudios sobre la Historieta*, vol. 1.

Berdan, F. et al (1996). *Aztec Imperial Strategies*. Washington DC: Dumbarton Oaks Publishing Service.
Herner, I. (1979). *Mitos y monitos, historietas y fotonovelas en México*. México: Universidad Nacional Autónoma de México / Nueva Imagen.
Chun, W. H. K. (2016.) *Updating to Remain the Same: Habitual New Media*. Cambridge: The MIT Press.
Groensteen, T. (2009). *The System of Comics. Translated by Bart Beaty and Nick Nguyen*. Jackson: University Press of Mississippi.
Leslie, E. (2004). *Hollywood flatlands. Animation, critical theory and the avant-garde*. London / New York: Verso.

Beyond Image Search: Computer Vision in Western Art History

Leonardo Laurence Impett
leonardo.impett@epfl.ch
EPFL and Bibliotheca Hertziana, Max Planck Institute for Art History, Italy

Peter Bell
bell@uni-heidelberg.de
University of Erlangen-Nürnberg, Germany

Benoit Auguste Seguin
benoit.seguin@epfl.ch
EPFL, Switzerland

Bjorn Ommer
ommer@uni-heidelberg.de
University of Heidelberg, Germany

The Digital Humanities is largely a textual field. In part, this is a reflection of the supremacy of the written word in wider academic production; in part, of the history of the Digital Humanities as an interdiscipline in the study of textual sources, from poems to administrative records. As a community, we have become proficient in the creation of images - thinking and even programming through diagrams, visualising the results. The study of images, on the other hand, has had no such computational revival.

Many of these impasses are at least as much a product of technical difficulty as of intellectual habit or institutional inertia. Nelson Goodman (1968) noted that allographic arts follow kinds of notational systems - from poetry in the Japanese alphabets to modernist dance in Labanotation. The dynamics of such cultural phenomena may be far from the statistics of these notations, but these symbolic abbreviations give a first way in for the use of computational techniques in cultural history. The visual arts, on the whole, have no such notational projection. Nonetheless, recent advances in Computer Vision techniques (explored by last year's AVinDH SIG workshop) allow us to confront the anti-notational head on, fuelled by the genuinely 'big' data from mass cultural digitisation programs - PHAROS alone will soon publish 31 million images of Western art and architecture.

The question of symbols in the computational study of images opens up interesting questions with profound epistemological consequences. Should the question of the symbolic be avoided (through image search engines), or re-confronted (the invention of new notational systems)? Our panel aims to bridge a range of views on such matters, exhibiting radical work which is at once critically provocative and technically cutting-edge. Such dissonances echo historiographical debates over the role of the symbolic and the iconic.

Our panel attempts to examine such questions precisely in the contexts in which they are least accepted: in Medieval and Renaissance art history - a highly institutionalised and specialised discipline, compared to cinema and visual studies or Bildwissenschaft. Despite the high degree of technological involvement in digitisation, archiving and exhibition, art history has had almost no computational interventions in criticism - controversial and pioneering work in the computational analysis of Renaissance images by Robert Tavernor (1995) and Martin Kemp (Criminisi, Kemp and Zisserman 2005) are the exceptions that highlight the rule.

Much 'Digital Art History' concerns questions of linked archival metadata, digital publishing, or image digitisation; important questions, but which already have active DH communities. In an institutional sense, then, this panel implies a shift of the computational, from the world of GLAMs to those of university research departments.

The use of computational techniques in art history further opens up important questions in the blurring of relationships between art history, artistic criticism, artistic research and artistic production. Here again, the panel seeks to offer the whole horizon of intellectual opinions, none of which are to be merely dismissed as traditionalist: from the creation of ready-made software to be used as a tool by individual art historians, to algorithmic criticism presented as artistic practice in its own right.

Digital Gesturing in Early Renaissance Italy

Leonardo Laurence Impett

The gestures of early Renaissance Italian art are largely read through three lenses: iconography (Garnier 1982 or Barasch 1987), classical rhetoric (van Eck 2007), or universalist theories of expression (Freedberg 2007). This work focuses on a fourth view, complementary rather than contradictory: that of social life, of theatre and sermons, of dance and jesting, of insults and educational manuals.

The computer vision techniques I have developed over the past two years make it possible to automatical-

ly identify gesture from images of paintings and reliefs: from precise hand-shapes (e.g. the 'corni'), to body poses ('genuflection') and more subtle body language ('slouching'). We can visualise patterns, clusters and trends within 'gesture-space'. Such computations, through their blindness to non-gestural properties (gender, age, musculature, iconography), often propose novel, radical links that are both morphologically precise and visually estranged - leading to an epistemological understanding of this work from Brecht's Verfremdungseffekt and theory of Gestus.

The period of enquiry, 1300-1480, is chosen to bridge the gap between the two longue durée historical anthropologies of gesture: that of Jean-Claude Schmitt (1990) ending in c.1300, and of Peter Burke starting from c.1500 (1991). However, through such gesture-computational techniques, I give a primary role to the image, both as object of study and source of historical evidence. Such an approach is particularly appropriate to the gestural culture of the early Renaissance; as Dilwyn Knox (1991) has noted, this period is notable for its relative lack of universalist theories on gesture.

My digital approach uniquely allows for the inclusion of a great number of minor works. To study a gesture amongst tens of thousands of images is not just to operate on a different scale to what is possible 'by hand', but also to change the object of study: to include a critical mass of works for humbler patrons. My approach, however, is not a kind of 'distant reading' in a statistical sense; as I have argued in detail elsewhere, broad statistical statements on such collections are scientifically unjustified. Rather, I use my tools to identify groups of images containing specific gestures or poses: visualising them, curating a small linked database of gesture-images, and examining (including through primary texts) conventional aspects of context, purpose and connotation in detail.

In particular, I focus on gestural implications of the Trecento plagues, drawing on scholarship in its implications for general and medical thinking of the body. Specifically, I consider parallels between the gestures found in Trionfi della Morte and the extreme gestures described by Barasch (1976) in depictions of sorrow or apocalypse. The study of the gestures of the plague, especially in its considerations of post-Galenic / pre-Cartesian understandings of mental and physical health through humourism, will significantly inform more general questions of 'emotion' vs. 'character' in XIV-XV century gesturality, an opposition which I seek to problematise.

Finally, on a methodological point, I hope to demonstrate the possibility of computer-aided art history which isn't based on rigid universalist textual taxonomies (which I see as repeating early 1990s mistakes in 'Expert Systems' AI), neo-formalist computational iconography, or automaton connoisseurship.

Exploring Large Art Historical Photo Collections

Benoit Auguste Seguin
EPFL, Switzerland

In recent years, museums and institutions have spearheaded global open-access efforts leading to the digitization of many artworks (paintings, engravings, sketches, old photographs etc.). Leading that trend is the PHAROS Consortium, which includes the biggest photographic-archives in the world, aiming to place online in the upcoming years, roughly 31 million images. Through the IIIF standard (International Image Interoperability Framework), these images are now available online in an easily accessible manner, with each institution/museum opening its content with standard APIs. This brings an unprecedented opportunity for interactions and global search across collections.

Using the case of the digitized photo-collection of the Cini (330'000 elements), we will focus on two cases where Computer Vision is very beneficial: duplicate detection and pattern tracking. The first task is extremely beneficial to reconcile different collections and detect conflicting metadata, while the second one is more akin to a form of visual search. For both tasks, we found crucial however, to have a good interface to explore the collection and acquire training examples in order to solve both problems.

An important concern when designing our search interface was to bring as much freedom as possible to the researcher. In addition to a traditional textual query based on metadata, we have two types of visual query, which themselves can be combined with metadata filters. Additionally, two types of visualization allow to explore search results in different ways, and save connections between images.

A visual search system is tightly coupled with an underlying visual similarity. However, an iconographic similarity is not the same as the similarity of two paintings made by the same artist, but representing different subjects. The central question of what sort of results the researcher is looking should be part of every project involving visual search. Here, we propose to tackle it by having the users edit a graph of visual connections between the images, which allows the search system to train itself on what to retrieve from these very large databases. For example, such a system allowed us to be able to retrieve drawings or engravings from a picture of the original painting, or the reuse of a pattern by followers of an artist, all without looking at any single metadata entry.

Iconography, Pose Recognition and a Grammar of Gestures

Peter Bell

Description and iconography are first and main steps of an art historical analysis. Art history has forever created

taxonomies and encyclopedias for iconography. Therefore, many images of Western Art can be easily identified and categorized. In case of very common representations, like the prominent scenes of the gospels, this amounts to iconographic categorization - these are present as metadata in art historical image repositories, but remain too broad to represent the variations of motifs and historical change. Diachronic links (for example between medieval art and the 'Nazarener') are not necessarily connected.

To visualize these differences and connections, we follow a strictly form-oriented anthropocentric approach, especially focused on the positions, gestures and interactions of figures. State of the art algorithms are able to handle this anthropocentric approach by detecting the position of the main figures of an image truthfully. From about 60 illustrated gospel scenes of the life of Jesus, we chose two very prominent narratives as case studies for this approach: the annunciation and the baptism. Whereas the baptism is normally presented in one central moment, Michael Baxandall has shown that in the annunciation at least five different phases can be differed. Therefore we can show the variations in the composition and gestures of one central moment and the movements within a certain plot. Details like appearance, the use of different objects (i.e. shell or bowl for baptism), and the composition of the images can be analysed easily via this method. We do this comparison with about a thousand images for each prominent scene, using neural-network-based full body recognition algorithms and nonlinear dimensionality reduction.

The goal of the project is to analyse a huge corpus of Christian art, detecting gospel scenes from the life of Jesus, and to compare each single scene in its own tradition, within its plot and with every other scene. This method helps us to visualize what 'iconography' really consists of, and how it changes in time. It helps to reconstruct the inner relationship between motifs and narrative structures, and finally opens a close reading of gospels, apocrypha and other theologian writings (and concepts) which are the basis of the representations. Instead of being an image search, the aim of the approach is to visualize the similarities in a plot, so that structures and clusters can be interpreted from a distance and closely connected images can be analysed in detail. Finally, the approach helps to understand the use of nonverbal language in art and reconstruct the changes of the concepts of human body and interaction.

Understanding Art: Distant Viewing Meets Close Reading

Björn Ommer

Computer vision and object retrieval are well suited for understudied, large, and, until now, untagged image datasets containing repetitive or similar motifs. Visual elements can be directly retrieved by means of computer vision, enabling specific search functions and reducing the need for textual annotation of the digitized data, which can be a costly part of database projects. Especially in the case of large numbers of equal or similar subjects, computer-based algorithms can probe images and provide surveys and information via statistical visualizations that shed new light on the material. In addition, by assembling series of images based on similarities, computer-based analysis can facilitate attribution to an artist, a date or dependent art works.

Furthermore, the content and composition within the images can be analyzed by detecting objects and their respective locations. The development of computer-based algorithmic image analysis will force us to reconsider the role of human connoisseurship; but first and foremost computer vision can assist by processing the enormous visual data of cultural heritage. On the one hand the variety of cultural heritage across societies, various media and materials requires the flexible visual search. On the other hand cultural heritage is however based on many canonic pictorial subjects, common symbols, scriptures, and icons as well as concrete objects and standardized ornaments. Therefore there is also a strong necessity for a search algorithm which can retrieve specific objects, subjects and characters.

The paper explores the opportunities to support the research of scholars from various fields. Art history as well as archeology, visual studies and other image oriented research fields can benefit from visual retrieval not only to find details and objects but also to analyze their similarity, strategies of variation and reproduction and visual as well as semantic connections. Computer vision benefits from a challenging problem in the humanities that stimulates advances in computer vision, leading to improvements in the understanding of visual representations.

Nevertheless there are restrictions in the computer vision of art. The semantic gap and elaborate concepts of iconography and iconology as well as the various shades of attribution in the field of connoisseurship are challenging problems. Visuals similarity itself has incommensurable and contradictory definitions: through motif, composition, style, technique. The paper points out the boundaries of computer vision concerning art and reflects on possible solutions.

References

Barasch, M. (1976). *Gestures of despair in medieval and early Renaissance art*. New York: New York University Press.

Barasch, M. (1987). *Giotto and the Language of Gesture*. Cambridge: Cambridge University Press.

Burke, P. (1991). The language of gesture in early modern Italy. In Bremmer, J and Roodenburg, H. (eds), *A cultural history of gesture*, pp.71-83. Cornell University Press.

Criminisi, A., Kemp, M. and Zisserman, A. (2005). Bringing pictorial space to life: computer techniques for

the analysis of paintings. In *Digital art history: A subject in transition*, pp. 77-100, Intellect.
Freedberg, D.A. (2007). Empathy, motion and emotion. In: Herding, C and Krause-Wahl A. (eds.), *Wie sich Gefühle Ausdruck verschaffen: Emotionen in Nahsicht*, pp.17-51. Driesen.
Garnier, F. (1982). *Le langage de l'image au Moyen Âge. II. Grammaire des gestes*. Le Leopard d'or.
Goodman, N. (1968). *Languages of art: An approach to a theory of symbols*. Hackett publishing.
Knox, D. (1990) Ideas on Gesture and Universal Languages, c. 1550-c. 1650. In: Henry, J and Hutton, S, (eds.) *New Perspectives on Renaissance Thought. Essays in the History of Science, Education and Philosophy in Memory of Charles B. Schmitt*. pp. 101-136. Duckworth: London.
Schmitt, Jean-Claude (1990). L*a raison des gestes dans l'occident médiéval*. Editions Gallimard: Paris.
Tavernor, R. (1995). Architectural history and computing. arq: *Architectural Research Quarterly*, 1(1), pp.56-61.
Van Eck, C. (2007). Classical rhetoric and the arts in early modern Europe. Cambridge and New York: Cambridge University Press.

Building Bridges With Interactive Visual Technologies

Adeline Joffres
adeline.joffres@huma-num.fr
Huma-Num, France

Rocio Ruiz Rodarte
rocioruizrodarte@gmail.com
Tecnológico de Monterrey, Mexico

Roberto Scopigno
roberto.scopigno@isti.cnr.it
CNR-ISTI, Italy

George Bruseker
bruseker@ics.forth.gr
ICS-FORTH, Greece

Anaïs Guillem
anais.guillem@gmail.com
UC Merced, United States of America

Marie Puren
marie.puren@inria.fr
INRIA, France

Charles Riondet
charles.riondet@inria.fr
INRIA, France

Pierre Alliez
pierre.alliez@inria.fr
INRIA, France

Franco Niccolucci
franco.niccolucci@gmail.com
PIN, Italy

General presentation of the panel

Digital technologies offer modern instruments to create strong and persistent bridges between different cultures, countries, disciplines or communities. Among those, interactive visual technologies (3D graphics, or the several incarnations of 2D images) offer unprecedented capabilities to link different contexts: - visual technologies bridge the scholar or the professional to the reality, by providing high fidelity digital clones of the works under study; - they offer excellent methods to bridge different scientific domains, by providing tools for integrating different data and show their interplay, thus strongly supporting multidisciplinary investigation; - they help bridging different cultures and communities, since the visual presentation of different heritage is the first step towards improved mutual knowledge and discovery of commonalities; - they help bridging academic research with the public, since those visual and interactive media are extremely powerful instruments to disseminate to the public in museums or on the web; - and, finally, they help bridging the past with the future, since visual representations together with linked semantic data are the key resource to preserve our current knowledge for the future (documenting not just the conservation status but also the reasoning process that leaded us to some insight) and to support future work on the same subjects.

Such benefits are secured by research infrastructures allowing experts to foster and share their best practices and allowing less experienced scholars to acquire knowledge on 3D data management. Another ambition of these infrastructures is to promote interoperability in research practices by giving access to guidance and services. Furthermore, these infrastructures also provide tools through iterative and collaborative processes. In that perspective, as far as 3D standards and practices are concerned, the H2020 project PARTHENOS (http://www.parthenos-project.eu/) is building a service called "Standardization Survival Kit (SSK)" which is an example of such a collaborative approach.

The panel will review the status of a few visual technologies and related CH/DH workflows, mostly focusing on the technologies supporting: - the digitization of CH assets; - the visual analysis and comprehension; - the semantic enrichment and preservation of knowledge. It will also demonstrate how these technologies' best practices can be secured by the PARTHENOS standardization policy materialized by the SSK.

The panelists will present the status of the different sub-domains with very short talks (10 minutes each), highlighting opportunities, consolidated approaches and open issues. This state of the art review will build a common space for the further Q&A discussion with the audience on the perceived strengths and weaknesses of current digital instruments used in the DH domain.

Panelists

- Rocio Ruiz Rodarte (Tecnológico de Monterrey, Mexico, https://tec.mx/es) - Pierre Alliez (Inria, France, https://team.inria.fr/titane/pierre-alliez/) - Roberto Scopigno (CNR-ISTI, Italy, PARTHENOS member, http://vcg.isti.cnr.it/~scopigno/) - George Bruseker (ICS-FORTH, Greece, http://www.ics.forth.gr/isl/index_main.php?l=e&c=452) & Anais Guillem (UC Merced, USA) - Marie Puren or Charles Riondet, Inria, France, PARTHENOS members

Chairs

- Franco Niccolucci, PIN, PARTHENOS's project coordinator. - Adeline Joffres, CNRS Huma-Num, PARTHENOS member.

Reconstruction 3D, from Archaeological Reports to Digital Museographies

Rocio Ruiz Rodarte

3D visualization has served as a bridge between the world of archaeologists, academics and general public in a country like Mexico where geography and budget do not always help this union to arise in a simple way.

Seventeen years ago we combined the knowledge of the archeologist of Calakmul, Ramon Carrasco, with those of students and professors of Architecture and Robotics to develop digital reproductions of the findings of this ancient site hidden within the Mayan jungle.

We used simple methods such as AutoCAD, 3DMax, Unreal Engine and ARToolKit to reconstruct buildings and burials with the purpose of making virtual reality tours, augmented reality installations and VRML models for web pages. All these three-dimensional archives served the archaeologists and restorers of Calakmul to make presentations at conferences and present work reports for several years. Some, the least, were placed in a museum near Calakmul.

The project also served to promote the interest among Mexican researchers in cultural heritage themes. Some of the initiatives that are presented today in national forums as DH2018, arose from projects developed back in 2000 in institutions such as CNA, UAM, UNAM and Tecnológico de Monterrey.

For this event, we have made an upgrade of our projects with the same intention of making them accessible to be viewed by general public. The simple methods we used for those projects still serve us today for virtual tours with the 4th version of Unreal Engine and more commercial applications of augmented reality such as Aurasma. Nowadays we also have the possibility to embed those 3D models inside eBooks for tablets.

There is always people who have the knowledge, the content, as well as developers with extensive technological skills and advanced tools. These forums allow us to get to know each other and establish the bridges to communicate and work together.

3D Digitization and Reconstruction

Pierre Alliez

The technological advances of geometric measurement devices have revolutionized our ability to digitize the world in 3D. This revolution has made possible the development of many new technologies to digitize cultural heritage artefacts and scenes. In addition to sensor technologies, a key issue at the heart of this revolution is that of surface reconstruction, which consists in converting the raw measurements into a computerized surface representation. Beyond faithful topology and geometry, cultural heritage artefacts requires capturing additional properties such as complex interactions between light and materials. Key issues to generate meaningful cultural digital resources include the documentation and quality assessment of the acquisition process, the capturing of conservation status and the recording of provenance information. Open issues are related to the continuous update and consolidation of cultural digital resources. Instead of centralized and static acquisition by a single expert, a dire need is to shift to collaborative and dynamic approaches where communities, active sensor networks and active data resources cooperate to continuously create meaningful cultural resources and generate new knowledge with high relevance to cultural heritage practitioners.

Digital technologies are now mature for producing high quality digital replicas of Cultural Heritage (CH) artefacts. Those digital clones are becoming important assets in many DH activities

Roberto Scopigno

Many applications require to share the models produced, to support the cooperative work of professionals or scholars. This is an emerging need in the DH/CH community: results of digitisation should not remain of restricted use of the scholar or museum who commissioned the digitisation, but should be open to the large community of

experts and practitioners. Therefore, the web is the ideal distribution and sharing context. But publishing on the web a complex 3D model was not easy until recently.

The focus of the talk will be to show and discuss practical solutions for the easy publication on the web and visualisation of high-fidelity 3D models. The talk will show some practical examples where high-quality 3D models have been transformed in web-compliant format to be used in CH research, restoration and conservation. The examples will include tools developed to enrich the visual data or to produce insight from the interactive visual analysis.

3D Models, Humanities & Ontologies

George Bruseker

The versatility that 3D modelling techniques offer scholars to pursue new and old lines of research, has led to an intense interest in its application in fields across the humanities. The enthusiastic uptake of these techniques raises important meta-methodological questions in terms of the correct scientific documentation of 3D models and the long term ability to objectively test their validity. Facing these questions demands interdisciplinary cooperation between computer science, information management, and the implicated humanities themselves foremost. A key question to answer is how to elaborate a methodological and technical means to trace the provenance of digital objects, from the point of their original creation through digitization or hypothesis, through their various iterations and adaptations. We argue that only a formal ontology solution, such as CRMdig, that generalizes over a series of popular metadata formats, together with technically aided documentation of process by scholars during their 3D research can achieve these goals.

Investment in developing and applying ontologies to the 3D modelling problem is justified by its increasing importance in research. The range of the application scenario for the use of 3D modelling has long moved past simple visualization for presentations to become an analytic research tool in its own right that can aid in primary research. 3D models can be used to carry out research on physical structures and spaces at a micro or macro level. Accurately capturing physical dimensions in detail can help in the study of design, style, function, provenance and decay of objects. Such models provide an empirical base for pursuing broader research questions such as understanding intention, continuity and change. Virtual reconstruction work, which extends measurement of extant partial objects with evidence based hypothetical reconstructions of past states, allows for complex arguments to be pursued in relation to scholarly sound digital models of partially lost past works.

Given the extant investment of time and thought in such models, it is imperative that this knowledge be preserved. The challenge of doing so arises due to the plethora of tools, institutions and methods which are brought to bear in applying 3D techniques. There is a dizzying area of products and tools already in the 'market' and these increase every day. Meanwhile, owing to a lack of standards and a general meta-methodology for the treatment of such models, work done in the past that involved important capital and human investment, is now unusable. The situation calls for a remedy.

We argue that there can be no top-down solution to this problem, imposing certain metadata standards or certain techniques for all. Because of the on-going and developing nature of this research approach, such approaches have in the past and will continue to fail. Rather, what is needed is, on the one hand, the creation of an awareness of the provenance tracking requirement, to create adequate metadata in the first place, and then the application of sufficiently broad ontological models, in order to allow technical solutions for the integration of heterogeneous data from multiple repositories into long term viable storage.

PARTHENOS, a European Project Building and Disseminating Collaborative Tools: the Example of 3D Standards within the Standardization Survival Kit

Charles Riondet & Marie Puren

The lifecycle of 3D objects, from their production to their reuse, including processing, description and visualization, involves many highly technical steps and the use of a wide range of technologies, methods and tools, evolving quickly. For each step, the diversity of practices and protocols is a hindrance to agree on a unique standardized solution that could fit all users' needs and solve all the possible problems. Consequently, a crucial task consists in producing guidelines and documenting research practices, in particular at both ends of the 3D objects' lifecycle: - the digitization/modelling phase, in the course of which these objects are created and thus have to be documented properly, - and the data reuse phase, in the course of which the availability of proper metadata and accessible archives is a pre-condition for further reuse of digital assets.

PARTHENOS, in particular with the development of the Standardization Survival Kit, aims at being the place where these good practices can be recorded and presented, by means of specific research scenarios where the handling of 3D objects is the core activity, and that would be presented together with documentation, literature and technical resources.

The Standardization Survival Kit can be seen as the host for the state of art documentation of standards re-

lated to 3D (amongst others). It is currently under development and will be soft launched in the beginning of 2018, so that a demonstration could be provided during the conference in June 2018. The 3D scenarios will be created by domain experts - partners of the PARTHENOS project or external scholars -, and based on the white paper "Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation", result of a 2016 PARTHENOS Workshop (Alliez, P. et al., 2017).

References

Alliez, P. and Bergerot, L. and Bernard, J. F. and Boust, C. and Bruseker G., et al. (2017) *Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation*. White paper: A result of the PARTHENOS Workshop held in Bordeaux at the Maison des Sciences de l'Homme d'Aquitaine and at Archeovision Lab. (France), November 30th - December 2nd, 2016. PARTHENOS. Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation, Nov 2016, Bordeaux, France. pp.71, 2017. https://hal.inria.fr/hal-01526713v2

The Impact of FAIR Principles on Scientific Communities in (Digital) Humanities. An Example of French Research Consortia in Archaeology, Ethnology, Literature and Linguistics

Adeline Joffres
adeline.joffres@huma-num.fr
CNRS, Huma-Num, France

Nicolas Larrousse
nicolas.larrousse@huma-num.fr
CNRS, Huma-Num, France

Stéphane Pouyllau
stephane.pouyllau@huma-num.fr
CNRS, Huma-Num, France

Olivier Baude
olivier.baude@huma-num.fr
CNRS, Huma-Num, France

Fatiha Idmhand
fatihaidmhand@yahoo.es
Université de Poitiers, France

Xavier Rodier
xavier.rodier@univ-tours.fr
Université de Tours, France

Véronique Ginouvès
veronique.ginouves@univ-amu.fr
Université d'Aix-Marseille, MMSH/phonotèque, France

Michel Jacobson
michel.jacobson@huma-num.fr
BNF, France; CNRS, Huma-Num, France

The French TGIR Huma-Num, whose mission is to facilitate the digital turn in Humanities and Social Sciences research, offers services dedicated to the production and reuse of data. These services aim at avoiding loss and facilitating the reuse of scientific data in Social Sciences and Humanities. To do this, Huma-Num supports research teams and disciplinary consortia throughout their digital projects to allow the sharing, reuse and preservation of data thanks to a chain of devices focused on interoperability. Through these processes, Huma-Num also encourages compliance with the FAIR data principles.

Huma-Num's tools connect normalised metadata to the Linked Open Data cloud and give them an extended visibility. By labelling consortia (through scientific, financial and technical support when needed), Huma-Num drives and supports them in this initiative: data producers are encouraged to clean and normalise their data, they benefit from Huma-Num's services to store, share, expose, signal and enrich their data. What's more, at the end of the chain, these processes also allow the data to be "ready" or, at least, "prepared" for archiving, which is a real issue for research data.

But what is the effect induced on the data producers by sharing their resources, particularly on the metadata? Do they realize that the metadata produced for a specific project are not suitable for more generic needs and need some polishing?

Does this virtuous circle produce in turn better metadata and also better practices?

The panel will try to answer these questions by presenting feedback from different disciplines and communities in order to trigger discussion. More specifically, through the panel, four of Huma-Num's consortia - in archaeology, ethnology, literature and linguistics – will present their experience, practices and some tools they have produced in order to measure the impact of this supposedly virtuous circle on the quality of the data and metadata they have produced and exposed in the LOD. Additionally, this will allow us to discuss the role of a national research infrastructure like Huma-Num in France and the collaborative and expert networks developed through the creation of Huma-Num's consortia.

By sharing the points of view of various disciplinary but multi-approach consortia in Social Sciences and Humanities, the panel will aim at proposing a reflexive approach to the impact of the FAIR principles.

The panelists will present the status of the different subdomains with very brief talks (15 minutes each), hi-

ghlighting opportunities, consolidated approaches and open issues. The different perspectives and experiences presented, on various types of data and in different disciplines, will build a common space for the further Q&A discussion with the audience on the application of the FAIR principles in the DH domain.

Panelists

- Michel Jacobson, CORLI consortium, CoCOon, https://cocoon.huma-num.fr/exist/crdo/ - Fatiha Idmhand, CAHIER consortium, CNRS/Huma-Num, http://cahier.hypotheses.org/ - Xavier Rodier, MASA consortium, CNRS/Huma-Num, https://masa.hypotheses.org/ - Véronique Ginouvès, AdE consortium, CNRS/Huma-Num, https://ethnologia.hypotheses.org/

Chairs

- Adeline Joffres, TGIR Huma-Num, CNRS, France - Nicolas LARROUSSE, TGIR Huma-Num, CNRS, France

Archives des Ethnologues" Consortium, France. Anthropologists, Archivists and Fieldwork Materials: Best Practices of a French Consortium

Véronique Ginouvès

For the resource centers that compose it, the creation of the Consortium "Archives des ethnologues" within the Huma-Num TGIR in 2012 was key to the acquisition of new methodological reflexes in the digital domain. The aim of its eight partners is to store, process, share and publish the documents produced by anthropologists in their fields. The scientific and heritage importance of these survey materials, the richness and diversity of the societies studied, oblige us to take the singularity of these data better into account.

We will show how the consortium is working towards greater convergence of the practices of description, structuring and access to ethnological data, notably through the FAIR Data principles ("findable, accessible, interoperable, reusable").

These practices give us the opportunity to adapt these data to documentary projects or scientific research and their objectives.

Thus, in order to enhance the discovery and availability of these archives, the metadata used to describe them are not only produced in DC (Dublin Core) but also, for example, in EDM (Europeana data model) or EAD (Encoding archival description).

Similarly, their transfer to different data archives (ODSAS, Kinsources, Portal of oral heritage, MediHal, ...) and their access via different platforms (Calames, Clarin, Europeana, Isidore) are essential for their availability, research, identification and increased interoperability.

We will also focus more specifically on the use of vocabularies, which are essential to improve the search for these data on major platforms.

The first example will be that of the authorities. We will show how the use of tools such as ISNI, VIAF or IDREF "propel" the members of the societies studied by ethnologists into international standard name systems. These witnesses, who have remained anonymous for a long time, then become truly authors of their word and their families can thus find traces of their names through the devices put in place.

The second example of a vocabulary that will be presented is the creation of a thesaurus (in SKOS format and produced with OpenTheso software) on uniform titles of tales in order to enrich the research on oral literature.

Finally, we will also address the issues of data access and reuse.

From the beginning of the Consortium, discussion has been ongoing to address the ethical and legal issues related to the dissemination of humanities and social sciences data. A blog (https://ethiquedroit.hypotheses.org) provides answers to the concrete questions that arise during online publication and a guide of good practices will be published in September 2018. We will also present some examples of the re-use, on our platforms, of some of the data processed by the different centers: a crowdsourcing project with Transcrire, a work space for research with ODSAS or the provision of specific data (the kinship data) with Kinsources.

These principles are also the strength of national institutions (TGIR Huma-Num, CINES) that support the implementation of projects over the very long term. They provide us with a solid framework for organizing data life, access and sharing.

"MASA" consortium, France. What is the Cost and the Efficiency of Exposing Archaeological Data in the LOD?

Xavier Rodier

There is a very great disparity in the organization of archaeological data and their management and archiving systems, when they exist. The immediate consequence of this state of affairs is the risk of the irreversible loss of a significant amount of inaccessible archaeological data. The MASA consortium has therefore set itself the objective of digitally sharing archaeological data by proposing guidelines to the archaeological community. There is an urgent need both to safeguard existing archival collections and ensure the re-use of old databases, and to ensure that emerging new databases use sustainable systems that will provide interoperability and the long-term

reuse of data. The consortium's work therefore focuses on the classification and digitization of old archaeological archives, the documentation of its collections according to an appropriate structure, the re-use of old databases, the alignment of vocabularies used with standards, the matching of archaeological information systems with the domain ontology for cultural heritage (CIDOC-CRM, ISO 21127:2014), on the online publication of archives, data and syntheses linked with data.

The difficulties to be overcome are many, depending on whether structured databases or more informal batches are being processed, and must reconcile various situations such as: - The digitization of old archives (often resulting from the work of only one archaeologist) which must be classified, safeguarded, made available and documented according to the archaeological value added. This involves producing an overlay to compensate for the absence of a data structure. - The transformation of old databases developed with systems and in formats that are no longer accessible or disappearing, in order to preserve the data themselves and their structure when it exists. - The interoperability of old, structured and still maintained systems, whose redeployment in standard and open formats is necessary but beyond the means available. While this may seem simpler, it is not the case and care must be taken not to delay those who have had advance notice. - The development of new systems which must be designed directly in accordance with existing interoperability standards so that they do not have to be rethought in mid- term.

All the experiments carried out show the heuristic value of the operations necessary for the digital sharing of data, which is a reflective step in terms of both content and information structure. However, the final quality of the information shared according to these four processes varies. The creation of metadata on loosely structured corpora is a definite added value but never achieves the finesse of description of structured information systems. In addition, the use of metadata description standards alone does not offer the semantic enrichment that can be achieved by mapping with reference ontologies, which constitutes a production of knowledge in itself. In fine, exposing archaeological data in the LOD will help to build bridges with other heritage data but also with other themes.

These different processes and their consequences will be explored using a few examples.

"CAHIER" Consortium, France. CAHIER: "Bridges / Puentes" Between Text Sciences

Fatiha Indman

CAHIER is a French consortium whose mission is to promote good digital practices in text sciences and to build a network of expertise in the SSH scientific community. The particularity of CAHIER is that it does not bring together research centers but projects. Project members aim at collectively finding or sharing solutions to digitize, edit, display and process their data. The purpose of the consortium is not to register data in a specific field but to create links between disciplines that use texts as their scientific objects and subjects: Literature, Linguistics, History, Philosophy, ICT, Computer science, etc.

CAHIER's approach propagates and disseminates practices that comply with the "FAIR principles", by acting upstream of the projects, in order to guarantee and promote the quality of the data and metadata that are produced.

Through the example of the "WebOai" metadata exhibit tool, developed by the Cahier consortium with the help of Huma-Num, we will show how the Cahier consortium prepares its digital corpus of sources (teiHeader) for dissemination, exhibition and research. WebOai implements an OAI-PMH repository from XML-TEI encoded data sources. We will show how the confrontation of methods and the exchange of solutions, within the consortium, have allowed researchers to reflect on their data quality and how we are contributing to building the "Digital humanities" community in France.

Platform-"CORLI" Consortium, France. The Benefits of Data Linking and Use of the CoCOon Repository

Michel Jacobson

The first advantage of data linking is that it requires cleaning the data to make them sufficiently homogeneous for mass treatment, either automatic or assisted. In a repository where the applicant provides the description with little or no help and with little moderation, it leads quite quickly to alternative forms for the same resource. For instance, the identification of a person by name is not always normalized but is often subject to variants (case variant, order of elements, changes in civil status, use of a pseudonym, abbreviation, typing error, etc.). Moderation is sometimes made difficult also because of the use of foreign scripts or conventions.

Linking to a repository means that identification and description needs can be separated. For example, in a documentary resources repository, the actors involved in creating the document have to be identified in the document description. It will be advantageous to describe the actors in a distinct and specialized repository since the description templates for actors and documents won't necessarily be the same. Moreover, an actor exists independently of his documentary production and other repositories (of events, of objects, etc.) may have the same requirement for identification, making it interesting to share the service.

The criteria of coverage, governance and interoperability need to be taken into account when choosing a

vocabulary. This choice is important because linking the data also means in some cases, deporting the task of description. In particular, the "collaborative" modes of governance of projects such as Geonames or Dbpedia have the advantage that one can directly enrich the repository to cover missing needs. One can also, as we have tried to do with the CoCOon (https://cocoon.huma-num.fr) platform dedicated to digital oral corpora, make producers or depositors responsible for enriching these vocabularies themselves. The choice of repositories rapidly proves to be strategic because they will be the linchpin in decompartmentalizing the data: either the repositories share common vocabularies, or their separate vocabularies are interlinked.

As part of the work on the CoCOon platform, for example, we have indexed a collection of speech recordings ("Speech Treasures") by aligning the themes present in their metadata with a Thesaurus (RAMEAU). This allowed us to 1) offer a new axis of navigation in the data, 2) bring these records closer to other cultural data (those of the BnF) which are indexed with the same thesaurus, 3) potentially bring together other cultural data through the alignment of RAMEAU with other reference systems such as the Dewey classification, the thesaurus of the German National Library, the National Library of Spain and the Library of Congress, 4) envision multilingualism with no added cost by exploiting this alignment between repositories, 5) facilitate the reuse of data and their discovery.

Normalized vocabularies are a bridge between repositories, making it possible to bring together isolated data and thus to give them a richer context, improving their readability. For example, the use of the Lexvo vocabulary - which includes all the codes of the ISO-639-3 standard - makes it possible to reconcile recordings, scientific documentation, geopolitical information, etc. for a given language.

DH in 3D: Multidimensional Research and Education in the Digital Humanities

Rachel Hendery
r.hendery@westernsydney.edu.au
University of Western Sydney, Australia

Steven Jones
s3jones1@gmail.com
University of South Florida at Tampa, United States of America

Micki Kaufman
mickikaufman@gmail.com
Graduate Center of the City University of New York, United States of America

Amanda Licastro
amanda.licastro@gmail.com
Stevenson University, United States of America

Angel David Nieves
angel.nieves@yale.edu
Yale University, United States of America

Kate Richards
k.richards@westernsydney.edu.au
University of Western Sydney, Australia

Geoffrey Rockwell
grockwel@ualberta.ca
University of Alberta, Canada

Lisa M. Snyder
lms@idre.ucla.edu
University of California, United States of America

Computing capabilities for rendering high-quality three-dimensional graphics have progressed remarkably in recent years, largely in response to competition in the gaming and defense industries. While public awareness and engagement with virtual reality (VR) and/or augmented reality (AR) platforms has risen sharply, scholars are taking a measured and thoughtful approach, engaging with the new technology while remaining meta-critical about how high-speed computational capabilities like 3D, VR and AR can effectively represent the multiple dimensions in their digital humanities research.

An increasing number of multidimensional projects by digital humanities scholars focus on the modeling and simulation of real, historical physical spaces, and/or the articulation of imaginary or data-derived spaces for pedagogy and research in the humanities. A common thread of the use of three-dimensional representations and techniques is that they are at once both extremely complex and stunningly intuitive, both to render *and* to interpret. The same paradoxicality can be said of some aspects of digital humanities research. Using algorithms to approach questions of subjectivity and distance, employing visualization to explore voice and genre, and leveraging the virtual to explore the real, multidimensional scholarship likewise applies the rigid logic of computation to understand deeply subjective aspects of the human experience, in an immersive application of "thick mapping" (c.f. Presner et al 2014). The ability for DH to flourish while comprising such internal contradictions suggests the capabilities of multidimensional technology to distill and refine the essential points of complexity by articulating them in those dimensions. In this manner, multidimensional scholarship seeks to reveal the underlying essence of DH projects by employing rich, deep and immersive experiences in pedagogy, data visualization, modeling and simulation.

This panel brings a diverse host of scholars together to demonstrate and discuss their exploration of three-dimensionality, including virtual and augmented reality, in Digital Humanities research. Rachel Hendery and Kate Richards, of Western Sydney University, will describe their group's experiences of co-designing virtual reality and other 3D experiences with members of Australian First Peoples' communities, Steven Jones of the University of South Florida at Tampa will present on his research and simulation of the first dedicated humanities computing center. Amanda Licastro will discuss her work in critiquing and building VR applications with undergraduate students. Micki Kaufman will show the results of her utilization of three-dimensional interactive spaces for data visualization and storytelling of the Kissinger Correspondence, Angel David Nieves, of Yale University, will discuss the ways in which 3D historical reconstructions can be used as tools for the promotion of social justice advocacy in digital humanities. In addition, Geoffrey Rockwell of the University of Alberta will report on recent experimentations with augmented reality and the role of play in pedagogy.

References

Anning, B. (2010). Embedding an Indigenous graduate attribute into University of Western Sydney's courses. *The Australian Journal of Indigenous Education*, *39*(S1), 40-52.

Foster, D., Williams, R., Campbell, D., Davis, V., & Pepperill, L. (2006). 'Researching ourselves back to life': new ways of conducting Aboriginal alcohol research. *Drug and alcohol review*, *25*(3), 213-217.

Heppell, M., & Wigley, J. J. (2017). *Black out in Alice: A history of the establishment and development of town camps in Alice Springs*. Canberra, ACT: Development Studies Centre, The Australian National University.

Kukutai, T., & Taylor, J. (Eds.). (2016). *Indigenous Data Sovereignty: Toward an Agenda* (Vol. 38). ANU Press.

Merlan, F. (2014). Recent rituals of Indigenous recognition in Australia: Welcome to country. *American Anthropologist*, *116*(2), 296-309.

Trescak, T., Esteva, M., & Rodriguez, I. (2010). A virtual world grammar for automatic generation of virtual worlds. *The Visual Computer*, *26*(6), 521-531.

NEH-funded project: Reconstructing The First Humanities Computing Center: https://recaal.org [to go live spring 2018].

Emerson, Lori. (2014). *Reading Writing Interfaces: From the Digital to the Bookboun*d. Minneapolis: University of Minnesota Press.

Kraus, Kari. (2012.) Introduction to *Rough Cuts: Media and Design in Process*: http://mediacommons.futureofthebook.org/tne/pieces/introduction.

Parikka, Jussi. (2012). What Is Media Archaeology? Cambridge, UK: Polity.

Sayers, Jentery. (2015). "Kits for Cultural History." *Hyperrhiz* 13 (Fall 2015), http://hyperrhiz.io/hyperrhiz13/workshops-kits/early-wearables-essay.html.

Bowman, Denvey (2011). "The Empathy Experiment." Capital University.

Davis, Mark H. (1983). "Measuring the Individual Differences in Empathy: Evidence for a Multidimensional Approach." *Journal of Personality and Social Psychology*. Vol. 44, No. 1:113-126.

O'Brien, Ed, Sara H. Konrath, Daniel Grühn, Anna Linda Hagen (2013). "Empathic Concern and Perspective Taking: Linear and Quadratic Effects of Age Across the Adult Life Span." *The Journals of Gerontology: Series B*, Vol. 68, Issue 2: 168–175.

Galeazzia, Fabrizio, Marco Callieri, Matteo Dellepiane, Michael Charno, Julian Richards, and Roberto Scopigno (2016). "Web-based visualization for 3D data in archaeology: The ADS 3D viewer." *Journal of Archaeological Science Reports* vol. 9: 1-11.

Potenziani, Marco, Marco Callieri, Matteo Dellepiane, Massimiliano Corsini, Federico Ponchio, Roberto Scopigno (2015). "3DHOP: 3D Heritage Online Presenter." *Computers & Graphics* 52:129-141.

Risam, Roopika (2015). "Beyond the Margins: Intersectionality and the Digital Humanities." *Digital Humanities Quarterly* vol. 9, no. 2.

Sullivan, Elaine, Angel D. Nieves, and Lisa M. Snyder (2017). "Making the Model: Scholarship and Rhetoric in 3D Historical Reconstructions." *Making Things and Drawing Boundaries: Experiments in the Digital Humanities*. ed., Jentery Sayers. University of Minnesota Press.

Abt, Clark C. Serious Games. Lanham, MD: University Press of America, 1987.

McGonigal, J. (2011). Reality is Broken: Why Games Make Us Better and How They Can Change the World. New York, Penguin.

Rockwell, G., K. Uszkalo, C. Henry, E. deJong, S. Lucky, M. Illovan, L. Gutierrez, S. Gouglas, P. Boechler and E. Stroulia (2013) "Campus Mysteries: Serious Walking Around." Vol. 7. No. 12. Loading... Winter 2013. http://journals.sfu.ca/loading/index.php/loading/article/view/115

Szulborski, David. This is Not A Game. Lulu, 2006.

Si las humanidades digitales fueran un círculo estaríamos hablando de la circunferencia digital

Tália Méndez Mahecha
tata.mendez.m@gmail.com
Biblioteca Nacional de Colombia, Colombia

Javier Beltrán
jrbeltran@bibliotecanacional.gov.co
Biblioteca Nacional de Colombia, Colombia

Stephanie Sarmiento
ssarmiento@bibliotecanacional.gov.co
Biblioteca Nacional de Colombia, Colombia

Duván Barrera
dbarrera@bibliotecanacional.gov.co
Biblioteca Nacional de Colombia, Colombia

Sara del Mar Castiblanco
scastiblanco@bibliotecanacional.gov.co
Biblioteca Nacional de Colombia, Colombia

María Helena Vargas
mvargas@bibliotecanacional.gov.co
Biblioteca Nacional de Colombia, Colombia

Natalia Restrepo Saldarriaga
nrestrepo@mincultura.gov.co
Ministerio de Cultura, Colombia

Camilo Martinez
gemartin@uniandes.edu.co
Universidad de los Andes, Colombia

Juan Camilo Chavez
juan-cha@uniandes.edu.co
Universidad de los Andes, Colombia

Resumen

El desarrollo de productos dentro las Humanidades Digitales podría verse como la diferencia entre círculo y circunferencia. La circunferencia es lo que rodea al círculo, y el círculo es todo lo que contiene la circunferencia. En este orden de ideas, el objetivo de este panel es plantear las diferentes posiciones desde lo que rodea al producto como: la estrategia digital, el diseño emocional, las herramientas para producirlas, y cómo esto ha pasado de la teoría a la práctica.

Con ocasión de la Conferencia de Humanidades Digitales de la ADHO, queremos proponer esta conversación desde algunos escenarios y casos de estudio colombianos para integrarnos en las discusiones regionales e internacionales sobre lo que significa para nosotros estudiar las Humanidades Digitales desde América Latina; y, además, porque el espíritu del libro encaja muy bien dentro de la temática de la Conferencia.

Esta conversación nos permitirá plantear una postura desde el Grupo de Investigación sobre lo que entendemos por Humanidades Digitales en Colombia a partir de un contexto de producción "circunferencia". Reconocemos que desde el mundo hispanohablante ha habido un trabajo arduo por alcanzar un consenso, pero la diversidad de prácticas ha señalado particularidades que hacen que continuamente se revalúen las tentativas definiciones.

Adicionalmente, consideramos que la conformación del grupo "De Punto a Pixel" y su participación en el fortalecimiento de las Humanidades Digitales ha tenido una historia particular, desde que surgió la idea de conformarlo en la Biblioteca Nacional de Colombia. Relatar ese proceso nos permitirá contrastar nuestra experiencia con la latinoamericana y la de otras latitudes.

Estrategia digital

Natalia Restrepo Saldarriaga

> ... para que la ciencia avance, no basta concebir ideas fructíferas, elaborar nuevos experimentos, formular nuevos problemas o establecer nuevos métodos. Las innovaciones deben ser efectivamente comunicadas a otros. A fin de cuentas, esto es lo que entendemos por contribución a la ciencia: es algo que se da al fondo común del conocimiento.
> Robert K. Merton, *La sociología de la ciencia*.

El fin de toda producción científica es que sea conocida por otros. No estamos eximidos de ello los humanistas y mucho menos cuando se tiene intenciones de exponer un acervo, un resultado de investigación o proponer una forma novedosa de hacer algo en el sector de las humanidades. A pesar de que lo anterior es claro para los humanistas, es evidente que hoy en día los proyectos propios de las humanidades digitales no tienen estrategias de divulgación y circulación que garanticen que los públicos interesados estén enterados de las nuevas producciones.

El espacio digital propicia escenarios en donde las audiencias no solo consumen información, sino que también la producen y la reconstruyen, lo que abre un amplio espectro para nuevos procesos de innovación en la sociedad. La idea de divulgar las humanidades a través de lo digital necesita ir acompañada de estrategias de construcción colectiva de conocimiento; bien sea durante o posteriormente, esa participación crea nuevas relaciones y significados a partir del reconocimiento y apropiación de los contenidos científicos por parte de la comunidad. La misma Internet hoy facilita estas dinámicas de colaboración a través de los medios sociales como herramientas para que cualquiera pueda crear, compartir, publicar o reconstruir un contenido. Sin embargo, los humanistas digitales están haciendo poco uso de estas herramientas. Romero-Frías (2014) pone de manifiesto el escaso uso que los humanistas digitales le están dando a los medios sociales. En su investigación encontró que "la presencia en redes sociales es moderada (Twitter , 50 %; Facebook , 30,8 %)... evidenciando cómo aún hay camino por recorrer en las humanidades digitales para asumir las ideas de las culturas digitales…".

Esta propuesta de presentación para el panel de nuestro grupo de investigación tomará como caso de estudio la Estrategia Digital de Apoyo a la Formación Musical 'Viajeros del Pentagrama' del Ministerio de Cultura de Colombia, la Fundación Nacional Batuta y la OEI, como un proyecto cultural que desde su gestación se ha construido en torno a dos objetivos. El primero, y oficial, es que to-

dos los niños del país adquieran habilidades y competencias musicales antes de terminar su educación primaria. El segundo es que cada contenido de la Estrategia sea al mismo tiempo un insumo para campañas de divulgación y apropiación de la misma. Esto significa que los productos realizados, como videos, podcast o infografías, fueron pensados en términos académicos y culturales y también en términos de divulgación de la estrategia.

Esta presentación muestra al humanista digital una propuesta de generación de estrategias digitales de divulgación y apropiación proyectadas desde el origen del proyecto, de manera que se optimicen recursos produciendo contenidos y formatos que cumplan al mismo tiempo con la misión de exponer el tema y de ser usados en pro de la divulgación del proyecto. Además de que puedan hacer uso de herramientas digitales interactivas, como los medios sociales, que permiten a las audiencias interactuar y apropiar el contenido.

Diseño emocional

Tália Méndez Mahecha

El objetivo de esta presentación es evidenciar la importancia de adaptar metodologías como el *Design Thinking* a los proyectos de Humanidades Digitales y reflexionar sobre el hecho de generar nuestras propias formas de empatizar digitalmente, para lograr la comprensión profunda del usuario y sus necesidades desde una perspectiva creativa que genere desarrollos digitales con un componente emocional que origine una relación cercana entre nuestros usuarios y el acervo patrimonial.

Pensar empáticamente desde la creación de producto es ponernos en el lugar del otro, es anteponer las necesidades de las personas, entender su mundo y comprender lo que sienten y, al mismo tiempo, equilibrar los requerimientos que se tengan como entidad, organización o institución. Si reflexionamos sobre esto en nuestro diario vivir podríamos decir que nos sentimos más vinculados a aquellos productos que nos son más cercanos, por tanto customizar logrará la gran diferencia para generar un compromiso o involucración respecto al producto.

Como Biblioteca Nacional de Colombia y colectividad cultural, exigirnos apropiar y transformar metodologías como el *Design thinking* para cautivar a nuestros usuarios, nos permitirá generar argumentos y contextos humanos tan fuertes desde las disciplinas involucradas que, a la hora de crear desarrollos digitales, podremos deshacernos de todo lo que no es esencial y lograr un compromiso emocional humanizando el resultado, en nuestro caso el patrimonio de la nación.

Las herramientas como facilitadoras en las HD

Duván Barrera

La profundización sobre las herramientas en las HD es el de la optimización de contenidos y productos digitales para mejorar la experiencia de usuario (UX), hace una década probablemente el aspecto visual no era tan atractivo como lo es ahora y hace veinte años era tan escaso como precario, en los años recientes se ha hablado bastante sobre mejorar la experiencia de usuario y esta suele ser la hoja de ruta para una gran cantidad de proyectos y productos digitales que emergen y proliferan cada día más. Pero no es solo hablar del aspecto visual ya que una experiencia de usuario destacada está dada en buena medida por la arquitectura propuesta para su contenido y la metodología que se usó desde su planificación, no es solo la usabilidad y navegabilidad, es un compendio de elementos que en conjunto nos dan un resultado para ser publicado como sitio web, app o *software*.

Afortunadamente en la actualidad existen una gran cantidad de herramientas que facilitan tanto a humanistas como a personas con conocimiento netamente técnico a poner en marcha sus proyectos digitales, abarcaremos con ejemplos actuales las ventajas que nos proporcionan algunas de dichas herramientas y su rol en el diseño de una experiencia de usuario óptima.

Más que una guía definitiva sobre las herramientas para usar o no usar este capítulo, se piensa como una guía práctica para aportar con ejemplos algunas herramientas que ayuden en la conceptualización, ejecución y puesta en marcha de proyectos de Humanidades Digitales que encuentran un bache en los aspectos técnicos inherentes al producto o contenido final.

Lenguajes de marcado para las humanidades digitales

Camilo Martínez

Los lenguajes de marcado son herramientas útiles para el trabajo con contenido textual que debe ser manipulado tanto por máquinas cómo por seres humanos, sin embargo el problema de los lenguajes de marcado más populares, como XML y sus derivados, tiene que ver con que el uso de etiquetas de apertura y cierre y atributos tiende a crear archivos con gran extensión, que son confusos de leer por seres humanos. La reciente introducción de lenguajes de marcado livianos como Markdown y YAML está reduciendo la complejidad del trabajo con texto en el contexto de las Humanidades Digitales. Estas herramientas permiten escribir y leer textos estructurados más fácilmente y pero al mismo tiempo pueden ser procesados automáticamente por sistemas digitales. Esta característica hace de estos lenguajes una serie de herramientas una eficiente a la hora de hacer trabajo investigativo en el

campo de las Humanidades Digitales. Sin embargo, más allá de la optimización los procesos de trabajo con corpus textuales, estos nuevos lenguajes de marcado pueden facilitar prácticas colaborativas de anotación y estructuración de textos. Esta posibilidad permite pensar en formas de abrir los procesos de investigación y construcción de conocimiento propios de las Humanidades Digitales a la participación de comunidades no especializadas. En este ensayo se analizarán los lenguajes de marcado livianos más populares del momento y algunas herramientas con las que es posible crear contenido, asignar metadatos o estructurar un texto semánticamente para su posterior procesamiento por sistemas digitales. Finalmente, se plantearán posibles usos de estos lenguajes, evidenciando su potencial para la práctica abierta y colaborativa de las humanidades digitales

La conservación y la restauración

María Helena Vargas y Sara del Mar Castiblanco

Para el campo de la conservación y restauración del patrimonio bibliográfico y documental, la nueva era digital ha traído consigo retos y tareas importantes en cuanto la implementación de nuevas metodologías, procedimientos y técnicas, tanto para el análisis de la materialidad de los soportes físicos nacientes, como para la preservación a largo plazo de los soportes y la producción de contenidos en formatos digitales. Tal como lo plantea Lafuente (2014) no se trata de que figuras como la del conservador, curador, el bibliotecario o el mismo archivista ya no se necesiten. El asunto es que tanto, bibliotecas,museos, casas de cultura ycentros culturales en general, tienen que reinventarse en un nuevo contexto en el que el acceso a la información no sólo es fácil y económico, sino que implica prácticas informales, tecnologías distribuidas y procesos deslocalizados.

Hace ya varios años, la conservación y restauración se encuentra en una encrucijada conceptual. Lo que hoy en día conocemos como patrimonio, organizado y almacenado en bibliotecas, museos, e instituciones culturales, no siempre refleja lo que diferentes grupos sociales consideran como propio y cada vez es mayor la necesidad de que estas instituciones se acerquen a las personas con nuevas formas de presentar la información y el conocimiento. Insurralde (2010) plantea que el objeto restaurable ya no es sólo un objeto histórico o artístico que vale por sí mismo, sino necesariamente un objeto que adquiere valor por los significados que los sujetos vierten sobre este. Vale la pena preguntarse frente a esta afirmación ¿qué objetos se valoran o se valorarán ahora en la era digital?¿Cómo aprovechar las humanidades digitales para generar o reactivar diferentes significados sobre los objetos?

Tanto la encrucijada conceptual como este último cuestionamiento llevan a pensar que,si las humanidades digitales son un campo de trabajo en donde las áreas de conocimiento tienen nuevos horizontes donde poder evolucionar de formas inesperadas y donde se pueden desarrollar, gestionando su información de forma más compleja, necesariamente los profesionales en conservación y restauración deben dirigir su mirada hacia esa dirección.

Partiendo de autores como Arsenio Sánchez, Javier Tacón, Luis Crespo y Alberto Campagnolo, entre otros teóricos de la conservación del patrimonio en bibliotecas, pero también de autores como Charles Faulhaber, Piscitelli, Antonio Lafuente, Gimena del Río, Helena Blanco, entre otros,esta propuesta de presentación para el panel de nuestro grupo de investigación abordará las diferentes relaciones entre la conservación y restauración del patrimonio bibliográfico y documental como disciplina y su desarrollo necesario en el campo de las humanidades digitales.

La propuesta buscará resaltar el gran potencial y variedad de posibilidades y perspectivas que se están desarrollando con el tránsito lento pero constante en este sentido.Este es sólo un paso más para afianzar la reunión de estas dos áreas que requiere también, integrarse con otras áreas de conocimiento —que se entiende desde la conservación y restauración, deben tener presencia pero que aún las vemos escasamente articuladas— como por ejemplo las ciencias de la documentación, ciencias de la información y la bibliotecología, la bibliografía, la historia y las ciencias y técnicas historiográficas, la lingüística, entre otras (Vargas, 2017).

Siendo la conservación y restauración de bienes bibliográficos y documentales una disciplina incipiente en Colombia que ha tenido muy pocos espacios en nuestro país para desarrollarse, evaluarse, validarse, etc., se busca además resaltar el trabajo del grupo de conservación de la Biblioteca Nacional en la mencionada articulación entre disciplinas

La edición digital

Javier Beltrán

Cómo guardar versus *guardar como*: la producción de contenidos digitales para la preservación bibliográfica y documental

Más allá del trasnochado debate del supuesto enfrentamiento entre el libro impreso y el libro electrónico, de la aparente novedad de los formatos de las publicaciones digitales y de la utópica innovación de las plataformas que las contienen, existe una realidad incuestionable en el mundo de hoy, hiperconectado e hipercomunicado: la digitalización y la producción de contenidos nacidos digitales puede salvar la memoria de la humanidad. Sí, *salvar*, *memoria* y *humanidad*, en el sentido más literal de esas palabras.

Es muy posible que los bits acaben por imponerse definitivamente sobre el papel en un futuro todavía indeterminado, más posible aún que los formatos electrónicos y los análogos convivan como complementos los unos de los otros en un futuro cercano; pero lo que resulta verdaderamente imposible es que los bits, los metadatos, la arquitectura de la información, la web semántica, los html5, los ePub2 o los ePub3 reemplacen la impronta que la historia y la memoria colectiva de la humanidad han dejado en millones de folios y superficies de papel que hoy empiezan a desvanecerse por no ser preservados, o por no estar en la lista de espera o de priorización de lo que debe ser digitalizado.

Y hoy no hay mejor lugar para constatar esa cruda realidad que una Biblioteca Nacional latinoamericana: ante la escasez presupuestal para adquirir una tecnología que garantice la total y perfecta salvaguarda del patrimonio documental y bibliográfico de una nación, una institución con esa misión se ve en la situación paradójica de poner en cola el patrimonio que tal vez se pueda digitalizar y preservar en un futuro (o no), y simultáneamente producir contenidos digitales que satisfagan la demanda de millones de usuarios que los piden o necesitan en todos los rincones de una complicada geografía, mientras los formatos que produjo hace menos de diez años ya entraron también en la obsolescencia informática y tecnológica y pasan a una cola de espera, más atrás de la otra cola de espera.

¿Aliarse lo público con lo privado para salvar la memoria de la humanidad? ¿Dejar de innovar en la producción de contenidos para reforzar la preservación y la restauración digital? ¿Diseñar una política digital responsable que pueda garantizar la preservación democrática de los documentos en papel y al mismo tiempo la creación y gestión de los contenidos digitales?

Hoy más que nunca se necesita una ingeniería de la edición, la mezcla idónea de humanismo, ciencia y arte para salvar nuestra memoria, escribirla, reescribirla y hacerla circular.

Existimos en la sociedad de la información en donde la tecnología moldea las nuevas formas de conocimiento, cultura y sociedad (Manuel Castells, 1996). Esto propone un reto para la construcción de proyectos digitales cohesivos, responsables y relevantes en el mundo de las Humanidades Digitales.

En este orden de ideas, desde el diseño como una disciplina y una herramienta amplia y ramificada que propone y participa activamente en los procesos de estructuración y creación de proyectos en Humanidades Digitales, particularmente desde su capacidad para reunir y entrelazar varias facetas, derivando en soluciones creativas o diferentes, por medio metodologías articuladas como *Design Synthesis* que es el proceso de manipular, organizar y filtrar datos de un contexto para producir soluciones o conocimiento por medio de varios métodos (Jon Kolko, 2010), al igual que el *Design Thinking* que manifiesta el diseño desde un centro humano (IDEO, 2016).

No basta únicamente con ser un experto en esta área ni tener un equipo de trabajo para desarrollar proyectos multidisciplinares, es necesario crear puentes y conexiones para producir desde *la emoción*, por esto es vital empatizar humana, emocional, conceptual y digitalmente con cada etapa del proyecto: "La empatía digital es un proceso en el cual una persona puede analizar > reflexionar > proyectar > predecir > sentir mediante la comunicación con lo digital" (Friesem, 2105).

La experiencia de crear productos digitales en la Biblioteca Nacional de Colombia hace parte de la transformación en la forma en como se hacen y se muestran los productos digitales para un contexto latinoamericano y específicamente colombiano y particularmente desde uno de nuestros proyectos llamado Piedra y Cielo, un movimiento poético alternativo de finales de los años 30, integrado por Integrado por Jorge Rojas, Carlos Martín, Arturo Camacho Ramírez, Eduardo Carranza, Tomás Vargas Osorio, Gerardo Valencia y Darío Samper, en el que deciden publicar "su entrañable verdad", la poesía en sí misma sin mensajes políticos ni segundas intenciones.

Varias cosas fueron transversales en la construcción de Piedra y cielo, como la conformación de un equipo multidisciplinar compuesto por un coordinador, una investigadora, cuatro editores, una diseñadora, dos ingenieros y una estratega digital que empatizó con los contenidos y con la relevancia del proyecto. Así mismo, representa una forma de trabajo no piramidal, podríamos decir que *circular* en la distribución y liderazgo de tareas.

Referencias

Lange, Josua (2015). *Rise of the Digitized Public Intellectual: Death of the Professor in the Network Neutral Internet Age*. DOI 10.1007/s10780-014-9225-3.

Meza, Aurelio. "Decolonizar las humanidades digitales: cómo diseñar un repositorio digital de sur a norte". *Intervenciones en estudios culturales* volumen 4, 2017: 109-131. https://intervencioneseecc.files.wordpress.com/2017/07/n4_art07_meza.pdf.

Pons, Anaclet. *El desorden digital: guía para historiadores y humanistas*, Siglo XXI de España Editores, S.A., 2014.

Sánchez Hernampérez, Arsenio. "Paradigmas Conceptuales En Conservación". CoOL Documents. 23 de noviembre del 2008. 15 de enero del 2017. http://www.cool.conservation-us.org/byauth/hernampez/canarias.html

Vargas M., Maria (2017). *Las reparaciones 'de época' en libros medievales*. Tesis. Lleida, España.

Gil, Manuel y Joaquín Rodríguez. *El paradigma digital y sostenible del libro*, Trama editorial, Madrid, 2011.

Gil, Manuel y Martín Gómez. *Manual de edición. Guía para estos tiempos convulsos*, Cerlalc, 2016.

Digital Humanities meets Digital Cultural Heritage

Sander Münster
sander.muenster@tu-dresden.de
TU Dresden, Germany

Fulvio Rinaudo
fulvio.rinaudo@polito.it
I-Change, Politecnico di Torino, Italy

Rosa Tamborrino
rosa.tamborrino@polito.it
I-Change, Politecnico di Torino, Italy

Fabrizio Apollonio
fabrizio.apollonio@unibo.it
Department of Architecture, Alma Mater Studiorum – Università di Bologna, Italy

Marinos Ioannides
marinos.ioannides@cut.ac.cy
Digital Heritage Lab, Cyprus University of Technology, Limassol, Cyprus

Lisa Snyder
lma@ucla.edu
University of Southern California, United States of America

Introduction

As a main characteristic of digital humanities their objects are cultural heritage – according to Panofsky "'the records left [by] my man' — works of literature, art, architecture, and other products and traces of human intellectual labor" (Alvarado, 2011). While digital humanities focus on the application of digital technologies to support research in the humanities (c.f. e.g. Schreibman et al., 2004, Waters, 2013, Gibbs, 2011), the scholarly community on (digital) cultural heritage concentrates on tangible and intangible cultural heritage objects and their preservation, education and research (c.f. e.g. (UNESCO, 2003). Even if cultural heritage may be an agora (Ch'ng et al., 2013), there are some central topics adressed as

> Documentation (Geometric, Architectural, Historic etc.), involving 2D and/or 3D for archiving, for studies, for planning protective interventions etc.
> Accurate measurements, suitable for restoration actions, reconstructions, structural studies, protection etc.
> Monitoring of its state, involving recording deformations, state of materials, assessing pathology etc.
> Proper Management of its data for sustainability, risk management etc.
> Preservation possibilities specially suited for fragile objects (e.g. libraries etc.)
> Public Outreach, which involves visualization, dissemination, raising awareness of the public and many more (cited according to Georgopoulos, in print).

Concerning an application of digital methods, numerous associations were funded and a lively scholarly community has arisen during the last decades. One of the most renowned associations worldwide is the CIPA Heritage Documentation, an International Scientific Committee (ISC) of ICOMOS and ISPRS (International Society for Photogrammetry and Remote Sensing). It was founded in 1964 and has the responsibilities to keeping up with technology and ensuring its usefulness for cultural heritage conservation, education and dissemination (Münster, 2017a).

In addition, there are numerous conferences and journals focusing on digital cultural heritage, and various specific topics can be traced. Most prominent research areas are data acquisition and management, visualization or analysis. Recent topics are for instance unmanned airborne vehicle (UAV)-based 3D surveying technologies, augmented and virtual reality visualization, metadata and paradata standards for documentation or virtual museums (Münster, 2017b). In addition, data access seems to be a crucial point in terms of legal admission, annotation and semantics, technical as well as infrastructures.

Finally, there are some characteristics shared with digital humanities. A scholarly discourse is closely related to practical applications within projects (Münster, 2017b) and often takes place within cross-disciplinary cooperation.

Against this background, the idea of this panel is to sketch an outline of current research topics, challenges and practices in the field of digital cultural heritage. Our overarching interest is to initiate a fruitful discussion about communalities and differences between digital humanities and digital cultural heritage as well as to assess, to which extend they are two sides of the same medal.

Contributions

A scholarly community on digital cultural heritage

Within this paragraph, I will outline some characteristics of the scholarly field of digital heritage. In particular, a scientific community, usage-related challenges and demands as well as epistemic cultures will be mentioned.

A community perspective: *Who are stakeholders of cultural heritage? What are topics of scientific discourse?* We studied these aspects by analysing ~5000 publications in that field. Even if digital cultural heritage is a relatively new subject and a still emergent community, several protagonists, both individuals and institutions, are visible and have continuously been involved in an academic discourse since decades. Most of the researchers in the field of cultural heritage are Europeans and have a disciplinary background in the humanities and in

particular archaeology. As mentioned in the introduction, a discourse is primarily driven by technologies and in particular data acquisition and management, visualization or analysis.

A usage perspective: What are challenges and demands?

To examine current challenges and demands we conducted an online survey with more than 1000 participants. At a glance, money is named as biggest obstacle, including lacking funding opportunities for digital activities, costs for hardware and software as well as budget priorities for non-digital activities within organizations. Another big problem is a missing awareness such as a generation gap or digital divide in terms of digital literacy and frequency of use of digital tools as well as a general fear of or resistance to digital methods or – vice versa – missing awareness of limitations and requirements in the digital world. Moreover, the lack of competency and skills especially in technical domains is frequently named and vice versa would be seen as most important prospection. Finally, several participants see no obstacles for employing digital approaches in their organization.

An epistemic view: *How does digitization change research approaches in the field of cultural heritage? What marks a disciplinary culture of digital cultural heritage?*

For that research, we employed various in-depth research methods such as qualitative interviews and workshops. Similar to digital humanities, also for cultural heritage the use of digital technologies and approaches is currently estimated between another sub-domain of humanities studies and to "redefine traditional humanities scholarship through digital means" (Adams and Gunn, 2013). Beside the "technology-enabled" use of computational technologies to answer new types of research questions and the "technology-facilitated" employment of computational technologies as medium "for new research practices without necessarily transforming researchers' methods" (Long and Schonfeld, 2014, p. 42), a third type gets visible: "humanities-enabled" research as trading in humanities techniques to answer technology related questions like user-engagement, research ethic or to perform a comprehensive explanation of technical results. A key aspect of digital cultural heritage is cross-disciplinary cooperation. With regards to De Solla Price, digital cultural heritage could be seen as a mode 2 science (De Solla Price, 1963) with an emphasis on cross-disciplinary teamwork, the use of machines and a joint intellectual property. Consequently, a disciplinary culture is widely common to engineering but less to humanities. That may explain why humanities scholars report more frequently than engineers about the need to gain qualifications in order to enter the field of digital cultural heritage.

Space and time in Digital approach to Cultural Heritage

Digitisation changed the field of disciplines by creating a hybrid of their methodologies with ICTs. At the same times the complexity of new research, the quantity, and the quality of data available and the potential of new tools for managing and using data, ask for new cross-disciplinary approaches.

Digital Humanities and Digital Cultural Heritage could have same or different goals in using digital technologies for developing researches and interacting with a wide range of stakeholders, both in the aim of improving *keys of interpretation* of the Past (Digital Humanities) and/or of improving the understanding of the Past through the fruition of Cultural Heritage. A certain link between these methodologies has to be reached to avoid repetition of digitisation efforts and to reuse data in the aim of an open access of research outcomes.

The proposed contribution will show in which way Digital Urban History and Digital Cultural Heritage meet, by adopting digital strategies of two experts, in history and in geomatics, to obtain research results useful *to document* a Cultural Heritage asset, *to give information* to assess cultural meanings and for the subsequent actions of valorisation, conservation, restoration, and at the same times *to implement* the knowledge of historical assets.

The proposed interdisciplinary approach shares the focus on 'spatialising' historical information that have different significance in the two disciplines: space is a fact for geomatics as well as times is a fact for historians. Essential relationships between history and the space emerged since 1970s (Lefebvre, 1991) unless digital history emerged in the late 1990s (Brügger, 2010) and recent awareness confirms the role played by technologies for improving this trend by visualisation of relationships between space and time (Bodenhamer, 2013).

The contribution of Digital Urban History to Cultural Heritage documentation is the description of urban landscape changes by assessing the modifications and their possible causes, by linking material changes to human and natural actions such as political decisions, economic factors, earthquakes, climate changes, etc.

The study of those phenomena requires the use and interpretation of historical sources, the mutual validation of the quality of the used data and the filling of the "information gaps" that sometimes the archive contents do not allow to solve properly without an historical interpretation. Usually the digitisation, required to develop an historical research, has to put in evidence the spatial meaning of the sources in term of cartographic localisation and 3D modelling at the needed scales. Those solutions have to be shared with experts to avoid misunderstanding of the historical sources and their interpretation.

Digital Cultural Heritage experts need not only the results of the researches but also the used primary data

in digital sharable formats to be distributed among the experts involved in documentation, management, valorisation, and restoration design.

Digital technology changed the way to collect, share, and manage information. Digital technologies are instruments and, as instruments, they do not have to overpass the aim of the different research. They are tools and not the focus of the Humanities and the Cultural Heritage.

A challenge: formalizing semantic knowledge and new forms of representation

The availability of new and more effective digital technologies, applied to Cultural Heritage studies, does not represent only the move from analogue to digital source material (Brügger, 2016), but obviously, other factors also play a role (Svensson, 2011, p. 42). Digital technologies introduce the possibility of interchangeable media able to offer multiple nodes of access to a given term or object, and enable a multidimensional approach to knowledge on several levels (Stefani et al., 2013).

Even though the Digital Humanities open up an array of possibilities either for doing what was previously done in new ways, or for rethinking well-known practices of the humanities, for instance by integrating software-supported methods and by using digital research infrastructures ((Brügger, 2016)), the inescapable problem remains the need to make retrievable the documentation process (Münster et al., 2016) behind the production of any digitized, born-digital, and reborn-digital material, as well as that concerning the cultural asset and the preservation of the data during the whole lifecycle of CH. Therefore, besides spatial modeling and its representation the digital humanities, as well as digital heritage, open to the temporal dimension (diachronic and synchronic) - which allows to know artifact not only in its evolution and transformation during its life cycle, but also through its analysis - and to the extrapolation of various possible models from fragmentary pieces of information (remains), which imply of portraying uncertainty in a digital imagery, and defining an inventory of new forms of representation for indicating distinctions between known and projected or imagined evidence.

Thanks to the development of the ICT technologies and infrastructures, and their application to research on architectural and urban cultural heritage, the semantic virtual environment platforms can become the engine for dissemination of different and customized level of knowledge (Apollonio, in print).

According to theoretical humanities approaches to knowledge as knowing, observer dependent, emergent, and process-driven rather than entity-defined, next challenges will be focused on defining appropriate methodology able to ensure, through descriptive metadata jointly the connection of the data-sources and the knowledge processes involved in creating digital objects (knowledge provenance by means of semantic database) ((Brügger, 2016); (Bruseker G. et al., 2015)), the possibility to modeling the human processes of understanding and interpreting the digitized sources (paradata) in order to produce the 3D digital outputs, semantically enriched.

As humanistic methods are necessarily probabilistic rather than deterministic, performative rather than declarative, more advanced models of simulation than the literal techniques of current visualization will need to be designed in order to incorporate these methods within Digital Humanities.

Even though web-based ICT systems can offer increasingly updated tools for the Cultural Heritage management, providing a smart 3D navigation system, always accessible to the users via internet, we need to define standardized methodology of source or reality-based 3D reconstruction of tangible Cultural Heritage, able to ensure, throughout a (i) a transparent reconstruction workflow, (ii) 3D modeling qualified by readable quality/properties, (iii) a proper semantic structure of the 3D digital model, and (iv) a retrievable knowledge reconstruction and formalization process (Apollonio, in print), the interoperability of data sets by referring to recognized standard reference ontologies.

The challenge, as hoped by Johanna Drucker (Drucker, 2012) due to shifting humanistic study to a humanistically informed theory of the making of technology, consists in developing a new web philological toolbox (Brügger, 2016) that can help the scholar gain as much information as possible about the object of study. This approach, in fact, should be able to develop applicable working techniques, to define valid strategies, and to apply classifications useful to supporting scientific work besides the conveyance of knowledge to its extraction, elicitation and representation.

The Virtual Multimodal Museum Network (ViMM)

Virtual Multimodal Museum (ViMM) is a major Coordination and Support Action across the field of Virtual Museums (VM), within the overall context of European policy and practice on Digital Cultural Heritage (DCH), funded under the Horizon 2020 program of the European Union. A highly-expert seven (7) partner consortium, coordinated by Cyprus University of Technology (CUT) leverages the support of a unique and powerful Advisory Group, consisting of many of the Europe and the world's leading public and private sector organisations in the field, to define and support high quality policies, strategic and day-to-day decision making, the utilisation of breakthrough technological developments such as VR/AR and to nurture an evidence-based view of growth and development impacted by VM, supported by a set of case studies in culturally-rich regions of South Europe affected by economic recession.

Its work will be founded on building a consensual framework directly involving Europe's leading VM decision-makers and practitioners in defining and resolving existing issues spread across 7 interlinked Thematic Areas ('the 7 Ds'): Definitions – Directions – Documentation – Dimensions – Demand- Discovery - Decisions and aims for wide-reaching stakeholder participation and very high visibility. The latter will be achieved through organisation of key events at policy and practitioner/ stakeholder levels, extensive use of the media, and by the introduction of an interactive and wide-reaching communication platform, deploying social media and novel approaches to enable focused debate by all interested parties, supported by access to representations of excellence and a decision-support tool for stakeholders.

An initially broad and open approach will be refined through a process of definition, consolidation and resolution activities to arrive at a clear Manifesto and Roadmap for Action on VM/DCH, validated at a final ViMM international conference. Measurable impacts will be achieved on the role and capability of DCH – and VM in particular – to meet their enormous potential in society and the economy.

References

Adams, J. L. & Gunn, K. B. 2013. Keeping Up With...Digital Humanities. *American Library Association* [Online], April 5, 2013.

Alvarado, R. 2011. The Digital Humanities Situation. *The Transducer,* May 11th, 2011.

APOLLONIO, F. I. in print. The production of 3D Digital Archives and the methodologies for digitally supporting research in architectural and urban cultural heritage. CIPA's Perspectives on Cultural Heritage. *In:* Münster, S., Friedrichs, K., Niebling, F. & Seidel-Grzinska, A. (eds.) *Urban Heritage in the Age of Digital Libraries.* Springer.

Bodenhamer, D. J. 2013. Beyond GIS: Geospatial Technologies and the Future of History. *In:* VON LÜNEN, A. & TRAVIS, C. (eds.) *History and GIS: Epistemologies, Considerations and Reflections.* Dordrecht: Springer Netherlands.

Brügger, N. 2010. *Web History,* New York.

Brügger, N. 2016. Digital Humanities in the 21st Century: Digital Material as a Driving Force. *Digital humanities quarterly [Online],* 10.

Bruseker G., Guillem, A. & Carboni, N. 2015. Semantically documenting virtual reconstruction: building a path to knowledge provenance. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* II-5/W3, 33-40.

Ch'ng, E., Gaffney, V. & Chapman, H. 2013. *Visual Heritage in the Digital Age,* London.

De Solla Price, D. 1963. *Little Science - Big Science,* New York, Columbia Univ. Press.

Drucker, J. 2012. Humanistic Theory and Digital Scholarship. Debates in the Digital Humanities. Part II. Theorizing the Digital Humanities [Online], Available: http://dhdebates.gc.cuny.edu/debates/text/34 [Accessed: November 22, 2017].

Georgopoulos, A. in print. CIPA's Perspectives on Cultural Heritage. *In:* MÜNSTER, S., FRIEDRICHS, K., NIEBLING, F. & SEIDEL-GRZINSKA, A. (eds.) *Urban Heritage in the Age of Digital Libraries.* Springer.

Gibbs, F. W. 2011. *Digital humanities definitions by type* [Online]. Available: https://moodle.ucl.ac.uk/course/view.php?id=11859&section=3 [Accessed 19 Sept. 2011].

Lefebvre, H. 1991. *The Production of Space,* Cambridge, Blackwell.

Long, M. P. & Schonfeld, R. C. 2014. *Supporting the Changing Research Practices of Art Historians*, Ithaka S+R.

Münster, S. 2017a. Employing bibliometric methods to identify a community, topics and protagonists of digital 3D reconstruction in the humanities. *In:* STERZER, W. (ed.) *iConference 2017 Proceedings.* iSchools.

Münster, S. 2017b. A Survey on Topics, Researchers and Cultures in the Field of Digital Heritage. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* IV-2/W2, 157-162.

Münster, S., Hegel, W. & Kröber, C. 2016. A classification model for digital reconstruction in context of humanities research. *In:* Münster, S., Pfarr-Harfst, M., Kuroczyński, P. & Ioannides, M. (eds.) *3D Research Challenges in Cultural Heritage II.* Cham: Springer LNCS.

Schreibman, S., Siemens, R. & Unsworth, J. 2004. *A Companion to Digital Humanities,* Oxford, Blackwell.

Stefani, C., Busayarat, C., Lom-Bardo, J. & De Luca, L. 2013. A web platform for the consultation of spatialized and semantically enriched iconographic sources on cultural heritage buildings. *International Journal on Computing and Cultural Heritage,* 6.

Svensson, P. 2011. The digital humanities as a humanities project. *Arts & Humanities in Higher Education,* 11, 42-60.

UNESCO 2003. *Charta zur Bewahrung des Digitalen Kulturerbes, verabschiedet von der 32. UNESCO-Generalkonferenz am 17. Oktober 2003 in Paris*.

Waters, D. J. 2013. An overview of the digital humanities. *Research Library Issues,* 284, 3-22.

Digital Chicago: #DH As A Bridge To A City's Past

Emily Mace
mace@lakeforest.edu
Lake Forest College, United States of America

Rebecca Graff
graff@lakeforest.edu
Lake Forest College, United States of America

Richard Pettengill
pettengi@lakeforest.edu
Lake Forest College, United States of America

Desmond Odugu
odugu@lakeforest.edu
Lake Forest College, United States of America

Benjamin Zeller
zeller@lakeforest.edu
Lake Forest College, United States of America

Panel Overview

The Mellon-funded grant project "Digital Chicago: Unearthing History and Culture" represents the fruits of three years of collaboration and digital humanities learning by faculty and student researchers at Lake Forest College, drawing on work by faculty from across the humanities and humanistic social sciences. Our work has centered on a particular theme: the history of Chicago, as explored through diverse digital humanities approaches and tools that tell stories of Chicago's forgotten or at-risk past. Our full project draws on work done in several disciplines, and through this approach, the city of Chicago becomes a bridge between disciplines, traversed by means of digital humanities tools, even as Chicago's river is itself connected by many bridges.

We have selected several representative projects and one overview presentation from the Digital Chicago's collaboration to share with attendees at the ADHO 2018 conference. These specific projects include a mapped timeline of racial restrictive housing covenants in Chicago's Cook County; a set of 360° immersive, educational tours of Chicagoland sacred spaces, a history of Shakespeare in Chicago as reflective of the city's cultural development, and a map reflecting points of origin for artifacts from an archaeological dig.

A final presentation links together the themes of the Digital Chicago project as a whole, and will emphasize scaffolding to meet the capabilities the small liberal arts college environment, which often presents challenges in terms of staffing, student support, and scope of research.

Restricted Chicago

Desmond Odugu

This presentation exploresthe history of educational inequities in Chicago through the lens of Chicago and Cook County's history of discriminatory housing practices such as racial restrictive covenants.This presentation will focus on the archival and digital work that undergirds the project, as well as research conclusions drawn from the project.

The digital project represents the results of archival research into official housing records, which provided ample evidence of restrictive practices. Student research assistants, working with the faculty member, located over 200 affected "subdivisions" on a Neatline map, aligning those locations with a Neatline timeline to indicate when each restriction was enacted and eventually removed. Accessible for a broad scholarly and general audience, the timeline and map permits the user to visualize the changing shape of Cook County's housing practices during the twentieth century.

These restrictive housing policies had a strong effect on the educational disparities that still characterize schools in Chicago and Cook County. By aligning subdivisions with United States government census tracts, this project reveals the connections between racial restrictive covenants and their consequences in terms of socioeconomic status, school rankings, and educational outcomes. The presenter, a professor of education, will also discuss how the fruits of digital humanities research such as this can be used to inform local citizens about their own history.

Sacred Spaces in 360°

Ben Zeller

This panel presentation features a project creating educational virtual reality walkthroughs of three churches in the Chicago area, all of which have historic, architectural, and religious value. The tours are intended to be used by high school or undergraduate students in lieu of in-person field trips to the sites, and the paper addresses the process of creating tours using tools such as the Ricoh Theta S camera, Panotour software by Kolor, and student research assistance.

Two of the tours utilize historical digital images overlaid on digital images of the contemporary structure. This method proved particularly relevant in the case of a Baptist church and former synagogue which burned in 2006. Photographs of what is now the burned-out shell of a historically significant building becomes a device for travel through time.

In particular, this project reveals the potential for readily available tools and software to create informative, educational, and scholarly tours of historic sites. Because the tours are optimized for desktop as well as smartphone (and therefore Google Cardboard, etc.), they translate easily to classroom experiences, and offer users a way of touring sites that might be otherwise inaccessible.

The presenter will also address the incorporation of 360° digitalization projects into the undergraduate curriculum. After an initial pilot project, the instructor extended the creation of these tours into a class about sacred spaces, which presents some pedagogical advantages and disadvantages, but nonetheless serves multiple learning goals and their assessment.

Mapping Historical Consumerism through Urban Archaeology

Rebecca Graff

This presentation reveals how historical archaeology coupled with digital tools can disclose historic patterns of consumption and consumerism in a mixed class neighborhood in of turn-of-the-twentieth-century Chicago. The project locates the manufacturing or point of sale origins—Chicago and worldwide—of artifacts excavated from the area surrounding a well-known historic Chicago home, using historical and anthropological methods to identify the origins of specific items. The student-faculty team created two digital maps that trace the fragments to their points of origin, alongside images of the historical advertising and other textual records that help to identify an item.

The Charnley-Persky House was designed by architect Louis Sullivan, with architect Frank Lloyd Wright serving as Sullivan's chief assistant. Completed in 1892, it offers an important example of American architecture, and the House is now a museum as well as the headquarters for the Society of Architectural Historians (SAH). Extensive renovation and construction work at the site revealed a rich deposit of nineteenth-century refuse which excavated in 2010 and 2015.

The digital maps allow for bridges between the architectural scholarship about the Charnley-Persky House to that which was found archaeologically. By clicking through the digital maps of products eaten, worn, used, and ultimately discarded adjacent to the Charnley House, site visitors can engage with the globalizing tastes and trends of American urban dwellers at the turn of the twentieth century.

Shakespeare in Chicago

Richard Pettengill

Shakespeare's plays have been an integral part of Chicago's history ever since the city's incorporation in 1837. This project traces the history of Shakespeare in Chicago as part and parcel of the history of the city itself, with a particular focus on the city's cultural development.

The project uses several digital tools to present this historical narrative: first, a timeline suggests the ubiquity of Shakespeare on the American frontier, detailing the shift from "low-brow" to "high-brow" culture (as described by Lawrence Levine) by the end of the nineteenth century, and leading to Chicago's world-renowned twentieth and twenty-first-century theater scene.

Second, photographic time sliders suggest the dramatic changes that have taken place in Chicago's built theatrical environment. These demonstrations rely on historic imagery and modern digital photography to overlay images of Chicago's theaters of the past onto the hustle and bustle of their contemporary locations.

Finally, a map plots the locations of Chicago's now-vanished theaters on the contemporary streetscape of the city, highlighting which areas of Chicago's former incarnations brought stagecraft to the city's scene.

Taken together, this collection of digital tools informs both a public and scholarly audience about the forgotten history of Shakespeare as representative of the city's theater scene. The three tools are eminently accessible, and will appeal particularly to a non-scholarly audience. They offer a way of engaging with history that bridges that gap between public and scholarly users of the site.

Building a Small-Scale Bridge via #DH

Emily Mace

The Mellon-funded grant project "Digital Chicago: Unearthing History and Culture" represents the fruit of three years of collaboration and digital humanities learning by faculty and student researchers at a small liberal arts college, drawing on work from across the humanities and humanistic social sciences.

Our work has centered on a particular theme: the history of Chicago, as explored through diverse #DH approaches and tools that tell stories of Chicago's forgotten or at-risk past. The project as a whole supported digital humanities work in many disciplines––including theater, English, religious studies, anthropology, music, communications, history, education, and politics––and with this cross-disciplinary approach, the city of Chicago becomes a bridge between disciplines, traversed by means of digital humanities tools, even as Chicago's river is itself connected by many bridges.

Although the Digital Chicago suite of projects is Mellon-funded, we nevertheless faced the challenges of doing digital humanities on the smaller scale of the liberal arts college. To this end, this final presentation will reflect on the value of scaffolding our understanding of "digital humanities" to meet the capabilities of individual faculty as well as of the institution. Our projects relied on undergraduate student research assistance, usually with one research assistant per faculty member. We relied, also, on pre-existing digital humanities tools as vehicles for our projects, and focused on finding the best tool to present each project to a broad audience of scholars and non-scholars alike.

Taken together, this approach outlines a scope of work attainable at smaller schools which also bridges the gap between scholarly audience and the general public, through the creation of an engaging and broadly useful digital humanities website.

Bridging Between The Spaces: Cultural Representation Within Digital Collaboration and Production

Stephanie Mahnke
mahnkes1@msu.edu
Michigan State University, United States of America

Shewonda Leger
legershe@msu.edu
Michigan State University, United States of America

Suban Nur Cooley
cooleys7@msu.edu
Michigan State University, United States of America

Victor Del Hierro
vjdel@utep.edu
University of Texas, El Paso

Laura Gonzales
ljgonzales3@utep.edu
University of Texas, El Paso

Since the work of digital humanities has illuminated rhetoricians to technology as a form of literacy and means of representing cultural community advocacy in dynamic ways (Enoch & Gold, 2013), cultural rhetoricians have transferred their discipline's concerns of ethical methodology to community-engaged digital production (Ridolfo, 2013; Smith, 2016). One of the questions we ask is how do we read or create digital projects and platforms that engage with cultural groups and stakeholder communities on their own terms? As a developing subject area in Rhetoric and Writing Studies, the confluence of digital and cultural rhetorics has yet to produce a robust set of research around ethical strategies for community-oriented and participatory digital work with underrepresented groups. Scholars in DH have already begun to lead the way on developing ethical research methods using digital tools for research (Underberg and Zorn, 2013; Gubrium and Harper, 2013; Gubrium, Harper, and Otañez, 2016). Furthermore, as scholars who represent the communities we are engaged in we hope to follow the lead of community engaged projects like the Cree Cultural Institute, The Houston Hip Hop Archive, and the Densho Digital Repository. This panel samples digital work from Somali, Haitian, Filipinx, and Latinx communities to explore how digital production can be re-conceptualized and utilizedto accommodate global epistemologies. Conference-goers may walk away with knowledge of digital strategies cultural groups use to navigate complicated positionalities, methodological considerations for cultural collaboration, and why such digital representation and responsible researcher reflexivity are crucial for digital cultural heritage work.

The Collective in the Individual: Digital Collaboration and the Filipinx Community

Presenter 1 will discuss Filipinx-American communities' production of digital content through the negotiation of collective Filipinx consciousness and western audiences. Consideration of collective consciousness as an intrinsic Filipinx value has shaped how researchers and practitioners collaborate with Filipinx communities (Strobel, 1997; Pe-Pua & Protacio-Marcelino, 2000). The Filipinx Indigenous worldview and ontology draws from the cultural and spiritual connection to one's fellow beings (*kapwa*). Other concepts such as *pakikiramdam* (shared perception) and *pakikipagkapwa-tao* (shared identity) reinforce the idea that knowledge is created by communal interaction. With these particular core Filipinx values at center, the process toward digital representation becomes complicated by the difficulty of achieving community validation while also needing visible community advocacy amidst western audiences. By reflecting on her collaboration with a Filipinx cultural center and Filipino American National Historical Society (FANHS) in the creation of a website and digital map around underrepresented Filipinx history and heritage in the Midwest, the presenter explores the distinct methodological challenges associated with ethical cultural heritage and digital work. In this presentation, she introduces the application of cultural rhetorics methodologies as a means of respecting the views, needs, and culture of the Filipinx community during digital collaboration, and demonstrates the unique considerations that manifest during digital production, as a result. Gathering and adhering to cultural input around design, content, and functionality transforms (and in many ways, slows) feedback loops to accommodate key instances and spaces of cultural negotiation and participation. Based on the collaborative experiences around these two projects, the presentation hopes to then propose strategies for working with underrepresented cultural groups, as well as generate discussion and thought around the cultural sensitivities involved with digital humanities and cultural heritage work.

Multimodal Counterstories: The Circulation of Cultural Knowledges Through a Haitian Lens

In a world where colonial narratives are dominating digital and media spaces, Haitian women continue to find ways to challenge epistemic oppressions in the ways they compose. Often this means (re)claiming practices towards creating, sharing, and transferring knowledge. Therefore, presenter 2 will discuss how multimodal composition practices make space to produce and share knowledge(s), where Haitian women are able to compose, construct, build, and make meaning in various forms that

embrace cultural identity and practices embedded in making and knowledge production.

We can understand multimodality composition as "communication using multiple modes that work purposely to create meaning" (Lutkewitte 2). Looking at Lutkewitte's definition of what multimodality composition aims to do, we can see multimodal composition as a way of disrupting colonial practices that generate epistemic exclusion, which limits and/or undermines the ways in which marginalized communities create meaning. Considering the importance of multimodal composing, finding ways to address epistemic exclusion, and understanding the lived experiences of Haitian women, presenter 2 will focus on the ways multimodal composing takes place in film production and how these ways of composing connect to her own identities as a Haitian American woman, as well as, the identities of other Haitian women.

In dominant practices of film representations, colonial powers more often have the upper hand, because of their control of media and digital spaces, funding, and resources for composing film. But, if Haitian women produce films and take part in films around their lived experiences, they can take some of that agency. By taking that agency Haitian women will interrogate the colonial gaze through films that talk back and reclaim colonial representations of Haitian lived experiences.

Further, presenter 2 will discuss how positions from which Haitian women speak or write in film production or any form of visual or digital representation is important, because of how these films will be circulated and interpreted through media spaces and the connection that occurs between film and viewer during circulation and sharing. When Haitian women produce and/or are part of the production process of filmmaking, they agitate stereotypes of former visual and digital depictions around Haitian women's lived experiences. Presenter 2 will further discuss how Womanist and Black feminist filmmaking practices are primary elements used by Haitian women filmmakers, because the practices involved in Womanist and Black feminist filmmaking practices contribute to the production of films centered around Haitian women's lived experiences. Film production then becomes an act of activism because the narratives that Haitian women want to tell are the ones that circulate—and this is important towards digital work that can be reconceptualized and utilized to accommodate global epistemologies.

Reclaiming Stories: Somali Women Using Social Media to Build Gendered Counternarratives

In the West, complexities of identity emerge for Somali womyn who are struggling to maintain a semblance of Somali nationalism, cultural continuity, and self-identity, while also navigating trauma histories and the newness of assimilation in geographic communities outside of Somalia – and the impact all of this has on identity formation. These emergences then develop counternarratives/counterstories and constructions within the Somali refugee and migrant community, as they slowly become diaspora beings with diaspora identities.

As Langellier explains, for those in the diasporic Somali community, identity formation is an embodied and situated dialogue that is enveloped by discourses about refugees, Somalis, and Muslims. Thus, "identity as an unfolding performative accomplishment challenges static and essentialized notions of differences and thus joins postmodern trends that emphasize hybridized, transnational, in-between, and other 'third space' conceptualizations." (p. 67, Langellier, 2010).

Presenter 3 will discuss how these 'third space' conceptualizations of identity differ between what womyn in the Somali diaspora do within their heritage communities, versus their public engagement in online social media spaces.

To define the term diaspora, I use Furusa, Little, and Vasquez's concept of Diaspora – a dynamic process; one that aims to explain the actions of people of "African, Asian, and Latin American descent who have been forcibly or voluntarily dispersed throughout the world" (1). The authors also accept the term's complexities – its cultural dynamism involving "substantive intercultural exchanges between peoples from different ethnic groups. It frequently requires the reconstitution or reconstruction of a culture and society by people's uprooted from homes that borrow and adopt new cultural elements as the plaster to patch up the fissures in their cultural foundation" (3).

Presenter 3 will discuss how these fissures in cultural foundations form an opening and opportunity for diasporic individuals to interpret and understand the world and its complexities through their lived experiences; one that does not reside solely in a particular culture, place, or space – but blends them in ways that enable for the existence of a diasporic orientation – and how Somali womyn display these blends on social platforms and in their communities.

Presenter 3 will examine how womyn from the North American Somali diaspora, with a high-profile digital presence, navigate and negotiate these fissures and fusions of identity to express and proclaim who they are, both against their 'home' communities and norms of the West in public digital spaces.

Designing a Translation and User-Experience Research Center for Technology Innovation on the Mexico/US Borderland

This paper will consist of a case study that outlines the design and development of a research center focused on multilingual technology innovation. As two faculty members teaching in El Paso, Texas on the Mexico/US border, we will present our framework for establishing a collabo-

rative research center that facilitates the design of multilingual tools and technologies (e.g., websites, software, applications) for a wide range of organizations. These organizations include local hospitals and government agencies, nonprofit organizations, and schools, all of which are located in the El Paso, Texas and Cuidad Juarez, Chihuahua borderland. By collaborating with local organizations and by training students to design, test, and disseminate technologies in multiple languages (including but not limited to Spanish and English), this research center is a site of multilingual technology innovation that lead to smart learning, both in and outside of the classroom.

We designed Sites of Translation User-Experience Research Center (http://www.utep.edu/liberalarts/translationux/about/index.html) as a nonprofit, interdisciplinary, community and University-driven resource that supports student development and local community organizations. Developed as a partnership among community organizations, academic researchers, and technology industry professionals, *Sites of Translation User- Experience Research Center* is envisioned as the place where social-justice oriented organizations come to request help in creating and disseminating their bi– or multilingual content (e.g., websites, web applications, informational tools) aiming to meet the needs and highlight the assets of linguistically diverse users. Local businesses and organizations come to this research center to request help in creating and disseminating their bi– or multilingual content. As faculty members, we pair local organizations with students and researchers who then help to design and test and translate tools and technologies. This collaboration results in the development of tools and technologies that are useful in multiple languages. In this presentation, we outline how the development of this research center has helped us build partnerships between University campuses and local organizations to provide valuable professional experiences for students interested in user-experience and technology design. We will then share implications and strategies for other faculty seeking to build similar initiatives to foster multilingual technology design at their institutions and in their communities.

References

Enoch J. & Gold, D. (2013). Introduction: Seizing the Methodological Moment: The Digital Humanities and Historiography in Rhetoric and Composition,*College English,*76(2): 105-114.

Gubrium, A., & Harper, K. (2013). *Participatory visual and digital methods*. Left Coast Press.

Gubrium, A., Harper, K., & Otañez, M. (Eds.). (2016). *Participatory visual and digital research in action*. Routledge.

Langellier, Kristin. (2010). Performing Somali Identity in the Diaspora.*Cultural Studies*, 24(1): 66-94.

Little, William A., et al.(2006). The Borders in all of Us: New Approaches to Global Diasporic Societies. Northridge, CA: New World African Press.

Lutkewitte, Claire, (Ed.). (2013). Introduction. *Multimodal Composition: A Critical Sourcebook*. Bedford/St. Martin's.

Pe-Pua, R. and Protacio-Marcelino, E. A. (2000). Sikolohiyang pilipino (filipino psychology): A legacy of Virgilio G. Enriquez.*Asian Journal of Social Psychology,3*(1): 49-71.

Ridolfo, J. (2013). Delivering Textual Diaspora: Building Digital Cultural Repositories as Rhetoric Research. *College English,*76(2): 136.

Smith, K.G. (2016). Negotiating community literacy practice: Public memory work and the Boston Marathon Bombing Digital Archive. *Computers and Composition,*40: 115-130.

Strobel, L.M. (1997). Coming full circle: Narratives of decolonization among post-1965 Filipino Americans. In M.P.P. Root (Ed.) *Filipino Americans: Transformation and Identity*. Sage Publications, pp. 62-79.

Underberg, N. M., & Zorn, E. (2013). Digital ethnography: Anthropology, narrative, and new media. University of Texas Press.

Pensar filosóficamente las humanidades digitales

Marat Ocampo Gutiérrez de Velasco
eljabberwocky@gmail.com
Universidad Nacional Autónoma de México, México

Francisco Barrón Tovar
barronar@gmail.com
Universidad Nacional Autónoma de México, México

Ana María Guzmán Olmos
sositap@gmail.com
Universidad Nacional Autónoma de México, México

Sandra Reyes Álvarez
sandroide.filos@gmail.com
Universidad Nacional Autónoma de México, México

Elena León Magaña
leonelenna@gmail.com
Universidad Nacional Autónoma de México, México

Ethel Rueda Hernández
alzilei@gmail.com
Universidad Nacional Autónoma de México, México

Cuando hablamos de pensar filosóficamente las humanidades digitales, nos referimos a la producción de un tipo de aproximación crítica que revisa la configuración tecnológica, así como la producción discursiva que la acompaña. Dicha configuración en algún momento se pensó

gozaba de cierta neutralidad, pero el análisis de la tecnología ha conducido, cada vez más, a reconocer que la producción de estas herramientas tiene supuestos teóricos que, a partir de un examen filosófico, permiten encontrar parte de sus problemas y posibilidades. Por otra parte, si los discursos se reconocen técnicos pueden contribuir a fortalecer la tradición de investigación dirigida por la lectura y escritura de textos, a partir de las posibilidades de visualización, producción de datos, acumulación y procesamiento de información. Pensar filosóficamente las humanidades digitales se trata no sólo de una oportunidad de hacer consciente el carácter técnico del trabajo académico en las humanidades, sino de mostrar un desplazamiento en la producción del saber humanístico, y el sentido que éste tiene, a partir de la intervención tecnológica.

Nuestra aproximación busca hacer un examen crítico de estas prácticas, como ya lo hacen algunos estudios de teoría de las humanidades digitales, las ontologías tecnológicas, los estudios de retórica computacional y otros campos. Este análisis, por otro lado, nos permite reflexionar sobre las técnicas de producción del pensamiento filosófico. A partir de este examen queremos mostrar las dificultades prácticas y discursivas que surgen dentro de las humanidades digitales, que requieren pensarse para delinear sus discursos y mejorar sus prácticas.

Distant reading y el ejercicio tecnológico de la filosofía

Francisco Barrón

Embistamos al ejercicio de la filosofía, tal como lo conocemos y llevamos a cabo aún el día de hoy, con un ejercicio de cuestionamiento tecnológico de las condiciones de su producción. Las Humanidades Digitales no sólo son la oportunidad de discutir los procedimientos académicos de producción de saber humanístico, permiten formular la pregunta sobre una modificación tecnológica del ejercicio del pensamiento.

El actual ejercicio de la filosofía comienza en unas épocas tecnológicas bien determinadas y anteriores a la nuestra. Las técnicas y tecnologías del ejercicio de la filosofía pertenecen a formas de producción artesanales. Sin embargo, vivimos la experiencia del malfuncionamiento, la alteración de la figura, las técnicas y las funciones de lo que llamamos el ejercicio del pensamiento. No es para nadie una noticia que el mecanismo político-académico moderno del pensamiento, su sentido, junto con sus prácticas, sus instituciones y las figuras que la encarnaban, parece ya no funcionar adecuadamente para ciertos acontecimientos contemporáneos. Sería arduo hacer el recuento de las condiciones que han hecho eso. La tecnología parece tener algo que ver también con ello, junto con políticas de muerte, transformación de las formas de escritura, caídas de instituciones de distribución del saber, alteración de los modos de reproducción de la subsistencia, y muchos más acontecimientos a enumerar.

El ejercicio de la filosofía, tal como se nos ha heredado en una tradición ilustrada-romántica, europea, parece ya no funcionar para nosotros aquí y ahora. Lo tecnológico, sobre todo su avatar digital contemporáneo, tiene efectos radicales en la producción y organización del trabajo intelectual. Hasta el día de hoy, en el ejercicio de la filosofía se reniega de su carácter técnico. Pero, quizás más nos valdría aventurarnos en la experiencia tecnológica del ejercicio de la filosofía. Más valdría hacer experimentos con tecnologías y metodologías de las Humanidades Digitales para saber, al menos, si se puede ejercer el pensamiento filosófico en estas condiciones digitales de producción de saber.

Franco Moretti (*Literatura vista desde lejos*, 2005 y *Lectura distante*, 2013) ha discutido, al analizar enormes conjuntos de datos para estudiar la literatura, supuestos teológicos de los actos de lectura y de la crítica literaria. Lo ha hecho así para producir saber y crítica de la literatura usando tecnologías computacionales. Quizás sería posible discutir si la *distant reading* nos permitiría ejercer tecnológicamente el pensamiento filosófico.

Frankenstein y Cthulhu. Crítica en las (in)humanidades digitales

Ana María Guzmán Olmos

En *Where Is Cultural Criticism in the Digital Humanities?* (2012) Alan Liu cuestionó el potencial crítico de las humanidades digitales y su perspectiva a futuro como disciplina. El texto se sitúa dentro de los clásicos que cuestionan la idea de que los procesos de digitalización del contenido humanístico son suficientes para generar saber, un saber mediado por lo digital. Sirviéndose del binomio que compone el concepto "humanidades digitales", en su diagnóstico Liu hace énfasis en la crítica como centro de la producción humanística. Según la tesis de Liu, si hay algún potencial a futuro en las humanidades digitales, este debería ser buscado en la noción de crítica; en mi presentación voy a discutir esta idea. Voy a señalar que el modelo de la crítica está vinculado a un mecanismo de representación que nubla la posibilidad de pensar la diferencia. La crítica, según el modelo presentado por Liu, invisibiliza la diferencia.

Para discutir el modelo crítico del proyecto humanístico seguiré el texto *Loving the Alien* (2017), en el cual Lisa Blackman se sirve de la categoría *alien* como herramienta para pensar aquello que no es comprensible dentro del campo de lo humano. Lo humano se construye como una categoría excluyente constituida mediante marcas de género, raza o modelos de corporalidad; este mecanismo produce un límite donde el exterior queda atravesado por condiciones de vulnerabilidad y explotación. Mediante las figuras de lo inhumano Blackman discute la idea de una política no centrada en el cuerpo. Siguiendo a Blackman voy a explorar el concepto del "inhumanismo de lo hu-

mano y el humanismo de lo inhumano" (Blackman, 2017), esto en el campo abierto por las humanidades digitales.

Mi tesis es que la noción de crítica en las humanidades, sean estas digitales o no, depende de hacer visible la multiplicidad en un cuadro homogéneo que permite a los sujetos reflexionar sobre la unidad; este es el modelo de la representación. Dicho modelo puede ser observado en diversas metodologías usadas en las humanidades digitales, las cuales voy a ejemplificar con el método de visualización de datos. Frente a la visualización de datos, que traduce el lenguaje computacional en lenguaje "humano", me interesa pensar, siguiendo a Luciana Parisi (*Reprogramming Decisionism*, 2017), el razonamiento de las máquinas que se elabora en procesos como la recombinación algorítmica autónoma, o el aprendizaje que hacen los algoritmos de otros algoritmos, y que dan lugar a espacios de indeterminación. Estos factores del razonamiento maquínico permiten elaborar un concepto de diferencia que depende de la repetición y el *performance* algorítmico, y no de la unificación. El potencial de las humanidades digitales estaría, en este sentido, ligado a su condición inhumana.

Filosofar la tecnología: un camino educativo

Sandra Reyes Alvarez

Los enfoques educativos actuales han enfatizado la importancia de la educación digital. Incluso es uno de los ejes de dichos enfoques, se considera pieza clave dentro de las competencias a desarrollar en los estudiantes y se califica como necesaria para "los retos del siglo XXI". Voy a referirme, concretamente, al Nuevo Modelo Educativo mexicano dentro de la Educación Media Superior, partiendo de mi experiencia docente en el área de las Humanidades en las asignaturas de filosofía.

Cuando se habla de educación digital se parte, en primer lugar, de una definición parcial de lo que se concibe como tecnología. En segundo lugar, se considera una brecha generacional que produce un déficit entre maestros y alumnos respecto a dicha tecnología, generalmente estas consideraciones suelen desprestigiar la figura del docente y exaltar la eficacia que implica la intervención tecnológica en los procesos de enseñanza-aprendizaje. Estos aspectos manifiestan la ausencia de una reflexión sobre lo tecnológico en relación con los enfoques psicopedagógicos que regulan la educación general y particularmente digital.

Lo que me interesa señalar es cómo a partir de una reflexión filosófica es posible mostrar las deficiencias y problemáticas que se derivan de la implementación de una educación digital que no contempla las prácticas tecnológicas de profesores y alumnos, y los discursos bajo los cuales éstas se han interpretado y bajo los que se gestionan. Me propongo elaborar una crítica filosófica al respecto, con la finalidad de esbozar una propuesta que encamine la implementación de dicha educación dentro de un contexto lo más aproximado a la realidad del bachillerato mexicano, específicamente en el ámbito de las humanidades y desde un pensamiento filosófico que reflexiona continuamente sobre la relación entre su saber -incluidas sus formas de producción- y lo tecnológico y cómo esto desplaza el sentido tradicional de la enseñanza filosófica actual, sin descalificar los procesos tradicionales ni la importancia de la enseñanza de la filosofía, pero mostrando cómo al desplazarse se generan nuevas formas, no sólo de enseñanza filosófica, sino de producción filosófica que intervienen el ámbito educativo y la profesionalización del filósofo como maestro que reflexiona en la producción de su saber, y de las maneras en que transmite y gestiona el mismo, ante la intervención de la tecnología.

Pienso que, frente a la importancia que se imprime a la educación en general, y particularmente a la digital, es necesario analizar los discursos y prácticas que sostienen su relevancia, así como los factores y actores que intervienen en su aplicación con la finalidad de trazar posibles caminos que consideren las variables que pudieran presentarse, así como los contextos y espacios concretos en donde se lleva a cabo la misma. Me interesa, por último, señalar cómo las dificultades de implementación del Nuevo Modelo Educativo intervienen en la aplicación concreta de una educación tecnológica que va desde la infraestructura que la hace posible, hasta el modo de concebir dicha educación.

Ejercer maquínicamente la Filosofía. Pensamiento técnico vs Pensamiento filosófico

Elena León Magaña

¿Puede un robot desarrollar pensamiento filosófico? El Seminario Tecnologías Filosóficas se planteó el problema del pensamiento maquínico en octubre del 2015. Durante el tiempo de vida del seminario hemos discutido la modificación tecnológica de la experiencia y de la práctica de pensamiento; en este sentido, hemos intentado desarrollar una genealogía de la relación entre el pensamiento filosófico y la tecnología. Derivado de ello hemos encontrado discursos que problematizan la articulación tecnología/experiencia. En este sentido, la idea de desarrollar una "herramienta digital" que ponga a prueba los discursos que enuncian la modificación tecnológica de la experiencia en la que se relaciona, tanto la práctica del pensamiento filosófico, como de la tecnología.

Dicho lo anterior, algunas de las preguntas que fueron surgiendo durante la planeación y problematización de dicha herramienta fueron: ¿Sería posible determinar una tecnología filosófica? ¿Determinarla fuera de esos mitos metafísicos de la herramienta, de un sujeto que la usa, de la creación del pensamiento o de la inspiración genial? ¿Sería posible determinar elementos y procesos del ejercicio del pensamiento filosófico? Después de un

largo camino, donde, por ejemplo, se planteó el uso de un *bot* basado en el pensamiento del filósofo Gilbert Simondon, que interactuase en Twitter con otros bots a partir del procesamiento en una herramienta de *Machine Learning*; o bien, un *bot* de Twitter que fuese capaz de discernir entre una postura negativa, positiva o neutra respecto de lo tecnológico, evaluando los *tweets*.

Finalmente, se desarrolló un *bot* filosófico en dos etapas:

Un *chatbot* basado en el Modelo oculto de Markov para la toma de decisiones; es decir, basado en el contenido proporcionado: una curaduría de citas sobre textos de los filósofos: Gilbert Simondon, Walter Benjamin, Unabomber, Tiqqun, Jean Louis Deottè, Günther Anders y Michael Foucault. Con base en estas citas el bot decidirá entre copiar el pensamiento para interactuar con *tweets* con las palabras clave seleccionadas, o bien, si a partir de la información proporcionada, desarrollará nuevos *tweets*. Las palabras que serán elegidas para iniciar la interacción son: tecnología, experiencia, técnica, filosofía, máquina, maquinaria, maquinal, maquinizado (a), artificial, artificialidad, artificio, herramienta, herramientas, automático, automatización, autómata, automatismo, función, funcionamiento, funcionalización, funcional, político (a), politización, políticas, revolución, revolucionario, aparato, aparatos, dispositivo, tecnológico, usuario (a), usuarios (as), tecnologizado, cibernético (a).

Deep learning; en esta etapa se utilizó una plataforma *deep learning*. A partir de ella el robot aprenderá y desarrollará nuevo pensamiento basado en sus interacciones en Twitter. La apuesta es desarrollar una discusión filosófica a partir de los resultados de la interacción para determinar si se ha *innovado* pensamiento filosófico, y si el producto de las interacciones del *bot* puede ser considerado pensamiento filosófico.

Filosofía y Tecnología: Proyectos y Teoría en el Seminario de Tecnologías Filosóficas

Efrén Marat Ocampo Gutiérrez de Velasco

La relación entre filosofía y tecnología permite un examen crítico sobre cómo se afectan ambos. Por una parte, la filosofía que reconoce sus medios tecnológicos permite explorar su producción en términos ontotécnicos y éticos. En el caso de la tecnología se puede explorar como ésta también pertenece a una forma de pensamiento y no implica una suspensión reflexiva ante su hacer como lo ha discutido Simondon. Pensar filosofía y tecnología aporta a las humanidades digitales una perspectiva crítica para el trabajo en estos campos.

En específico se busca revisar los problemas teóricos entre filosofía y tecnología que ha abordado el Seminario de Tecnologías Filosóficas. Esta agrupación ha optado por revisar la aproximación teórica encontrada en los procesos de producción de sus proyectos. Entre estos se encuentran el análisis de la reconsideración de la biblioteca pública en la Biblioteca Vasconcelos de la Ciudad de México, el análisis de la base de datos de las tesis de filosofía producidas en la Universidad Nacional Autónoma de México, la discusión de textos sobre filosofía de la tecnología y filosofía de los medios y la producción de un *bot* a partir de contenido filosófico. Esta muestra de proyectos ha ayudado a producir un nexo de problemas teóricos sobre la práctica de las humanidades digitales.

Las discusiones llevadas a cabo han permitido observar de qué forma la tecnología de la producción teórica, y la configuración de sus prácticas, conducen a una discusión sobre cómo se posibilitan, a partir de la investigación de las humanidades, las herramientas tecnológicas. Algunas de las aproximaciones que se han explorado son:

La modificación política del espacio público a partir de la digitalización de los archivos de las bibliotecas públicas señalando a la especialidad de las prácticas de las humanidades digitales. La noción de experiencia en términos del uso tecnológico.

La transformación de la práctica discursiva a partir de los mecanismos de automatización y programación de las humanidades. Incluida la pregunta sobre la diferenciación de su moralización, su carga de opinión y su relación con la producción de datos.

Los cambios en la producción de las prácticas y de la comprensión de su desarrollo. El problema de las nociones de lectura, escritura y visualización en términos digitales como una cuestión política, ontológica y ética.

La revisión del lugar de la teoría por su mediación tecnológica. ¿Qué hace que se necesiten mantener diferenciadas? ¿Por qué el pensamiento ha encontrado en su consideración técnica una forma de replantearse?

De esta forma, se piensa que reconociendo estas dinámicas se puede hacer una referencia a su impacto en estudio filosófico de las humanidades digitales. La filosofía en las humanidades digitales todavía tiene un papel a ser considerado como un desarrollo sobre la revisión crítica de prácticas cada vez más sofisticadas técnicamente, que revisando sus producciones entienden sus efectos.

Los laboratorios de humanidades: tecnologías del error

Ethel Zaira Rueda Hernández

Las humanidades actuales se hacen en el laboratorio. La proliferación de espacios de investigación humanística que se llaman a sí mismos, o se asumen de una u otra manera, como laboratorios, o como espacios de experimentación tecnológica y científica, indica la institucionalización de un mecanismo que hasta hace poco parecía vetado a las disciplinas que se ubican fuera del espectro de las llamadas "ciencias duras". Si bien el laboratorio no

es un dispositivo nuevo como modo de producción de saberes, sí tiene, en el ámbito de las humanidades, un halo de novedad que lo vuelve una opción atractiva y popular, frente a los ya consabidos formatos de seminarios, talleres, mesas de debate o discusión, etc.

¿Qué es, qué hace un laboratorio de humanidades? ¿Por qué se asocia esta figura, y la de la experimentación humanística, con lo digital? Para Bruno Latour el laboratorio es una máquina privilegiada para cometer errores. Esto es importante porque son justo esos mecanismos de aceleración de los procesos de ensayo y error los que permiten obtener el poder de modificar el mundo. Si a las humanidades les interesa tener alguna incidencia en lo real, parece que apropiarse de la maquinaria del laboratorio es una estrategia que, al menos a los ojos de Latour, podría resultar conveniente.

¿Qué es "cometer un error" para las humanidades? ¿Qué implica que el laboratorio se esté convirtiendo en una de las figuras dominantes de los campos de investigación humanísticos? ¿Funcionan los laboratorios de humanidades de manera similar o análoga a los laboratorios científicos? ¿Cuál es la relación con el discurso científico que se hace posible a partir de la ampliación del concepto de laboratorio? Si las tecnologías digitales abren la posibilidad al quehacer humanístico de tomar la forma de los tubos de ensayo ¿qué ensayamos?, ¿qué es este ensayar, este equivocarse sistemáticamente, este acumular errores, del que las humanidades digitales y las artes parecen tan interesadas en apoderarse?

¿Dónde radica el poder de un error? ¿En recordarlo para no cometerlo de nuevo? ¿En señalarlo para marcar un límite? ¿En incorporarlo a una base de datos que mapea todas las combinaciones posibles de lo real? Se trata de pensar el laboratorio como dispositivo, más allá de las disputas sobre el valor de la ciencia frente a las artes liberales, se trata de pensar nuestras tecnologías digitales experimentales más allá de la perversa ingenuidad de la tecnofilia y la tecnofobia, se trata de preguntarnos por el funcionamiento del laboratorio de humanidades en tanto que máquina de producción de conocimiento, de abrir la caja negra y adentrarnos en el entramado de cables, circuitos y piezas que constituyen las entrañas de nuestro labor cotidiana.

References

Arendt, H. (2016). *La condición humana*. Barcelona: Paidos.

Barrón, F. (2013). Individuación tecnológica. Apuntes para pensar la educación tecnológica. *Blog de la Red de Humanidades Digitales*. [en línea]. 16 de febrero del 2013. Disponible en: http://humanidadesdigitales.net/blog/2013/02/16/individuacion-tecnologica-apuntes-para-pensar-la-educacion-tecnologica/ [26/04/18].

Bawa.Cavia. (2018). "The Inclosure of Reason" en *Technosphere Magazine*. Dossier *Human;* 2016, visitado el 1 de febrero de 2018, <https://technosphere-magazine.hkw.de/article1/6aefb210-0ee6-11e7-a253-d9923802c14e >.

Blackman, L. (2016). *Loving the alien. A Post-Post Human Manifesto*. Miami:Institute of Contemporary Art Miami: Fall Semester.

Blatt, B. (2017). *Nabokov's Favorite Word Is Mauve*. *What the Numbers Reveal About the Classics, Bestsellers, and Our Own Writing*. New York: Simon and Schuster.

Brussa, V. (2017). Otros laboratorios: discutiendo la extitución y democratización tecnocultural en los laboratorios de Humanidades Digitales iberoamericanos. *Revista Virtualis*, 7(13).

Ernst, W. (2013). *Digital memory and the archive.* Minnesota:Minnesota University Press.

Estalella, A., Lafuente, A. y Rocha, J. (2013). Laboratorios de procomún: experimentación, recursividad y activismo. *Revista Teknocultura*, 10(1): 21-48. Gutierrez, E. (2013). Buenas prácticas en la docencia digital. *Blog de la Red de Humanidades Digitales*. [en línea]. 20 de noviembre del 2013. Disponible en: http://humanidadesdigitales.net/blog/2013/11/20/buenas-practicas-en-la-docencia-digital/ [26/04/18].

Gutierrez, E. (2012). La educación permanente y las nuevas tecnologías de la información y la comunicación. *Blog de la Red de Humanidades Digitales*. [en línea]. Disponible en: http://humanidadesdigitales.net/blog/2012/12/02/la-educacion-permanente-y-las-nuevas-tecnologias-de-la-informacion-y-la-comunicacion/ [26/04/18].

Haraway, D. (2016). *Staying with the trouble*. Estados Unidos: Duke University Press.

Herbón, D. (2014). Reflexiones contra el ciberfetichismo académico. *Blog de la Red de Humanidades Digitales*. [en línea]. 7 de mayo del 2014. Disponible en: http://humanidadesdigitales.net/blog/2014/05/07/ciberfetichismo-academico/ [26/04/18].

Ihde D. (2009). *Postfenomenología y tecnociencia*. España: Sello Arsgames.

Instituto de Investigaciones sobre la Universidad y la Educación. [IISUE UNAM oficial]. (2017/03/29). ¿El Nuevo Modelo Educativo promueve el uso de las TIC?. [Archivo de video]. Disponible en: https://youtu.be/hltn0oXFhaM [Recuperado 26/04/18]

Kittler, F. (2014). *The truth of the technological world.* Stanford: Stanford University Press.

Krzywkowski, I. (2010). *Machines à écrire : Littérature et technologies du xixe au xxie siècle*. UGA Éditions.

Laruelle, F. (1994). *The concept of 'first technology': a 'unified theory' of technics and technology*. trans. Nadita Biswas Mellamphy.

Latour, B. (2003). The World Wide Lab. Wired. http://www.wired.com/2003/06/research-spc/ Accesado el 1 de febrero de 2018.

Latour, B. y Woolgar, S. (1995). *La vida en el laboratorio*. Madrid: Alianza.

Liu, A. (2012): "Where is Cultural Criticism in the Digital Humanities?" en *Debates in the Digital Humanities*. University of Minnesota Press: visitado el 1 de febrero de 2018, < http://dhdebates.gc.cuny.edu/debates/text/20 >.

Liu, A. (2018). *Digital Humanities Diversity as Technical Problem*. *Alan Liu;* 15 de enero 2018. doi:10.5072/FK2222ZR81, visitado el 1 de febrero de 2018 <http://liu.english.ucsb.edu/digital-humanities-diversity-as-technical-problem/ >.
Mirowski, P. (2017). Against citizen science. https://aeon.co/essays/is-grassroots-citizen-science-a-front-for-big-business Accesado el 1 de febrero de 2018.
Moretti, F. (2007). *La literatura vista desde lejos*. Barcelona: Marbot Ediciones.
Moretti, F.(2016). *Lectura distante*, Buenos Aires: Fondo de Cultura Económica.
Negrestani, Reza (2014). "The labor of the inhuman (Parte I y II)" en *E-Flux.* Journal #52; febrero 2014, visitado el 1 de febrero de 2018, <http://www.e-flux.com/journal/52/59920/the-labor-of-the-inhuman-part-i-human/>.
Parisi, L. (2017). "Reprograming Decisionism" en *E-flux*. Journal #85; octubre 2017, Accessed on 20th december; 2017, visitado el 1 de febrero de 2018, <http://www.e-flux.com/journal/85/155472/reprogramming-decisionism/ >.
Parisi, L. (2013), *Contagious architecture : computation, aesthetics, and space.* Cambridge, Mass: MIT Press.
Piscitelli, A. (2005). Tecnologías Educativas. Una letanía sin ton ni son. *Scielo. Revista de estudios sociales.* [en línea]. N° 22. Disponible en: http://www.scielo.org.co/scielo.php?pid=S0123-885X2005000300012&script=sci_arttext&tlng=en [Recuperado 26/04/18].
Power, N. (2017) "Inhumanism, Reason, Blackness, Feminism" en *Glass - Bead*. Site 1: Logic gate, The Politics of the Artifactual Mind; 2017, visitado el 1 de febrero de 2018,<http://www.glass-bead.org/article/inhumanism-reason-blackness-feminism/?lang=enview >.
Ramírez, A. (2016). La (~~presencia~~) ausencia de las TIC en el Modelo Educativo 2016. *Blog de la Red de Humanidades Digitales*. [en línea]. 27 de septiembre del 2016. Disponible en: http://humanidadesdigitales.net/blog/2016/09/27/ticmodelo2016/ [26/04/18].
Ramírez, A. [armartinell]. (2016/09/24). Presencia de las TIC en el Modelo Educativo de la SEP 2016. [Archivo de video]. Disponible en: https://youtu.be/109Eeg-1fXWU [Recuperado 26/04/18].
Rangel, D. (2018). Educación futura. (n.d.) [online] Available from: http://www.educacionfutura.org/lo-que-el-modelo-educativo-no-contemplo-la-realidad-social/?platform=hootsuite [Recuperado 26/04/18].
Scolari, C. Alfabetismo transmedia: un programa de investigación. *Blog de Carlos Scolari*. [en línea]. 26 de septiembre del 2014. Disponible en: https://hipermediaciones.com/2014/09/26/transalfabetismos/ [26/04/18].
Secretaría de Educación Pública. (2017). Modelo Educativo Para la Educación Obligatoria. Disponible en: https://www.gob.mx/cms/uploads/attachment/file/198738/Modelo_Educativo_para_la_Educacio_n_Obligatoria.pdf [Recuperado 26/04/18].
Serra, A. (2013). Tres problemas sobre los laboratorios ciudadanos. Una mirada desde Europa. *Revista Iberoamericana de Ciencia, Tecnología y Sociedad*. 8(23): 283-298.
Shaviro, S. (2012) *Without Criteria.* Massachusetts: MIT Press.
Simondon, G. (2007). *El modo de existencia de los objetos técnicos*. Buenos Aires: Prometeo.
Simondon, G. (2015). *La individuación a la luz de las nociones de forma y de información*. Buenos Aires: Cactus.
The Point (2014). The New Humanities. *Revista The Point*. https://thepointmag.com/2014/criticism/the-new-humanities Accesado el 1 de febrero de 2018.

Perspectivas Digitales y a Gran Escala en el Estudio de Revistas Culturales de los Espacios Hispánico y Lusófono

Ventsislav Ikoff
vikoff@uoc.edu
Open University of Catalonia, Spain

Laura Fólica
lfolica@uoc.edu
Open University of Catalonia, Spain

Diana Roig Sanz
dsanzr@uoc.edu
Open University of Catalonia, Spain

Hanno Ehrlicher
hanno.ehrlicher@uni-tuebingen.de
University of Tübingen, Germany

Teresa Herzgsell
teresaherzgsell1@yahoo.de
University of Tübingen, Germany

Claudia Cedeño
claudiaceba@gmail.com
University of Tübingen, Germany

Rocío Ortuño
rocio.ortuno@uantwerpen.be
University of Antwerp, Belgium

Joana Malta
joanavmalta@gmail.com
SLHI - CHAM-FCSH/NOVA-UAc, Portugal

Pedro Lisboa
plisboa@gmail.com
SLHI - CHAM-FCSH/NOVA-UAc, Portugal

Este panel se propone avanzar, desde un punto de vista metodológico, en el análisis a gran escala de la revista como institución cultural, una aproximación que puede

cuestionar centros de producción literaria y revelar dinámicas de relación hasta ahora desconocidas entre las mal denominadas literaturas periféricas y los centros culturales hegemónicos. Para ello, los coordinadores de este panel proponemos cuatro presentaciones que abordan el estudio de revistas culturales históricas del ámbito hispánico y lusófono a través de diversos estudios de caso que utilizan herramientas digitales y combinan intereses disciplinares en los campos de la historia de las ideas, la historia cultural, los estudios de traducción, y la literatura comparada desde una perspectiva transnacional. El panel se propone dar muestra del estado actual del estudio de revistas culturales en los espacios hispánico y lusófono a través del uso de herramientas digitales que permitan avanzar en la discusión metodológica.

A este respecto, las propuestas de comunicación que se presentan se enmarcan en proyectos de investigación en curso vinculados a distintas universidades de Bélgica, Alemania, Portugal y España (Universiteit Antwerpen, Universität Tübingen, Universidade Nova de Lisboa y Universitat Oberta de Catalunya, respectivamente). Estos proyectos privilegian a la revista cultural como objeto de estudio y emplean distintas herramientas y metodologías digitales, dando sobradas muestras de la riqueza de perspectivas analíticas dentro del campo de las Humanidades Digitales (digitalización de materiales impresos y POS-tagging, construcción de repositorios electrónicos y de portales de investigación, bases de datos relacionales, geolocalización y visualización). Los autores de las distintas presentaciones comparten un compromiso similar en la colaboración científica entre pares, gracias a la publicación en abierto de los datos recogidos en sus investigaciones (*open source*). Las comunicaciones presentarán los respectivos proyectos en curso, ejemplificando con estudios de caso que de ellos se derivan.

En concreto, las comunicaciones del panel abordarán los siguientes objetos:

La creación de un repositorio de textos a partir de revistas en las Filipinas entre 1850 y 1945, dentro del marco del proyecto "Digitization of Philippine Rare Periodicals and Training in DH", con el propósito de facilitar el futuro estudio de textos históricos a través de herramientas digitales. Por medio de un análisis textual computacional, esta comunicación ejemplificará la utilidad de este repositorio con un estudio sobre la actitud que adopta la sociedad de habla filipina respecto de otros países en tres momentos concretos del siglo XX.

La presentación del entorno digital "Revistas culturales 2.0" al servicio de investigadores de revistas culturales históricas en lengua española. En base a este portal, la comunicación presentará un estudio de redes sociales entre los autores, revistas y géneros literarios con el objetivo de centrarse en textos programáticos (editoriales, prólogos o manifiestos).

La presentación de la base de datos de "Revistas de Ideias e Cultura" portuguesas del siglo XX que combina aproximaciones a partir de la historia de las ideas, la biblioteconomía y la ciencia de la información. La comunicación abordará las redes de recepción en revistas portuguesas en base a las obras y los nombres citados en ellas.

La identificación y análisis de las traducciones literarias publicadas en revistas hispánicas en el primer tercio del siglo XX con el objetivo de descubrir publicaciones hasta ahora desconocidas y revelar las relaciones literarias y editoriales entre distintos órganos de la prensa cultural hispánica a escala transnacional. Este estudio se realizará a partir de los datos recogidos en el VRE "MapModern" sobre revistas clave españolas e hispanoamericanas, mediadores culturales, y su participación en eventos y organizaciones culturales internacionales.

Philippines at the crossroads: enhancing research on Philippine periodicals and finding transnational attitudes in them

Rocío Ortuño Casanova

Key Words: Philippines, online repository, OMEKA, IIIF

The Philippines has been historically the centre of intercontinental, cultural, and economic relations: between Asia and Europe (Spain) both, in the time of the Spanish invasion (1565-1898) and nowadays (Montobbio, 2004: 11, 13); between America and Asia since the Manila galleon (1565-1815) (Giráldez 2015); and during the US invasion of the country (1902-1941,1946) (Kramer 398-407, San Juan 2000). However, the scarce research performed so far on the Philippines, and the difficulty of access to textual materials from the country have became two major problems for the study of these relations, in which the Philippines constitutes a blind spot. In order to address these problems, the AC/DC research group of the University of Antwerp is developing a project in partnership with the University of the Philippines and funded by VLIR-UOS to create an online repository of periodical publications in the Philippines, and to offer training in DH to potential users of this repository.

This talk is structured into two parts. The first one will provide an overview of the digitization scene in the Philippines, and will present the VLIRUOS TEAM project "Strengthening Digital Research at the University of the Philippines System: Digitization of Philippine rare newspapers and magazines (1850-1945), and training in Digital Humanities". The second part will offer an example of what kind of research results we expect to achieve with it.

The project has the initial objectives of (1) making written documentation available online for perusal of researchers both, from the Philippines and abroad. (2) Increasing academic research on humanities in the Philippines by the diffusion of DH methodologies. Therefore, two actions will be implemented:

The creation of an online repository of Philippine periodicals published between 1850 and 1945 and hosted at the University of the Philippines. Although the University of Santo Tomás is also uploading their rare periodicals collection, and there are other incipient projects for digitization in the Philippines, this repository will differentiate itself by considering three aspects:

A social aim: how can this repository be useful to a wide Filipino public?

Becoming useful to a range of researchers: how can we process the texts and what metadata are necessary to facilitate research for scholars from different disciplines such as linguistics, history or literature?

Facing the challenges of the Philippine context such as slow internet or multilingualism in periodical publications.

Organization of training session in DH and implementation of projects in four campuses of the University of the Philippines.

One of the main objectives of the project is producing interdisciplinary research on the Philippines, based on the digitized materials, using digital tools. In this talk, one example of the kind of research results that we expect to achieve will be provided. We will show an analysis of adjectives related to Spain, China and the US in 1918 (end of World War I), 1930 (after the Crack) and 1936 (between the declaration of the Philippines as a Commonwealth state and the beginning of the Spanish Civil war) in the Philippine cultural magazine *Excelsior*, obtained with POS tagging of the text. It aims to find the attitude of the Philippine speaking society towards other countries at the beginning of 20th century with computational text analysis (Computer linguistics). The data obtained allows to reach conclusions on historical and literary trends.

References

Giráldez, A. (2015). *The Age of Trade: The Manila Galleons and the Dawn of the Global Economy*. Lanham, Maryland: Rowman & Littlefield.

Kramer, P. (2006). *The Blood of Government: Race, Empire, the United States, and the Philippines*. Quezon City: Ateneo de Manila University Press.

Montobbio, M. (2004). *Triangulando la triangulación: España/Europa-América Latina-Asia Pacífico*. Barcelona: Cidob/ Casa Asia.

Roque, A. (2012). Towards a computational approach to literary text analysis. *Workshop on Computational Linguistics for Literature*. Montréal, Canada: Association for Computational Linguistics, June 8, 2012, pp. 97–104.

San Juan, E. (2000). *After Postcolonialism: Remapping Philippines-United States Confrontations*. Lanham, Maryland: Rowman & Littlefield.

Las Humanidades Digitales en la Mixteca de Oaxaca: reflexiones y proyecciones sobre la Herencia Viva o Patrimonio

Emmanuel Posselt Santoyo
eps537@hotmail.com
Universidad de Leiden, Netherlands

Liana Ivette Jiménez Osorio
lianaji@hotmail.com
Universidad de Leiden, Netherlands

Laura Brenda Jiménez Osorio
vreosorio@gmail.com
Universidad Autónoma Metropolitana, Unidad Xochimilco, México

Roberto Carlos Reyes Espinosa
robcar_67@hotmail.com
Universidad Tecnológica de la Mixteca, México

Eruvid Cortés Camacho
eruvid@mixteco.utm.mx
Universidad Tecnológica de la Mixteca, México

José Aníbal Arias Aguilar
anibal@mixteco.utm.mx
Universidad Tecnológica de la Mixteca, México

José Abel Martínez Guzmán
jam_e4@hotmail.com
Universidad Tecnológica de la Mixteca, México

Las Humanidades Digitales son una disciplina que comienza a tomar auge en el mundo, sin embargo, existen muchas regiones en las que aún se encuentran en una etapa inicial, como es el caso de la Mixteca ubicada en el estado de Oaxaca (figura 1). Los temas que se pueden tratar con esta disciplina son diversos y es difícil delimitar su enfoque, por eso se promueve que se quede en su forma plural "Humanidades Digitales" y que no se singularice (Fitzpatrick 2012). Ésta se ha definido como la interrelación entre las ciencias de humanidades con las digitales, también se ha optado por mencionar que es la conjunción de las humanidades con las posibilidades digitales, o la conjunción de lo digital con las posibilidades humanísticas (Fitzpatrick 2012 y Kirschenbaum 2012).

Asimismo, una definición que encontramos relevante es la que señala que las Humanidades Digitales tratan de modelar el mundo alrededor de nosotros a través del éxito y falla (en el trabajo interdisciplinario) para que tengamos un mejor entendimiento de lo que conocemos y no conocemos sobre la humanidad, sus actividades, artefactos, y registros (Vanhoutte 2013).

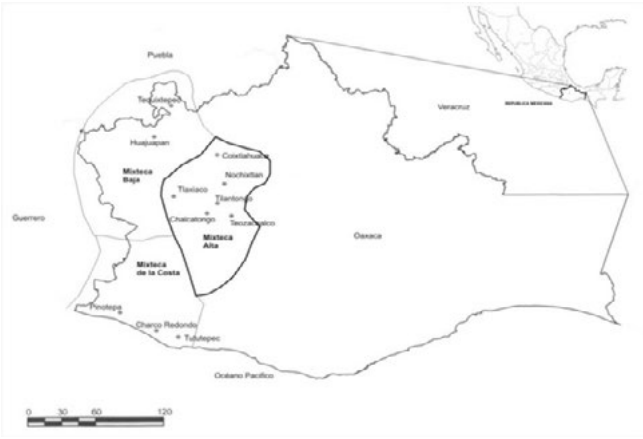


Figura 1.- Ubicación de la región Mixteca y de las poblaciones más importantes

En este panel dialogaremos y reflexionaremos sobre las proyecciones que ofrecen las Humanidades Digitales en la región de la Mixteca. Lo que nos reúne a los académicos de este panel es nuestro interés en reforzar, de manera creativa, los procesos educativos, la herencia viva y la identidad de Ñuu *Savi* o Pueblo Mixteco a partir de trabajos colaborativos y la aplicación de medios interactivos.

En el campo de las Humanidades nos enfocaremos en temas arqueológicos y de Herencia Viva o Patrimonio, y desde lo digital nos centraremos en el Diseño como disciplina e ingeniería, encausando los esfuerzos hacia: reconstrucciones virtuales de sitios arqueológicos, una impresión en 3d, una Interfaz Gráfica, un videojuego y un robot humanoide.

Los proyectos que aquí se presentan parten del diálogo entre los diferentes especialistas y el que se ha entablado con los habitantes de las poblaciones en donde se está trabajando. Esto ha generado que los proyectos sean abiertos, es decir, que si bien persiguen objetivos concretos, durante su desarrollo van surgiendo nuevas alternativas y metas. La constante es la búsqueda de mejores realidades, tanto en las investigaciones como en lo social.

A partir de los diferentes proyectos particulares se promueven también aspectos teóricos para reforzar cada vez más las Humanidades Digitales en los ámbitos que esta mesa se enfoca: se plantea una metodología sensible para la reconstrucción en 3D, se expone la relevancia de la experiencia vivencial en el proceso creativo digital, se discute sobre la digitalización y la materialización de las ideas, también sobre las formas de comunicación y enseñanza, se propone el trabajo con y para las comunidades, así como la importancia de considerar los contextos y los usuarios.

Un aspecto fundamental que consideramos como académicos de diversas disciplinas son los valores que promueven las Humanidades Digitales (Spiro 2012), que son:

1) apertura y transparencia, esto fomenta que las disciplinas participantes se nutran y compartan entre ellas y que no se reduzcan a una disciplina dominante, es decir, que se busque la democratización del conocimiento, 2) colaboración, es pensarse como parte de un equipo, en donde el aprender y la contribución es entre todos los participantes a través de un diálogo continuo, 3) apoyo y conectividad, es decir, compartir los conocimientos que se logran para todos aquellos que lo necesiten (incluyendo especialistas y no especialistas), 4) diversidad, tanto en las especializaciones como en las personas (diferentes bagajes culturales y género), esto hace que las Humanidades Digitales sean más vibrantes, con discusiones más ricas y multiplicidad de perspectivas, y 5) experimentación, con la cual no solamente se sugiere un método de prueba de ideas y la creación de conocimiento sino también su involucramiento en la transformación de enfoques tradicionales a la enseñanza e investigación.

De tal forma, los proyectos que se presentan se fundamentan en estos valores así como en los diferentes diálogos, y están en correspondencia con las realidades sociales que se presentan en la región (figuras 2 y 3). La Mixteca, al igual que otras regiones en México, ha estado marginada por años, y en el ámbito que nos compete no existe apoyo por parte del gobierno (estatal y federal) para el uso de las tecnologías digitales en pro de la educación y el fortalecimiento de la Herencia Viva. Por eso proponemos que es en el ámbito académico donde se pueden generar alternativas para cubrir estas necesidades de la población.

Figura 2.- El paisaje de la comunidad de Yoso Notu en la Mixteca Alta

Figura 3.- Un ejemplo de Herencia Viva en la Mixteca Baja: tapetes de aserrín para la conmemoración del cese del sitio de la ciudad de Huajuapan de León durante la Guerra de Independencia, atribuido al Señor de los Corazones

Finalmente, cabe señalar que los trabajos que integran la mesa se enfocan en comunidades Indígenas y zonas urbanas de la Mixteca, pero bien pueden ser aplicados en otras partes de México y el mundo que tengan una realidad social similar.

Bibliografía

Fitzpatrick, K. (2012). The humanities, done digitally. En: Debates in the Digital Humanities. pp 12-15, Matthew K. Gold (ed.), University of Minnesota Press, Estados Unidos de America.

Kirschenbaum, M. (2012). What is Digital Humanities and what´s doing in English departments?. En: Debates in the Digital Humanities. pp 3-11, Matthew K. Gold (ed.), University of Minnesota Press. Estados Unidos de America.

Spiro, L. (2012). "This is why we figth": Defining the values of the Digital Humanities. En: Debates in the Digital Humanities. pp 16-35, Matthew K. Gold (ed.), University of Minnesota Press. Estados Unidos de America.

Vanhoutte, E. (2013). "The gates of hell: history and definition of Digital | Humanities | Computing. En: Defining Digital Humanities. pp 119-156, editado por Terras, Melissa, Julianne Nyhan y Edward Vanhoutte, ASHGATE. Reino Unido.

Herramientas para el desarrollo de aplicaciones interactivas para promoción de la cultura y la historia de Huajuapan de león, Oaxaca

José Aníbal Arias Aguilar

Usando tecnologías interactivas se puede atraer a los usuarios para facilitar el desarrollo de ciertas tareas educativas y culturales. Ejemplos de estas tecnologías son los robots con capacidades de comunicación con personas (Breazeal 2014) y los videojuegos serios (Díaz, et al. 2015).

En la ponencia se describirán los procesos más importantes y algunas propuestas de herramientas necesarias para desarrollar aplicaciones interactivas que ayuden a difundir y apreciar la historia y la cultura de la ciudad de Huajuapan de León, Oaxaca, entre los niños y jóvenes. Se dará una visión general del tema, planteando que la enseñanza de la historia y las tradiciones a los jóvenes se está volviendo difícil utilizando sólo textos, imágenes e historias orales; y se abordarán dos casos de aplicación de herramientas interactivas en la educación: la enseñanza del Jarabe Mixteco utilizando un robot humanoide y el desarrollo de un videojuego serio basado en eventos históricos y elementos culturales de la ciudad.

Abordaremos entonces los procedimientos necesarios para plantear las aplicaciones mencionadas (entrevistas, encuestas, planeaciones didácticas) y para desarrollarlas (diseño interactivo, programación, modelado 3D, diseño 2D, realidad aumentada, uso de audio/video, etc.)

Bibliografía

Breazeal, C. (2004). *Designing sociable robots. Bradford Books, colección: Intelligent Robotics and Autonomous Agents Series*, MIT Press.

Díaz, J., C. Queiruga, C. Banchoff, L. Fava y V. Harari (2015). *Educational robotics and videogames in the classroom*, conferencia: 10th Iberian Conference on Information Systems and Technologies, 17-junio-2015, publicado por IEEE, Aveiro, Portugal

Una metodología sensible para la reconstrucción digital de sitios precoloniales en el contexto de la Nación Mixteca

Emmanuel Posselt Santoyo
Liana Ivette Jiménez Osorio

En esta presentación hablaremos sobre los aspectos que se deben considerar como parte de una metodología para la reconstrucción virtual de sitios precoloniales. Nos enfocaremos principalmente en el contexto de las comunidades Indígenas contemporáneas de la Mixteca Alta de Oaxaca.

En esta región se han hecho ya algunas reconstrucciones de sitios arqueológicos (dibujos a mano y virtuales) para complementar la información de las diversas publicaciones. Un primer esfuerzo fueron los dibujos (aproximados) de varios sitios que presentó Manuel Martínez Gracida a principios del siglo XX, los cuales no solamente son de la Mixteca sino de otras partes de Oaxa-

ca (Martínez 1883). Más tarde, entre 1937 y 1940, Jorge Acosta realizó las reconstrucciones de los montículos y exteriores e interiores de habitaciones del sitio de Monte Negro (1992:33, 43 y 50). A partir de ese momento los dibujos se basan en investigaciones arqueológicas, dándole otra proyección a los sitios precoloniales.

Más recientemente se vuelven a retomar los dibujos reconstructivos para apoyar las investigaciones arqueológicas: está el de Martijn Wijnhoven, quien trabaja un conjunto arquitectónico del sitio ubicado en el Cerro de la Corona en la comunidad de Tamazulapan (2001: 61) y el que realiza Leonardo López Zarate del sitio de Ñuyagua en Tilantongo (Hermann, 2015:54).

Mención especial merece Ronald Spores, quien ha realizado varias reconstrucciones en las que propone los elementos que por el paso del tiempo han desaparecido, creando una imagen más realista de los sitios al momento en que estuvieron habitados. Ejemplificó tres asentamientos de la Mixteca: del Preclásico (1500-500 a.C.), del Clásico (500-950 d.C.) y del Posclásico (950-1521) respectivamente, éstos no son de algún sitio en particular sino que muestran atributos del periodo al que hacen referencia (2008:29-30). Además, para reforzar la explicación sobre el sitio de Yuku Ndaa ubicado en Teposcolula, el mismo autor realizó una reconstrucción virtual (Spores, 2014: 318 y 319 y Spores y Robles, 2014: 26) y presentó un dibujo a mano elaborado por Enrique Martorell (Spores y Balkansky, 2013:31).

A nivel nacional, un trabajo importante es el que ha llevado a cabo el grupo Tlamachqui ya que reconstruyó virtualmente seis ciudades antiguas de Mesoamérica. Esto se logró a partir de la conjunción de información por parte de un equipo de profesionales. El objetivo de este proyecto fue mostrar la imponente imagen que tenían estas ciudades al estar en su momento de apogeo (Monsivais 2013).

Como se observa, la reconstrucción virtual de sitios arqueológicos en la Mixteca se encuentra en una fase inicial. Consideramos que esto se debe, por un lado, a la falta de una colaboración conjunta entre la arqueología y las ciencias digitales y, por el otro, a que este tipo de trabajos se dirigen a un público especializado en el campo arqueológico. Aunado a esto, bajo un marco institucional gubernamental, los sitios que se eligen para las reconstrucciones en 3D son los monumentales, excavados como parte de un gran proyecto arqueológico y que se han destinado para un turismo nacional e internacional, como Monte Albán, Teotihuacán y Palenque, entre otros.

De tal forma, la región de la Mixteca Alta es un campo fértil para desarrollar esta temática sobre reconstrucciones digitales ya que nos ofrece, al mismo tiempo, grandes posibilidades y retos. Por un lado, se tiene una riqueza arqueológica expresada en la diversidad de sitios (ciudades y santuarios) de diferentes temporalidades (desde el Preclásico Temprano hasta el Posclásico), también hay información invaluable en los libros antiguos (códices) realizados por los propios mixtecos en tiempos precoloniales, asimismo, se cuenta con el conocimiento de los pobladores contemporáneos sobre temas religiosos e históricos que nos dan cuenta de la importancia que tienen hoy en día los sitios arqueológicos (figura 1). En este sentido, nuestra intención es mostrar las posibilidades de acción en los lugares del paisaje (Ingold, 2000).

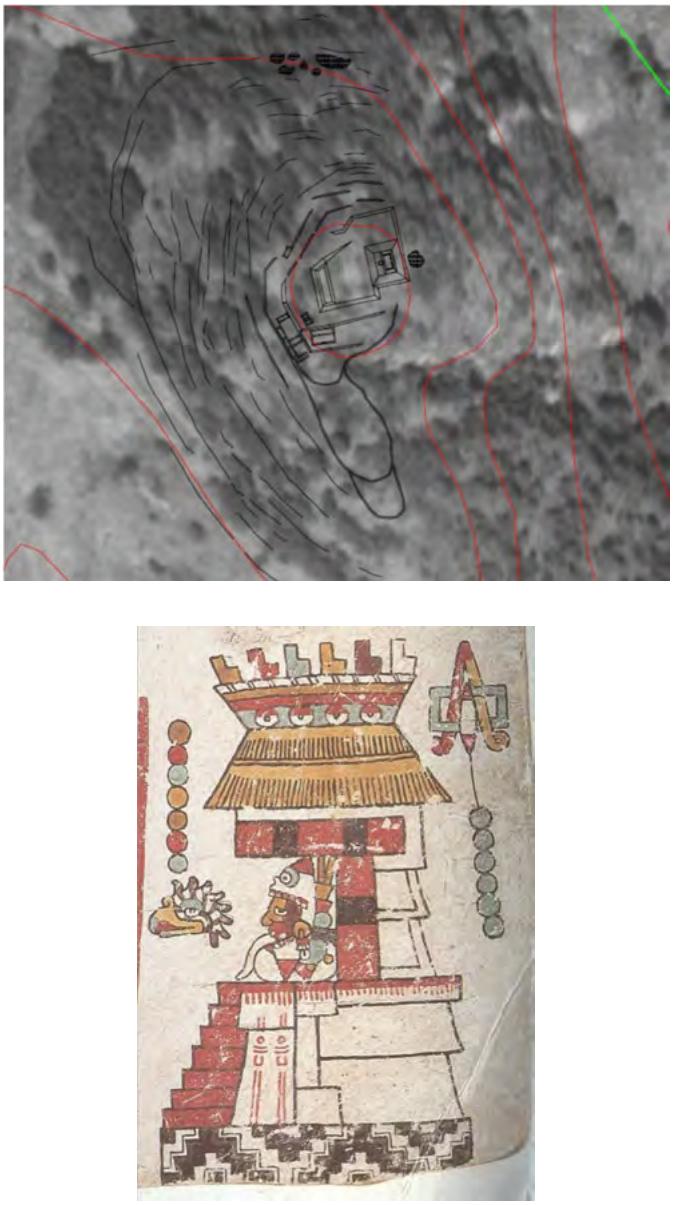


Figura 1.- Uno de los sitios arqueológicos que existen en la comunidad de Yoso Notu y la representación de un templo con un envoltorio sagrado en un códice mixteco

Por el otro lado, el reto es la inclusión e involucramiento de un público no especializado y que tiene otras formas de relacionarse con estos lugares antiguos (Atalay, 2014), las cuales van más allá de lo turístico, en este caso nos referimos a las comunidades de la región (figura 2). Estas posibilidades y retos son algunos de los aspectos que nosotros proponemos integrar como parte de una metodología a seguir en los trabajos de reconstrucción digital.

Figura 2.- Celebración del Segundo Viernes de Cuaresma en un santuario de origen precolonial y visita a un sitio arqueológico con las autoridades de la comunidad de San Miguel, quienes explicaron parte de la historia del sitio

Además, los proyectos en los que trabajamos siguen los principios señalados en la Carta de Londres (2009) y en Los Principios de Sevilla (Foro Internacional de Sevilla 2012), de manera general ambos tienen como objetivo reglamentar los trabajos digitales sobre patrimonio.

Asimismo, debido a que nuestro interés es sobre la Herencia Viva consideramos relevante los cuatro puntos que proponen Anusas e Ingold en relación a la Antropología del Diseño que definen como: la combinación del conocimiento fundamentado en la metodología y conocimiento antropológicos con la praxis imaginativa de las habilidades y procesos del diseño. Estos puntos son: 1) reflexiva hacia las propias creaciones disciplinarias, 2) participativa en su entendimiento de vida, 3) experta en las relaciones entre percepción, cultura y materiales y 4) activa en el involucramiento creativo para fomentar una vida mejor (2013:68-69).

Así, las reconstrucciones digitales de dos sitios precoloniales que trabajamos junto con Laura Brenda Jiménez Osorio y Roberto Carlos Reyes Espinoza (Jiménez, et al., 2017), respectivamente, así como una interfaz que realizamos con Eruvid Cortés Camacho tuvieron como objetivos: 1) mostrar la configuración arquitectónica y el uso del sitio en su época de esplendor para generar un mayor entendimiento de éste en relación a lo que se ve hoy en día. Lo que se busca es que el modelo complemente la realidad física y que ésta complemente al modelo, 2) integrar en la reconstrucción del sitio las diferentes voces respecto a su valor y significado y 3) tender diferentes puentes: entre el presente y el pasado precolonial, entre la arqueología, las comunidades de la región y las ciencias digitales, entre la herencia intangible y tangible, y entre los datos y las interpretaciones (figura 3).

Los tres trabajos los consideramos el resultado de un proceso de modelado, entendido como un proceso creativo de pensamiento y razonamiento en donde el significado es hecho y negociado a través de la creación y manipulación de representaciones externas (Ciula y Eide 2017:i34).

Figura 3.- Divulgación del trabajo interdisciplinario sobre la reconstrucción virtual de un sitio precolonial, Universidad de Chalcatongo

Esta propuesta metodológica se mostrará a partir de dos asentamientos precoloniales: el de Yuku Chayo en la comunidad de Chalcatongo de Hidalgo y el de Cerro de Pedimento en las comunidades de Santa Catarina Yoso Notu y San Miguel el Grande. El primero corresponde con una ciudad monumental del Preclásico Tardío (400 a.C.-300 d.C.) y el segundo con un santuario milenario que fue ocupado desde el Preclásico Temprano, con relevancia en el Posclásico (900-1521 d.C.) y que en la actualidad es un centro de peregrinación de importancia regional. La información antropológica y arqueológica que se retoma para estos ejemplos se deriva de tres investigaciones realizadas en la Mixteca Alta de Oaxaca (Jiménez y Posselt 2008, 2016 y s/f).

Por último, cabe mencionar que la investigación que produjo estos resultados forma parte del proyecto'Time in Intercultural Context',dirigido por el prof. dr. Maarten E.R.G.N. Jansen (Facultad de Arqueología, Universidad de Leiden)y fue financiada por el European Research Council en el marco del European Union's Seventh Framework Programme(FP/2007-2013) / ERC Grant Agreement No. 295434.

Bibliografía

Acosta, J. R. y J. Romero (1992). *Exploraciones en Monte Negro, Oaxaca. 1937-38, 1938-39 y 1939-40*. Antologías Serie Arqueológica. Instituto Nacional de Antropología e Historia, México.

Anusas, M. y T. Ingold (2013). Designing environmental relations: from opacity to textility. En: *Design Issues Journal* 29(4), pp:58-69.

Atalay, S. (2014). Engaging Archaeology: positivism, objectivity, and rigor in activist archaeology. En: *Transforming Archaeology: activist practices and prospects*, S. Atalay, Lee Rains, Clauss, Randall H. MacGuire y J. R. Welch (eds.). pp. 45-59. Left Coast Press.

Ciula, A. y Ø. Eide (2017). Modelling in digital humanities: signs in context, En: *Digital scholarship in the humanities*, vol. 32, Supplement 1, pp i33-i46.

Foro Internacional de Arqueología (2012). Los principios de Sevilla. Principios Internacionales de la Arqueología Virtual.

Hermann, L. M. A. (2015). El territorio de Tilantongo en el siglo XVI. Algunas consideraciones sobre su geografía histórica. En: *Configuraciones territoriales en la Mixteca*, Hermann Lejarazu, Manuel A. (ed.), pp. 37-91. vol. I. Centro de Investigaciones y Estudios Superiores en Antropología Social y Casa Chata México.

Ingold, T. (2000). Culture, perception and cognition. En: *Perception of the environment. Essays on livelihood, dwelling and skill*, pp 157-171, Ingold, Tim (ed.), Routledge, London.

Jiménez, O. L. I. y E. Posselt S. (2009). *Dinámica del patrón de asentamiento en el municipio de Chalcatongo de Hidalgo. Informe.* Instituto Nacional de Antropología e Historia, Oaxaca, México.

Jiménez, O. L. I. y E. Posselt S. (2016). *Archaeological survey in the Mixtec Highlands, Mexico: the capital and ceremonial centre of Ñuu Ndaya. Report*. Leiden University, Faculty of Archaeology, Leiden University Founds.

Jiménez, O. L. I. y E. Posselt S. (s/f). *Las "Líneas de Vida" como una alternativa en el quehacer arqueológico: La percepción del Tiempo en el Paisaje de Ñuu Savi.* Universidad de Leiden Holanda, Facultad de Arqueología.

Jiménez, O. L. I., E. Cortés C., R. C. Reyes E. y E. Posselt S. (2017). El Cerro de Pedimento en Santa Catarina Yoso Notu, la reconstrucción en 3D de un santuario de origen precolonial. En: *Tierras y Dioses en la Mixteca*, Escamilla, María Concepción Reina Ortiz (ed.), pp. 285-326. vol. 16. Universidad Tecnológica de la Mixteca, Oaxaca, México.

Martínez, G. M. (1883). *Cuadros sinópticos de los pueblos, haciendas y ranchos del Estado Libre y Soberano de Oaxaca*, México.

Monsivais, J. (2013). La Arqueología Virtual en México, en: *Virtual Archaeology Review*, vol. 8, pp 59-64.

Spores, R. (2008). La Mixteca y los mixtecos, 3,000 años de adaptación cultural, en: *Arqueología Mexicana. La Mixteca, Tres mil años de cultura en Oaxaca, Puebla y Guerrero*, Vol. XV-núm. 90, pp 28-33.

Spores, R. (2014). La conquista española y sus consecuencias. La traza urbana europea y el propuesto viejo cabildo indígena de Yucundaa-Teposcolula. En: *Yucundaa. La ciudad mixteca y su transformación prehispánica-colonial I*, pp 315-327, Spores, Ronald y Nelly M. Robles García (eds.). Instituto Nacional de Antropología e Historia y Fundación Alfredo Harp Helú Oaxaca, A.C., México.

Spores, R. y A. K. Balkansky (2013). *The mixtecs of Oaxaca: ancient times to the present. The civilization of the American Indian 267*. University of Oklahoma Press.

Spores, R. y N. M. Robles G. (2014). Introducción. La transformación cultural de Yucundaa-Pueblo Viejo de Teposcolula, Oaxaca, México. En: *Yucundaa. La ciudad mixteca y su transformación prehispánica-colonial I*, pp 23-51, Spores, Ronald y Nelly M. Robles García (eds.). Instituto Nacional de Antropología e Historia y Fundación Alfredo Harp Helú Oaxaca, A.C., México.

Wijnhove, M. (2001). *Un estudio de las casas de los númenes en Ñuu Sau*, Oaxaca, Universidad de Leiden Holanda, Facultad de Arqueología.

Londoncharter (2009). *Carta de Londres. Para la visualización computarizada del patrimonio cultural. londoncharter, for the computer-based visualisation of cultural heritage*. Web.

La materialización de las ideas sobre el sitio pre colonial de Yuku Chayo en Chalcatongo, Mixteca Alta de Oaxaca (reconstrucción virtual tridimensional, representación fotorrealista e impresión 3D)

Laura Brenda Jiménez Osorio

En la actualidad, la difusión de estudios e investigaciones así como de sus correspondientes resultados no debería limitarse a la elaboración de un trabajo escrito. La necesidad de información del ser humano está sujeta al acelerado cambio tecnológico; la palabra escrita y la hablada no son suficientes para comprender los nuevos conocimientos, para procesarlos e incorporarlos en nuestro existir.

Al respecto, Otl Aicher menciona que:

> Con el descubrimiento de la imagen tomamos consciencia de haber entrado en la época de la comunicación. La sociedad deviene un fenómeno de la comunicación; aquella sólo se hace verdaderamente comprensible a partir de ésta. Lo social de la sociedad es su continuo intercambio de informaciones, la producción de contenidos de conciencia siempre nuevos. (2001:56)

Habitamos la era digital y formamos parte de su transformación. Lejos de negarnos a lo evidente, debemos aprovechar las posibilidades que esta etapa digital nos brinda para sumar a la construcción del conocimiento, es decir, pasar los límites de lo textual a lo visual, de lo visual a lo táctil, de lo táctil a la administración y la reconstrucción del saber.

En este sentido, las disciplinas del diseño, de la arquitectura y de la arqueología se integran para dar paso a una colaboración interdisciplinaria. A partir de la investigación previa del arqueólogo, el diseñador y el arquitecto intervienen en la mejora y el enriquecimiento del nivel de presentación de resultados habitual; es decir, a través de propuestas más realistas, de perfeccionamiento de acabados, de representación de contextos y de la materialización de la ideas, la comprensión del espacio es más accesible (figura 1).

Figura 1.- Reconstrucción
de templos del Yuku Chayo, 2018

Para ejemplificar lo antes mencionado: a continuación se describe brevemente el proceso llevado a cabo para el desarrollo y producción digital del sitio arqueológico Yuku Chayo:

Etapa de recreación: Lo figurativo idealizado, tiene que ver con los supuestos, interpretaciones y planteamientos del arqueólogo a partir de una investigación previa y del trabajo de campo.

Etapa de dibujo: El arqueólogo plasma y comunica sus resultados en papel, de forma manual, en 1 o 2 dimensiones.

Etapa descriptiva: El arqueólogo detalla por escrito las características esenciales de los espacios y entornos que componen el sitio de estudio.

Etapa de levantamiento y modelación tridimensional: Por medio de un software para modelado 3D, el diseñador realiza el trazo del sitio hasta llevarlo a una tercera dimensión (figura 2).

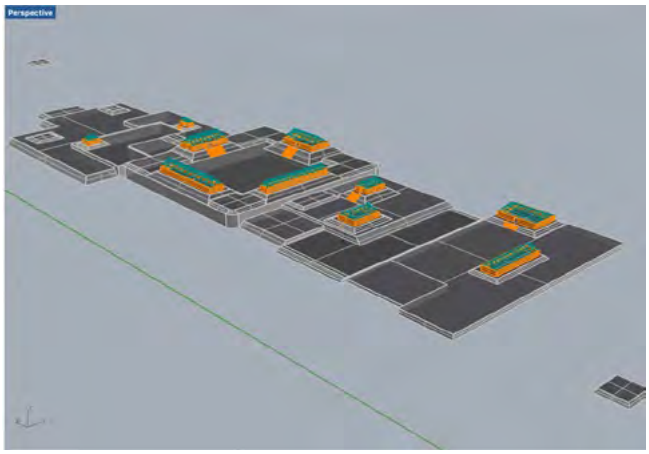


Figura 2.- Modelación tridimensional, 2018

Etapa de renderización - representación fotorrelista: el arquitecto asigna acabados o características específicas al modelo 3d a fin de generar las primeras *imágenes* realistas de nuestro modelo (también conocidas como renders).

Etapa de postproducción: por medio de un programa de edición fotográfica, el diseñador se encargará de perfeccionar el render antes obtenido, hasta llevarlo a un nivel de detalle y representación más apegado a la realidad (figura 3).

Figura 3.- Reconstrucción del sitio en su entorno, 2018

Etapa de impresión 3D: También conocida como *fabricación capa por capa* o *fabricación aditiva*, se define como un sistema de fabricación a base de superposición de capas sucesivas a partir de un material (Berchon y Luyt, 2016). Etapa final que comprende el siguiente nivel de representación que tiene que ver con la materialización de las ideas, sumamente valiosa debido a la posibilidad del investigador o del espectador de observar y tocar aquello de lo que se está hablando. El proceso aprendizaje en el ser humano siempre será mayor en la medida

que éste tenga la posibilidad de confrontar aquello que percibe visualmente con lo que percibe a través del tacto (figura 4).

Aunado a lo anterior, existe un punto fundamental respecto a la impresión 3D, este tiene que ver con el Patrimonio y su conservación. Un ejemplo de ello es el Museo Smithsonian de Washington, el primero en recurrir a esta tecnología para la conservación de sus obras. Por medio de la digitalización del l % de piezas de su catálogo, fue posible hacer una reproducción exacta en 3D que podría contemplarse, imprimirse y tocarse sin afectar la original (Berchon y Luyt, 2016).

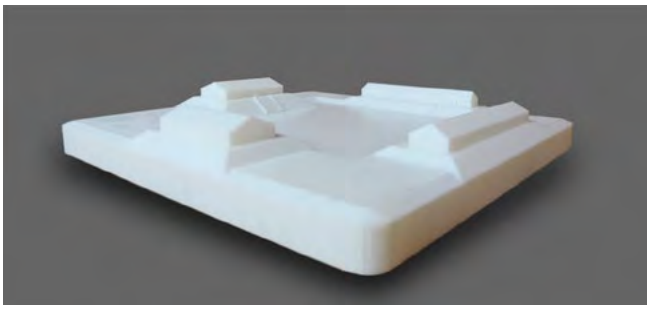


Figura 4.- Impresión del Modelo 3D, 2018

Se considera necesario que la arqueología involucre la reconstrucción y la impresión 3D como una herramienta de apoyo en la reproducción total o parcial de sitios arqueológicos. La investigación, el estudio, así como la preservación de restos arquitectónicos y de la cultura misma, tienen potencialidad gracias a la existencia de estas tecnologías.

En el ámbito del diseño, el alcance de los resultados obtenidos ha sido posible gracias a la utilización de distintos programas de visualización y representación así como de la continua experimentación a base de prueba y error. La popularización de la impresión 3D ha tenido gran impacto en el círculo creativo y por ende en nuestra sociedad, deviene en la democratización del conocimiento; la apertura, la evolución y la difusión del saber (figura 5).

Figura 5.- Presentación de este proceso creativo para la impresión 3D en una escuela de la comunidad de Chalcatongo, asistieron alumnos, maestros y autoridades, 2018

Bibliografía

Aicher, O. (2001). *Analógico y digital*. Barcelona: Gustavo Gili.

Berchon, M. y Luyt, B. (2016). *La impresión 3D. La guía definitiva para makers, diseñadores, estudiantes, profesionales y manitas en general.* Barcelona: Gustavo Gili.

Reflexiones y experiencias sobre la recreación de paisaje de sitios arqueológicos: Santuario de Pedimento de Santa Catarina Yoso Notu, Mixteca Alta de Oaxaca

Roberto Carlos Reyes Espinosa

En el presente trabajo se expone la relevancia de la experiencia vivencial por parte del investigador como parte fundamental para el proceso creativo en la reconstrucción tridimensional de un santuario milenario, se tomó como caso de estudio el Santuario de Pedimento de Santa Catarina Yoso Notu, ubicado en la Mixteca Alta del estado de Oaxaca, México.

Con referencia en algunos proyectos de similar índole, se nota que existe un aspecto que no se considera o no es considerado a menudo en la reconstrucción de sitios arqueológicos o santuarios, dicho aspecto es la experiencia que se vive en el lugar, tomándose en cuenta regularmente solo los aspectos materiales a reconstruir, como lo son arquitectura del lugar y una ambientación básica.

De este modo, en este proyecto se consideró significativa la experiencia vivencial que tiene el investigador para proponer la reconstrucción tridimensional del Santuario, en la cual existen elementos que se tomaron en cuenta para transmitir la esencia sagrada del mismo, siendo los siguientes: ser partícipe de la peregrinación y ritual de pedimento que se llevan a cabo el segundo viernes de cuaresma, realizar una visita guiada con arqueólogos para conocer la historia del lugar; a través de estas visitas se pudo apreciar el entorno del lugar (figura 1).

Figura 1.- De la capilla al cerro sagrado, viviendo el ritual de pedimento en Yoso Notu

Como aspectos importantes para transmitir la esencia sagrada del Santuario fueron considerados el ambiente (clima, vegetación, iluminación), arquitectura del sitio y decorado del mismo, así como la inclusión de ofrendas en altares y figuras religiosas.

Con la experiencia vivencial obtenida en las visitas al lugar se buscó: 1) recrear tridimensionalmente el entorno apreciado, de modo que fuera posible transmitir la esencia sagrada de un santuario milenario a peregrinos, pobladores aledaños y público en general (figuras 2 y 3), asimismo, 2) incentivar a otros investigadores a tener la experiencia vivencial al representar o reconstruir sitios arqueológicos y además 3) crear un puente entre el investigador y el usuario final a través de las experiencias que ofrece el lugar, para ello fue necesario también presentar este modelo en las comunidades de la Mixteca que estuvieron implicadas en esta investigación (figura 4).

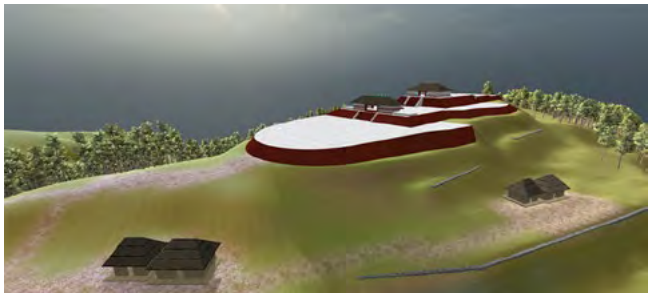


Figura 2.- Atardecer en el santuario precolonial del Cerro de Pedimento (reconstrucción virtual)

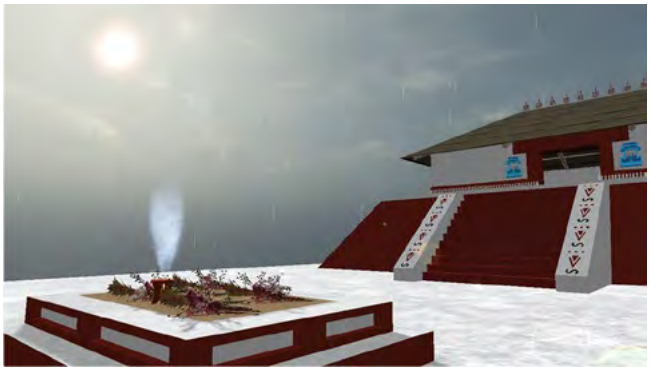


Figura 3.- Vista desde la plaza principal del santuario precolonial (a la izquierda la reconstrucción de un altar con su ofrenda y a la derecha el templo principal)

Es importante resaltar que para la realización de proyectos de este tipo se requiere de diversas disciplinas que se apoyen entre sí, de modo que se pueda tener un aprovechamiento de las herramientas tecnológicas y del pleno conocimiento de las ventajas y desventajas con las que se cuenta durante el proceso de trabajo.

El asistir al sitio arqueológico o santuario que se pretende reconstruir en algún proyecto es indispensable para la realización del mismo la experiencia vivencial ya que proporciona un enfoque distinto en el proceso creativo de reconstrucción y conecta al investigador con el entorno que rodea al lugar, se obtiene información relevante y que no se puede apreciar mediante datos o fotografías. Es información que se obtiene mediante los sentidos al involucrarse en el contexto que se da en el lugar y es un reto que como investigadores se debería asumir al intentar plasmar esas sensaciones experimentadas.

Figura 4.- Recorrido virtual en el santuario precolonial del Cerro de Pedimento, pruebas realizadas por los pobladores de la comunidad de Yoso Notu

Diseño de interfaz gráfica para facilitar y reforzar la Cultura Viva en los Pueblos de la Mixteca Alta de Oaxaca

Eruvid Cortés Camacho

El diseño de Interfaz Gráfica de Usuario (GUI) por sus siglas en inglés o Interfaz de Usuario (UI) es el diseño de un sistema (menús, navegación, mecanismos de control) que comunica al usuario con la aplicación o software (Salmond y Ambrose 2014), es decir, el entorno visual en que se desarrolla la interacción entre la persona y el dispositivo.

En esta presentación se abordarán temas que muestran la importancia de usar interfaces gráficas en las comunidades Indígenas de Oaxaca, las cuales deben ser diseñadas tomando en cuenta, tanto el contexto social y cultural como las necesidades y características específicas de los usuarios (Pratt y Nunes 2013). En este caso, la interfaz fue diseñada para los habitantes de Santa Catarina *Yoso Notu* y San Miguel el Grande, municipios de la Mixteca Alta de Oaxaca (figura 1).

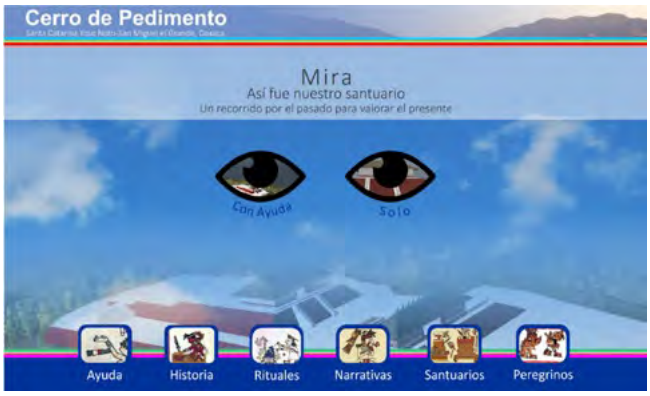


Figura 1.- Diseño de la interfaz gráfica propuesta para las comunidades de Yoso Notu y San Miguel el Grande, Oaxaca

El modelo de interfaz que se usará es el denominado Objeto-Acción o (OAI) Objet-Action-Interface, que está basado en la representación visual de objetos y acciones de la tarea del usuario, para ello se requiere de buscar analogías con el mundo real del usuario respecto a la acción para apelar a su intuición. El proceso de análisis del usuario, desarrollo metodológico y evaluación de la propuesta final, se presentarán en el presente trabajo.

En muchos municipios del estado de Oaxaca, el contexto físico y los escasos medios de comunicación (en sentido amplio), han dificultado el desarrollo de los Pueblos Indígenas. Sin embargo, al mismo tiempo esto ha permitido la conservación de una larga tradición cultural, haciendo visible el fuerte vínculo entre el pasado y el presente. Este vínculo representa una oportunidad para los especialistas en Arqueología y Herencia Viva, ya que pueden dar seguimiento a través del tiempo de las manifestaciones culturales. De acuerdo a las investigaciones arqueológicas en esta área un ejemplo de esta continuidad es la peregrinación al Santuario del Cerro de Pedimento, ubicado en los municipios señalados. En este santuario se enfocara la interfaz gráfica.

En este contexto, uno de los retos es como comunicar los resultados de las investigaciones aprovechando las nuevas tecnologías, rompiendo la brecha impuesta a estas comunidades por la falta de políticas incluyentes por parte del estado y el gobierno, lo que ha generado un aislamiento tecnológico. Este reto tiene que ver también con el hecho de no afectar las costumbres religiosas sino reforzarlas para las futuras generaciones. Para ello se analizan todas las posibilidades físicas y técnicas, buscando las más adecuadas, abordando las ventajas y desventajas de cada una para llegar a una solución centrada en el usuario. Asimismo otro elemento importante fue el diálogo y evaluación de este medio con las comunidades involucradas (figuras 2 y 3).

Figura 2.- Presentación sobre la investigación y el método empleado en la interfaz, en la comunidad de

Figura 3.- Pruebas de usabilidad de la interfaz en San Miguel el Grande

Bibliografía

Pratt, A. and J. Nunes (2013). *Diseño Interactivo*. Océano, España.

Salmond, M. and A. Gavin (2014). *Los Fundamentos del Diseño Interactivo*, Blume, Barcelona España.

Recorridos virtuales interactivos en tiempo real: una alternativa de presentación de información del sitio arqueológico Cerro de las Minas, Mixteca Baja, Oaxaca

José Abel Martínez Guzmán

En este trabajo se pretende dar a conocer la aplicación y el uso de la tecnología para la difusión y presentación de información de sitios arqueológicos en este caso del cerro de las minas, ubicado en la heroica ciudad de Huajuapan de león, Oaxaca.

Hoy en día de estos asentamientos humanos que en el pasado tuvieron lugar, solo quedan restos materiales de lo que antes fue una gran civilización. Estos hechos históricos tomaron lugar en el pasado y es por ello que a través de la investigación arqueológica podemos estudiarlos.

Es gracias a este tipo de investigación arqueológica realizados a partir de 1960 en la Mixteca Baja que hoy podemos conocer la manifestación de una gran cultura

denominada “ñuiñe” que tuvo apogeo en el periodo clásico del año 400 d. C al año 800 d. C.

Es sabido que la gran mayoría de personas de la sociedad mexicana actual, tienen una enorme falta de conocimientos de sus raíces históricas. Más aún, desconocen datos de importancia de aquellos sitios que de una u otra forma fueron importantes para formar el país que se tiene ahora. Esto sucede debido a que la mayoría de la información sobre ese pasado se limita a libros y documentos almacenados en bibliotecas y centros educativos; además de la forma en que se presenta, pues esta información suele ser muy especializada, extensa y escasamente ilustrada, de manera que acceder a ella resulta ser una tarea tediosa y aburrida.

En la actualidad el desarrollo de nuevas tecnologías ha permitido que gran cantidad de información que generalmente era contenida en libros se divulgue de una manera más fácil, en el caso de la aplicación de la tecnología en la difusión de información de zonas arqueológicas, podemos observar que hoy en día podemos ver más tours virtuales o reconstrucciones virtuales que comienzan a ser expuestos en internet.

Sin embargo podemos señalar que la fidelidad de estas reconstrucciones referente al uso de texturas y representación arquitectónica no ha sido la adecuada debido a las técnicas e informaciones que se utilizan como fuente, además de que la presentación de información se limita a ser presentada de manera lineal.

Es por ello que el siguiente trabajo trata del uso de recorridos virtuales e interactivos en tiempo real para presentar la información del sitio arqueológico el cerro de las minas y las principales ventajas que podemos encontrar son las siguientes:

1. Total realismo e inmersión en la visita mediante una reconstrucción 3D.
2. Absoluta innovación, se trata de un formato sorprendente para el usuario.
3. Gran cantidad de información que puede ser introducida en la visita, el usuario conocerá todos los detalles, mientras se entretiene en la visita.
4. Las imágenes y vídeos son la información más demandada por los usuarios en internet además de que al tenerlo publicado en internet, es una información disponible 24h/365 días del año.

A continuación se presentan algunas capturas de pantalla, de una aplicación diseñada para sistemas operativos Windows, dicha aplicación está desarrollada en motor de juego denominada Unity 3d, mediante el uso de esta herramienta se puede crear recorridos virtuales interactivos en tiempo real, a partir de una reconstrucción virtual de un Sitio Arqueológico, para este caso de estudio se creó una aplicación para el Sitio Arqueológico el Cerro de las Minas, ubicado en la Heroica Ciudad de Huajuapan de León, Oaxaca (figuras 1-4).

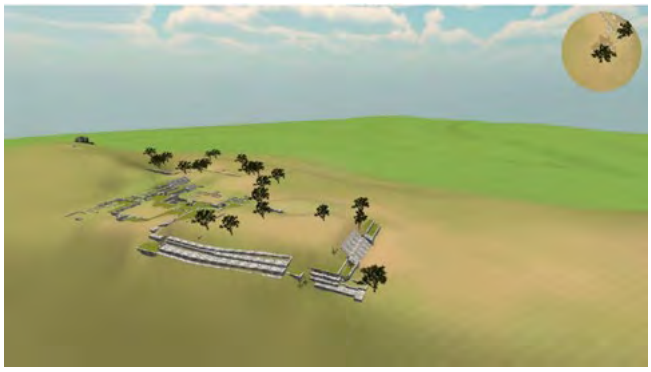


Figura 1.- Se muestra una vista aérea de la reconstrucción del Sitio Arqueológico Cerro de las minas. En dicha imagen se puede apreciar los diferentes accidentes topográficos del Sitio Arqueológico, lo que permite poner en un contexto los diferentes elementos arquitectónicos que conforman el Sitio Cerro de las Minas

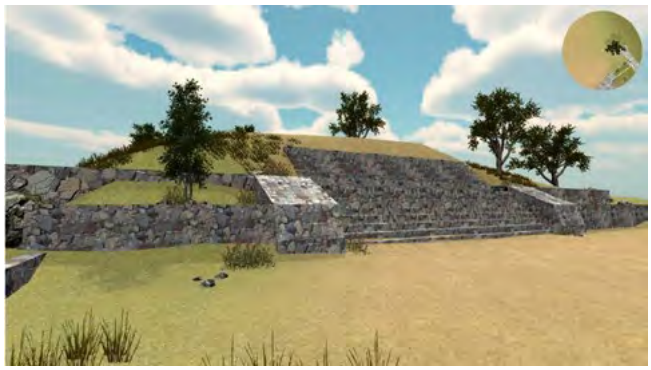


Figura 2.- Se muestra una vista hacia una escalinata desde la plaza 2. En esta imagen se puede apreciar el uso de texturas, iluminación y elementos 3d para ambientar el Sitio Arqueológico, y con ello dar realismo al Sitio

Figura 3.- Se muestra una vista al Sur-Oeste del Sitio, desde la plaza 1. En esta imagen se puede apreciar que las reconstrucciones virtuales, desarrolladas en un motor de juego como lo es Unity 3d, es una alternativa innovadora para presentar y difundir información del Sitio Arqueológico

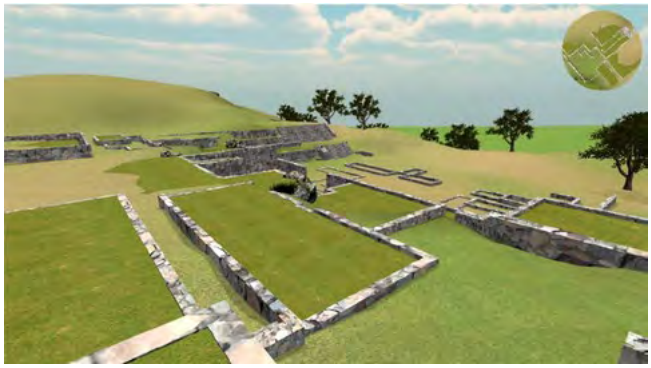


Figura 4.- Se muestra una vista al Sur del Sitio, desde la plaza 3. Las reconstrucciones virtuales e interactivas en tiempo real, permiten establecer recorridos guiados al usuario, además de tener total libertad para poder desplazarse libremente en la aplicación

Figura 5.- Se muestra la misma vista de la imagen anterior pero desde el lado norte. Es una foto del sitio Cerro de las Minas que nos permite comparar entre lo real y lo virtual

Project Management For The Digital Humanities

Natalia Ermolaev
nataliae@princeton.edu
Princeton University, United States of America

Rebecca Munson
rmunson@princeton.edu
Princeton University, United States of America

Xinyi Li
xinyili@princeton.edu
Princeton University, United States of America

Lynne Siemens
siemensl@uvic.ca
University of Victoria, Canada

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Micki Kaufman
mickikaufman@gmail.com
The City University New York, United States of America

Jason Boyd
jason.boyd@ryerson.ca
Ryerson University, Canada

Many projects taken on by humanists -- whether large scale with many team members and substantial budgets, or smaller, such as editing books and journals -- require management. Regardless of size, scope, and budget, projects members must coordinate tasks, responsibilities, budgets and achieve objectives (Boyd and Siemens 2014; Siemens 2009). Project management (PM) with its accompanying methods, tools and techniques provides one way to accomplish this. PM can help manage common issues related to risks, obstacles and tasks which might be unanticipated, team member turnover, timelines, scope creep, and budget overspending (Siemens 2016). To facilitate skill development in this area, there are training opportunities within Digital Humanities courses (Bailar and Spiro 2013), stand alone courses (University of Alberta 2015), training programs (Scholars' Lab 2011), workshops (DHSI 2015; HILT 2015; The European Summer University in Digital Humanities 2015) and, finally, supporting websites (Appleford and Guiliano 2013).

While it has proven extremely useful to apply PM methods and tools to DH work, this panel examines the reverse: how do the principles, methods, and concerns of DH inform our PM methods and techniques? How do we adapt PM frameworks to address issues specific to DH projects - such as complex scholarly research agendas, or interest in topics such as community engagement, design thinking, open source development, activism, etc.? This panel is concerned with what it means to incorporate project management into a DH project, looking particularly from the perspective of individuals who shape and implement these methods and tools. While scholars have reflected on aspects such as teamwork and collaboration (Siemens and INKE Research Group 2015; 2016; Ruecker and Radzikowska 2007) and project process and outcomes (Causer, Tonra and Wallace 2012; Simeone *et al.* 2011; National Endowment for the Humanities Office of Digital Humanities 2010; Guiliano 2012), surprisingly little attention has been paid to how DH transforms project management implementation.

This panel seeks to address this gap with papers that demonstrate various project management methodologies specifically for the DH context. Panelists will discuss: how design can be integrated into the project management process, how PM can support the creation of a distributed and emergent open source development model, how PM can facilitate rigorous and satisfying interpersonal scholarly exchange, and how PM has been used to manage a multi-year, large scale DH project with over 35 partners. The panel's goal is to showcase solutions to

issues that arise in DH work, and to see if we can derive a set of general principles or processes inherent in project management for the digital humanities.

Project Management and INKE

The Implementing New Knowledge Environments (INKE) is a large-scale, long-term interdisciplinary research project that has been researching the future of books, e-books and reading. To coordinate tasks, budget and a research team with over 35 members, research assistants, postdoctoral fellows, and partner organizations, the collaboration used a combination of project management tools. In particular, INKE incorporated governance documents and a yearly planning cycle with associated research plans.

INKE's governance documents were designed to guide the collaboration and support accountability by providing a foundation of common understandings. At the start of funded research activity, the administrative team jointly developed these and laid out the working relationship between researchers, the sub-research areas, the administrative team, partners, and the executive committee, and outlined an authorship convention, intellectual property clause, and decision-making and dispute resolution processes, among other things. An important part of these documents was a researcher agreement that all team members signed before receiving research funds. To further accountability, a copy of the governance documents were also posted on the online project planning workspaces as well as published and updated as necessary. These documents also proved useful for incorporating new team members and sustaining the working relationships. As a sign of their strength, these have served as models for other team projects (Nowviskie 2011; The Praxis Program at the Scholars' Lab 2011; nd).

Another important project management tool was the annual project plans where each sub-research area needed to develop to receive research funds. These outlined research tasks, outcomes, responsibilities and accountabilities, timelines and required resources. With approval of these documents, funds were then distributed to the sub-research areas and their research started for the year. To ensure accountability, the team reported at multiple points of the year and compared actual activities against those planned. The administrative team realized that this was not something to which they were accustomed and required skills that are not typically developed in graduate school and were often the equivalent to writing an article in terms of intellectual effort. Having said that, the administrative team realized that this planning process provided an important foundation to create cohesion, and underpin the project's working culture and serve to ensure that research was still completed even when researchers were busy with other responsibilities. Finally, given the pace of technological change, the yearly planning cycle made it easier to plan tasks that could be accomplished within a shorter timeframe while still addressing the overall research question which had a seven year mandate.

Overall, INKE has been a successful research endeavor as measured in terms of conference presentations, articles, and prototypes. This project management framework contributed to that success.

Design for Digital Humanities Project Management

The project management workflow at the Center for Digital Humanities at Princeton (CDH) changed significantly when we hired a User Experience Designer in 2016. Though the CDH had only been developing DH projects for two years, we had established a robust project management process in consultation with our institution's OIT Project Management Office, and with insights from literature and models from PM resources within the DH community (Siemens, 2016; Leon, 2011). But a designer's input helped surface aspects of the process that are crucial for DH work, and our revised workflows have enriched both research outcomes and product deliverables.

In this talk we will discuss why and how design can be integrated into the DH project management process. Visualization and design are becoming increasingly important in DH projects, and major points of intersection between design and DH have emerged. And we feel that DH project management would benefit from more engagement with the perspectives of theorists and practitioners in the design disciplines (Blauvelt 2008; Maurer et al, 2008).Design can play a key role in the "thinking-through-practice" (Burdick et al, 2012) ethos of DH work, and can contribute to the research process by shaping communication and argumentation. The addition of a "designerly way of knowing" (Archer, 1979) into the DH project management process can enhance research approaches by fostering productive synthesis in teams with diverse expertise and content knowledge. When design thinking and tools are integrated into the co-creation of research, tool- and resource-building, new methods of inquiry emerge that deepen collaboration and enhance knowledge-making.

We will open by discussing the role of design in the current non-DH Project Management field. We will then outline our own PM methodology, and describe the interventions of design in this process: tools and strategies such as creating sitemaps, siteflows, interface wireframing, art direction, design mockup and acceptance testing. Our presentation will be supported by examples from the major projects we have developed at our Center, which feature performative interfaces that go beyond pre-conceptualized interpretation and arguments to encourage discovery. These examples will demonstrate how design practice and ideation can inform each other in a iterative process of synthesis and refinement, and can facilitate diagrammatic thinking to help those unfamiliar with visual thinking adopt new approaches and perspectives. Im-

portantly, we will conclude by discussing the challenges of integrating design into the project management process, and offer suggestions for how to overcome common roadblocks and misunderstandings.

DH Project Management as Scholarly Exchange

When considering the ways in which the principles, methods, and concerns of Digital Humanities (DH) can usefully inform and adapt established Project Management (PM) methods and techniques, it is helpful to observe at the outset that "management" is not a particularly valourized term in the (digital) humanities. Regarded as a (cold, profit-driven) business mechanism rather than an important aspect of scholarly practice, (digital) humanities faculty typically bristle at the idea that their research (individual or collaborative) should be subject to managerial protocols—an attitude only exacerbated by the pervasive administrative control exercised over their professional lives as a result of the "corporatization of the university." "Digital humanities" has even been accused (along with other transgressions against the humanities) of really being the "managerial humanities" (Allington) because DH projects can require substantial grants (usually to employ research assistants and technicians) that require strict administration and reporting, thus turning researchers into project managers (which, Allington presumes, is a bad thing).

However, in this presentation I will argue, based on my own experience as a DH scholar and project manager that project management, as it adapts to the particularities of (digital) humanities project requirements and personnel is, at its best, a collegial facilitation of rigorous and satisfying interpersonal scholarly exchange that is not available in familiar modes such as the conference presentation, the academic journal monograph review or the blind peer review process. The key practices of project management as they function in DH, I suggest, contribute to optimizing a sustained, substantive, and productive dialogue that can both "get things done" and contribute to intellectual and professional growth.

Themes of Community-Driven Project Management

The purpose of project management is to leverage and coordinate the creativity and effort of human beings in a common commitment to accomplish a shared goal. Having managed a wide range of projects in music composition, performance and production, print and film/video production, commercial software development and, most recently, open source digital humanities projects, the core skills required from a project manager remain essentially those of a skilled conductor, irrespective of medium. Nevertheless, projects in the digital humanities present their own unique challenges for effective management.

The essential challenge facing the project manager – to coordinate actions across a team's heterogeneous backgrounds, requirements and skill sets – is foregrounded in Digital Humanities projects. Academic projects span a vast gamut of human interest, far more than commercial efforts tailored for profit, and they are evaluated not by market penetration or sales but on the value of their scholarly contribution. Unlike many commercial projects, in an academic project context, team members often cannot readily be hired or replaced for the purposes of fulfilling the needed skill sets at hand. In addition, academic projects can often depend on only a fraction of the work hours otherwise available from each team member, causing additional barriers to effective team interactions.

Accompanying and informed by the project's efforts to overcome challenges of staffing are matters of process. Within a digital humanities team, one often finds mismatching working hours between staff, significant differences in prior professional experience and other team issues caused by the conflicting demands of the academic context. As a result, process often evolves as a crutch or mitigation of the team's shortcomings rather than an emergent behavior that maximize their strengths. Good DH project and project management 'hygiene' requires that the manager(s) and team properly apply, and effectively cultivate, a team ethic that empowers all to provide ongoing input on the team structure and dynamics, a dialogue that informs and engages the appropriate levels of process.

To accomplish success in a DH context, every project manager must compose the appropriate team according to the available talent, identify and understand the goals and requirements of the project, and coordinate the team's activities according to the needs of a diverse user community. In my prior project management career, I have been gratified to be able to help compose and cultivate a team of talented collaborators, and to help a team process emerge from within the diverse team's core strengths and identities. By carefully evaluating various distributed and feature-specific team models, and by taking a minimalistic approach to job- and issue-tracking, a team can often emerge its own process, grounded in the scholarly context of its genesis and eventual reception that truly speaks to the scholarly intent of the project.

References

Allington, Daniel. "The Managerial Humanities; or, Why the Digital Humanities Don't Exist." Daniel Allington (blog). 31 Mar. 2013. http://www.danielallington.net/2013/03/the-managerial-humanities-or-why-the-digital-humanities-dont-exist/

Appleford, Simon, and Jennifer Guiliano. "Devdh: Development for the Digital Humanities". 2013. March 5, 2015. <http://devdh.org>.

Archer. L.B. "Whatever Became of Design Methodology." Design Studies, 1.1 (1979): pgs. 17–18. Quoted in Cross. N. "Forty Years of Design Research." Design Research Quarterly, 2.1

Bailar, Melissa, and Lisa Spiro. "Introduction to Digital Humanities". 2013. March 5, 2015. <http://digitalhumanities.rice.edu/fall-2013-syllabus/>.
Blauvelt, A. "Towards Relational Design," Design Observer, 3 November, 2008: http://designobserver.com/feature/towards-relational-design/7557
Boyd, Jason, and Lynne Siemens. "Project Management." DHSI@Congress 2014. 2014.
Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J. "Humanities To Digital Humanities." In *Digital_Humanities*, p. 13. Cambridge, MA, USA: MIT Press, 2012.
Causer, Tim, Justin Tonra, and Valerie Wallace. "Transcription Maximized; Expense Minimized? Crowdsourcing and Editing the Collected Works of Jeremy Bentham." *Literary & Linguistic Computing* 27.2 (2012): 119-37.
DHSI. "Digital Humanitites Summer Institute". 2015. March 5, 2015. <http://dhsi.org/>.
Guiliano, J. "Neh Project Director's Meeting: Lessons for Promoting Your Project." MITH Blog2012. Vol. October 3, 2012.
HILT. "Courses". 2015. March 5, 2015. <http://www.dhtraining.org/hilt2015/>.
INKE. "Implementing New Knowledge Environments". 2012. September 22, 2012. <http://inke.ca>.
Leon, Sharon M. "Project Management for Humanists: Preparing Future Primary Investigators". 2011. June 24, 2011. http://mediacommons.futureofthebook.org/alt-ac/pieces/project-management-humanists.
Maurer, L., E. Paulus, J. Puckey, and R. Wouters. "Manifesto - Conditional Design" Conditional Design, 2008: http://conditionaldesign.org/manifesto (Accessed 5 October 2014)
National Endowment for the Humanities Office of Digital Humanities. *Summary Findings of Neh Digital Humanities Start-up Grants (2007-2010)*. Washington, D.C.: National Endowment for the Humanities, 2010.
Nowviskie, Bethany. "Where Credit Is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship." *Profession* 13 (2011): 169–81.
The Praxis Program at the Scholars' Lab. "2011-12 Praxis Program Charter". 2011. September 19, 2012. <http://praxis.scholarslab.org/charter.html>.
The Praxis Program at the Scholars' Lab. "Toward a Project Charter". nd. October 19, 2017. <http://praxis.scholarslab.org/resources/toward-a-project-charter/>.
Ruecker, Stan, and Milena Radzikowska. "The Iterative Design of a Project Charter for Interdisciplinary Research." DIS 2007. 2007.
Scholars' Lab. "The Praxis Program at the Scholars' Lab". 2011. September 12, 2011. <http://praxis.scholarslab.org/>.
Siemens, Lynne. "Dhsi Project Planning Course Pack". 2012. March 5, 2015. <http://dhsi.org/content/2012Curriculum/12.ProjectPlanning.pdf>.
Siemens, Lynne. "'It's a Team If You Use "Reply All"': An Exploration of Research Teams in Digital Humanities Environments." *Literary & Linguistic Computing* 24.2 (2009): 225-33.
Siemens, Lynne. "Project Management." *Digital Pedagogy in the Humanities: Concepts, Models, and Experiments*. Eds. Rebecca Frost Davis, et al. New York: MLA Commons, forthcoming.
Siemens, Lynne. "Project Management and the Digital Humanist." *Doing Digital Humanities: Practice, Training, Research*. Eds. Constance Crompton, Richard J. Lane and R.G. Siemens. New York: Routledge, 2016. 343-57.
Siemens, Lynne, and INKE Research Group. "Faster Alone, Further Together: Reflections on Inke's Year Six." *Scholarly and Research Communication* 7.2 (2016): 1-8.
Siemens, Lynne, and INKE Research Group. ""Inke-Cubating" Research Networks, Projects, and Partnerships: Reflections on Inke's Fifth Year." *Scholarly and Research Communication* 6.4 (2015).
Simeone, M., et al. "Digging into Data Using New Collaborative Infrastructures Supporting Humanities-Based Computer Science Research." *First Monday* 16.5 (2011).
The European Summer University in Digital Humanities. ""Culture & Technology" - the European Summer University in Digital Humanities". 2015. March 5, 2015. <http://www.culingtec.uni-leipzig.de/ESU_C_T/node/97>.
Stanford Humanities + Design Research Lab (http://hdlab.stanford.edu).
University of Alberta. "Courses". 2015. October 13, 2017. <https://www.ualberta.ca/interdisciplinary-studies/humanities-computing/huco-courses>.

Can Non-Representational Space Be Mapped? The Case of Black Geographies

Jonathan David Schroeder
jdss@uchicago.edu
University of Warwick, United Kingdom

Clare Eileen Callahan
ccallahan@utexas.edu
University of Texas, Austin, United States of America

Kevin Modestino
kevin.modestino@howard.edu
Howard University, United States of America

Tyechia Lynn Thompson
tlthompson@bison.howard.edu
Howard University, United States of America

Description: This panel examines the ways spatial and geographical formation in African American Studies have integrated with or failed to integrate with Digital Humani-

ties scholarship and the growing use of mapping technology. Literary and historical scholarship on black geographies has grappled with the epistemological, political, and ethical problems of recovering the locations and routes of black resistance in both the antebellum U.S. and Jim Crow South. Katherine McKittrick has criticized the resulting tendency of scholars to translate blackness as "ungeographic" and Saidiya Hartman similarly writes about the historian's struggle "within and against the constraints and silences imposed by the nature of the archive."[1] Literature is, therefore, an important medium for African American Studies because it does not need to be verifiable and, in this way, speaks to "those qualities of spatial and geographical formations that are most difficult to detect from within the established, formalized explanatory frameworks of the physical and social sciences." Literature, in other words, has an advantage in representing spaces that have failed or have refused to be precisely represented.

What does the unmappability—the precise imprecision of the African-American Archive of resistance--have to say to our employment of mapping tools in the digital humanities? This panel seeks to address not only the possibilities of employing GIS technology to engage with the challenges posed by mapping African American literature (such as location vagueness or deliberate obfuscation), but also how African American Studies scholarship might help us rethink the development of mapping technologies. This panel will (1) prompt a theoretical reflection on how the digital, in its own ephemerality, might offer a privileged medium for thinking and visualizing such spaces and more broadly, (2) reflect on how DH can interact with the fragmentary archive upon which much of African American Studies relies. While literary mapping projects have engaged with the problem of spatial uncertainty in fiction and have developed methods in which to represent that uncertainty, we are interested in exploring how the limits of GIS can allow us to engage critically with deliberate obfuscation, that is, more generally, with the question of what it means to map a space that fostered black agency because of its original unmappability. How can GIS not merely represent spatial uncertainty but also critically engage with the absences and silences of the archive in a way that maintains the integrity of those silences?

Clare Callahan will discuss her work-in-progress, "Not-Quite Digital Cartography," which employs digital mapping tools to examine the geography of "not-quite" spaces in black "Post- Bellum, Pre-Harlem" literature. This paper will focus specifically on the fiction of W.E.B Du Bois and Pauline Hopkins. African American literature of this period is marked, Callahan argues, by an ambivalence toward representation and the subject to which it is tethered. The novels this project examines reimagine black subjectivity as a departure, in the dual sense of that term, from the history of black fugitivity of the antebellum south and of the Reconstruction period. This project maps, in other words, the literary landscapes characterized by a simultaneous resistance to and demand for representation—the swamp-settlement in W.E.B. Du Bois's *The Quest of the Silver Fleece* (1911), the hidden city of Telassar in Hopkins *Of One Blood* (1902), for example.

Recent scholarship on digital literary cartography has sought to address digital methods for mapping of literary spaces that only vaguely or tenuously correspond to actually existing geography. In "Mapping Literature: Towards a Geography of Fiction," Barbara Piatti et al. explicate the possibilities for mapping "imprecise geography" in fiction on multiple spatial levels. Similarly, in "Mapping Literature: Visualisation of Spatial Uncertainty in Fiction," Ann- Kathrin Reuschel and Lorenz Hurni propose a methodology for mapping a work of literature when "determining the location is only possible imprecisely."[2] But few, if any, such scholarly articles propose a *theoretical* inquiry into the stakes of mapping spaces that originated in and through concealment and which were sustained only insofar as they eluded mappability. Indeed, such spaces underpin and, in many cases, make possible the black narratives that speak to and of them.

This paper argues, first, that a more sustained theoretical reflection on what it means to map, however imprecisely, what Hortense Spillers refers to as "not-quite" spaces—the hidden geographies of black resistance—must be the first step toward determining a methodology for mapping African-American literary geographies more generally. Second, this paper proposes, through a digital cartography of the fiction of Du Bois and Hopkins, that theoretical reflection on the black spaces that confound the economy of representation prompts a reframing of how digital humanists understand the function of literary mapping in the digital age. The digital is itself, in many respects, a "not-quite" space, neither there nor not there, and may, therefore, open up new possibilities for visualizing and engaging with the literary aesthetic of ambivalence toward representation, toward becoming representable, that Callahan identifies as characteristic of Post- Bellum, Pre-Harlem literature.

Kevin Modestino will talk about his geocoding project, "William C. Nell's Revisionist Revolution," which maps out an important abolitionist history, Nell's *The Colored Patriots of the American Revolution* (1854). This mapping

1 See McKittrick, Katherine. *Demonic Grounds: Black Women and the Cartographies of Struggle*. (Minneapolis: U of Minnesota), 5; Hartman, Saidiya. *Scenes of Subjection: Terror, Slavery, and Self-Making in Nineteenth-Century America* (New York: Oxford University Press, 1997), 11.

2 Barbara Piatti, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, William Cartwright, "Mapping Literature: Towards a Geography of Fiction," in *Cartography and Art,* eds. William Cartwright, Georg Gartner, and Antje Lehn (Berlin: Springer-Verlag, 2009), 1-16; Ann-Kathrin Reuschel and Lorenz Hurni, "Mapping Literature: Visualisation of Spatial Uncertainty in Fiction" *The Cartographic Journal* 48, no. 4 (Nov. 2011), 293-308; see also, Eric Prieto, "Geocriticism, Geopoetics, Geophilosophy and Beyond." Ed. Robert Tally. *Geographical Explorations: Space, Place, and Mapping in Literary and Cultural Studies* (Basingstoke: Palgrave Macmillan, 2011).

project is an attempt not only to investigate and visualize what Martha Schoolman has identified as the key political interventions of abolitionist geographies but also to ask how digital humanities tools might respond to and be transformed by the African-American literary history.[3]

African-American texts, Modestino argues, present unique challenges to DH tools. In the case of geocoding, a critical awareness of how mapping served as a tool for the surveillance and conquest of enslaved, maroon, and indigenous populations has to remain at the center of any investigation. But just as black historians transformed the imperialist narratives of nineteenth- century nationalist histories from within, it is also possible to imagine a creative mapping of American space that would transform the false totalizations of the base map images utilized in typical DH projects through the overlaying of disfiguring lines of abolitionist flight and revision.[4]

By mapping various episodes from Nell's text, we can see a North America overlaid with a least six modes of C19 black movement: 1) two modes of Black Atlantic movement (Middle Pass and Transatlantic); 2) fugitive slave movement; 3) internal slave trade movement; 4) maroonage, and slave revolt geographies; 5) once-occluded geographies of black soldiers at the sites of national memory; and 6) centers of black abolitionist action and organizing. This mapping allows us to see how in even a single text of black American history encodes a multiplicity of critical geographical interventions that can be used to layer the visual narratives of maps with subversive and disfiguring traces.

Jonathan Schroeder will discuss "Passages to Freedom: Worlding the North American Narrative," a digital mapping project that he created with Douglas Duhaime at the Yale Digital Humanities Laboratory in 2017. In its current iteration, *Passages* maps the routes taken out of slavery by the authors of 22 of the 103 extant pre-emancipation slave narratives. When completed, it will map the 286 pre- and post-emancipation texts that make up the University of North Carolina's North American Slave Narrative corpus. The aim is to study black mobility in a different light, first by demonstrating that the slave narrative is empirically and emphatically a global genre. Authors from Olaudah Equiano to Henry "Box" Brown traveled by rail, sail, and even mail to escape slavery. Yet despite the surge of interest in questions of black fugitivity, no comprehensive study of black mobility in this genre exists today, and in fact very few surveys of the genre have been performed since landmark studies of the 1970s like Frances Smith Foster's *Witnessing Slavery*. What types of mobility can be distinguished within the genre, which constitutes perhaps the richest and most important source of descriptions of black mobility? Are there characteristic forms of mobility that correspond to the three phases of the narrative – slavery, flight, and fugitive freedom?

In light of the setting for this talk, Schroeder will devote specific attention to the question of how digital methods can both help and obscure the various knowledges that humanists have teased from these narratives. For example, while Frederick Douglass's *The Narrative of the Life of Frederick Douglass* (1845) is quite clearly a masterful instance of political rhetoric, it is only when we begin to think in geographic terms that we can see that it is the most location-specific narratives in the entire genre, providing street-level information about 1830s Baltimore to help demonstrate the most granular effects of slavery.

At the same time, the effort to map these narratives also raises highly important questions about symbolic representation, as the data extracted from these narratives differ radically in terms of degree of specificity, frequency, and related factors. For example, the repetitive movements associated with slave labor and shipping routes (frequently the occupation of fugitive slaves) are difficult to represent using standard mapping techniques and point to alternate forms of representation like digital animation.

In exploring the tensions that arise when black geographies and mobilities are translated into the digital research environment, Schroeder will give an account of the findings, insights, and problems raised by the project, while also inviting audience members to participate in the shaping of the future of the project.

Tyechia Lynn Thompson will discuss the creation and revision of her project, "Baldwin's Paris," which quantifies, visualizes, and analyzes over four decades of James Baldwin's writings about Paris. Her talk will open with a consideration of the method she used to create the project and conclude with a consideration of her current focus on UI/UX design in its recreation.

In its formative stages, *Baldwin's Paris* steered close to Matthew Jockers's claim that "a good deal of computational work is specifically aimed at testing, rejecting, and confirming, what we think we already know."[5] Yet if Jockers used R to demonstrate empirically that *Moby-Dick* is an aberration from 1,000 contemporaneous American novels, Thompson's project initially sought to gauge the significance of Baldwin's post-1963 work by testing Baldwin's representations of Paris via Google Earth. Despite limitations to this method, which necessitated significant guesswork due to placemarking inaccuracies in aerial and street view photography, it served as a launch pad to begin a critical analysis of Baldwin's work.

In the project's further elaborations and revisions, Thompson critiques the use of geographic information system tools to "map" a writer who had experienced extensive surveillance by the FBI. This critique is informed

3 Schoolman, Martha. *Abolitionist Geographies* (Minneapolis: U of Minnesota P, 2014).

4 Ernest, John. *Liberation Historiography: African American Writers and the Challenge of History, 1794-1861* (Chapel Hill: U of North Carolina P, 2004).

5 Jockers, Matthew. *Text Analysis with R for Students of Literature* (New York: Springer, 2014)

by African American literary theories of place, especially Toni Morrison's "literary archeology" and Baldwin's notions of being in "contact" and being a "witness."

The next stage in the development of "Baldwin's Paris" will move beyond the initial work of tagging locations in Baldwin's texts to embedding the theories of literary archeology, being in contact and being a witness, into Carto and the interface of the site. For it is in UI/UX design that the connections between data, tool, and critique ultimately lie.

***

This panel ultimately seeks to think through a hybrid empirical-theoretical approach; that is, a hybrid of the empirical methodologies that have characterized DH and of the more theoretical concerns of traditional humanities scholarship. The panel will ask what new methodologies, for both working with existing digital tools but also for directing the conceptualization of new tools informed by the unique demands of black literary geographies, emerge from the above hybrid frameworks: the mapping of the aesthetic of ambivalence toward representation in post-bellum black literature, mappings that subvert totalizing narratives of abolitionist history through disfiguring lines, the geographic tensions of black mobility in slave narrative, and the cartographic exploration of the connections between data, tool, and literary critique.

Producción y Difusión de la investigación de las colecciones de archivos gráficos y fotográficos en el Archivo Histórico Riva-Agüero (AHRA)

Rita Segovia Rojas
segovia.ra@pucp.pe
Pontificia Universidad Católica del Perú, Peru

Ada Arrieta Álvarez
aearriet@pucp.edu.pe
Pontificia Universidad Católica del Perú, Peru

Daphne Cornejo Retamozo
daphne.cornejo@pucp.edu.pe
Pontificia Universidad Católica del Perú, Peru

Patricio Alvarado Luna
patricio.alvaradol@pucp.pe
Pontificia Universidad Católica del Perú, Peru

Ivonne Macazana Galdos
ivonnemacazana@gmail.com
Pontificia Universidad Católica del Perú, Peru

Paula Benites Mendoza
paula.benitesm@gmail.com
Pontificia Universidad Católica del Perú, Peru

Fernando Contreras Zanabria
fernando.contreras@pucp.pe
Pontificia Universidad Católica del Perú, Peru

Melissa Boza Palacios
meboza@pucp.pe
Pontificia Universidad Católica del Perú, Peru

Enrique Urteaga Araujo
enurteaga@pucp.edu.pe
Pontificia Universidad Católica del Perú, Peru

El Archivo Histórico Riva-Agüero (AHRA), unidad académica del Instituto Riva-Agüero (IRA) de la Pontificia Universidad Católica del Perú (PUCP), guarda entre su acervo documental, no sólo documentos manuscritos e impresos sino un enorme conjunto de documentos gráficos reunidos en las diferentes colecciones que lo conforman. Dentro de las colecciones gráficas y fotográficas de épocas diversas que custodia; en la actualidad se quiere poner en valor y difundir proyectos que vinculen a las humanidades con lo digital, presentando el caso de 3 proyectos realizados con dicho fin. En primer lugar "Postales de Guerra. Centenario de la primera guerra mundial", luego "Balcones de Lima: Centro Histórico" y finalmente "Arzobispos de Lima y religiosidad en el Perú". Estos proyectos vinculan propuestas humanísticas que requieren de plataformas digitales y móviles para su difusión.

Postales de Guerra. Centenario de la Primera Guerra Mundial

Patricio Alvarado Luna
Paula Benites Mendoza
Rita Segovia Rojas

Dentro de las actividades desarrolladas respecto a este proyecto, entre octubre y noviembre del 2014 se llevó a cabo una exposición itinerante en los jardines de la Facultad de Estudios Generales Letras de la Pontificia Universidad Católica del Perú, cuyo público objetivo fueron los alumnos de los primeros años de la Universidad y entre el 01 y el 18 de diciembre del 2014 la exposición se trasladó al local del Instituto Riva-Agüero. Para dicho proyecto se realizó una selección de las postales por categorías, además de un análisis y contextualización de las mismas a fin de poder identificar a los personajes, temas y principales objetivos para su realización. Por otro lado, se realizó un video en el cual se presentan diversas recomendaciones para el análisis de una fuente primaria de carácter visual.

Asimismo, se desarrolló un aplicativo en formato Flash en el cual se puede encontrar -en formato de libro virtual- la información contenida en la exposición. Aquí se puede navegar a través del cursor por la pantalla y seleccionar la información que se desee ver. Ya sea en las primeras páginas donde se detalla las partes de una postal, la línea de tiempo de duración de la Gran Guerra, o dentro de las páginas dedicadas a cada temática de la postal, en las que, al colocar el cursor sobre las diferentes imágenes, las podemos ver con toda su información y detalle gráfico.

Como aporte a la exposición y al soporte digital ubicado en la página del IRA, el material gráfico presentado incluyó un código QR para el uso de *smartphones* que llevaba a la página web del Repositorio Institucional de la PUCP, espacio virtual donde se están colocando las imágenes de las postales para conocimiento público.

Dentro de los proyectos para los años 2018-2019, se realizará una nueva exposición itinerante de otras postales de la colección Kieffer-Marchand conmemorando el final de la Primera Guerra Mundial y la firma del Tratado de Versailles. En este caso, se enfatizará en las postales correspondientes a los últimos años de la Gran Guerra manteniendo los mismos grupos de categorías. Asimismo, se pretende incluir las postales más representativas en un aplicativo móvil que permita a los usuarios poder analizar a fondo la imagen, donde se pueda explicar -de forma más detallada y vinculándolas con otras páginas académicas- el contexto de la postal, los elementos presentes, personajes y principales elementos incluidos.

Balcones de Lima. Centro Histórico

Rita Segovia Rojas
Ada Arrieta Álvarez
Melisa Boza Palacios

El proyecto, desarrollado durante el año 2014- 2015 apuntó a generar información nueva y actualizada sobre los Balcones de Lima, principalmente los ubicados en el damero de Pizarro, en el centro histórico. Para ello, se tomó como base las colecciones de fotografías de balcones del Archivo Histórico Riva-Agüero, y a partir de estas imágenes, vincular temas como la importancia del patrimonio monumental del Centro Histórico de Lima, la revaloración del "Balcón" como objeto de decoración y uso de las antiguas casonas limeñas, la catalogación arquitectónica de cada tipo de balcón con características principales de acuerdo a estilos y época, además de una mirada a la comparación de qué balcones existían en la ciudad y cuáles han desaparecido, dentro del cuadrante del damero de Pizarro, donde hoy se ubica el Centro Histórico de Lima.

Este proyecto se desarrolló con dos tipos de metodología: un primer trabajo de campo para la ubicación de los balcones catalogados en las colecciones mencionadas y realizar la toma fotográfica de actualización y de registro, lo que resultó en un catálogo comparado. Por otro lado, se desarrollaron artículos de investigación sobre los elementos arquitectónicos del balcón, las características de las casonas limeñas y un breve recuento del cambio de Lima desde el siglo XVII hasta nuestros días, que dio como resultado la publicación *"Miradas en el aire. Los balcones limeños en la memoria fotográfica. Archivo Histórico Riva-Agüero"*.

Dicho proyecto sin embargo, se inicia en el interés que despertó el desarrollo de un primer CD ROM, llamado "De Calles Balcones y Plazuelas", que incluyó fotografías en formato original y con retoque fotográfico de las siguientes categorías: Arquitectura Civil: Alamedas y paseos, Balcones, Calles, Edificios Comerciales, Edificios Públicos, Plazas y Parques, Puentes, Vistas Panorámicas, Viviendas y Arquitectura Religiosa, cuya primera edición se desarrolló el año 2001, bajo la supervisión de la profesora Ada Arrieta Álvarez, coordinadora en ese entonces del Archivo Histórico Riva-Agüero, y de los profesores Carlos Chávez y Isaac Cazorla que se ocuparon del trabajo digital. El año 2010, se realizó la reactualización de dicho formato, contando nuevamente con Ada Arrieta Álvarez en la coordinación general, Rita Segovia Rojas en la coordinación, diseño y desarrollo de la plataforma actualizada y Luis Dulanto Carbajal en el desarrollo de la plataforma actualizada.

Actualmente, a partir del cierre de estos proyectos, la propuesta incluye generar una aplicación web o móvil que permita la geolocalización de los balcones en el espacio del cuadrante del Centro Histórico de Lima, apoyados por la web del Repositorio Institucional de la Pontificia Universidad Católica del Perú, en la que ya se digitalizaron la totalidad de las imágenes de las dos colecciones y que se convierte en nuestro aliado estratégico para generar este y todos los proyectos con plataformas virtuales y en la conservación y difusión de las colecciones del IRA.

http://repositorio.pucp.edu.pe/index/handle/123456789/35135

Arzobispos de Lima y religiosidad en el Perú

Daphne Cornejo Retamozo
Ivonne Macazana Galdós
Enrique Urteaga Araujo

En el marco de la visita de Su Santidad, el Papa Francisco, al Perú, se desarrolló el proyecto "Arzobispos de Lima en el tiempo". El objetivo fue poner en valor la colección Miranda Alzamora, que reúne material sobre la historia de la Iglesia de Lima y que está custodiada por el Archivo Histórico Riva-Agüero, del Instituto del mismo nombre, de la Pontificia Universidad Católica del Perú (PUCP).

Nuestro primer producto fue una exposición itinerante que mostraba los retratos de todos los arzobispos de Lima, una breve biografía de los mismos y la firma de cada uno de ellos. Se incluyó, además, la transcripción

paleográfica de dos documentos correspondientes a los dos primeros arzobispos de Lima; es decir, Gerónimo de Loayza y Toribio de Mogrovejo.

En el trascurso de la investigación para dicha exposición notamos que no existía una plataforma virtual que condensara información sobre la religiosidad en el Perú (trabajos académicos, fuentes históricas, actividades, etc.). Esto nos condujo a plantear este proyecto que consiste en una red dinámica y colaborativa de trabajos académicos sobre religión y religiosidad en el Perú, que nos permita mostrar, de manera rápida y sencilla, investigaciones relacionadas a la religión y la religiosidad en el Perú, generar alianzas estratégicas con instituciones que estén relacionadas con la investigación sobre la temática en cuestión, acercar a los investigadores de humanidades a las nuevas tecnologías, enseñándoles sus ventajas y herramientas; además de generar una red de colaboración que permita a los usuarios ampliar sus conocimientos sobre temas religiosos y vincularse con las humanidades digitales no sólo como una plataforma, sino también como un recurso para futuras investigaciones.

Se propone la realización de una web 3.0 que permita la difusión e intercambio de conocimientos, tomando como inspiración plataformas de difusión como the programming historian, wikis, internet archive, etc. se podrá generar una red colaborativa (crowdsourcing) en donde los investigadores podrán mostrar sus trabajos académicos, las instituciones y aliados estratégicos nos proporcionarán las fuentes para futuras investigaciones y el público en general podrá acceder a esta información a través de materiales didácticos elaborados por los usuarios que se registren en nuestra plataforma.

Unanticipated Afterlives: Resurrecting Dead Projects and Research Data for Pedagogical Use

Megan Finn Senseney
mfsense2@illinois.edu
University of Illinois, United States of America

Paige Morgan
p.morgan@miami.edu, University of Miami
United States of America

Miriam Posner
mposner@humnet.ucla.edu
University of California, United States of America

Andrea Thomer
athomer@umich.edu
University of Michigan, United States of America

Helene Williams
helenew@uw.edu
University of Washington, United States of America

Overview

Pedagogical exercises in the digital humanities rely on student access to humanities data. While strategies range from instructor-prepared datasets (Sinclair and Rockwell, 2012) to having students digitize texts directly from print materials (Croxall, 2017), data repositories and web-based DH projects are two of the most attractive sources for identifying, appraising, and accessing data for classroom use.

Yet data for teaching is rarely cited as a prime motivation or rationale for sharing research data. In "The Conundrum of Sharing Research Data," Christine Borgman examines four rationales for sharing research data: (1) to reproduce or to verify research, (2) to make results of publicly funded research available to the public, (3) to enable others to ask new questions of extant data, and (4) to advance the state of research and innovation (2012). Pedagogy may be included implicitly in the third rationale, but by foregrounding pedagogical intentions, we can more readily operationalize a process for how we enable others to ask new questions of our data, which, in turn, will inform our motivations for sharing as well as the manner in which we do so.

Web-based DH projects are often conceived and developed for public consumption with short-term support through grant funding. While initiatives such as these have proliferated since the 1990s, they often languish as legacy projects on institutional servers without clear plans for sustainability or sunsetting (Rockwell et al., 2014). Rather than construe long dormant projects as an institutional burden, these artifacts may continue to function as object lessons and raw materials for use in the DH classroom. Evaluating early digital projects based on their fitness for use as pedagogical datasets distinguishes the project from its component parts and allows aspects of the project to live on in new contexts.

This panel will include representatives from five public research universities across the United States. We will begin with a brief overview, followed by four case studies. Each panelist will speak for fifteen to twenty minutes, leaving time for questions from—and conversations with—the audience. Cases are drawn from the DH 101 course at UCLA, the DH Librarianship course at the University of Washington, the University of Miami Libraries' Legacy Site Adoption Project, and the Humanities Data workshop at DHOxSS. Our goal is to explore the intersection of data sharing and digital pedagogy to interrogate how past projects (whether formally archived or otherwise) are adopted as data sets for teaching and training; propose evaluation criteria for selecting these data sets; discuss what these classroom efforts indicate about the

sustainability of DH projects (and their data); and examine how our knowledge of these classroom cases might inform curatorial decisions in active DH projects.

Learning from our mistakes: Using old projects to create better library/faculty collaborations

The Legacy Sites Adoption Project (LSAP) developed in response to what the library administration saw as a significant problem: the library website hosted nearly 40 digital projects built 5-20 years earlier by a former library faculty member, now malingering in various states of brokenness, but still placed prominently on the website. Retiring and removing the sites would erase the memory of the library's institutional history, but repairing them would create an impossible burden for the web & application development team; and would reinforce the idea of the library playing a service-and-support role in DH, rather than an active partnership.

The solution that we are currently implementing is to experiment with making the legacy sites "adoptable": the content and metadata of each site are made available as a zip file containing CSVs of data and metadata and accompanying images/audio/visual files, along with a readme pointing both to the current site on the library servers and an archived (and often more functional) version of the site in the Internet Archive's Wayback Machine. Faculty and students are able to use the zip files as base material for creating their own version(s) of the original sites, either carrying on the original concept as stated or taking it in a new direction. The original versions of the sites present opportunities for classes to think about developing DH projects with a direct focus on revision -- potentially reading and critiquing the original sites through the lenses of recent scholarly essays, or considering the choices made by the original creators in the light of how DH practices and tools have changed since the sites were built.

LSAP engages with ongoing questions about what makes a good entry point into digital humanities work. Instead of building entry points around particular tools (Omeka, Voyant, etc.); or around a particular research question or collection of material that is not a project yet, adopting legacy sites centers and foregrounds the iterative nature and inevitable fragility of project webpages, while making explicit the relationship between the websites and the flat files of their content.

With LSAP, we are also attempting a positive intervention into collaborative relationships between departmental faculty and librarians. Frequently, faculty come to librarians to ask for support for a particular idea for a digital project; or to incorporate digital methodologies into a classroom setting. In such instances, the faculty member may have little experience or knowledge with various key factors, including scoping and scaling project milestones, the availability of digitized objects, copyright/permissions restrictions, and the affordances of out-of-the-box tools. Our hope is that by offering projects that are ripe for revision, and focusing on areas that are frequently taught and studied at the university, we can provide an entry-point for collaboration that is more appropriately bounded, resulting in less uncertainty and less labor-intensive experiences for faculty, students, and librarians.

Awakening sleeping data for the DH classroom

As anyone who teaches digital humanities knows, humanities-related datasets are as hard to find as they are desirable. Since the closure of the Arts and Humanities Data Service in 2008, no centralized repository for humanities data has emerged. The DH instructor is faced with the necessity of scouring the web for data to share with students so that they can practice data-cleaning, -manipulation, and -visualization. Sometimes this data comes from libraries, archives, and museums, but it comes just as often from scholars' long-hibernating research projects. Indeed, scholars are often surprised to learn that their data has taken on a new life as the basis for student projects.

The last several decades have seen explosive growth in flexible, accessible tools for working with data. These new platforms offer possibilities for visualization and analysis that would only have been possible with custom programming just 10 or 15 years ago. Because of this palette of tools, even relatively inexperienced students can breathe new life into data left mostly untouched for years.

This presentation offers some case studies of student projects built on "dormant" data, explaining how students are trained to analyze, contextualize, visualize, and make sense of data they had no involvement in collecting. It discusses best practices for providing this data, as well as a scaffolded approach to helping students become conversant in techniques for understanding and working with data. It suggests a "toolkit" of off-the-shelf platforms that are affordable and easy for students to grasp and shows how one can build on the other until even novice students are able to create full-fledged, sophisticated digital humanities projects in the space of a semester.

For those who have collected data they wish to share with students, this presentation offers some suggestions for documenting, packaging, and contextualizing research data so that it is not only technically sound, but in a format that students can understand. It also offers a set of best practices for collaborating with students on a data-based research project, including methods for sharing, documenting, citing, and reusing data.

Fit for use: Repurposing research data, reconstructing provenance, and refining "clean" data

When it comes to teaching materials, data curation education may have become a victim of its own success: finding "dirty" data for classroom use is persistently difficult, in part because most published datasets have already been cleaned and curated! However, there are teachable moments to be found even when working with relatively "clean" data. Published data can be mined, re-structured, re-formatted and otherwise curated for new uses. Additionally, the process of tracking down and contextualizing already published datasets can prove instructive in and of itself. The detective work needed to understand someo-

ne else's project, and to reconstruct its provenance, can reveal unexpected idiosyncrasies about the dataset, and thereby reveal useful data wrangling skills to be taught.

In this talk, we describe our work finding, curating and reconstructing the provenance of "The Pettigrew Papers," a published (and relatively clean) dataset we have used over two years of teaching week-long workshops on digital humanities data curation at the Digital Humanities at Oxford Summer School (DHOxSS). Thomas J. Pettigrew (also known as Thomas "Mummy" Pettigrew) was a Victorian surgeon, antiquarian, and Egypotologist. Pettigrew wrote several early texts on Egyptian mummies and was the founding treasurer of the British Archaeological Association. Though his correspondence is archived at Yale University's Beinecke Rare Book and Manuscript Library, it came to our attention via a "data paper" published in the Journal of Open Archaeology Data (Moshenska, 2012), containing transcriptions of select letters.

In our first year teaching with the Pettigrew dataset, we wrote simple Python scripts to mine named entities from the letters, and to pull out header information about the letters as a spreadsheet for cleaning in OpenRefine. In hands-on sessions, we asked students to consider how they would clean and curate the dataset for new uses: what steps would need to be taken to create a network diagram of the entities named in his letters? To create a map of his correspondents? To create a timeline?

In our second year teaching with this dataset, we spent more time reconstructing the original provenance of the Pettigrew letters themselves. In addition to the hands-on sessions from the first year, we asked students to consider how they might improve the metadata for the original data paper, and how they might resolve discrepancies between the data paper and the original finding aids created by the Beinecke (Ducharme, 2010). We additionally discussed how they might incorporate copies of Pettigrew's publications available in the HathiTrust Digital Library in their work.

Overall, we found that asking students to clean and re-curate this already published dataset was only the starting point in our teaching; as we found further connections in digital libraries and archives beyond the original data paper, we identified subtle and important issues in the digital humanities and digital curation that guided our workshop design. In addition to teaching hands-on data cleaning and manipulation skills, we found it important to teach students a nuanced understanding of provenance: both in the sense of the archival "chain of custody" that contextualizes and validates a fonds, and in the sense of the processes that led to a dataset's current form.

Training DH librarians: Using old DH projects to move forward

The DH Librarianship course at the University of Washington Information School investigates the multiple roles librarians play in DH scholarship and prepares students for a wide range of career options in libraries, DH centers, and academic departments. DH librarian roles range from fully-credited collaborator with faculty to last-minute data cleaner, and everything in between. DH librarians also need to be prepared to support projects and research across the spectrum of disciplines, so we examine varying research methods across the humanities. The final project for the course asks that students locate an abandoned, or complete but aging DH project, and insert themselves as a librarian; they provide an evaluation of the content as well as the technology of the project and suggest ways to improve or update both.

The data sets in these projects varies and examples include: hand-collated quotations by a famous author on a fan site; census numbers provided in a project about London families in the 17th century; a list of shooting locations for a television show; metadata for photos of logging camps in the Pacific Northwest; multimedia elements in a documentary film; boxes of music programs from a summer camp; and quilting patterns.

Some projects also include the more typical (and larger) type of data set, such as those from HathiTrust or Google-generated Ngrams, but they have proven to be the exception. Working with small data sets means that cleaning doesn't occupy much time during a 10-week quarter, and they can be rearranged quickly to utilize multiple visualization or data processing options.

Students evaluate the data sets early in the process; in nearly all cases, data sets are either incomplete or inaccurate, and for some, updated data or other content is available. This is where the multi-disciplinary expertise of librarians comes in, as MLIS students are trained in searching out valid information sources from multiple perspectives, whether that's using vendor-supplied databases, open web search engines, or (gasp) sources in print or microform. This is also where students begin to see the striation of roles between true collaborators, project leaders, subject specialists, technical consultants, or data-wranglers.

In reviewing aging or abandoned projects, students learn how easily the data, other content, and the functionality of the site/project can be lost. This gives them the added perspective they need to start thinking about curation and preservation, rather than tackling those issues as add-ons if they have time.

Through these immersive projects, students have a chance to see DH through multiple lenses: those of a potential user, a collaborator, and a disciplinary specialist. They learn how to re-create and improve on a project. In doing so, they gain experience in evaluating and collecting data as well as in multiple platforms and software that are prominent in DH (some current, some defunct). Some students also reach out to the original site or project owner, and in a few cases have worked with that person to update the project, putting preservation or stabilizing features in place for future users.

References

Borgman, C. L. (2012). The conundrum of sharing research data, *Journal of the Association for Information Science and Technology*, 63(6): 1059-78.

Croxall, B. (2017). Digital humanities from scratch: A pedagogy-driven investigation of an in-copyright corpus*, Digital Humanities 2017: Conference Abstracts*, Montreal: McGill University, pp. 206-7.

Ducharme, D. J. (2010). *Guide to the Pettigrew Papers OSB MSS 113*. New Haven: Beinecke Rare Book and Manuscript Library. http://hdl.handle.net/10079/fa/beinecke.pettis1 (accessed 27 April 2018).

Moshenska, G. (2012). Selected correspondence from the papers of Thomas Pettigrew (1791-1865), surgeon and antiquary. *Journal of Open Archaeology Data*, 1(0). https://doi.org/10.5334/4f913ca0cbb89 (accessed 27 April 2018).

Rockwell, G., Day, S., Yu., J., and Engel, M. (2014). Burying dead projects: depositing the Globalization Compendium. *Digital Humanities Quarterly*, 8(2). Retrieved from http://www.digitalhumanities.org/dhq/vol/8/2/000179/000179.html (accessed 27 April 2018).

Sinclair, S. and Rockwell, G. (2012). Teaching computer-assisted text analysis. In Hirsch, B. (ed) *Digital Humanities Pedagogy: Practices, Principles, Politics*. Open Book Publishers, pp. 241-54. Retrieved from https://www.openbookpublishers.com/product.php/161 (accessed 27 April 2018).

Global Perspectives On Decolonizing Digital Pedagogy

Anelise Hanson Shrout
ashrout@fullerton.edu
California State University Fullerton, United States of America

Jamila Moore-Pewu
jmoorepewu@fullerton.edu
California State University Fullerton, United States of America

Gimena del Rio Riande
gdelrio.riande@gmail.com
IIBICRIT, Argentina

Susanna Allés
susanna_alles@miami.edu
University of Miami, United States of America

Kajsa Hallberg Adu
khadu@ashesi.edu.gh
Ashesi University College, Ghana

Panel Abstract: Global Perspectives on Decolonizing Digital Pedagogy

Digital pedagogy is often heralded as a way to undercut the "digital divide," combat structural inequality and "disrupt" the status quo. However, when English-language DH writing conjures a "typical" DH student, he or she (but most often he) fits a fairly limited mold. He is enrolled in school full-time (usually at a school in North America), and is planning to finish his B.A., B.S., or B.F.A. in four years. He attends a school that comes with a DH center, academic programmers, and the funding needed to execute a plethora of student-driven projects every year, facilitated by low student-faculty ratios, bespoke seminars and intensive faculty guidance for DH projects. In this imagined academic context, students familiar with academic norms, practiced in their use of technologies, conversant in the vernaculars of online communication, and eager to "hack" the academy flourish and thrive. As Matt Gold pointed out in 2012 and Anne McGrail reinforced in 2016 digital humanities is most represented at elite institutions, serving "traditional" undergraduates and research-oriented graduate students. (Gold, 2012; McGrail, 2016) These assumptions do not reflect the diversity of students who occupy digitally-inflected classrooms.

DH centers and programs both within and beyond the United States and non-Anglophone spaces are increasingly serving wider and more diverse communities of students. ([CSL STYLE ERROR: reference with no printed form.]; [CSL STYLE ERROR: reference with no printed form.]) For example, in the United States, nearly half of all undergraduates attend community college, and more than half are the first in their families to pursue a degree beyond high school. (The Penn Center for Minority Serving Institutions; Montenegro and Jankowski, 2015) This means that many U.S. students come to higher education without the educational, cultural or technological capital often assumed by DH syllabi. Students in Sub-Saharan Africa often do not have access to wired internet or personal computers.(Robison and Crenshaw, 2010; Lechman, Ewa, 2015) This has meant that students developed a robust engagement with mobile technology, and that pedagogy followed their cues. Students in Latin America often do not engage with Anglophone digital humanities. This means that the humanidades digitales are exploring the ways in which Spanish linguistics have shaped digital thought and practice.

Despite this diversity, scholarship on digital pedagogy largely remains focused on a small subset of institutions within North America. (Fiormonte, Domenico, 2014) This roundtable makes space for discussions of whether "centers" (that is, programs in the United States and Anglophone world) are "ready to learn from peripheries," or if we must jettison frameworks of central and peripheral knowledge altogether. It takes seriously Élika Ortega's contention that "all DH is local pedagogy," as well as Padmini Ray Murray's admonition that "your [Western]

DH is not my DH." (Quoted in (Risam, Roopika, 2016)) It features panelists who work with students traditionally excluded from this scholarship, both within the United States and beyond the Anglophone world. Each panelist will speak to the process of building DH curricula that center, rather than merely accommodating, "non-traditional" DH students.

Each panelist will speak briefly about the DH programs or classes at their institution. The remainder of the session will be devoted to a mix of moderated discussion and audience-generated Q and A. Panelists will be asked to describe particular challenges, assignments, or pedagogical tools, in order to ground the session in pedagogical practice. At its conclusion, we hope that the roundtable panelists and audience members will explore digital curricula and pedagogy that are necessarily decolonized, global, and anti-neoliberal.

This roundtable is not the first to make claims about the U.S.- and Anlgophone-centered nature of DH or DH pedagogy. It's framework owes much to the work of the GO:DH Special Interest Group, and the scholarship of Moya Z. Bailey, Isabel Galina, Alex Gil, Jessica Marie Johnson, Dorothy Kim, Elizabeth LaPensé, and Élika Ortega. (Bailey, 2012; Galina, Isabel; Johnson, Jessica Marie and Neal, Mark Anthony, 2017; 2014)

Paper 1: Digital Humanities Pedagogy from The Global South & Ghana

Ashesi University College is located in the village of Berekuso in the peri-urban Eastern Region of Ghana and its student body is intentionally diverse: Pan-African with students from Southern, Eastern and Western Africa and with a handful of exchange students from the Global North, economically diverse with 50 percent on scholarship and the other half full fee paying, and almost equal numbers of men to women. The philosophy behind the 15 years young university is training ethical leaders for the continent in engineering, business, and computer science, but with a broadly defined liberal arts foundation of Written and Oral Communication, Mathematics, Statistics, Programming, English Literature, Social Sciences, Statistics, and Leadership. All students of different majors are taking the liberal arts core as well as African studies electives together. Teaching this multiverse group can be both challenging and rewarding. Using digital tools can exacerbate differences if relying heavily on hardware (not everyone owns a smartphone, for instance), but it can also allow for a level playing field where the technology allows us to focus on ideas and shared experiences.

The three examples from Ashesi University use digital tools like Wiki, Virtual Reality and WhatsApp, all to no additional or very low cost to the university and student. As the decolonization of the academy requires rethinking everything from content, textbooks, assignments, and assumptions we bring into the classroom, the first example shaking up the status quo, is by teaching students how to edit and contribute to Wikipedia. The first trial was done with the Wikimedia Foundation for the Social Theory core class in spring of 2017. After completing the training, students read and edited articles relevant to the class topics such as Slave Forts in Ghana. Students also set up their own Wiki club to continue creating knowledge.

In the French as a Foreign Language class, a couple of classroom meetings were devoted to travel and students "traveled" to foreign places using virtual reality technology (VR). Students took turns using the phone holder (a box of 20 units for USD 20 each has been acquired by the university) with a student smartphone running a VR app inside, allowing them to "visit" other places and, using vocabulary just absorbed, then tell their colleagues about it in French.

For the African Philosophical Thought elective, a WhatsApp list was created where students either use their smartphones or a friend's smartphone to join the pre-class conversation group and share ideas on a specific weekly topic. The lecturer would come in to add comments and questions. This use of this digital tool connects to the immense popularity of the WhatsApp-app in Africa and builds a community around the course, creating a new center of knowledge.

The three examples from Ashesi University all empower students with knowledge and tools that give them agency, moving away from a passive absorption of texts written in the Global North, hence decolonizing the classroom.

Papers 2 and 3: Teaching Humanidades Digitales (HD)

Part 1: "HD from scratch". Gimena del Rio Riande (IIBICRIT Argentina)

Part 2: "HD and the other linguistic divide". Susanna Allés-Torrent (University of Miami)

Our contribution faces the digital divide from the point of view of heterogeneity (Cornejo-Polar, 2003) and transculturation (Ortiz, 1963) processes of the Anglophone Digital Humanities in Spanish speaking communities, where the institutionalization of the discipline is experiencing the dawning of associations -Asociación Argentina de Humanidades Digitales (AAHD), Humanidades Digitales Hispánicas (HDH), RedHD (Red Humani), Red Colombiana de Humanidades Digitales, Humanidades Digitales en Cuba, among others-, research groups and pedagogical initiatives. Certificates, MA and specialized courses are appearing in a crucial moment while DH in Spanish is still defining itself and thus is still pondering how to create a curriculum in DH in Spanish (or in HD, Humanidades Digitales).

We conceive HD from a cognitive approach, in which language and thought work as a whole (Lakoff, 1986). Thus, we understand HD as a set of digital and emerging methodologies and practices in Spanish, that covers different curricular topics, within a different level of institutionalization. We are not only dealing with a geolo-

calization issue, which goes far beyond Spain and Latin America, or even with a set of different cultural and academic backgrounds, but with the central role of language as builder and communicator of knowledge.

The linked presentations described in this abstract aim to offer possible leads from real and different experiences in curricular design and pedagogical materials creation, in which both panelists have participated as faculty. A previous and conjointly experience consisted of the design of two certificates *Experto en Humanidades Digitales* y *Edición Digital* offered at LINHD (UNED, Spain), where we dealt with the challenge of online education by approaching a field that can be defined as a "hands-on and guided experience" and to a wide community (mostly Spain, but also Latin America or Spanish speakers as a whole, from graduate to postgraduate students).

The first presentation ("HD from scratch") describes a context lacking an extensive academic history of DH and solid digital infrastructures (Argentina). The courses taught here are held in a discontinuous timeframe: semester courses not always offered nor exclusively devoted to HD. The second presentation describes teaching DH in Spanish in a quite satisfying context as far as infrastructure is concerned, the University of Miami (FL), but not yet under a planned sequence of DH courses at undergraduate level, meaning that students do not have any DH previous exposure, and there is not yet a DH follow-up or continuation in their curriculum. In addition, due to Miami's diversity and Hispanic population, they deal in a same classroom with students that or they are Spanish native speakers (mostly Caribbean) or they do not master Spanish language. DH pedagogy is here "the other linguistic divide".

We will share our experiences with the development of pedagogical materials in the language that we use in our teaching that can be understood and benefit outside the classroom experience. It is not worth just thinking on translation, but on the production of materials capable to build knowledge with the student during and after classes.

To that end, we work on some approaches towards a curriculum with a global perspective on DH including the local disciplines, going back consequently to concepts such as "situated practice" (Lave and Wenger, 1991), contextual pedagogies (Cabaluz Ducasse, (2015) and "local knowledge" (Grenier 1998; del Rio Riande 2016). We also revise the ways in which technology has been used to conduct research in the humanities to give an answer to its impact in a HD curriculum.

Paper 4: First Generation DH

While most work on American higher education focuses on the Research-1/small liberal arts college divide, most American students attend other types of institutions. Of the just more than seventeen million students enrolled in higher education in the United States, over two million are enrolled in for-profit colleges, and students enrolled in community colleges make up forty-six percent of all undergraduates.(Cottom, 2017; McGrail, 2016) Over fifty percent of students enrolled in all American higher education institutions – including community colleges, small liberal arts colleges, comprehensive regional universities and large research institutions – come from families where neither parent completed a baccalaureate degree. Despite this majority overall, only a minority of these first-generation students attend Research-1 universities or small liberal arts colleges.

A 2017 survey of Digital Humanities programs, found that the vast majority of American institutions offering DH degrees, certificates, minors and concentrations are classified either as "Doctoral Institution: Highest Research Activity" or "Baccalaureate Institution: Arts & Science Focus." (Hackney et al., 2017) The In short, while first generation students tend to be enrolled in regional comprehensive universities and community colleges, DH pedagogy writing tends to emphasize relatively elite students, and DH programs tend to be developed at relatively elite institutions.

This paper explores what it means for DH that a growing plurality and soon-to-be majority of students in U.S. classrooms do not enter college with the social capital that comes from college-educated parents, and attend neither small liberal arts colleges nor research institutions. It calls for us to think about the humanistic in new ways, and to build curricula that acknowledge that many of our students are not the imagined, prototypical college student.

We build on extant research on a "digital divide" premised on educational background. This research has found that first-generation students often have trouble navigating the unspoken norms of higher educational spaces.(Stephens et al., 2014; Stephens et al., 2012) They are less likely to use digital tools for "capital-enhancing online activities."(Hargittai, 2010) They are less likely to own personal computers, have multiple spaces to access the internet, and have the time to access the internet. They are often disinclined to seek out institutional resources. They are also less likely to take classes outside of a proscribed academic/vocational plan, and less likely to try classes in new fields that might be a risk to their GPA.(National Center for Education Statistics, 2005) Finally, they are more likely to succeed in interdependent, collaborative work environments.(Stephens et al., 2012)

This paper suggests several best practices for the development of DH pedagogy that centers first-generation students. Second, we need to remember that, despite these findings about first-generation students not using the internet in ways that researchers expect them to, they are already using the digital to "explore humanistic questions. First, we need to introduce DH in lower division courses, so that students can "test drive" digital humanities without the risk of committing to an entire

course. Finally, we need to adopt Roopika Risam's call to tell "alternate histories of the digital humanities… through intersectional lenses" and keep in mind the ways in which structural inequality has already conditioned the development of the digital humanities.(Risam, 2015)

References

Bailey, M. Z. (2012). All the Digital Humanists Are White, All the Nerds Are Men, but Some of Us Are Brave. *Journal of Digital Humanities* http://journalofdigitalhumanities.org/1-1/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-brave-by-moya-z-bailey/ (accessed 21 November 2017).

Cottom, T. M. (2017). *Lower Ed: The Troubling Rise of For-Profit Colleges in the New Economy*. New York, N.Y.: The New Press (accessed 1 August 2017).

Fiormonte, Domenico (2014). Digital Humanities from a global perspective. *Laboratorio Dell'ISPF*, XI–2014 doi:10.12862/ispf14L203. http://www.ispf-lab.cnr.it/2014_203.pdf.

Galina, Isabel Is There Anybody Out There? Building a global Digital Humanities community | Humanidades Digitales *RED de Humanidades Digitales* http://humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/ (accessed 21 November 2017).

Gold, M. (2012). Whose Revolution? Towards a More Equitable Digital Humanities http://blog.mkgold.net/2012/01/10/whose-revolution-toward-a-more-equitable-digital-humanities/ (accessed 2 August 2017).

Hackney, S. E., Cunningham, P. and Sula, C. A. (2017). A Survey of Digital Humanities Programs. *The Journal of Interactive Technology & Pedagogy*(11) https://jitp.commons.gc.cuny.edu/a-survey-of-digital-humanities-programs/ (accessed 1 August 2017).

Hargittai, E. (2010). Digital Na(t)ives? Variation in Internet Skills and Uses among Members of the 'Net Generation'*. *Sociological Inquiry*, 80(1): 92–113 doi:10.1111/j.1475-682X.2009.00317.x.

Johnson, Jessica Marie and Neal, Mark Anthony (eds). (2017). Black Code. *The Black Scholar*, 47(3).

Lechman, Ewa (2015). *ICT Diffusion in Developing Coutnries: Towards a New Concept of Technological Takeoff*. Springer.

McGrail, A. B. (2016). The 'Whole Game': Digital Humanities at Community Colleges. *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press.

Montenegro, E. and Jankowski, N. A. (2015). *Focused on What Matters: Assessment of Student Learning Outcomes at Minority-Serving Institutions*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

National Center for Education Statistics (2005). First-Generation Students in Postsecondary Education: A Look at Their College Transcripts.

Risam, R. (2015). Beyond the Margins: Intersectionality and the Digital Humanities. , 9(2) http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html (accessed 30 March 2017).

Risam, Roopika (2016). Navigating the Global Digital Humanities: Insights from Black Feminism. In Klein, Lauren F. and Gold, Matthew K. (eds), *Debates in the Digital Humanities*. 2016th ed. http://dhdebates.gc.cuny.edu/debates/text/80 (accessed 21 November 2017).

Robison, K. K. and Crenshaw, E. M. (2010). Reevaluating the Global Digital Divide: Socio-Demographic and Conflict Barriers to the Internet Revolution*. *Sociological Inquiry*, 80(1): 34–62 doi:10.1111/j.1475-682X.2009.00315.x.

Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S. and Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology*, 102(6): 1178–97 doi:10.1037/a0027143.

Stephens, N. M., Hamedani, M. G. and Destin, M. (2014). Closing the Social-Class Achievement Gap: A Difference-Education Intervention Improves First-Generation Students' Academic Performance and All Students' College Transition. *Psychological Science*, 25(4): 943–53 doi:10.1177/0956797613518349.

The Penn Center for Minority Serving Institutions Supporting Minority Serving Institutions (MSIs) by the numbers http://www2.gse.upenn.edu/cmsi/sites/gse.upenn.edu.cmsi/files/CMSI.Infographic.v4.pdf (accessed 2 August 2017).

(2014). Whispering/Translating during DH2014: Five Things We Learned *Élika Ortega* https://elikaortega.net/2014/07/21/dhwhisperer/ (accessed 21 November 2017).

DH Organizations around the world http://testing.elotroalex.com/dhorgs/ (accessed 21 November 2017a).

DH Course Registry https://registries.clarin-dariah.eu/courses/ (accessed 21 November 2017b).

Cabaluz Ducasse, F. (2015). *Entramando Pedagogías Críticas Latinoamericanas. Notas teóricas para potenciar el trabajo político-pedagógico comunitario*. Santiago de Chile: Editorial Quimantú.

Cornejo-Polar, A. (2003). *Escribir en el aire: Ensayo sobre la heterogeneidad socio-cultural en las literaturas andinas*. Lima: Centro De Estudios Literarios Antonio Cornejo Polar - CELACP.

Grenier, L. (1998), *Working with Indigenous Knowledge. A Guide for Researchers*.International Development Research Center: Ottawa

Lakoff, G. (1986). Cognitive Semantics. *Versus*, 44/45: 119–154.

Lave, J. and Wenger, E. (1991). *Situated Learning. Legitimate peripheral participation*, Cambridge: University of Cambridge Press.

Ortiz, F. (1963). *Contrapunteo cubano del tabaco y el azúcar*. La Habana: Consejo Nacional de Cultura.
Rio Riande, G. del (2016). ¿De qué hablamos cuando hablamos de Humanidades Digitales? *Actas I Jornadas de Humanidades Digitales de la AAHD*, Buenos Aires: Editorial de la Facultad de Filosofía y Letras, pp. 50-62. https://www.aacademica.org/jornadasaahd/3.pdf
Cottom, T. M. (2017). *Lower Ed: The Troubling Rise of For-Profit Colleges in the New Economy*. New York, N.Y.: The New Press (accessed 1 August 2017).
Hackney, S. E., Cunningham, P. and Sula, C. A. (2017). A Survey of Digital Humanities Programs. *The Journal of Interactive Technology & Pedagogy*(11) https://jitp.commons.gc.cuny.edu/a-survey-of-digital-humanities-programs/ (accessed 1 August 2017).
Hargittai, E. (2010). Digital Na(t)ives? Variation in Internet Skills and Uses among Members of the 'Net Generation'*. *Sociological Inquiry*, 80(1): 92–113 doi:10.1111/j.1475-682X.2009.00317.x.
McGrail, A. B. (2016). The 'Whole Game': Digital Humanities at Community Colleges. *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press.
National Center for Education Statistics (2005). First-Generation Students in Postsecondary Education: A Look at Their College Transcripts.
Risam, R. (2015). Beyond the Margins: Intersectionality and the Digital Humanities. , 9(2) http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html (accessed 30 March 2017).
Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S. and Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology*, 102(6): 1178–97 doi:10.1037/a0027143.
Stephens, N. M., Hamedani, M. G. and Destin, M. (2014). Closing the Social-Class Achievement Gap: A Difference-Education Intervention Improves First-Generation Students' Academic Performance and All Students' College Transition. *Psychological Science*, 25(4): 943–53 doi:10.1177/0956797613518349.

Computer Vision in DH

Lauren Tilton
ltilton@richmond.edu
University of Richmond, United States of America

Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Thomas Smits
t.smits@let.ru.nl
Radboud University, The Netherlands

Melvin Wevers
melvinwevers@gmail.com
Digital Humanities Group, KNAW Humanities Cluster, The Netherlands

Mark Williams
mark.j.williams@dartmouth.edu
Dartmouth College, University States of America

Lorenzo Torresani
lt@dartmouth.edu
Dartmouth College, University States of America

Maksim Bolonkin
mbolonkin@cs.dartmouth.edu
Dartmouth College, University States of America

John Bell
john.p.bell@dartmouth.edu
Dartmouth College, University States of America

Dimitrios Latsis
dlatsis@ryerson.ca
Ryerson University, Canada

Overview

Visual culture is often overlooked as an object of study within the digital humanities (DH). Yet, visual culture is central to fields such as art history, cultural studies, history, and media studies. Cultural forms such as drawings, film, video, painting, photography and drawing continue to shapes culture, politics and society. As the field begins to turns its attention to other forms of media, new methods and tools are necessary to study visual culture at scale. Recent advances in computer vision are proving a promising direction. The panel "Computer Vision in the Digital Humanities" will present three approaches to using computer vision in the Digital Humanities (DH).

"Distant Viewing: Analyzing Moving Images at Scale" argues that computer vision is a powerful tool for distant viewing time-based media. The authors will outline their method and then describe the Distant Viewing Toolkit, a set of machine learning computer vision algorithms for analyzing features such as color, shot and scene breaks, and object identification such as faces. They will then turn to a case study of the American Network Era (1952-1984) television to show how their method reveals the representational politics of gender during the era.

"Seeing History: Analyzing Large-Scale Historical Visual Datasets Using Deep Neural Networks" will then focus in on how convolutional neural networks (CNN) can be used for historical research. The authors focus on two case studies applied to two major Dutch national newspapers. They used CNNs to identify over 400,000 advertisements from 1945-1995 in the first study and over 110,000 photographs and drawings from 1860 to 1920 in the second

study. They then will explain the two tools they developed to support visual and textual search in the new corpuses.

Finally, The Media Ecology Project and Visual Learning Group are creating software that allows people to search in untagged films and videos in the same way that they search through the text of a document. The tool takes search queries expressed in textual form and automatically translates them into image recognition models that can identify the desired segments in the film. The image recognition results can be cached for quick searching. Initial prototype results, funded by The Knight Foundation, focused on educational films at The Dartmouth Library and The Internet Archive that are common to many libraries and archives.

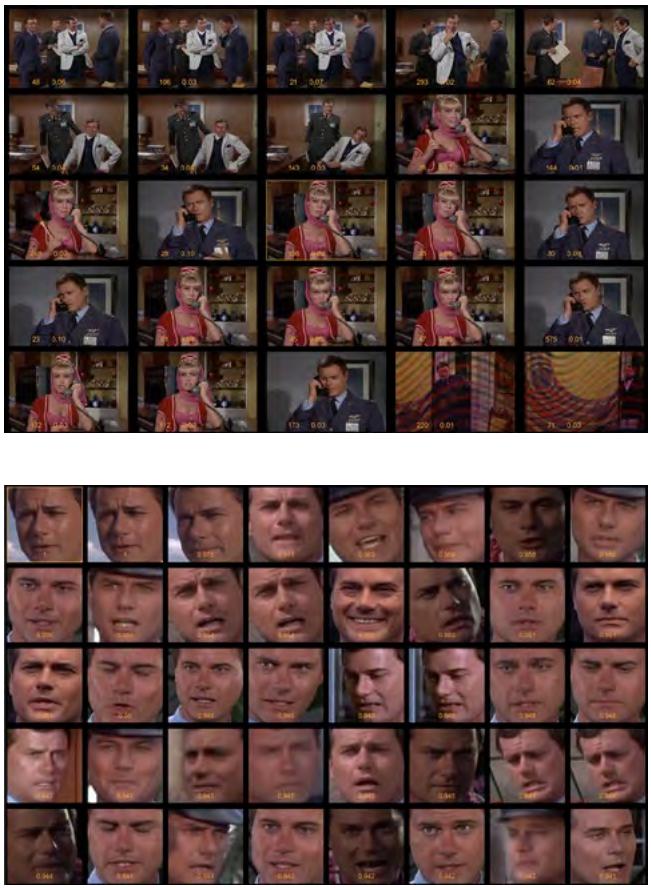
All of the papers will address the need to develop open access historical humanities data sets for developing computer vision as a DH technique.

Distant Viewing: Analyzing Moving Images at Scale

Digital humanities' (DH) focus on text and related methodologies such as distant reading and macroanalysis has produced exciting interventions (Jockers 2013; Moretti 2013). However, there is an increasing call to take seriously visual culture and moving images as objects of study in digital humanities (Posner 2013; Acland and Hoyt 2016; Manovich 2016; ADHO AVinDH Special Interest Group). In this paper, we will discuss how we are using computer vision and machine learning to distant view moving images.

The paper will begin by outlining our method of distant viewing and then turn to our Distant Viewing toolkit. Using and developing machine learning algorithms, the toolkit analyzes the following features: (1) the dominant colors and lighting over each shot; (2) time codes for shot and scene breaks; (3) bounding boxes for faces and other common objects; (4) consistent identifiers and descriptors for scenes, faces, and objects over time; (5) time codes and descriptions of diegetic and non-diegetic sound; and (6) a transcript of the spoken dialogue (see Figures 1 & 2 for examples of these analyses). These features serve as building blocks for analysis of moving images in the same way words are the foundation for text analysis. From these extracted elements, higher-level features such as camera movement, framing, blocking, and narrative style can then be derived and analyzed. These techniques then allow scholars to see content and style within and across moving images such as films, news broadcasts, and television episodes, revealing how moving images shape cultural norms.

To illustrate this approach, we have applied our Distant Viewing toolkit to a collection of series from the Network Era (1952-1984) of American television. The Network Era is often considered formulaic and uninteresting from a formal perspective despite how highly influential this era of TV was on U.S culture (Spiegel 1992). Our analysis challenges this characterization using computational methods by showing how the formal elements of the sitcoms serve to reflect, establish, and challenge cultural norms. In particular, we will focus on the representational politics of gender during the Network Era. For examples of how we are distant viewing TV, please see distanttv.org.



Shots detected by the Distant Viewing Toolkit on an episode of I Dream of Jeannie.

Seeing History: Analyzing Large-Scale Historical Visual Datasets Using Deep Neural Networks

Scholars are increasingly applying computational methods to analyze the visual aspects of large-scale digitized visual datasets (Ordelman et al., 2014). Inspiring examples are the work of Seguin (Seguin et al., 2017) on visual pattern discovery in large databases of paintings and Moretti's and Impett's (Moretti and Impett, 2017) large-scale analysis of body postures in Aby Warburg's Atlas Mnemosyne. In our paper, we will present two datasets of historical images and accompanying texts harvested from Dutch digitized newspapers and reflect on ways to improve existing neural networks for historical research. We will discuss how large historical visual datasets can

be used for historical research using neural networks. We will do this by describing two case studies, and will end our paper by arguing for the need for a benchmarked dataset with historical visual material.

The sets were produced during two researcher-in-residence projects at the National Library of the Netherlands. The first set consists of more than 400,000 advertisements published in two major national newspapers between 1945 and 1995. Using the penultimate layer in a Convolutional Neural Network (CNN), 2,048 visual aspects were abstracted from these images, which can be used to group images together (Seguin et al., 2017). The second dataset includes about 110,000 classified images from newspapers published between 1860 and 1920. The images were classified using a pipeline that consists of three classifiers. The first one detects images with faces (Geitgey, 2017), the second categorizes images according to eight different categories (buildings, cartoons, chess problems, crowds, logos, schematics, sheet music, and weather reports), and the last one sorts images as either photographs or drawings (Donahue et al., 2013).

We developed two tools to query these datasets. The first tool offers exploratory search in the advertisement dataset, which enables users to find images sharing a degree of visual similarity and can be used to detect visual trends in large visual datasets. The second one enables users to find images in the second set by searching for specific (combinations) of visual subjects and keywords. For example, images of 'buildings' with 'faces' and the keyword 'protest' in the text.

Finally, our paper discusses several challenges and possibilities of computer vision techniques for historical research. Most CNN's are trained on contemporary materials (ImageNet). As a result, these networks perform well in recognizing the categories of the ImageNet challenge. However, the fact that they were trained on contemporary data can cause problems when working with historical images. For example, detecting bicycles works relatively well because the design of the bicycle has remained more or less similar during the last century, while trains are much more difficult since they have changed significantly over the years. Also, models trained on ImageNet have difficulties detecting objects in illustrations, which are often used in newspapers. They are regularly classified within the uninformative category 'cartoon.' In short, we will discuss how to improve these models and argue for the development and benchmarking of datasets with visual historical material.

Unlocking Film Libraries for Discovery and Search

Where the library of the 20th century focused on texts, the 21st century library will be a rich mix of media, fully accessible to library patrons in digital form. Yet the tools that allow people to easily search film and video in the same way that they can search through the full text of a document are still beyond the reach of most libraries. How can we make the rich troves of film/video housed in thousands of libraries searchable and discoverable for the next generation?

Dartmouth College's Media Ecology Project, led by Prof. Mark Williams and architect John Bell, and the Visual Learning Group, led by Prof. Lorenzo Torresani, are applying computer vision and machine learning tools to a rich collection of films held by Dartmouth Library and the Internet Archive. Using existing algorithms, we describe what is happening and translate the resulting tags into open linked data using our Semantic Annotation Tool (SAT). SAT provides an easy-to-use and accessible interface for playing back time-based annotations (built upon W3C web annotation standards) in a web browser, allowing simple collection development that can be integrated with discovery and search in an exhibition. What was once a roll of film, indexed only by its card catalog description, will become searchable scene-by-scene, adding immense value for library patrons, scholars and the visually impaired.

Dartmouth College's Visual Learning Group is already a leader in computer vision and machine learning, developing new tools for object and action recognition. This project has brought together cross-curricular groups at Dartmouth to collaborate on applying modern artificial intelligence and machine learning to historic film collections held by libraries and archives.

Our tool takes search queries expressed in textual form and automatically translates them into image recognition models that can identify the desired segments in the film. The entire search takes only a fraction of a second on a regular computer. We have a working prototype of the search functionality and are creating a demonstration site that will be featured in the conference presentation. Our initial prototype results, funded by The Knight Foundation, focused on educational films at The Dartmouth Library and The Internet Archive that are common to many libraries and archives. Our software leverages image recognition algorithms to enable content-based search in video and film collections housed in libraries. By utilizing The Semantic Annotation Tool, the project also works to bring together human- and machine-generated metadata into a single, searchable format.

By improving the cutting edge algorithms used to create time-coded subject-tags (e.g. http://vlg.cs.dartmouth.edu/c3d/), we aim to lay the foundation for a fully-searchable visual encyclopedia and to share our methods and open source code with film libraries and archives everywhere. Our goal is to unlock the rich troves of film held by libraries and make them findable and more useable—scene by scene, and frame by frame--so future generations can discover new layers of meaning and impact.

References

Acland, C. R. and Eric Hoyt, Editors (2016).*The Arclight Guidebook to Media History and the Digital Humanities*. REFRAME Books.

Bertasius, G., Shi, J. and Torresani, L. (2015) "High-for-Low and Low-for-High: Efficient Boundary Detection from Deep Object Features and its Applications to High-Level Vision," in *IEEE International Conference On Computer Vision*, ICCV.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2013), "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ArXiv:1310.1531 [Cs], available at: http://arxiv.org/abs/1310.1531 (accessed 23 November 2017).

Geitgey, A. (2017), Face_recognition: The World's Simplest Facial Recognition Api for Python and the Command Line, Python, available at: https://github.com/ageitgey/face_recognition (accessed 23 November 2017).

Hediger, V. and Vonderau, P. (2009) *Films that Work: Industrial Film and the Productivity of Media (Film Culture in Transition)* Amsterdam: Amsterdam University Press.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

Manovich, L. (2016). *"The Science of Culture? Social Computing, Digital Humanities, and Cultural Analytics". The Datafied Society. Social Research in the Age of Big Data*, edited by Mirko Tobias Schaefer and Karin van Es. Amsterdam University Press.

Moretti, F. (2013) *Distant Reading*. Verso Books.

Moretti, F. and Impett, L. (2017), "Totentanz", *New Left Review*, No. 107, pp. 68–97.

Ordelman, R., Kleppe, M., Kemman, M. and De Jong, F. (2014), "Sound and (moving images) in focus – How to integrate audiovisual material in Digital Humanities research", presented at the *Digital Humanities 2014*, Lausanne, available at: (accessed 15 November 2017).

Posner, M. (2013). "Digital Humanities and Film and Media Studies: Staging an Encounter." *Workshop della Society for Cinema and Media Studies Annual Conference*, Chicago. Vol. 8.

Seguin, B., di Leonardo, I. and Kaplan, F. (2017), "Tracking Transmission of Details in Paintings", presented at the *Digital Humanities 2017*, Montreal.

Spigel, L. (1992). *Make Room for TV: Television and the Family Ideal in Postwar America*. University of Chicago Press.

Tran, D., Bourdev, L. , Fergus, L. , Torresani, L. and Paluri, M. (2015). "Learning Spatiotemporal Features with 3D Convolutional Networks," in *IEEE International Conference On Computer Vision*, ICCV.

Williams, M. (2016). "Networking Moving Image History: Archives, Scholars, and the Media Ecology Project" in *The Arclight Guidebook to Media History and the Digital Humanities*, Charles R. Acland and Eric Hoyt, eds.

Harnessing Emergent Digital Technologies to Facilitate North-South, Cross-Cultural, Interdisciplinary Conversations about Indigenous Community Identities and Cultural Heritage in Yucatán

Gabrielle Vail
vailg@email.unc.edu
University of North Carolina, United States of America

Sarah Buck Kachaluba
sbuckkachaluba@ucsd.edu
University of California, United States of Americ

Matilde Cordoba Azcarate
mcazcarate@ucsd.edu
University of California, United States of Americ

Samuel Francois Jouault
samuel.jouault@correo.uady.mx
Universidad Autónoma de Yucatán, Mexico

This panel brings together practitioners of digital humanities with backgrounds in history, anthropology, and film to discuss projects that embrace interdisciplinary perspectives in exploring cross-cultural conversations about the role of expert knowledge and digital technologies in community and tourism development, indigenous cultural heritage protection, and learning pedagogies. Using digital technologies and interfaces, the four projects bring to light historical and contemporary indigenous experience in various communities in Yucatán, Mexico.

The first makes use of an interactive website that documents historical and contemporary rural Yucatecan community experiences to explore evolving individual and communal identities through interviews with elderly community residents remembering social reform, economic restructuring, and state-building in the decades following the Mexican Revolution, as well as owners of current businesses and young adults. Of interest is how such identities are affected by globally-driven economic development, as the Yucatán is increasingly defined by commodity production for global export, and as historical haciendas that once served as the center of communal economic production are converted into luxury hotels to serve foreign tourists. The website aims to capture the voices of and facilitate communication between community members, visitors, scholars, and interested parties.

The second project employs an online format that allows diverse audiences ranging from lay people to academic specialists to explore the calendric, iconographic, hieroglyphic, and thematic content of prehispanic Maya screenfold books using a sophisticated search engine to perform queries. This tool enables contemporary

speakers of Mayan languages to engage, either on their own or as participants in instructor-led workshops, with valuable historical objects that constitute part of their cultural heritage, but ones that they would otherwise lack the means to access. Use of the website in Latin American and U.S. communities with large Maya populations helps to foster discussions across borders about Maya identity in the past and today.

The third presentation discusses *Co'ox Mayab*, an organization dedicated to sustainable tourism in Yucatán through practices that promote collaboration and solidarity, with a focus on social consciousness and justice. The organization has several goals, including the promotion of intercultural dialogue through the interaction of visitors and indigenous hosts. Digital technologies—comprising a webpage, documentary, and promotional videos—play an important role in evaluating the effectiveness of *Co'ox Mayab* as an organization, and in understanding how sustainable tourism practices are achieved in Yucatán.

The final presentation describes the preliminary steps in creating an ethnographic documentary that charts changes in the socio-cultural organization of the former henequen town of Tekit in inland Yucatán. The town has recently been transformed into a domestic factory town, and it produces textile souvenirs and uniforms for the hospitality industry in the region and beyond. The documentary aims to show how the domestic *maquila* (workshop) system of production has led to its differentiation from other Yucatecan communities that cannot provide enough work for their inhabitants, leading to extensive migration. This ethnographic documentary is part of a collaborative effort to link the social sciences and the humanities, digital media technologies, and documentary film through multimedia storytelling as a pedagogical practice.

These four projects each engage with and facilitate discussion about the role of digital interfacing, multimedia initiatives, databases, and documentary film in fostering more inclusive and participatory community development projects, cultural heritage protection practices, and learning pedagogies across the Americas.

The panel consists of the following presentations:

Pueblos Yucatecos: Buildingan Interactive Website to Create, Exhibit, and Examine Indigenous Community (Hi)stories

Sarah Buck Kachaluba

This presentation details the creation of a website exhibiting materials to tell community stories that originated with oral history work the presenter undertook in 2000 in villages in the state of Yucatán, Mexico. The interviews themselves (both audio versions and textual transcripts) and selected photographs will be archived in an academic digital library with the website pointing to such digital objects, while simultaneously providing access to supplemental material, including additional photographs, maps, records, and eventually interactive features such as chat and TEI-enabled tools that will facilitate further content creation, editing, and commentary. This project has expanded and changed significantly from the original oral history project defined by the goal of charting the experiences of women who were the members of women's leagues that played a role in securing Mexican women the right to vote and expanding the definition of Mexican female citizenship in the decades following the Mexican Revolution. The current project explores rural communal identities in the face of economic development driven by globalization, as the Yucatán is increasingly defined by commodity production for global export and historical haciendas that once served as the center of communal economic production are converted into luxury hotels (funded by multinational corporations such as Banamex, Citibank, and Sheraton) to serve foreign tourists. The website aims to capture the voices of community members, as well as other visitors, scholars, and interested parties. The project is simultaneously an exhibit of and a tool to shape community, community histories and contemporary stories, and identity formation. It also reveals and explores the use of various kinds of screen technologies, including smart phones and social media, to enable cross-cultural, multilingual, and North-South communication.

The Online Maya Codices Database: Fostering Community Conversations about Maya Heritage and Identity

Gabrielle Vail

The Maya Codices Database, developed with funding from the National Endowment for the Humanities, provides an online format for exploring the calendric, iconographic, hieroglyphic, and thematic content of the prehispanic Maya screenfold books dating to the Late Postclassic period (c. 1250-1521 CE). It was developed with dual user interfaces to maximize its utility for reaching a variety of audiences—the general public, students in grades 4-12 and above, codical specialists, and academics in other fields of study. The database is used as a teaching tool in hieroglyphic workshops held in both the U.S. and Latin America, often for Maya-speaking audiences consisting primarily of educators, students, and ritual specialists. The visual representations of rituals and daily activities enacted byancestors of contemporary Maya people, along with the written hieroglyphic texts describing what is pictured, generally trigger a series of recollections among indigenous participants, who then relate their stories, reminiscences, and descriptions of ritual activities, tying this information into the almanacs depicted in the codices.

The database was recently introduced to students of Maya descent in western North Carolina as part of an exchange program to build bridges of cultural understanding among international participants (in this case, with Yucatec Maya students of a similar age attending school in Valladolid, Yucatán). The North Carolina students, despite having parents who grew up in Maya communities in the highlands of Guatemala, felt that they knew little about their indigenous roots and were eager to participate in the workshops being offered and to interact with the archival material made available to them as part of the program. One of the primary goals of the "Maya from the Margins" project was to provide access to resources highlighting prehispanic and more recent indigenous history to give students an opportunity to explore their heritage and what it means to their understanding of their identity. As part of the workshops held with the students, the codices database served as an important avenue for initiating discussions about these topics within the communities involved. Exhibit panels created by the students further served this goal.

The many "born digital" materials from the Maya from the Margins exchange program will form the basis for a forthcoming digital humanities project that explores the lives of Maya descendant populations living in communities far from where they originated, as well as those (the Yucatec students) who still live in their natal communities while attending university in Valladolid. We envision this digital project as a way to continue to link communities whose members share a common heritage, despite the divides of distance and of shifting cultural norms and expectations.

Co´ox Mayab: Multimedia Experiences to Create Bridges

Samuel François Jouault

The results of the *Atlas de turismo alternativo en la Península de Yucatán* carried out by a group of social entrepreneurs dedicated to sustainable tourism in the state of Yucatán revealed the difficulties that individuals and groups providing alternative forms of tourism face in promoting and commercializing their services and attracting a sufficient market to support their business.

The term *Co´ox Mayab* means "Let's go to the Land of the Maya" ("Vamos al Mayab"). Today, *Co'ox Mayab* is also the name of a union of ten social entrepreneurs committed to sustainable tourism in Yucatán, who aim to demonstrate the value of responsible tourism while promoting solidarity and collaboration, social consciousness, and justice across public, private, and commercial sectors through the promotion, commercialization, and practice of alternative, community tourism in Yucatán. Such alternative tourist practices can be described as:

- **responsible** tourism that reduces the impact of tourist activity
- **supportive** and **reciprocal** encounters that allow visitors to engage with social reality and coexist with their hosts, promoting intercultural dialogue
- a **just** commercial industry in which public and private institutions cooperate with local organizations to ensure an equitable division of the benefits generated by tourism
- **conscientious** travel experiences contributing to personal growth through shared living experiences between visitors and local hosts, while also remaining sensitive toward natural and cultural heritage and resources

This presentation will focus on the role that various digital technologies, including a webpage, documentary, and promotional videos, play in facilitating and evaluating the *Co'ox Mayab* organization itself, as well as the experience and practice of alternative and sustainable tourism in Yucatán.

An Ethnographic Documentary on Sewing for Tourism at Home: Kinship Frictions, Souvenirs, and Debt in Inland Yucatán

Matilde Cordoba Azcárate

This paper discusses how Tekit, a former henequen village in Yucatán, Mexico has become invisibly but densely entangled with the tourism reality of the region through its specialization in the manufacture and distribution of *guayaberas* and work uniforms for the hospitality industry. Tekit's transformation has enforced the spatial and social re-organization of the village around a fragmented domestic *maquila* system of small factories and intermediaries working for major national and international textile corporations, mimicking the US-Mexico border *maquila* system of production.

The paper shows how this system of production has transformed Tekit into a modern factory town, or *desakota*, an in-between space, neither urban but not rural, with an extremely economically and socially vulnerable and fragile population but also, and paradoxically, into a space that is home to a thriving young population with increasing material wealth and a generalized sense of well-being. I follow the stories of three different families in the town and discuss how this system of production has deeply differentiated Tekit from nearby communities in which the lack of work has resulted in massive migration to Cancun, Mérida, and the United States. By comparing and contrasting the narratives and livelihood strategies of these families, the paper evinces how internal forms of kinship obligations, dependencies and reciprocities are reconfigured through external forms of debt. Individuals with weaker or smaller kinship networks are more prone to face liquidity constraints and remain in the lowest

steps of the production chain while those with larger ones are more likely to perform better.In all the cases, however, this domestic *maquila* system of production allows families to embrace urban cosmopolitan imaginaries, to redefine traditional gender roles while not migrating and practicing traditional Maya land and parenting practices in a region deeply marked by migration.

This paper is part of both the author's academic book manuscript in progress on the social, political, and ecological effects of tourism development in Yucatán, as well as part of a more recent collaborative educational ethnographic documentary on uneven forms of globalization in the region developed between the University of California San Diego and the Universidad Autónoma de Yucatán. This ethnographic documentary, still in its preliminary phases, will be part of a larger interdisciplinary and collaborative effort between the social sciences and the humanities, digital media technologies and documentary film and media to translate concepts and ethnographic research on global-local relations, uneven geographical development and cultural life into the world of audiovisual storytelling for research and pedagogic purposes. Our aim is to put together a grant proposal to fund an examination of historical and contemporary forms of everyday life in rural communities in the state of Yucatán, Mexico, using a combination of experimental media and historical and anthropological approaches to the study of culture.

Digital Humanities Pedagogy and Praxis Roundtable

Amanda Heinrichs
ahenrichs@amherst.edu
Five College Digital Humanities and Blended Learning, Amherst College, United States of America

James Malazita
malazj@rpi.edu
Rensselaer Polytechnic Institute, United States of America

Jim McGrath
james_mcgrath@brown.edu
Brown University, United States of America

Miriam Peña Pimentel
miriampenapimentel@gmail.com
Universidad Nacional Autónoma de México, Mexico

Lisa Rhody
lrhody@gc.cuny.edu
The Graduate Center, CUNY, United States of America

Paola Ricaurte Quijano
pricaurt@itesm.mx
Tecnológico de Monterrey, Mexico

Adriana Álvarez Sánchez
adralvsan@gmail.com
Universidad Nacional Autónoma de México, Mexico

Brandon Walsh
bmw9t@virginia.edu
University of Virginia, United States of America

Ethan Watrall
watrall@msu.edu
Michigan State University, United States of America

Matthew Gold
mgold@gc.cuny.edu
The Graduate Center, CUNY, United States of America

Overview/Panel Abstract

First developed in 2013, the Praxis Network (praxis-network.org) brought attention to the ways in which digital humanities was being used to rethink the nature of student training, campus partnerships, and pedagogy. The institutions profiled in the project aimed to reorient student training towards new, collaborative practices that would prepare students for forward-looking scholarship and meaningful careers in the humanities. Five years later, at DH 2018, we propose to reflect on these efforts, to assess the current state of digital humanities training and its relationship to and effects on praxis-oriented pedagogy.

The participants assembled for this roundtable session represent both the past and the future of these efforts to join theory and practice, classroom and public. Part of the roundtable will consist of participants that were part of the original Praxis Network, longstanding presences on their local campuses whose architects will report on the institutional challenges and successes they have faced in the five years since their profiling in the Network. The University of Virginia's Praxis Program Fellowship, for example, annually engages student cohorts in the development of collaborative research projects, but the program's success exposes limitations offered by this model even given attempts to expand it to a wider community with regional partners. Similarly, a representative from Michigan State University will discuss attempts to expand the collaborative nature of student work while also turning the projects increasingly towards the public. And the CUNY Graduate Center will share recent efforts by their digital fellows program to develop weeklong institutes in digital methods both locally and nationally.

The remaining institutions on the roundtable all represent new, like-minded initiatives with unique constituencies and concerns. Many of our programs encourage public-facing scholarship, and a representative from Brown University's Center for Public Humanities and Cultural Heritage will argue for experiential learning in the digital humanities in a public humanities context. The Five College Digital Humanities and Blended Learn-

ing Initiative offers lessons in how to address resistance to incorporating praxis-oriented methodologies in the classroom. Alt.code, an initiative funded by the National Endowment for the Humanities at Renssealaer Polytechnic Institute, offers a model for joining the instruction of technical skills with critical perspectives on technology as linked parts of the same epistemic domain. Finally, a trio of scholars from the Universidad Nacional Autónoma de México and Tecnológico de Monterrey will speak to the difficulties they have encountered developing pedagogical and institutional initiatives based in digital humanities in Mexico.

This roundtable thus hopes to draw participants new and old into a conversation about methodological training, which necessarily must become inflected differently in diverse local, institutional contexts. Collectively, we argue for a pedagogy that is public, collaborative, and that centers the student. In the spirit of the original Praxis Network, we hope that this collection of programs will offer models, lessons, and cautionary tales. Looking to the future, we hope that the roundtable will start new conversations and seed new ideas.

Michigan State University's Cultural Heritage Informatics Initiative: Methods and Models for Building Capacity in Digital Cultural Heritage

Ethan Watrall

As with many other domains, cultural heritage has entered a new age in which digital methods and computational approaches are having an unavoidable impact on research, teaching, preservation. public engagement, and all aspects of scholarly communication. The problem is that cultural heritage scholars and professionals who have not traditionally characterized themselves as being particularly digitally inclined are increasingly being asked to engage with issues, methods, models, and practices that are uniquely digital in nature. Unfortunately, while the need for innovative digital praxis exists, we are only starting to establish methods and models to build vital digital capacity among undergraduates, graduate students, and existing professionals and scholars.

It is within this context that this talk will explore Michigan State University's Cultural Heritage Informatics Initiative (chi.anthropology.msu.edu). Founded in 2010 and administered by the Michigan State University Department of Anthropology in partnership MATRIX: The Center for Digital Humanities & Social Sciences and the Lab for the Education and Advancement in Digital Research (LEADR), the Cultural Heritage Informatics Initiative was originally conceived as having two primary goals. First, it was intended to serve as a platform for interdisciplinary scholarly collaboration and communication in the domain of digital cultural heritage practice at Michigan State University. Second, it was intended to equip students and professionals with the skills to apply digital methods and computational approaches to cultural heritage materials and questions. Despite these two initial goals, the initiative has shifted over the years to focus almost exclusively on the second, providing cultural heritage students and professionals with an opportunity and environment to learn and build digital skills. The two most tangible expression of this goal are the Cultural Heritage Informatics Graduate Fellowship Program and the Digital Heritage Fieldschool.

The intention of this talk is to reflect upon the challenges the initiative has faced since it was founded, exploring successes and failures, and look forward as we move towards a decade of building capacity and community in digital heritage at Michigan State University and beyond.

HD pedagogy and praxis in Mexico: practices, platforms, institutions

Miriam Peña Pimentel
Adriana Álvarez Sánchez
Paola Ricaurte Quijano

In Mexico as elsewhere, universities differ in their human and financial resources, their physical, technological and administrative infrastructures, their educational models and their programs, among other variables. This multiplicity of conditions frame the dynamics and the spaces for knowledge production and student training. Within this heterogeneous institutional context, Digital Humanities are building an emerging field where different initiatives have been developed. Digital academic publications; DH events; activities in support of the open access movement (MOOC, data, digital libraries, Wikipedia); the design of new courses and DH projects; and the creation of networks and labs, are part of our strategy to consolidate the field. In our experience, implementing a DH curriculum in a higher education setting in Mexico will never be an easy task. We find multiple elements in the institutional structures that hinder the development of DH teaching and research. The training of professionals and the development of digital projects face a series of problems related to bad administrative decisions and lack of vision of DH as a field. There is not enough political will to include DH courses and methodologies within the curriculum of undergraduate programs or to create new DH graduate programs. There is resistance to support DH projects and platforms. At the praxis level, we are trying to subvert the traditional pedagogical model and developing alternatives to learn in situated contexts and with different methodological and technological tools. Several questions arise from this experience: Is our DH pedagogy embodied and embedded in its sociocultural context? How is DH pedagogy in Mexico different from the DH pedagogy elsewhe-

re? What does praxis mean in our local scenario? What do we bring to the reflection of DH pedagogy and praxis? This paper aims to reflect on these questions and describes two cases of DH pedagogy and praxis in two higher education institutions, UNAM and Tecnológico de Monterrey. We describe the purpose and methodologies of two DH labs: e-labora and Openlabs. We believe that these initiatives are trying to approach the question of how to teach and learn DH in our local setting. Labs have three obvious repercussions: a) promote the appropriation of new methodologies and the reformulation of the disciplines involved; b) allow participants to become involved with their environment by applying what they learn to the solution of everyday situations, and c) create favorable conditions for the development of multiple competencies.

Alt.code: DH as Reconfiguration of the Boundaries of Computer Science Education

James Malazita

"alt.code" is a National Endowment for the Humanities funded initiative that combines humanities, the arts, and computer science to teach critical digital literacy, the politics of technology, and technical skills to undergraduates. The initiative uses the digital humanities as a vehicle to articulate the intersections of technical knowledge and critical thought to both computer science and humanities undergraduate students in the same classroom. Though alt.code is directed by Humanities, Arts, and Social Sciences (HASS) faculty, collaboration with Computer Science (CS) faculty allows humanists to teach critical theory directly within foundational CS courses, including "Introduction to Computer Science," where CS students read critical humanities and social science texts while also learning the fundamentals of programming in Python. In turn, CS undergraduates are encouraged to use a portion of their required humanities credits to enroll in alt.code HASS courses, team-taught classes that encourage students to explore issues of social justice, epistemic bias, and politics of technology through prototyping and critical design exercises.

The alt.code initiative includes a sequence of four classes held across RPI's School of Humanities and School of Science, a Digital Humanities guest speaker series, and the development of RPI's "Tactical Humanities Lab" to support undergraduate, graduate, and faculty praxis-based research. The broader, long-term goals of the initiative include: the redevelopment of the Computer Science curriculum to teach critical theory across the majority of their technical courses; encouraging computer science undergraduates to frame critical perspectives on technology as a **core** part of their disciplinary expertise; and encouraging HASS undergraduates to frame digital and technical work as forms of political knowledge and action.

While other undergraduate programs have worked to bridge teaching about social "impacts" of technology with teaching technical skills, they often do so in a modularized fashion. That is, CS and STEM students often take a series of technical courses that teach the "fundamentals" of technical production (as in, abstract mathematics, programming, and decontextualized construction), which are supplemented with "politics of technology" classes or humanities electives. Ostensibly, students will combine their "technical" and "non-technical" coursework to practice socially just or conscientious technological development.

Science & Technology Studies (STS) and Engineering & Liberal Arts scholars, however, have noted that even hybridized curricular structures encourage the bifurcation of technical practice from social thought. STEM students begin defining the "epistemic object" of their studies as abstract, decontextualized technical production, with "social concerns" treated as external constraints that impact, but are never fully a part of, technical practice. Similarly, adjuncting technical practices into humanities classrooms as "methods" may encourage humanities and social science students to frame digital and technical tools as epistemically and politically neutral skills and practices, as opposed to social, political, and epistemological arguments in their own right.

Alt.code treats teaching critical theory and technical practice infrastructurally, encouraging students to synthesize these bifurcated domains by teaching them critical perspectives on technology and technical skill in the same classroom, framed as inseparable parts of the same epistemic domain. In the future, we hope to propagate that model across both the CS and humanities curriculum at RPI, and that the successes and failures of our program can serve as a curricular and pedagogical model for other institutions grappling with bridging humanistic and technical knowledge.

"'Yes, but…': Praxis in a Theory-Focused Environment"

Amanda Heinrichs

At the DH 2018: Pedagogy and Praxis session, I will report on a series of efforts to introduce faculty to pedagogical praxis at liberal arts institutions that have been historically resistant to praxis. Since 2014, The Five College Digital Humanities and Blended Learning Initiative has been supported by a combination of Mellon and Teagle Foundation grants, and until this current academic year digital humanities and blended learning were conceived of as separate entities with separate goals, audiences, and even directors. On the one hand: DH/research/theory. On the other: BL/teaching/praxis. Yet the founders of the DH/BL initiatives (and administrators across the Consortium) want more investment in DH praxis for undergraduate students; my job this year is to develop modules that

will lay the groundwork for a 5 College DH undergraduate certificate.In addition to a single director, under whom the program will necessarily merge DH and BL, theory and praxis, 2017-2018 is the last year of external funding for the 5CollDH/BL program. Thus the program enters a resource-poor environment, even as praxis and theory are (at least in name) integrated.

One particular challenge in a consortium of four small liberal arts colleges and one large public land-grant university is conflicting attitudes towards praxis across the Five Colleges. Faculty at Hampshire and Amherst in particular have argued praxis should be left to career development programs: and most praxis-oriented classes are taught at University of Massachusetts-Amherst, the only public institution of the five. Despite this, many faculty at all five colleges have taught themselves digital methods—or employed technical support staff—in order to develop robust DH projects.

In this way, faculty at prestigious undergraduate institutions like Amherst say "Yes, but…" to DH praxis. They are open to the idea that DH tools can address complex theoretical questions, but for various institutional, historical, and personal reasons do not often bring those tools to their classrooms. Therefore, if the goal of the roundtable is to discuss the state of praxis-oriented DH education, I would like to tweak the question a bit, and ask, "What happens when faculty say *yes* to DH praxis, *but* not for their undergraduates? What are the ethical stakes when professors at prestigious undergraduate institutions push praxis-oriented courses to the one public university in the consortium?"

In addition to drawing on principles of minimal computing, as well as scholars such as Ryan Cordell and Kalani Craig, I will report on a series of faculty seminars titled "A No-Tech Introduction to the Digital Humanities." Funded by the Center for Humanistic Inquiry at Amherst College, this series presented digital humanities tools in their intellectual context: network analysis as an intervention in Heideggerian phenomenology, for example. I pair this theoretical understanding of DH tools with my experience assisting with the practical course at UMass-Amherst where students learned the fundamentals of Python for text and network analysis. Ultimately, I argue that a focus on theory which aligns with minimal computing principles not only allows for praxis in a resource-poor environment as the Five College Digital Humanities Initiative moves away from external grant funding, but also provides a potential response to the ethical dilemma of elite small colleges outsourcing praxis-oriented courses to a public institution.

Public Works: Lessons in Experiential Learning from Digital Public Humanities Classrooms

Jim McGrath

In recent years, institutional desires for digital humanities projects and initiatives that are invested in ideas of "public humanities" have materialized in initiatives like the National Endowment for the Humanities' "Digital Projects for the Public" grants, the inclusion of chapters on public humanities by Sheila Brennan and Wendy Hsu in *Debates in the Digital Humanities*, and the creation of a graduate certificate in Digital Public Humanities at George Mason University (among other recent developments). While many digital humanities practitioners publish, debate and disseminate "in public" through the creation of open access scholarship, freely-available datasets, and content published on web sites and social media networks, these activities don't always interpellate -- or invite collaborations and engagement with -- publics who do not traditionally reside within or in the orbit of academic institutions and discourse communities. There are notable exceptions: Wendy Hsu's investments in "building at the rate of inclusion" on digital initiatives, Mitchell Whitelaw's call for "generous interfaces" that anticipate a wider range of users and needs, Lori Emerson's reminders of the limitations and restrictions of "invisible interfaces" that fail to acknowledge varied experiences with technology, Mia Ridge's careful considerations of crowdsourcing initiatives, the *Documenting The Now* initiative's investment in ethical uses of social media activism. How can we effectively teach digital humanities practitioners to imagine, engage, and collaborate with various publics and their various interests, needs, and uses of digital tools, networks, and resources?

Drawing on the decade-long history of Brown University's John Nicholas Brown Center for Public Humanities and Cultural Heritage, this panel presentation will discuss the ways digital tools, methodologies, and contexts have materialized in the modes of experiential learning valued by our Master's Program and Certificate Program in Public Humanities: its postdoctoral fellow in Digital Public Humanities, its courses, independent studies, community collaborations, exhibitions, and working groups. Through a survey of its teaching, collaborations, and projects, the presentation will highlight generative ways to take the challenge and opportunity of public-facing (and public-serving) digital initiatives seriously: by anticipating economic limitations and audience needs that might impact a project's design and accessibility, through collaborating directly with publics rather than assuming a familiarity with their needs, and in favoring, when appropriate, iterative and ephemeral approaches to project implementation that embrace limitations of resources and temporal conditions. In addition to highlighting the benefits of a digital pedagogy informed by public humanities concerns, methodologies, and professional contexts, this presentation will also consider what lessons and ideas public humanities programs, courses, projects, publications, institutional structures, and methodologies interested in digital contexts could and should take from digital humanities practitioners.

Praxis and Scale: On the Virtue of Small

Brandon Walsh

The University of Virginia's Praxis Program attempts to redefine graduate training by means of a targeted digital humanities intervention during the early years of graduate students' time in their programs. Each year, staff and faculty work alongside a student cohort to theorize a new digital intervention and train the students to carry it out alongside them in the spring. The result is that the students get an intensive introduction to a variety of digital humanities issues ranging from project management to technical training for their particular project. Drawing upon the pedagogical theories of Cathy Davidson, Paolo Freire, Bethany Nowviskie, and others, the program aims to equip graduate students with the skills and ethos necessary to thrive in collaborative, open work, the very things that can help prepare them for a variety of careers beyond the tenure track, and we do this by putting the students in charge as much as possible.

Now in its seventh year, the program has a proven track record of success based on exit interviews (both qualitative and quantitative), job placements, and future awards received by our alumni. This presentation will discuss one consistent criticism the program has received throughout its existence: scale. Each cohort consists of six students, and we consistently get requests to expand the program with more students or for new audiences, requests that are difficult to carry out. We have developed a rough stack of technical and administrative lessons that are consistent year-to-year, but, the program relies on flexible instructors that can respond to student interests as they develop over the course of the year. The program represents a significant investment of resources, and, given the emphasis on student-driven work, it can be difficult to predict exactly what staff will be the primary points of contact for the project.

We are limited in the number of cohorts that we can maintain at any one time, which may point to a fundamental limitation in student-centered programming: for better or worse, there is a limit to the scale at which these programs can be offered. At our institution, we have confronted this difficulty in scale by offering a diverse range of graduate programs with different structures, pedagogies, and audiences. We also engage alumni of the Praxis Program in other initiatives, folding their strengths into our pool of resources to allow recent students to quickly become able teachers. Since the Praxis Program's inception seven years ago, for example, we have piloted a program in digital humanities and library research methods for undergraduates that draws upon some of the lessons from the Praxis Program and engages Praxis alumni as instructional assistants.

While there are compelling examples of scaling up student-centered programs, most notably in the FutureEd Initiative, for our group, to adopt a program that centers project-based education is to make a strategic investment in the small scale as a meaningful point of intervention. Such investments pay off in big ways for the people involved, and this paper thus advocates for praxis-oriented pedagogical approaches as a means of centering the development of the people and relationships at the core of our work. Far from limiting our impact, I argue that cherishing the small means fully investing in the long-term, future effects that these students have as they become teachers in their own right.

The Digital GC in Praxis: Degrees, Fellowships, and Community-driven Support

Matthew Gold
Lisa Rhody

At The Graduate Center, City University of New York (CUNY), Digital Initiatives continue to grow and change in response to institutional investments in new modes of graduate training. This presentation will offer an overview of our multi-pronged approach to graduate education through degree programs, fellowships, and community-driven support.

The Masters of Arts in Liberal Studies program features a 2-course introduction to digital humanities through theory and practice. Students participate in workshops, explore datasets, and complete the Fall semester by proposing a potential project to be executed in the Spring. In the Spring's problem-based curriculum, students refine their proposals, learn to develop project and data management plans, evaluate project proposals, select proposals, and form working groups to bring two to three proposed projects to fruition. At the end of the year, students present their prototypes at an annual campus-wide event. The success of the MALS program and the DH Praxis Seminar sequence has led to the creation of 2 new master's degree programs to begin in Fall 2018.

The GC Digital Fellows Program operates as an in-house think-and-do tank for digital projects, connecting Fellows to digital initiatives throughout The Graduate Center. Digital Fellows work collaboratively to help build out "The Digital GC" — a vision of the Graduate Center that incorporates technology into its core research and teaching missions. The eleven fellows occupy a leadership role as mentors and advisors to peers who are interested in integrating digital methods into their scholarship. Fellows have developed extensive tutorials and workshops on topics ranging from establishing a digital scholarly identity to using Python and R to create data visualizations, and fellows serve as faculty during week-long digital institutes for faculty and students. Fellows develop project management and grant-writing skills while actively participating in on-going digital humanities

projects, such as the CUNY Academic Commons and Manifold Scholarship. Fellows have also taken on projects that make use of public humanities data, such as the NEH Impact Index.

Graduate Center students can access methods training and support throughout their academic career. Each January during the Digital Research Institute--a week-long digital methods intensive --participants with no prior technical experience learn foundational skills such as working from the command, version control with git and GitHub, Python, and database management, then choose from electives like natural language processing, machine learning, HTML/CSS, and APIs. The week begins and ends with discussions about project planning and digital research ethics.

Provost's Digital Innovation Grants provide doctoral students in good standing with funds to propose and prototype research projects. Students write proposals and budgets, execute their idea, and present ongoing work at the end-of-year, community showcase. Previous student projects have received external funding and awards, including the Lisa Lena Opas-Hänninen Prize for New and Young Scholars and the Paul Fortier Prize for New and Young Scholars.

Through examples of student projects, research, and professional development workshops, this presentation will offer an overview of the opportunities and challenges of scaling out graduate training in digital humanities methods to wider communities.

References

#FutureEd. Humanities, *Arts, Science, and Technology Alliand and Collaboratory*. https://www.hastac.org/initiatives/futureed

Bauwens, M. (2005). The political economy of peer production, *CTheory*, 1.

Berry, D. M. (2012). *Understanding Digital Humanities*. New York: Palgrave Macmillan.

Davidson, C. (2017). *The New Education: How to Revolutionize the University to Prepare Students for a World In Flux*. New York, NY: Basic Books.

D'Iori, P. & Barbera, M. (2011). Scholar Source: A Digital Infrastructure for the Humanities. In Bartscherer, T. (Ed.). *Switching Codes. Thinking Through Digital Technology in the Humanities and the Arts*. Michigan: University of Chicago Press.

Freire, P. (1970). *Pedagogy of the Oppressed*. New York, NY: Bloomsbury Academic.

Koh, A. (2014). Introducing Digital Humanities Work to Undergraduates: An Overview. *Hybrid Pedagogy*. http://www.hybridpedagogy.com/journal/introducing-digital-humanities-work-undergraduates-overview/

Lafuente, A. (2016). Laboratorios ciudadanos: ciencia ciudadana y ciencia común. CCCD. Página web.

Ostrom, E. (2007). *Understanding Knowledge as a Commons.* Cambridge, MA: MIT Press.

Manovich, L. (2013). *Software Takes Command*. New York: Bloomsbury.

Manovich, L. (2015). The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. Disponible en https://www.academia.edu/15328596/

Mignolo, W. (2010). *The geopolitics of knowledge and the colonial difference*. Praxis Publica http://praxispublica.org/wp-content/uploads/2010/10/WALTER-MIGNOLO-GEOPOLITICS-OF-KNOWLEDGE-DUKE-UNIVERSITY.pdf

Mignolo, W. (2013). Geopolítica de la sensibilidad y del conocimiento. Sobre (de) colonialidad, pensamiento fronterizo y desobediencia epistémica. *Revista de Filosofía*, 74(2), 7-23.

Mignolo, W. (2003). *Historias locales/diseños globales: colonialidad, conocimientos subalternos y pensamiento fronterizo*. Madrid: Akal.

Mignolo, W. (2002). The geopolitics of knowledge and the colonial difference. *The South Atlantic Quarterly*, 101(1), 57-96.

Moretti, F. (2004). Gráficos, mapas, árboles. Modelos abstractos para la historia literaria. *New Left Review*, 24, pp. 60-85. http://newleftreview.es/24

Nowviskie, B. (2011). "Praxis and Prism." http://nowviskie.org/2011/praxis-and-prism/

———. (2012). "Too Small To Fail." http://nowviskie.org/2012/too-small-to-fail/

Ramos, I. y Ricaurte, P. (2015). Análisis de redes sociales y comunidades virtuales. *Revista Virtualis,* No. 11.

Rogers, K. (2015). "Humanities Unbound: Supporting Careers and Scholarship Beyond the Tenure Track." *Digital Humanities Quarterly* 9.1.

Rogers, R. (2013). *Digital Methods*: Cambridge: MIT Press.

The Praxis Program. University of Virginia Library's Scholars' Lab. http://praxis-network.org/.

Justice-Based DH, Practice, and Communities

Vika Zafrin
vzafrin@bu.edu
Boston University, United States of America

Purdom Lindblad
purdom@umd.edu
University of Maryland, United States of America

Roopika Risam
rrisam@salemstate.edu
Salem State University, United States of America

Gabriela Baeza Ventura
gabines1022@gmail.com
University of Houston, United States of America

Carolina Villarroel
carolina@central.uh.edu
University of Houston, United States of America

Introduction

In the last 25 years, advocacy- and justice-based research has seen development in sociolinguistics, social policy, history, and many areas of digital humanities. We take cues from past literature (Cameron et al., 1992; Risam, 2015; Bailey, 2011; Wernimont and Losh 2016) to define this as research that aims to benefit not only its audience but also its subjects, and is conducted from an intersectional awareness of compounding oppressions.

What does justice- and advocacy-based work look like from different vantage points within the knowledge work/cultural heritage ecosystem — infrastructure, research, support, organizational change, community outreach, teaching, co-learning? What does being a practitioner look like in the current international political climate?

Supporting digital scholarly practice—at its best, an active research practice of its own—has recently seen new challenges. Discourse on the inherently political nature of what we do has picked up (Risam et al.; Bourg, 2016; Drake, 2017; Jules, 2018; Hathcock, 2015; Johnson and Dinsman, 2016; Nowviskie, 2015); and the recent upsurge in volatility of U.S. politics has reverberated across DH and digital libraries work. Institutions like Digital Library Federation and Association for Computers and the Humanities have published statements responding to U.S. actions; begun and enabled conversation around more inclusive representation in knowledge production; taken hard looks at our own complicity with perpetuating the socio-political status quo and made funding available toward doing better. Some organizations—notably Global Outlook::Digital Humanities—have questioned their own institutional contexts. Do we continue consolidated efforts? At what point is a break away from even our relatively young establishments warranted?

In fall 2016, two of the authors began a conversation with digital scholarship support practitioners in which emerged a need for a collaboratively created online space to help us work from a more consciously justice-based perspective. This space needs to enable anonymous and attributed conversation around vulnerable topics, contain a list of resources for a socio-politically active framing of our work, encourage mentoring, and enable us to build and consciously deploy institutional infrastructures in volatile times.

At DH2018, we will present first a prototype of this community resource, and then specific examples of our own justice-based digital scholarship work from a variety of institutional contexts. Represented will be large and small, public and private U.S. institutions; "miracle worker" DH support professionals working largely on their own, those working in established centers, and more traditionally situated scholar-academics.

We will discuss formal institutional structuresneeded to support these efforts, describing ways to take advantage of existing structures informally, flexibly, at times opportunistically—and the benefits, drawbacks, and risks of doing so. We will invite discussion that we hope will inform further development of the online resource.

The panel will offer a range of possibilities for justice-based work in different academic and cultural settings. We hope audience members will leave with new ideas for incorporating or enhancing a justice-or advocacy-based perspective in their work without leaving their professional context.

Presentations 1, 2, and 3 will be given in English. Presentation 4 will be given in Spanish. Translated outlines of each presentation will be available.

Presentation 1

Vika Zafrin will discuss Boston University Libraries' institutional thinking behind mobilizing their Research I resources and a rare opportunity for organizational redefinition to orient new projects towards working with an explicitly anti-colonialist mindset.

In the last year and a half, BU Libraries began several simultaneous conversations: some with faculty members who tend to work in small, sometimes personal, archives; others with area librarians about collaborating to amplify each other's existing and nascent efforts towards social justice.

BU Libraries have long supported digital scholarship informally, but the Digital Scholarship Services department (DiSc) is young. This allows it some freedom in setting its own direction. The charge is to help create better infrastructure for digital scholarship at BU — including a technically stronger and more sustainable institutional repository, increased and strategic digitization of our holdings, increasing the number of technically competent staff able to work on digital projects, and community education around both tools and issues of digital scholarship.

All this accommodates a variety of possible projects. DiSc has chosen to dedicate part of its efforts to building relationships that allow for discovery of small analog archives of social and political activism materials vulnerable to either disappearance or obscurity, and using what resources (power) the Libraries have toward beginning to assist in their digitization, curation with an eye to research, and preservation. DiSc staff are conscious of the perils of appropriating such materials, and are taking guidance from recent digital humanities and libraries work on anti-colonialism in digital collection creation (Risam et al.). In addition, the team is guided by the notion of slow DH (Hyman via Corona, 2017) and the work of our colleagues

in libraries and DH around representational belonging (Caswell et al., 2016), the tricky notion of empowerment (Mckesson, 2017), and community control of the narrative around digital objects (Christen, 2012; Cushman, 2017).

Zafrin will also discuss the ways in which a justice-based mindset changes daily work as a research library effects major organizational change. How does it inform the search for a new university librarian? How are we making decisions about what student employees and interns do and don't do? How are we treating our supervision of young or less experienced workers as pipeline work— and articulate that it deserves our scarce time regardless of what professional path they take later? How does a higher ed library balance serving its university constituency and serving the larger community in which it is situated? How do we combine forces with other institutions to do activist infrastructure work that requires relatively little effort from each participant and brings disproportionately positive results? How does justice-based work inform what resources we ask for, what do we do when most of the resources we have are people, how do we start a community outreach project without a history of such? How do we expand our thinking about the social aspects of knowledge creation and propagation by learning different ways of knowing?

Presentation 2

Purdom Lindblad'spresentation focuses on questions of how those new to digital humanities are introduced to advocacy and justice-based work? How do we collaboratively articulate and frame responses to the overlapping and compounding oppressions that are both inherent and tacit in digital work? What practices and approaches are needed to render the theoretical underpinnings of digital work more accessible (and how do we teach such practices)? Situating the collaborative design and teaching of Maryland Institute for Technology in the Humanities' (MITH) "Anatomy of Digital Research" course as a case study, Lindblad will describe the specific decisions and practices used to design and teach an advocacy and justice-based digital humanities course.

Drawing from the open, cohort-based model of the MITH digital humanities incubators — which are collaboratively designed with the African American History and Culture and Digital Humanities (AADHum) initiative — and Advocacy by Design practices outlined by Jeremy Boggs and Purdom Lindblad, the MITH 498 "Anatomy of Digital Research" course centers intersectionality and justice as foundational to digital work. The course modules move from a broad introduction to the varieties of common methodological and technical approaches to project management to literature reviews and finally to the social and political aspects of digital humanities work. The course modules act to frame discussions of what research is centered, technical choices — from data selection and cleaning to manipulation and analysis — and design can expose or render invisible the theoretical and ethical underpinnings of the work.

The MITH course development team leaned on the lessons learned from the MITH - AADHum collaboration, which explicitly centers black studies, blackness, and black culture as core to digital work. Such centering meant naming openly our shared commitment to black studies and referencing this concern through every decision made around the digital humanities incubators. Thus, the most consequential actions were those of imagination-specifically, our expectations of who the "right" people are to do the work of digital skills incubators within the framework of AADHum; the right people were those who centered black people and the concerns of black studies, who could have limited (or amazing) technical skills. Hiring and staffing for the AADHum project then shaped the research questions and approaches which guide the digital incubators. MITH 498 fronts how decisions around what is core to the research then shape subsequent decisions, from raw materials, citations, what collaboration looks like for this research, to how the final designs reflect or obscure what is center and why. Boggs and Lindblad's work on Advocacy by Design lends questions of how people are represented in, or are subjects of, academic work. MITH 498 borrows Advocacy by Design's emphasis on fronting the 'why' of research, articulating particular stances on interrelated concepts (principles) and using specific approaches (elements) to make visible these principles in the workflows, interactions, and research products of the course.

Presentation 3

Roopika Risam will discuss designing a network of digital humanities practitioners at regional public universities, as a justice-oriented intervention. While initiatives like the Institute for Liberal Arts Digital Scholarship and the Digital Liberal Arts Exchange promote engagement and shared infrastructure between research universities and small liberal arts colleges, there are no cooperative initiatives addressing the unique challenges of undertaking digital humanities pedagogy and scholarship at the poorly-funded lower-tier public universities that serve the majority of students in higher education in the United States. The demographics of students who attend these universities are diverse in race and ethnicity, socioeconomic status, age, career goals, and paths to college. Consequently, models and initiatives for digital humanities designed for elite university students at flagship research universities or small liberal arts colleges do not meet the needs of this vulnerable student population.

Digital humanities practitioners who work in this context face a number of barriers: the challenges of their

student population; high teaching loads; lack of research funding; and the imperative to serve as their own project managers, developers, designers, and IT support for digital humanities initiatives. Yet, they have two important, justice-based perspectives that those at other institutions do not: 1) the possibility of engaging students who typically do not understand themselves as producers of knowledge in digital humanities research and 2) a sense that their universities are "stewards of place" and thus have an ethical responsibility not only to their students but also to their local communities (Saltmarsh et al., 2014). Without a mechanism for collaboration, like a national network, individual initiatives cannot be properly leveraged and are limited by the already-strained fiscal and institutional constraints of these universities.

In response to these challenges, this paper explores the work undertaken by Risam and her colleagues to design a network that responds to the unique circumstances of digital humanities practitioners at regional comprehensive universities, supported by a grant from the National Endowment for the Humanities and the Institute for Museum and Library Services. It begins by discussing the results of a study of regional comprehensive digital humanities practitioners around the United States, identifying the successes, challenges, and barriers that these practitioners encounter when undertaking digital humanities initiatives with underserved student populations who stand to benefit tremendously from learning experience with digital humanities. The paper then considers the primary areas where a collaborative network can help support practitioners: place-based curricula and pedagogy, faculty-librarian-student collaboration, shared infrastructure, and professional development. It also discusses the challenges of maintaining and sustaining a distributed network of digital humanities practitioners while fostering connections to existing networks and communities of practice. Through this network, this paper suggests, digital humanities practitioners are better positioned to create learning experiences that meet the diverse needs of student populations at regional comprehensive universities, ensuring that opportunities to engage with digital humanities are not limited to elite students in higher education.

Presentation 4

The majority of digital projects (especially those funded by major granting institutions) tend to focus on canonical texts that reinforce Western epistemologies. Scholars at the University of Houston's Recovering the US Hispanic Literary Heritage are committed to calling attention to the gaps in digital scholarship and highlighting work in fields such as US Southwest studies, US Latina/o studies, Indigenous studies, border and transamerican archives, and racial and ethnic literatures and histories.

Gabriela Baeza Ventura and Carolina Villarroel will focus on transcultural and transamerican digital scholarship and the ways in which diverse projects contribute to and de-center the growing field. They will discuss the steps taken to establish the first Digital Humanities Research Center for US Latina/o Studies at the University of Houston. A process that begins with a Mellon Foundation planning grant to visit various DH centers in the US, which has been fundamental in understanding and seeing first-hand infrastructures and methodologies that document the need of a center such as theirs. The center is based on the foundational work of the Recovering the US Hispanic Heritage and aims to foster, produce and promote Latino scholarship in the digital humanities. Recovering the US Hispanic Literary Heritage is a program to locate, preserve and make available the written legacy of Latinas and Latinos in the United States since colonial times until 1960. Through 26 years of successful work Recovery has not only being able to inscribe the excluded history of Latinas/os, but also has created an inclusive, vast and interdisciplinary digital repository that facilitates scholarship in this area of studies. The foundation of a digital humanities center becomes then a natural step to continue what Gloria Anzaldúa refers to as "doing work that matters" and Chela Sandoval as enacting the methodology of the oppressed to revert the exclusion in the digital humanities and other humanistic discourses of a fundamental component of the culture and history of the United States, the Latino community.

Furthermore, in order to decentralize and decolonize the archive, they demonstrate the need to keep building on the work of Recovery scholars—researchers who recover the legacy Latinas and Latinos who live in the United States prior to 1980. The use of DH resources in Recovery scholarship will lead us (scholars, community, students, etc.) to rethink how research projects are conceived and to challenge archival traditional modes of representing knowledge. In this sense, their goal is to demonstrate that minority archives and digital humanities (DH) projects that highlight minority voices disrupt the mainstream perception of history and literary canon through these unacknowledged histories.

References

Bailey, M. (2011). All the digital humanists are white, all the nerds are men, but some of us are brave. *Journal of Digital Humanities* 1.1, http://journalofdigitalhumanities.org/1-1/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-brave-by-moya-z-bailey/ (accessed 26 April 2018).

Bourg, C. (2016). Libraries, technology, and social justice. Access 2016, Fredericton, NB, https://chrisbourg.wordpress.com/2016/10/07/libraries-technology-and-social-justice/ (accessed 26 April 2018).

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., and Richardson, K. (1992). *Researching Language: Issues of Power and Method*. London: Routledge.

Caswell, M., Cifor, M., and Ramirez, M. H. (2016). 'To suddenly discover yourself existing': uncovering the impact of community archives. *The American Archivist* 79(1): 56-81. https://doi.org/10.17723/0360-9081.79.1.56 (accessed 26 April 2018).

Christen, K. (2012). Does information really want to be free?: indigenous knowledge and the politics of open access. *The International Journal of Communication*, 6: 2870-2893.

Corona, C. (2017). Digital humanities keynote speaker highlights the importance of geographic information systems. *Daily Titian*, https://dailytitan.com/2017/11/digital-humanities-geographic-information-systems/ (accessed 26 April 2018).

Cushman, E. (2017). Supporting manuscript translation in library and archival collections: toward decolonial translation methods. Recording Lives: Libraries and Archives in the Digital Age, Boston University, Boston, MA.

Drake, J. (2017). How libraries can trump the trend to make America hate again. 2017 meeting of the British Columbia Library Association, Vancouver, BC, https://medium.com/on-archivy/how-libraries-can-trump-the-trend-to-make-america-hate-again-8a4170df1906 (accessed 26 April 2018).

Hathcock, A. (2015). White librarianship in blackface: diversity initiatives in LIS. *In the Library with the Lead Pipe*, http://www.inthelibrarywiththeleadpipe.org/2015/lis-diversity/ (accessed 26 April 2018).

Johnson J. M. and Dinsman, M. (2016). The digital in the humanities: an interview with Jessica Marie Johnson. *Los Angeles Review of Books*, https://lareviewofbooks.org/article/digital-humanities-interview-jessica-marie-johnson/ (accessed 26 April 2018)

Jules, B. (2018). We're all bona fide: Why preserving cultural heritage on the web/social media should be an inclusive and community centered effort. *On Archivy*, https://medium.com/on-archivy/were-all-bona-fide-f502bdaea029 (accessed 26 April 2018).

Mckesson, D. (2017). We don't know how this movie ends. Pod Save the People, https://crooked.com/podcast/we-dont-know-how-this-movie-ends/ (accessed 26 April 2018).

Nowviskie, B. (2015). Digital humanities in the Anthropocene. *Digital Scholarship in the* Humanities 30:suppl_1: i4-i15, https://doi.org/10.1093/llc/fqv015 (accessed 26 April 2018).

Risam, R. (2015). Beyond the margins: intersectionality and digital humanities. *Digital Humanities Quarterly* 9.2, http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html (accessed 26 April 2018).

Risam, R., et al. An invitation towards social justice in the digital humanities. http://criticaldh.roopikarisam.com/ (accessed 26 April 2018).

Saltmarsh, J., O'Meara, K., Sandmann, L., Giles Jr., D., Cowdery, K., Liang, J., and Buglione, S. (2014). Becoming a steward of place: lessons from AASCU Carnegie community engagement applications. Washington: American Association of State Colleges and Universities.

Wernimont, J. and Losh, E. (2016). Problems with white feminism: intersectionality and digital humanities. In Crompton, C., Lane, R., and Siemens, R. (eds), *Doing Digital Humanities*. New York: Routledge, pp. 35-46.

Long Papers

The Hidden Dictionary: Text Mining Eighteenth-Century Knowledge Networks

Mark Andrew Algee-Hewitt
mark.algee-hewitt@stanford.edu
Stanford University, United States of America

Introduction

While written works are often encountered by readers as linear phenomena, one of the most important conceptual advances offered by the Digital Humanities is the way that computational text analysis has permitted researchers to find non-linear patterns that speak to organizational principles embedded in even a single text. The methods developed to parse thousands, or millions, of texts can, in the context of a single work, reveal connections and patterns that are unavailable to a human reader.

Even in reference books, whose alphabetic order discourages the same kind of linearity found in novels, digital methods have proven effective at revealing alternative ordering principles. This has been particular important in eighteenth-century studies. For example, recent digital work on the French *Encyclopédie* has sought to assess the compatibility of the multiple ways in which the text was organized by its authors. In their 2002 article, Gilles Blanchard and Mark Olsen measure the knowledge domains described by Diderot in his introduction by counting the number of renvois, or "see alsos" between articles in each domain. Similarly, Heuser, Algee-Hewitt and Bender have also reconstructed the French *Encyclopédie* based on which articles are connected by renvois. In both cases, an alternative structure emerges: one that speaks to connections between domains of knowledge that are more meaningful than the alphabetic layout would suggest.

In this project, I employ a similar set of methodologies to explore the other foundational linguistic reference book of the eighteenth century, Samuel Johnson's 1755 *Dictionary of the English Language*. While it lacks a system of renvois to counter-balance the prevailing alphabetic order it shares with Diderot's work, it nevertheless contains a hidden system of connections between seemingly disparate articles, whose organization can only be revealed through quantitative analysis: the quotations used in the definitions of each word. These quotations are what separate Johnson's dictionary from other, earlier dictionaries. In providing a contextual basis for assessing meaning, Johnson grounds definitions in historical usage and contingent situations. Yet, by Johnson's own definition, the quotations have an *educational* and *referential* purpose that remains implicit within their use. And, by sheer volume, their presence is the most notable aspect of the dictionary. A given page of the 1775, second edition of the text, defines 17 words using 52 quotations. The typographical imbalance between the definitions and the quotations, which overwhelm the page, is striking, even while this is a fact that should come as no surprise to users of the OED, the spiritual successor to Johnson's Dictionary.

This project, therefore, seeks to answer three questions. First, who is cited in what contexts in the *Dictionary*? Here, a quantitative methodology should allow for unprecedented access to the fine-grained details of the text. Second, if Johnson's *Dictionary* were rearranged to group articles connected by shared quotations together, what patterns of relationship emerge? And finally, how does Johnson use his quotations to reflect back on the works that he cites?

Analysis

In order to answer these questions, this project begins with an html marked-up copy of Johnson's 1755 *Dictionary*. Parsing this text using regular expressions, I was able to compile a table of entries for the 42,400 words identified in the second edition of the Dictionary and the accompanying 115,354 unique quotations. Following Blanchard and Olsen, the most straightforward visualization of the *Dictionary* reordered by citations would be a network, where each word is a node, and each edge is a shared citation. In the *Dictionary*, however, the sheer number of terms and quotations renders any naïve network graph unusably complex. Johnson's own text, however, can be used to simplify the connections. For each word that he defines, Johnson also tags it as a part of speech (noun, verb, adjective and adverb). By reducing the articles to these parts of speech, and by consolidating the quotations within these groups, I am able to not only identify large-scale organizational patterns in the *Dictionary* but also uncover the different patterns of citation that lie behind these structures. For example, the works that Johnson cites most distinctively in his definitions of nouns are drawn from a variety of contemporaneous sources: from Dryden and Sidney, who are literary authors, to Boyle (a chemist) and Daniel (a historian) (Table 1).

Author	Work(s)	Number of Quotations
Dryden	Dramatick Poesy; Virgil's Georgics; Annus Mirabilis	480
South	Sermons	162
Hooker	Sermons	134
Sidney	Arcadia	85
Boyle	Colours; Chymical Principles	75
Atterbury	Sermons	64
Tillotson	Sermons	54

Author	Work(s)	Number of Quotations
Bentley	Sermons	49
Taylor	Rules for Holy Living; Guide to Devotions	42
Hammond	Fundamentals; Practical Catechism	34
Broom	Notes on Pope's Odyssey	22
Spratt	Sermons	22
Daniel	Civil War	19
Allestree	Government of the Tongue	18

Table 1: Most distinctive authors in the definitions of nouns

By contrast, in his definitions of verbs, Johnson employs predominately Shakespearean tragedy (Macbeth, King Lear, Hamlet, etc.) as well as biblical sources (Genesis, Ecclesiasticus and Job) (Table 2). These patterns indicate that Johnson's use of citation is not random: there is a clear logic to his choices of quotations for different parts of speech that reflects an implicit theory of the relationship between word meaning and knowledge creation. Here, science is the locus of objects, while tragedy is the locus of action. This not only reveals the ways in which the dictionary organizes language according to an implicit theory of textual meaning, but also how textual meaning is reflexively assigned by the *Dictionary*.

Author	Work(s)	Number of Quotations
Shakespeare	Macbeth; King Lear; Coriolanus; Othello; Hamlet; Henry IV	3716
Locke	Education; Understanding	838
Decay	Piety	146
Knolles	History of the Turks	144
Genesis		102
Ecclesiasticus		99
Sidney	Arcadia	84
Job		82
Rowe	Jane Shore; Royal Convert; Ambitious Stepmother	75
Deuteronomy		67
Addison	Cato; Ovid; Spectator	56
Acts		56
Maccabees		51
Ezekiel		50
Jeremiah		50
Psalms		50
Samson		49
Smith	Phaedrus Hippolytus	48
Proverbs		48
Philips	Briton	47

Table 2: Most distinctive authors in the definition of verbs

The parts of speech reveal an implicit structure to Johnson's patterns of quotation. Their simplification of the text also allows me to visualize this structure as a meaningful network (Figure 1). In this graph, words are connected by shared authors; however the authors are limited to the group of most distinctive authors for each part of speech group (nouns, adjectives, verbs, adverbs). Similarly, the colors correspond to these groups as well: blue points are adverbs, orange points are adjectives; purple, nouns; and green, verbs. In addition to the four-quadrant structure that corresponds to the four parts of speech, there are other macro-scale phenomena in these graphs as well. Note how nouns are connected to adjectives (as are some verbs), but verbs and adverbs are the least like each other in shared quotations. More importantly, a comparison between this network, and a word embedding analysis of a contemporaneous corpus (the Eighteenth Century Collections Online corpus) using the GloVe algorithm reveals that the association Johnson draws between various terms using quotations are partially reflected in word associations through usage across the century, suggesting that Johnson both drew upon, and influenced, the language use of his time.

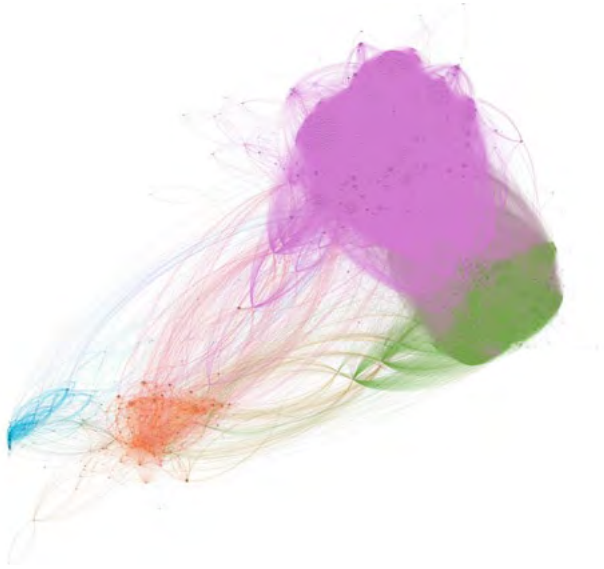


Figure 1: Network of Johnson's Dictionary limited to authors distinctive of parts of speech

At a much smaller scale, the conceptual patterns that bind together definitions become more apparent. A group of verbs near the edge of the graph captures an overarching thematic set of concerns (Figure 2). "Despise," "estrange," "oppress" and "devour" suggest both an affective, as well as thematic unity. There is no reason for these groups to cluster together, unless Johnson sought specifically to create these conceptual groupings. In such cases, the association of words and texts is mutual: just as Shakespeare's Macbeth provides an illustration for the act of despising, so too does the association of despise (or oppress, estrangement) with Macbeth color our reading of the play. As became apparent in the distinctive words that he uses for quotations, Johnson encodes interpretive theories into the apparently benign act of furnishing illustrative quotations.

Figure 2: Detail of "negative affect" cluster of verbs in part of speech network

Through a computational analysis, the underlying structure of the dictionary reveals the doubled work of lexicography. Quotations are both empirical proofs of contingent meaning, and, in turn, receive meaning themselves through accretion: slow layering of the conceptual unities between the words that they are used to define. The complexities of the dictionary reveal a set of organizing principles embedded in the quotations: this project allows us to extract these patterns and reveal their fundamental influence on the development of, not just lexicography, but language itself across that last three hundred years.

De la teoría a la práctica: Visualización digital de las comunidades en la frontera México-Estados Unidos

Maira E. Álvarez
mealvarez@uh.edu
University of Houston, United States of America

Sylvia A. Fernández
sferna109@gmail.com
University of Houston, United States of America

Las prácticas de humanidades digitales y la incorporación de archivos para analizar otras historias y saberes conllevan a lo que Roopika Risam describe como un entrecruce de la teoría con la práctica desde una perspectiva poscolonial que cuestiona la representación desde la producción del saber (2017). De acuerdo a Risam, "postcolonial digital humanities offers a language to ask of digital humanities important questions such as who is speaking, who is being spoken of, who is spoken for, which languages are being used, and what assumptions subtend its productions, distribution, and consumption" (2017). Esto es importante ya que el contenido de archivos visualiza otras historias que al incorporarse a plataformas digitales logran mayor alcance y difusión del material. Sin embargo, el acceso a los archivos varía según su estado ya que algunos pertenecen a instituciones gubernamentales, privadas, u organizaciones no lucrativas lo que afecta su disponibilidad y acceso en línea. Por otra parte, el estado físico del material también puede llegar a limitar el acceso dado a que su mantenimiento, por lo general, depende de su relevancia para la cultura dominante, el apoyo y la distribución de fondos.

Con una formación de Research Fellows en el programa de Recuperación del Legado Escrito Hispano de los Estados Unidos se trabajó con una base de datos de más de mil periódicos recuperados. Este material dio inicio a un corpus de periódicos de los siglos XIX y XX, que no solamente fueron publicados en el lado estadounidense, sino además en los estados del norte de México (Ver imagen 1). Como fronterizas, nuestra afinidad con las comunidades fronterizas nos llevó a seleccionar un grupo de periódicos que documentan la frontera México-Estados Unidos por medio de un legado escrito el cual documenta historias personales, locales, nacionales e internacionales de estas regiones. La necesidad de producir otras historias alternas a la oficial, sobre todo que representen la frontera desde sus propias comunidades, conlleva a que archivos de periódicos sean visualizados en una plataforma digital. Estas prácticas rompen con las formas tradicionales del archivo, transformándolo de un "static repository [to] an active site of knowledge production... [that re-]interpret, and even shapes knowledge from the ground [up]" (Cotera, 2015). Lo que incita a la creación de proyectos digitales desde una conciencia crítica que visualice otras historias desde las comunidades.

Borderlands Archives Cartography (BAC) surge como una resistencia a discursos que imponen estereotipos hacia los inmigrantes, mexicanos, Latinos, entre otros, ya que estos repercuten en comunidades fronterizas volviéndose forzosamente parte de su identidad. Estos discursos son problemáticos debido a que tienen un efecto a nivel local, nacional y global. Al mismo tiempo invisibiliza perspectivas como las nuestras en las que la frontera se entiende como un espacio donde coexisten diversas culturas bajo el control transnacional de hegemonías políticas, económicas y sociales; así como, un espacio donde las regiones se influyen entre sí, pero mantienen sus propias identidades.

De tal forma, el objetivo de BAC es dar visibilidad a periódicos de la frontera del siglo XIX y XX que reflejan historias de múltiples voces. BAC localiza, digitaliza, facilita y compila archivos periodísticos de ambos lados de la frontera antes y después del establecimiento de la actual línea divisoria. Con el uso de CARTO, un software de Sistema de Información Geográfica (SIG) se visualiza la ubicación de los periódicos a través de un mapa digital. Por lo tanto, el proyecto por medio de las humanidades digitales amplía las nociones de la frontera, metodologías, análisis de data y uso de archivo que conlleva a modos de investigación interdisciplinaria.

Se utiliza un mapa digital para visualizar la ubicación geográfica de los periódicos ya que "movement, manipulability, and specificity of the dynamic maps give us a glimpse of what deep contingency might look like over time. By allowing us to see space and time at a distance, in relatively abstract ways, the maps show us dissolving and crystallizing patterns otherwise invisible in rows of numbers or static maps based on the same data" (Ayers, 2010). El mapa digital permite añadir otras capas para cuestionar los discursos hegemónicos impuestos en los espacios geopolíticos. Por ejemplo, la integración de la data de periódicos fronterizos a un mapa digital desestabiliza la noción de frontera como una división reciente y estática además deconstruye los discursos que presentan y representan este espacio como una amenaza; ya que estas comunidades han practicado la producción literaria y auto-documentación continuamente por más de 200 años.

La visualización del contenido de este archivo digital de periódicos fronterizos desafía la perspectiva colonialista e imperialista de lo que es la frontera entre México y Estados Unidos. Por ejemplo, los periódicos disponibles, ya sea en formato digital, en microfilm, o en estado físico documentan las interacciones culturales que dieron lugar a nuevas identidades como resultado de la pérdida de territorio, inmigraciones, exilio, desterritorialización, flujos transfronterizos y dinámicas transnacionales. Por otra parte, como menciona Nicolás Kanellos, los periódicos documentan la solidaridad entre los individuos y residentes para proteger sus derechos mediante la lucha contra la segregación y la discriminación, especialmente después de la cesión de una parte del territorio mexicano a los Estados Unidos en 1848 (Kanellos and Martell, 2000).

La integración de estos archivos en plataformas digitales revela un continuum en el contenido de los periódicos a través de los cambios históricos, políticos, económicos y sociales. Esto permite la teorización de la frontera y sus dinámicas desde una aproximación poscolonial, al igual que la posibilidad de diversos análisis interdisciplinarios como estudios lingüísticos, literarios, históricos, económicos, políticos, de género, culturales, globales, entre otros. Trabajar con archivos transnacionales el proceso de ubicación e integración de estos puede llegar a ser desafiante por lo antes mencionado. Sin embargo, la data obtenida constantemente revela patrones antes invisibles y proyecta nuevos análisis y lecturas del material.

Borderlands Archives Cartography es un compromiso personal con las comunidades fronterizas. Su plataforma disponible a través de Wix, no sólo alberga el mapa digital, sino que también incorpora material relevante a la frontera y contextualiza la data por medio de gráficas utilizando RawGraphs para la visualización y análisis de ésta (Ver imagen 2). Además de ser una base de datos que está disponible al público con la función de crowdsourcing y uso de ella para futuros proyectos.

BAC, como un proyecto transnacional de periódicos fronterizos de los Estados Unidos y México, se resiste y seguirá desafiando los discursos contra las comunidades fronterizas. Por ende, como fronterizas, queremos traer este material a nuestras comunidades a través de las humanidades digitales y con esto deconstruir los discursos en contra de la frontera que han sido persistentes a través de los años.

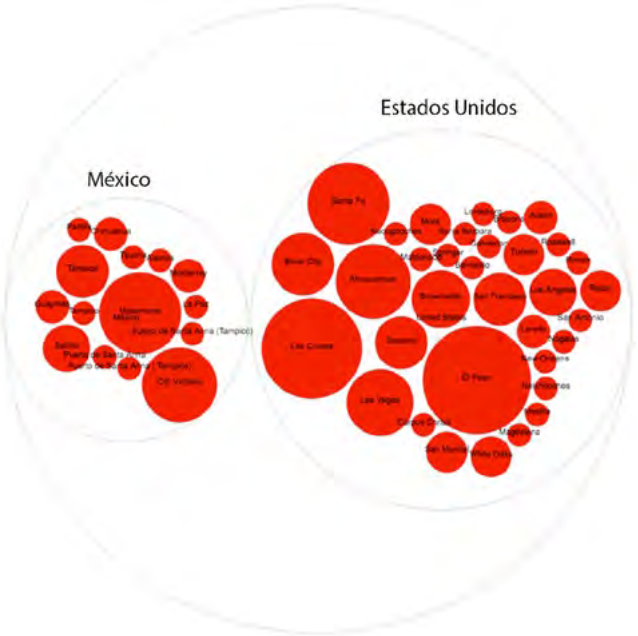


Imagen 1. Distribución de periódicos entre México y Estados Unidos pertenecientes a la data de BAC del 2017.

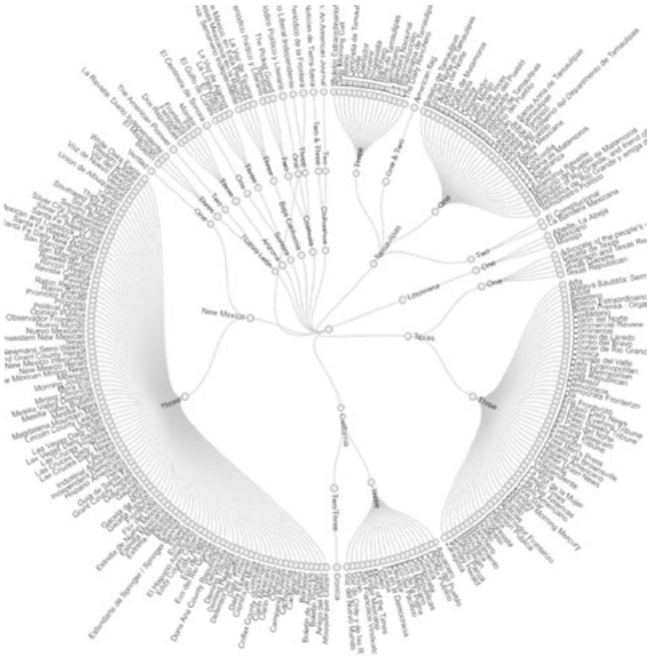


Imagen 2. Visualización del corpus de periódicos de la frontera México-Estados Unidos de acuerdo a la data recaudada en el 2017:
1.) Títulos de los periódicos 2.) Periódo histórico (One, Two, Three) 3.) Estado

References

Ayers, E. (2010). Turning toward place, space, and time. In Bodenhamer D., Corrigan J. and Harris T.M. (eds), *The Spatial Humanities: GIS and the Future of Humanities Scholarship*, pp. 1-12. Indiana University Press.

Cotera, M. (2015). Invisibility is an unnatural disaster: feminist archival praxis after the digital turn. *The South Atlantic Quarterly,* 114 (4): pp. 781-801.

Risam, R (2017). Breaking and building: the case of postcolonial digital humanities. In Singh, J.G. y Kim, D.D. (eds), *The Postcolonial World.* Routledge, pp. 345-362.

Kanellos, N, & Martell, H 2000, *Hispanic Periodicals in the United States, Origins to 1960: A Brief History and Comprehensive Bibliography*, Arte Público Press, Houston.

Comparing human and machine performances in transcribing 18th century handwritten Venetian script

Sofia Ares Oliveira
sofia.oliveiraares@epfl.ch
EPFL, Switzerland

Frederic Kaplan
frederic.kaplan@epfl.ch
EPFL, Switzerland

Introduction

Automatic transcription of handwritten texts has made important progress in the recent years (Sanchez et al., 2014; Sanchez et al., 2015, Sanchez et al., 2016). This increase in performance, essentially due to new architectures combining convolutional neural networks with recurrent neutral networks, opens new avenues for searching in large databases of archival and library records. This paper reports on our recent progress in making million digitized Venetian documents searchable, focusing on a first subset of 18th century fiscal documents from the Venetian State Archives (Condizione di Decima, Quaderni dei Trasporti, Catastici). For this study, about 23'000 image segments containing 55'000 Venetian names of persons and places were manually transcribed by archivists, trained to read such kind of handwritten script, during an annotation phase that lasted 2 years. This annotated dataset was used to train and test a deep learning architecture, with the objective of making the entire set of more than 2 million pages searchable. As described in the following paragraphs, performance levels (about 10% character error rate) are satisfactory for search use cases, which demonstrates that such kinds of approaches are viable at least for this typology of handwritten scripts. This paper compares this level of reading performance with the reading capabilities of Italian-speaking transcribers, preselected with a test based on 100 transcriptions. More than 8500 new human transcriptions were produced, confirming that the amateur transcribers were not as good as the expert. However, on average, the machine outperforms the amateur transcribers in this transcription tasks.

Machine performance

We developed a transcription system based on the combination of convolutional and recurrent neural networks as described in (Shi et al., 2017) for handwritten text (Fig.1a) (The code is implemented in python and is available at https://github.com/solivr/tf-crnn). On the one hand, convolutional neural networks (CNN) capture hierarchical spatial information, with the first layers capturing low level features and later ones capturing high level ones. On the other hand, recurrent neural networks (RNN) capture temporal data, with the ability to grab contextual information within a sequence of arbitrary length. Convolutional recurrent neural networks (CRNN) combine the best of both worlds to handle multi-dimensional data as sequences.

From an input image, the convolutional layers extract a sequence of compact representations which corresponds to the columns of the feature map. They are processed from the left to the right of the image to form a sequence of local image descriptors (Fig.1b).

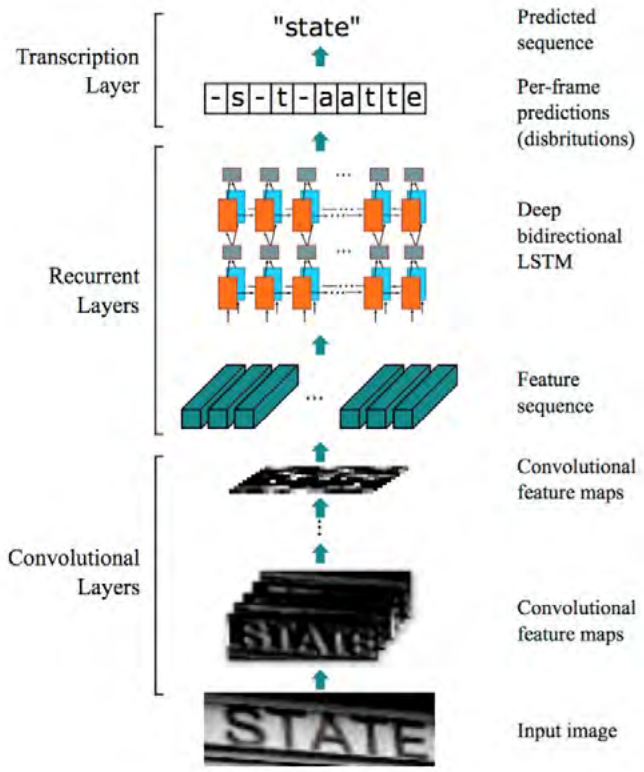


Fig 1 (a) Network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence. (Shi et al., 2017)

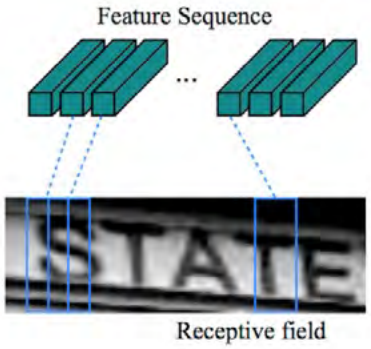


Fig 1 (b) Feature sequence (Shi et al., 2017)

The sequence is then input to the recurrent layers which consist of stacked bidirectional long short-term memory (LSTM) cells (Hochreiter et al., 1997). LSTM cells have the ability to capture long-range dependencies but are directional, and thus only use past contexts. Since in image-based sequences context from both directions are useful and complementary, one forward and one backward LSTM cells are combined to form bidirectional LSTMs which are then stacked to have several recurrent layers. The recurrent network outputs per-frame predictions (probabilities) that need to be converted into a label sequence.

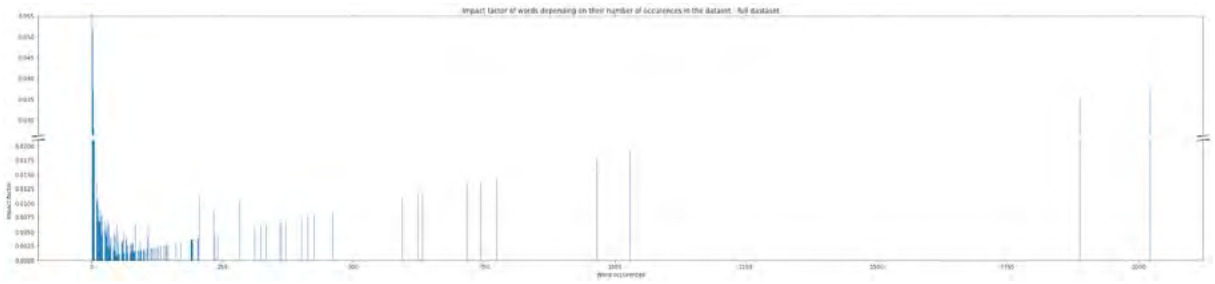
In the transcription layer, the connectionist temporal classification (CTC) (Graves et al., 2006) is used in order to obtain the "sequence with the highest probability conditioned on the per-frame predictions". The sequence label is found by taking the most probable label at each time step and mapping the separated labels to the correct sequence label (see (Graves et al., 2006) to have the detailed explanation on how the repeated and 'blanks' labels are dealt with).

The CRNN was trained on data coming from various types of Venetian handwritten documents. The dataset is composed of image segments of mainly names and places that have been transcribed by archivists in Venice. Image segments are used in order to reflect only the performance of the transcriber system, without introducing possible errors from the segmentation process. Thus, the segmentation step is not part of the proposed experiment. The set was randomly split into training and testing set and the content of the image segments ranges from one to several words (Tab.1).

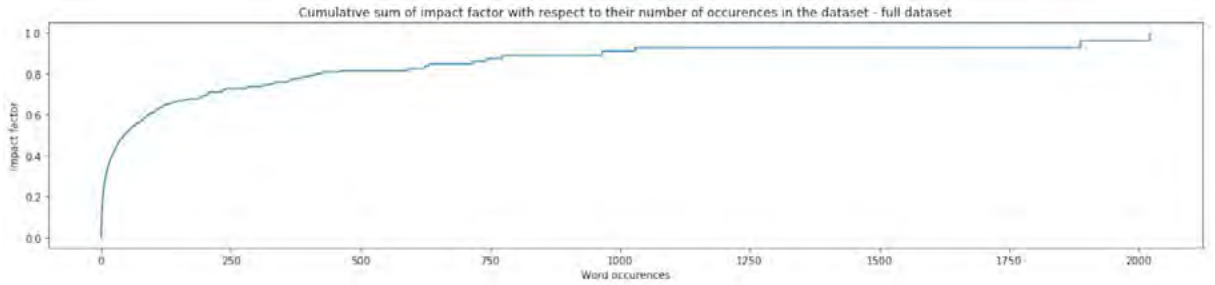
Set	# images segments	# total words	size of vocabulary
Training set	20712	48628	8848
Testing set	2317	5559	2157
Full set	23029	54187	9429

Table 1: Datasets used

We show in Fig.2a and 3a how words are distributed in the dataset. We de ne the vocabulary to be the set of different words. The impact factor IF is a measure of the words' distribution in the dataset and is defined as IF (i) = $^{c(n^{i})}$ *hist*(i; c), with c the vector of counts of each vocabulary word, n the total number of words, *hist* the histogram operation and *hist*(i; c) the number of vocabulary words that occur i times. The left part of these plots shows that most of the words do not appear commonly but a few are very present in the dataset as it can be seen on the right of the figures (those are mainly prepositions such as `di', `de', `in', etc). The cumulative sums (Fig.2b and Fig.3b) show that common words have limited impact, but also that the system does not suffer from overfitting to the vocabulary since most of the words used for training are 'rare' in the dataset.



(a)

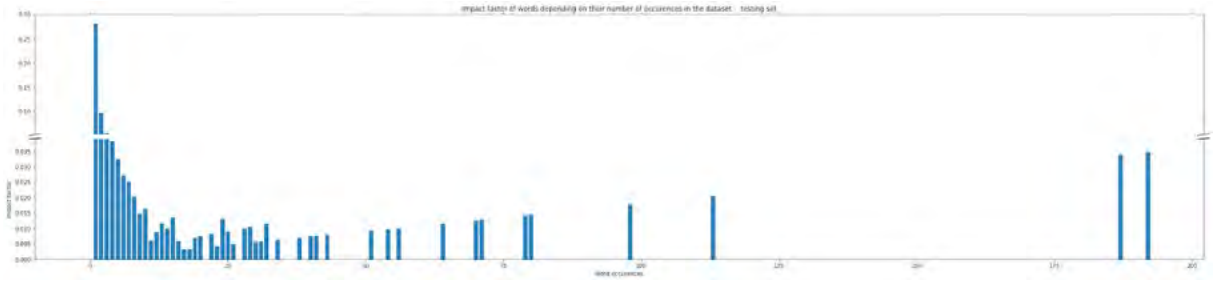


(b)

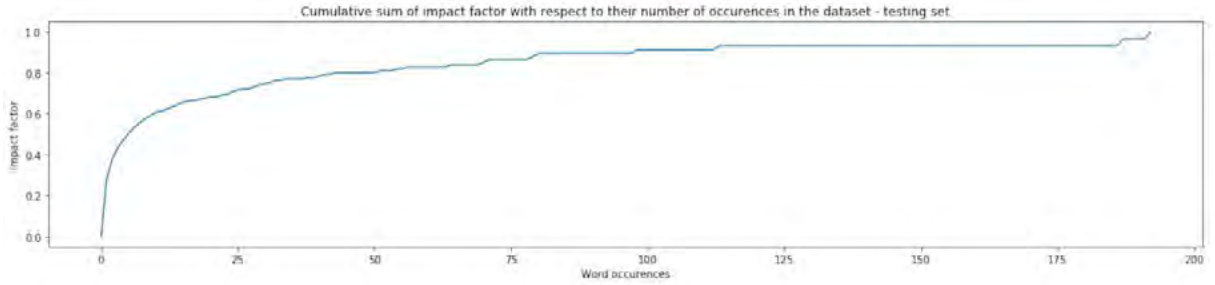
Figure 2: Word distribution and impact factor in full dataset. We observe that 70% of the dataset is represented by words appearing less than 250 times (out of 54187 words)

To evaluate the performance of the system we use the Character Error Rate (CER) measure on the test set defined as CER = (i + s + d)=n with i, s, d, n the number of character insertions, substitutions, deletions and total characters respectively. The numerical results are shown in Tab. 2. Several experiments were performed using different sets of characters (called 'Alphabet' hereafter) and resulted in one model per Alphabet. A few randomly selected examples can be seen in Appendix A.

On this dataset, our transcription system is below 10% CER, which is sufficiently good to be able to search for entities in documents using regular expressions and fuzzy matching. Moreover, we believe this performance is better than the human average and in order to verify our hypothesis, we conducted an experiment described in the following section.



(a)



(b)

Figure 3: Word distribution and impact factor in the testing dataset

Alphabet	Set of characters	# image segments	CER
Capital-lowercase-symbols	A-Za-z'.,: -=	24035	0.089
Capitals-lowercase-digits-symbols	A-Za-z0-9'.,:; =()[]/	96198	0.045
Digits	0-9	72326	0.013

Table 2: The Character Error Rate (CER) for each Alphabet

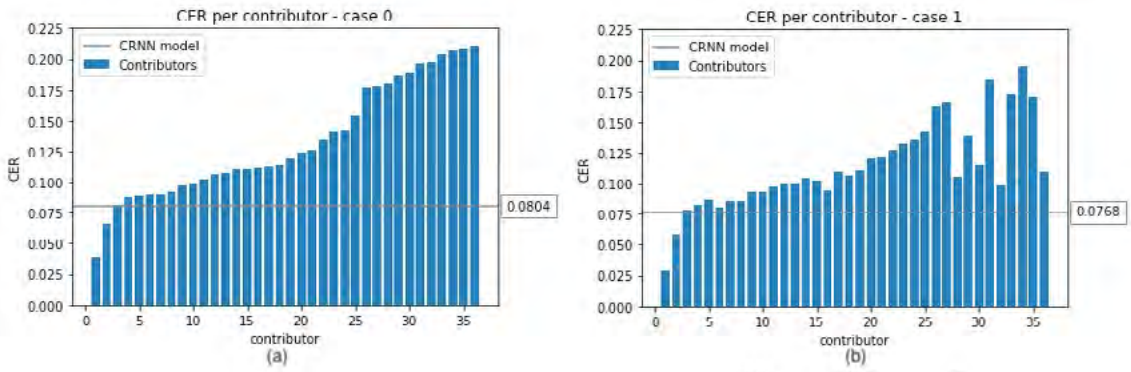
Human performance

In order to quantify the human average error rates on our dataset, we conducted an experiment on Crowdflower's platform, where Italian speaking persons were paid to transcribe image segments of the testing set (see examples in App. A). The contributors had to decipher a few units before being able to start the survey and during the experiment some of their transcriptions were evaluated. There were 103 evaluation questions that allowed to separate low accuracy contributors' answers from reliable ones. Each image segment was transcribed at least three times, and in total 11'727 units were transcribed. Only the answers of contributors maintaining at least 60% accuracy throughout the experiment and who transcribed at least 50 units were taken into account for the analysis. This resulted in a total of 8'674 valid transcriptions to analyze. The number of transcriptions (judgments) per contributor and its location can be seen in Fig.6.

We compare the performance of the system and the amateur transcribers in Tab.3 and Fig.4,5 (onesample t-test, $p < 0{:}005$). It is clear from the graphs that the CRNN system has a better CER and WER than the human average on this dataset, and only a few contributors have lower or comparable performance to the system but is not yet as good as the expert. It is interesting to notice that the performance of the best amateur transcriber almost doubles when capital letters and punctuation are not considered (case 3) whereas the CRNN makes little improvement. Indeed, although the system has inferred some sort of weak language model, we have seen it producing unlikely transcriptions whereas the best contributor uses its knowledge of Italian proper nouns to deduce the correct transcription when some characters are di cult to read. Thus, the system's CER and WER could be reduced by using a lexicon-based transcription, where the output of the neural network would be compared to a dictionary and the closest element would be chosen.

Case		CER		WER	
		CRNN	contributors	CRNN	contributors
0	: No modifications (Fig.4a)	0.0804	0.1328	-	-
1	: Capital letters replaced by lowercase (Fig.4b)	0.0768	0.1137	-	-
2	: All punctuation removed (Fig.4c, 5a)	0.0766	0.1241	0.2709	0.4318
3	: Combination of Case 1 and Case 2 (Fig.4d, 5b)	0.0718	0.1047	0.2551	0.3507

Table 3: Comparison of Character Error Rates (CER) and Word Error Rates (WER) considering different formatting cases of the transcriptions for our system and the mean of the contributors (ground-truth and predictions are formatted in the same way)



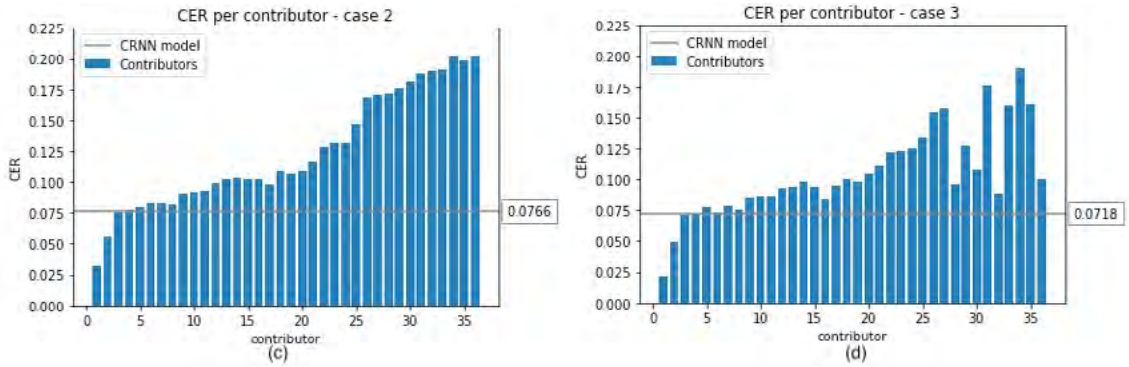


Figure 4: Character Error Rate per contributor for different cases (refer to Tab.3).

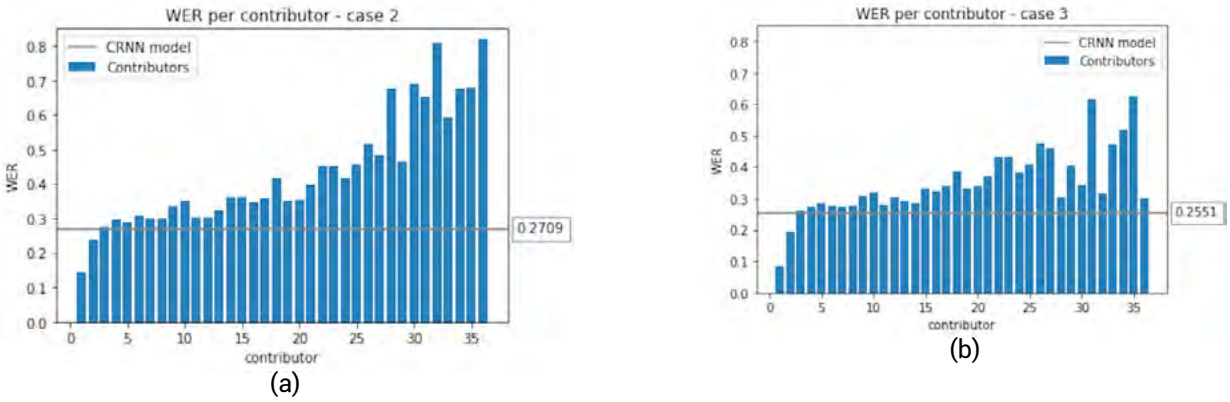


Figure 5: Word Error Rate per contributor for different cases (refer to Tab.3).

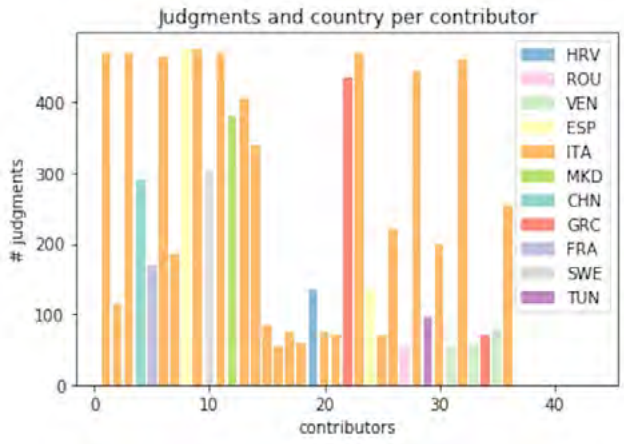


Figure 6: Number of judgements made (image segments transcriptions) by each contributor and its location. The contributors' ordering is the same as Fig.4a (by increasing CER)

Perspectives

The developed system shows promising results to make possible the textual search on digitized handwritten documents. These results open up new prospects for massive indexing, analyze and study of historical documents. We showed that the system had lower Character and Word Error Rates than the human average, thus being sufficiently reliable to use for searching purposes. Further work will focus on improving the architecture of the model, especially the CNN. We will also explore the possibility of lexicon- or rule-based transcription to decrease error rates.

More generally, it seems that the automatic transcription is currently passing a threshold in terms of performance, now giving better results than good amateur transcribers. Future research will show how far this level of performance depends on the expert initial training set or whether, after some exposition with dozens of different scripts, the automatic transcriber may be able to generalize by itself without further specific training.

Appendix A : Transcription examples

References

A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *Proceedings of the 23rd international conference on Machine learning*, pp. 369-376, ACM

S. Hochreiter and J. Schmidhuber (1997) Long short-term memory, *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

J. A. Sanchez, V. Romero, A. H. Toselli, and E. Vidal (2104). Icfhr2014 competition on handwritten text recognition on transcriptorium datasets (htrts),

Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pp. 785-790, IEEE
J. A. Sanchez, A. H. Toselli, V. Romero, and E. Vidal (2015). Icdar 2015 competition htrts: Hand-written text recognition on the transcriptorium dataset, *Document Analysis and Recognition (ICDAR)*, 2015 13th International Conference on, pp. 1166-1170, IEEE
J. A. Sanchez, V. Romero, A. H. Toselli, and E. Vidal, (2016). Icfhr2016 competition on handwritten text recognition on the read dataset, *Frontiers in Handwriting Recognition (ICFHR)*, 2016 15th International Conference on, pp. 630-635, IEEE
B. Shi, X. Bai, and C. Yao (2017) An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298-2304

Metadata Challenges to Discoverability in Children's Picture Book Publishing: The Diverse BookFinder Intervention

Kathi Inman Berens
kathiberens@gmail.com
Portland State University, United States of America

Christina Bell
cbell@bates.edu
Bates College, United States of America

37% of the United States population is non-white, but 90% of the books published for children during the last twenty-one years contain no multicultural content. This discrepancy has been called "The Diversity Gap" (Erlick, 2015) and, more starkly, the "Apartheid of Children's Literature" (Myers, 2014). Based on data gathered since 1985 by the Cooperative Children's Book Center, the representation gap has barely shifted over thirty years, with books by and about non-white people hovering between 10-14% of total children's book production (CCBC 2017). Panels and initiatives about diversity in book publishing have not actually produced more books by and about non-white people. Book discoverability is thus a significant challenge to parents, librarians, and teachers seeking picture books depicting the lives of non-white children.

As it's currently practiced in the North American book industry, "diversity" usually tallies "how many" rather than delving into the lived experience of non-white people. The Diverse BookFinder [DBF], a database and metadata project sponsored by Bates College and funded by the Institute for Museum and Library Services [IMLS], asks: how can metadata help to tackle these entangled problems? Can we build a network of information about children's picture books that trains users to search for and discover books using complex concepts related to their own communities rather than race or ethnicity as the sole marker?

Our long paper:

1. Surveys the problem of whiteness as the de facto point-of-view in children's books, and the populist social media movements that resist this phenomenon;
2. Examines the limitations of current metadata in k-3 books about non-white children;
3. Presents Diverse BookFinder as a strategic disruption of current metadata practices;
4. Conveys the pedagogical value of Diverse BookFinder in academic and public settings.

Readers have created online, massively participatory movements to prompt the predominantly white book publishing industry to publish more books by and about non-white people (Low, "Diversity Baseline Survey," 2016). #WeNeedDiverseBooks, #1000BlackGirlBooks, and #OwnVoices originated in Twitter hashtags then converted their social media capital (likes, shares, reposts, followers) into an array of recommendation services: published anthologies, book finder apps, a short story contest, even a granting agency. Populist interventions are welcome and useful, but they are insufficient to remedy the problem of classifying existing books using metadata that reinscribe white privilege.

One intervention in these human and machinic systems is a human-curated and -coded catalog, Diverse BookFinder (https://diversebookfinder.org/). Metadata and recommendation systems are not neutral. They operationalize cultural assumptions that the creators may not have intended or even be aware of. Critical code studies, the scholarship of platforms and software that examines computer source code hermeneutically, has charted useful ground in exploring how metaphors of containment and layers, for example, rationalize logics of racial exclusion (McPherson 2012). Many metadata schemas for books relate back to the physical structures libraries and classrooms use to organize books for readers. These systems create fixed and singular ways of relating items that construct contextualized exclusion (Drabinski 2013). The common cataloging systems used in the United States, including the Library of Congress and Dewey Decimal System, are centered in whiteness and maleness and reinforce the otherness of diverse titles. The separation of topics on women and gender (including queerness) from broader topics such as literature or history, for example, reinforce the notion that women and queers secondarily contribute to history and literature. This segregation repeats in the various forms of metadata where difference is replicated and continually defined by whiteness.

Exclusion is not specific to physical location like a library. Algorithmic "overfitting" is the phenomenon where recommendations are culled from a narrow spectrum

of a user's interests. Overfitting "can occur when a user is trying to be helpful by providing explicit feedback only about the content s/he strongly likes. This leads to the creation of a very specific model that knows the exact user preferences but is unable to detect any other types of interesting items since the user has not shown any interest in it" (Kunaver & Poztlz, 2017, 156). Under this system, the typical person would have difficulty in finding diverse titles online if those books did not already match to their past search behavior. Inconsistencies in the application of metadata compound this problem.

Metadata sometimes contain errors that hide or misrepresent the books, or don't classify the types of information that would be most relevant to communities seeking books about non-white children. Such books are "mirrors and windows," that reflect back or "mirror" one's own lived experience in the faces, bodies, customs and cultural milieu depicted in the book, and open "windows" onto new cultures different from one's own (Bishop 1990). Such books develop myriad literacies beyond reading comprehension, including conflict resolution, tolerance for the unfamiliar, and awareness of cultures beyond one's own. When used in the classroom as an intervention toward intergroup contact, diverse picture books can foster intercultural understanding among children (Aronson, et al. 2016). Unfortunately, existing metadata does not account for the intricacies of diverse titles and so these books remain difficult to comprehensively identify or locate. Hand-coding is a remedy to discoverability problem.

Without controlled language, books are simply not findable. There is no eschewing metadata; there is only writing better metadata, and theorizing the best practices that writers of descriptive metadata should follow in order not to reinscribe racist stereotypes and cultural marginalization. The purpose of this vocabulary is not to undo prior standards--which are each problematic in many ways-- but to contribute to the larger representation of diverse books and fill in the information holes. Systematic SEO work is underway to add language as it is used by the communities represented; for example, a user may enter "Boricua" into DBF and yield results about Puerto Rican characters. The goal is to write metadata that reflects the lived experience of people the books depict.

When books are entered into the Diverse BookFinder, they go through a multi-step, hand coding process to compile the metadata commonly missing from other sources. Book characters are coded for racial and/or ethnic identity, gender, setting, with additional tags such as tribal nation, immigration status, or religion where applicable. Most books fall into one of nine categories that capture the message conveyed by these books. The categories are: Beautiful Life, Oppression, Cross-group, Biography, Race/Culture Concepts, Folklore, Incidental [ensemble or background characters of color], and Informational [factual content unrelated to race or culture]. These categories arose from an application of grounded theory, and created by a rigorous analysis of commonalities in picture book stories. This analysis shows that the concept of Beautiful Life, stories about a particular racial or cultural group experience, dominates diverse book publishing, but such a message is commonly unavailable in existing metadata outside of DBF. African Americans are most likely to be depicted in situations of oppression; Native Americans are disproportionately represented in "folklore," and Hispanic and Latinx people are underrepresented generally in picture books. DBF has engaged students at all levels, at several institutions in thinking about representation and participating in research to better understand the role of picture books in children's development.

It's pedagogically valuable to give students a direct search experience of how imprecise book metadata impedes book discoverability. In a lab exercise designed by Bell and implemented by Inman Berens, master's students retrieved book metadata across three venues for two books in the DBF database. Those venues were: publisher website, retailer, library catalog. Students discovered significant errors in metadata, and notable variability from venue to venue. Ensuing class discussion allowed students to trace the interoperation of human classification errors and legacy systems such as Library of Congress Subject Headings with machinic processes. The students then reviewed a Library of Congress copyright form submitted to the LoC by our student-run trade press (Ooligan Press), and discovered ambiguities in the Library form's language that prompted misclassification of our press's just-released young adult novel. This exercise drove home that automated processes are framed by human judgment.

The Diverse BookFinder is unique precisely for the level of human labor that goes into the data entry and book coding process. The inconsistencies and inaccuracies in book metadata and the additional information added to each book's metadata could not be done by machine. This process serves to bridge the gaps in metadata, help users identify many more diverse titles than the average search, and provides new insights into what stories dominate in picture books. As public scholarship, this project seeks to move the diverse books discussion beyond a focus simply on the lack of numbers to also consider content and impact by translating research findings so that they are accessible and useful.

References

Aronson, K. M., Stefanile C., Matera C., Nerini A., Grisolaghi J., Romani G., ...Brown R. (2016). Telling tales in school: extended contact interventions in the classroom. *Journal of Applied Social Psychology,* 46, 229–241. doi: 10.1111/jasp.12358

Bishop, R.S. (1990). Windows and mirrors: children's books and parallel cultures. In M. Arwell and A. Klein (Eds.), *California State University San Bernardino*

Reading Conference: 14th Annual Conference Proceedings (pp.11-20). San Bernadino, CA: CSUSB Reading Conference. Retrieved from https://files.eric.ed.gov/fulltext/ED337744.pdf#page=11

Cooperative Children's Book Center (CCBC). (2017). Children's Books By and About People of Color Published in the United States. Retrieved from (http://ccbc.education.wisc.edu/books/pcstats.asp).

Drabinski, E. (2013). Queering the catalog: queer theory and the politics of correction. *Library Quarterly: Information, Community, Policy*, 83(2), 94-111. doi.org/10.1086/669547

Erlick, H. (2015, March 5). The diversity gap in children's publishing, 2015 (blog post). *Lee and Low Books: The Open Book Blog*. Retrived from http://blog.leeandlow.com/2015/03/05/the-diversity-gap-in-childrens-publishing-2015/

Jackson, Chris. (2017). Diversity in Book Publishing Doesn't Exist -- But Here's How It Can (blog post). Retrieved from http://lithub.com/diversity-in-publishing-doesnt-exist-but-heres-how-it-can/.

Kunaver & Poztlz (2017). Diversity in recommender systems -- A survey. *Knowledge-Based Systems* 123 (154-162). http://dx.doi.org/10.1016/j.knosys.2017.02.009

Low, Jason T.(2016, January 26). Where is the Diversity in Publishing? The 2015 Diversity Baseline Survey Results (blog post). *Lee and Low Books: The Open Book Blog.* Retrieved from http://blog.leeandlow.com/2016/01/26/where-is-the-diversity-in-publishing-the-2015-diversity-baseline-survey-results/

McPherson, Tara. (2012). Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation. *Debates in Digital Humanities*, ed. Matthew K. Gold. Retrieved from http://dhdebates.gc.cuny.edu/debates/text/29

Myers, Christopher. (2014, March 15). The Apartheid of Children's Literature. *New York Times*. https://www.nytimes.com/2014/03/16/opinion/sunday/the-apartheid-of-childrens-literature.html?_r=0

The Idea of a University in a Digital Age: Digital Humanities as a Bridge to the Future University

David M. Berry
d.m.berry@sussex.ac.uk
University of Sussex, United Kingdom

In this paper I examine the historical constellations of concepts that have made up the idea of an idea of a university. The aim is to provide a tentative genealogy that maps the changes in the sets of concepts and affects that were bound together at particular historical junctures to declare an idea of a university. By definition this idea changes over time, through the effects of social change, political contestation, or other social forces operating on the university. I explore the specific reasons for why an idea of a university has been thought historically to be useful, and why, perhaps, we should revisit the idea of an idea of a university in light of rapid changes taking place following new pressures on the university coming from both digital technologies and new social forces. Indeed, the idea of a university has served as an important way of discussing an institution that comes in a multiplicity of shapes and sizes, with differing national traditions, and different ways of understanding what a university is for. But one constant that remains important in the outlining of an idea of a university is that the idea of an idea of a university is a compass for theory and practice in the university itself, and often in the wider society. In this sense, the idea of a university, comes to stand for a method of scoping the function and direction of the institution which we call a university, and most importantly provides a framework for determining what a university should and should not do. The idea of a university is, then, a compass for decision-making, it is a signature, a distinctive pattern.

The idea of a university is in this sense a kind of boundary object, which allows distinctions to be made between those institutions which are, and those which are not universities. The notion of a boundary object is useful because it acknowledges that universities are heterogenous and requires cooperation between multiple actors in order to be successful (Star and Griesemer 1989). Here, I want to think about the "idea of a university" as something like a boundary object, that is as an ideal type, as an object that is abstracted from all domains, sometimes fairly vague, but that nonetheless offers a "good enough" road map for all parties (Bowker et al 2016: 191). The problem of adapting a university to changing historical and social forces was often viewed intellectually as "finding the correct 'idea' of a university" (Rothblatt 1997: 33). This is the notion that one needs to find a "pattern of orientation", that is, a conception is required to relate an actor, individual or collective to a manifold of objects in their situation of action, so that through internalisation for an individual, or through institutionalisation of a group – there is an organisation of the system of action.

The university as a form has never been frozen in aspic, it has continually adapted, grown, shrunk, expanded and shifted for all of its history. This draws attention to the way in which, at certain points in history, it was considered important that one should have an "idea" in mind in relation to the institutions of higher learning. By "idea" I mean a sense of what has been described variously by a number of thinkers as the "mission", "end", "soul", "aims", "principles", "models" and "ideals" of an institution of advanced or higher learning. This is a debate that has gone on, revived in every generation, concerning the role and

purpose of a university and the education it provides. We should note that the idea of a university paradoxically changes over time, through the effects of political contestation, social change or other social forces operating on the university. However, the notion of an idea of a university, as an institution requiring an essential core which is used to guide its operation and provide its raison d'être has remained in place until quite recently. John Henry Newman, of course, wrote perhaps the most famous idea of the university in 1859 when he argued, "a University.... is a place of *teaching* universal *knowledge*" (Newman 1996: 3). He maintained that the university had an essential function in the conservation of knowledge and ideas and their transmission to an elite body of largely undergraduates, a model he drew from Oxford. Similarly, Abraham Flexner writing in the 1930s with John Hopkins University in mind, argued that the university is "an institution consciously devoted to the pursuit of knowledge, the solution of problems, the critical appreciation of achievement, and the training of [students] at a very high level" (Flexner 1968: 42). But as the varieties of universities began to grow and their internal complexity multiplied, it became seemingly more difficult to identify an essential idea of a university.

By the late 1920s, for example, Robert Maynard Hutchins was remarking that the modern university was a set of schools and departments held together by a central heating system. Later in the 1960s, Clark Kerr described the modern university as "a series of individual faculty entrepreneurs held together by a common grievance" over car parking (Kerr 2001: 15). And today it does sometimes seem like the 21st century university is a set of schools and departments held together by a shared grievance over the IT support. However, in this talk, I will question Kerr's dismissal of the university as an idea whose time is over, contesting his notion of the university as a *multiversity*. Kerr believed that by force of circumstances, if not by choice, "administration everywhere... becomes a more dominant feature of the university" (Kerr 2001: 21).

I will bring these strands together to think about the challenges we face today in what Cathy Davidson (2017) has called "The New Education". Today we live in a digital age, and indeed around us we see the implications of digital exosomatization in all aspects of our lives and societies (see Stiegler 2016). From the pressures on the economy and work through new forms of automation, difficulties with our ability to concentrate from new techniques of attention control and manipulation, and effects on our sense of identity, our societies and our politics through the use of social media and Big Data, the digital presents new challenges for the 21st century. It is, therefore, not surprising that digital transformations should come to the university. Although, it is interesting to note how long digital forms took to effect change in research and teaching, even as university administration had been computerised for quite a while beforehand. The digital revolution, if we can call it that, is notable for confounding the critics, particularly internal to the university, who doubted that digital technology would have much of an effect on the structures, practices, processes and activities of the university. It seems that computation and the digital alone was not, in and of itself, enough to provide the step-change in the university, and we had to await the arrival of a number of different technologies, including radio networking, digital archives and tools, pocket computers and social media, combined with a number of corresponding social forces, such as digital homophily, a new political economy of data, and a generation entelechy that has never bought a paper newspaper, used a vinyl record or a CD, and finds the scholarly concentration required in the historical form of close reading arduous and unfamiliar. This is where the importance of the digital humanities as a field that can act as a bridge between past and future ideas of a university and could potentially contribute to a new idea of a university for a digital age –what we might call a contributory infrasomatization (see Berry 2016). In this talk I outline this research project, and the way in which I consider the digital humanities a crucial source of concepts for thinking about the idea of a university today.

References

Berry, D. M. (2016) Infrasomatization, *Stunlaw*, http://stunlaw.blogspot.co.uk/2016/12/infrasomatization.html

Steigler, B. (2016) The New Conflict of the Faculties and Functions: Quasi-Causality and Serendipity in the Anthropocene, *Qui Parle: Critical Humanities and Social Sciences*, Volume 26, Number 1, June 2017, pp. 79-99

Hierarchies Made to Be Broken: The Case of the Frankenstein Bicentennial Variorum Edition

Elisa Beshero-Bondar
ebb8@pitt.edu
University of Pittsburgh at Greensburg, United States of America

Raffaele Viglianti
rviglian@umd.edu
Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

Science fiction has been theorized as a laboratory in which text serves as the medium for experimentation with perspective and epistemology.[1] Yet scientific methods are more practicably applicable to the systematic efforts of textual scholars. Computationally assisted collation demands continual refinements to verify the accuracy of textual data and metadata and challenges a singular view of any documentary edition. Moreover, collation can test hypotheses about change over time, and the output of machine collation can serve as an experiment to identify, quantify, survey, and analyze the data of textual change. Digital collation of science fiction seems to combine the practical with the theoretical in its lab space.

An early form of modern science fiction, the 19th-century novel *Frankenstein* has itself been the subject of digital variorum experiment since the mid-1990s production of the Pennsylvania Electronic Edition (PAEE) by Stuart Curran and Jack Lynch, a daring effort to prioritize the critical apparatus, pulling it from the obscurity of small type at the bottom of printed pages to make it front and center on screen displays.[2] The PAEE challenges us to find new ways to tell the variorum narrative of change over time. Much like Victor Frankenstein's composition of the Creature from multiple bodies, the effort to aggregate the distinct versions of this novel into a variorum might succeed in communicating a multi-dimensional narrative of its own composition and decomposition, inviting the reader to evaluate its successive stages just as the reader is invited to evaluate the three storytellers within the novel.

In the history of preparing digital texts with markup languages, whether in early HTML, SGML, or XML, markup standards tensed between two poles: a) the acknowledgement of a coexistence of multiple hierarchical structures and b) the need to prioritize a single document hierarchy in the interests of machine-readability, while permitting signposts of overlapping or conflicting hierarchies as of secondary importance.[3] In this paper we present a view of texts that emerges from the experience of comparing documents encoded in conflicting ways. Like the makers of the genetic *Faust* edition, we find that multiple encoding structures must co-exist and correlate to achieve a meaningful comparison of editions.[4] Further, hierarchies need to be reconceived in dynamic terms—where are their flex points for conversion from containment structures to *loci* of intersection? In the process of collation, hierarchies must be dismantled and flattened in order for meaningful multiplicity to be represented, and in order for us to understand a dialogic relationship among textual variants. To study variation over time vexes the organizing principle of any singular hierarchy, but hierarchy in the context of collation may nevertheless build a robust architecture that *bridges* distinct encodings rather than isolating them. In this architecture, arches and connecting spans are more viable than monoliths.[5]

This paper addresses the serious issues of collating digital editions made at different times by different editors, and it discusses the bicentennial *Frankenstein* variorum project as a challenging, illustrative case in point. We are preparing a variorum edition of Frankenstein in TEI P5 based on the 1818 and 1831 *Frankenstein* digital texts due to be released in 2018 in celebration of the bicentennial of *Frankenstein*'s first publication. Our collation source documents are adapted from the 1990s encoding of the PAEE (for the 1818 and 1831 editions), and the Shelley-Godwin Archive's diplomatic edition of the manuscript notebooks.[6] We are also newly incorporating a little-known edition of 1823 produced from corrected OCR. Our collation should yield a meta-narrative of how *Frankenstein* changed over time in four versions that passed through multiple editorial hands. It is widely understood that the 1831 edition diverges sharply from the first print edition of 1818, adding new material and changing the relationships of characters. Less known is how the notebook material compares with the print editions, and how much we can identify of the *persistence* of various hands involved in composing, amending, and substantially revising the novel over the three editions. For example, to build on Charlie Robinson's identification of Percy Bysshe Shelley's hand in the notebooks,[7] our co-

1 Jones, Gwyneth, *Deconstructing Starships: Science, Fiction and Reality* (Liverpool UP, 1999) 4.

2 See a representative page at http://knarf.english.upenn.edu/Colv1/f1101.html. Curran, Stuart and Jack Lynch. *Frankenstein; or, the Modern Prometheus*. The Pennsylvania Electronic Edition. Est. 1994. http://knarf.english.upenn.edu/.

3 See for example, P. M. W. Robinson, "The Collation and Textual Criticism of Icelandic Manuscripts" *Literary and Linguistic Computing*, Volume 4, Issue 2, 1 January 1989, 99–105, https://doi.org/10.1093/llc/4.2.99; Dekker, Ronald Haentjens, Dirk van Hulle, Gregor Middell, Vincent Neyt, and Joris van Zundert, "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project" Digital Scholarship in the Humanities.
Volume 30, Issue 3, 1 September 2015, 452–470, https://doi.org/10.1093/llc/fqu007; Eggert, Paul, "The reader-oriented scholarly edition" *Digital Scholarship in the Humanities*, Volume 31, Issue 4, 1 December 2016, 797–810, https://doi.org/10.1093/llc/fqw043; and Holmes, Martin, "Whatever happened to interchange?" *Digital Scholarship in the Humanities*, Volume 32, Issue suppl_1, 1 April 2017, 163–168, https://doi.org/10.1093/llc/fqw048.

4 See Gerrit Brüning, Katrin Henzel, and Dietmar Pravida, "Multiple Encoding in Genetic Editions: The Case of 'Faust'", Journal of the Text Encoding Initiative (4: March 2013).

5 An inspiration for the bridging concept are the visualizations in Haentjens Dekker, Ronald, and David J. Birnbaum, "It's more than just overlap: Text As Graph," Balisage: The Markup Conference 2017, Washington, DC, August 1 - 4, 2017; in *Proceedings of Balisage: The Markup Conference 2017*: https://www.balisage.net/Proceedings/vol19/html/Dekker01/BalisageVol19-Dekker01.html#d11284e1180 . The authors conceptualize an ideal model of texts in a graph structure organized primarily by their semantic sequencing and in which structural features are a matter of descriptive annotation rather than elemental hierarchy.

6 The Shelley-Godwin Archive's edition of the manuscript notebooks of *Frankenstein* builds on decades of intensive scholarly research to create TEI diplomatic encoding: http://shelleygodwinarchive.org/contents/frankenstein/.

7 See Charlie Robinson's Introduction to the Frankenstein Notebooks

llation can reveal how much of Percy's insertions and deletions survive in the later print editions. Our work should permit us to survey when and how the major changes of the 1831 text (for example, to Victor Frankenstein's family members and the compression and reduction of a chapter in part I) occurred. We preserve information about hands, insertions, and deletions in the output collation, to serve as the basis for better quantifying, characterizing, and surveying the contexts of collaboration and revision in textual scholarship.

The three print editions and extant material from three manuscripts are compared in parallel, to indicate the presence of variants in the other texts and to be able to highlight them based on intensity of variance, to be displayed like the highlighted passages in each visible edition of *The Origin of Species* in Darwin Online.[8] Rather than any edition serving as the lemma or grounds for collation comparison, we hold the collation information in stand-off markup, in its own XML hierarchy. That XML "bridge" expresses relationships among the distinct encodings of diplomatic manuscript markup in which the highest level of hierarchy is a unit leaf of the notebook, with the structural encoding of print editions organized in chapters, letters, and volumes. While the apparently nested structure of these divisions might seem the most meaningful way to model *Frankenstein*, these pose a challenge to textual scholarship in their own right. As Wendell Piez has discussed, *Frankenstein*'s overlapping hierarchies of framing letters and chapters have led to inconsistencies in the novel's print production. Piez deploys a non-hierarchical encoding of the novel on which he constructs an SVG modeling (in ordered XML syntax) of the overlap itself.[9] Our work with collation depends on a similar interdependency of structurally inconsistent encoding.

Our method involves three stages of structural transformation, each of which disrupts the hierarchies of its source documents:

1. Preparing texts for collation with CollateX[10],
2. Collating a new "braided" structure in CollateX XML output, which positions each variant in its own reading witness.
3. Transforming the collation output to survey the extents and kinds of variation, and to build a digital variorum edition.

In the first stage, we adapt the original code from the Shelley-Godwin Archive and from the PA-EE to create new forms of XML to carry predictable markers to assist in alignment. These new, pre-collation editions are resequenced (as when we move marginal annotations from the end of the XML document into their marked places *as they would be read* in the manuscript notebook). They are also differently "chunked" than their source texts, resizing the unit file so that each represents an equivalent portion small enough to collate meaningfully and large enough that each document demonstrably aligns with the others at its start and end points.

Stage two weaves these surrogate editions together and transfers information from tags that we want to preserve for the variorum. Interacting with the angle brackets as patterned strings with Python, we mask several elements from the diplomatic code of the ms notebooks so that they are not processed in terms of comparison but are nevertheless output to preserve their distinct information. In CollateX's informationally-rich XML output, these tags render as flattened text with character entities replacing angle brackets so as not to introduce overlap problems with its critical apparatus. In Stage three, we work delicately with strings that represent flattened composite of preserved tag information and representations of the text, using XSLT string-manipulation functions to construct new files for analysis. We can then study, for example, where the strings associated with Percy Shelley are repeated in the later editions, and how many were preserved by 1831. We also build a scaffolding in stand-off markup for the digital variorum that bridges multiple editions, as modelled in Figure 1.

This example shows how the stand-off collation identifies variant readings between texts by grouping pointers as opposed to grouping strings of text according to the parallel segmentation technique described in Chapter 12 of the TEI Guidelines.[11] The TEI offers a stand-off method for encoding variants, called "double-end-point-attachment", in which variants can be encoded separately from the base text by specifying the start and end point of the lemma of which they are a variant. This allows encoders to refer to overlapping areas on the base text, but despite its flexibility, this method still requires choosing a base text to which anchor variant readings. While choosing a lemma for each variant may be necessary for a critical edition, it is not ideal for a variorum edition that, by design, does not choose a base text.[12] Our approach, therefore, simply identifies variance and groups readings from multiple sources without conflating them into one document and with accommodation of multiple hierarchies.

(Garland 1996), reproduced here: http://shelleygodwinarchive.org/contents/frankenstein/the-frankenstein-notebooks-introduction/.

8 Barbara Bordalejo, ed. *Darwin Online*. See for example the illumination of variant passages in *The Origin of Species* here: http://darwin-online.org.uk/Variorum/1859/1859-1-dns.html

9 These hierarchical issues provided an application for Piez's invented LMNL "sawtooth" syntax to highlight overlap as semantically important to the novel; see Wendell Piez, "Hierarchies within range space: From LMNL to OHCO" *Balisage: The Markup Conference Proceedings* (2014): https://www.balisage.net/Proceedings/vol13/html/Piez01/BalisageVol13-Piez01.html

10 CollateX software applies a graph-based model of text to locate variants in documents. See https://collatex.net/doc/

11 See especially the TEI P5 Guidelines, 12.2.3 and 12.2.4: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS

12 For a related example, see Viglianti, R. Music and Words: reconciling libretto and score editions in the digital medium. *"Ei, dem alten Herrn zoll' ich Achtung gern"*, ed. Kristina Richts and Peter Stadler, 2016, 727-755.

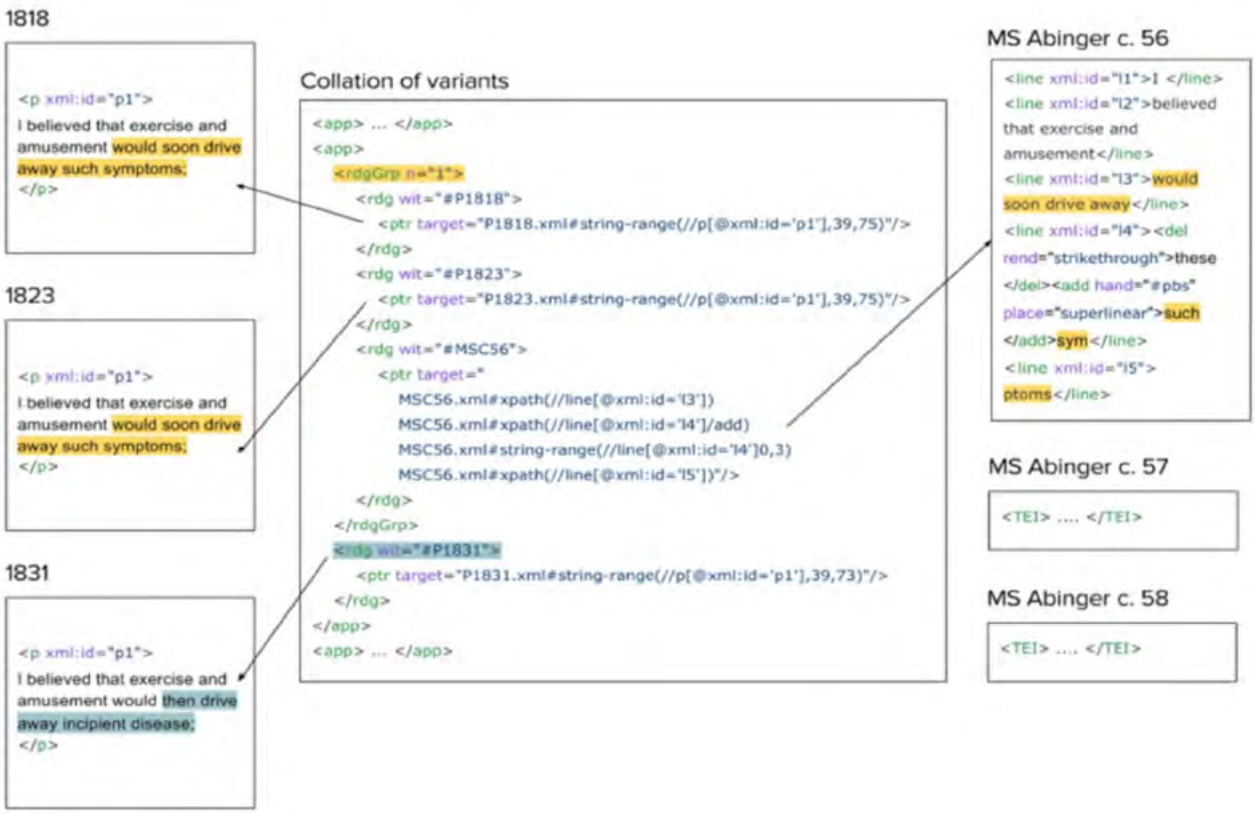


Fig. 1 An example variant with two different readings, showing Percy Bysshe Shelley's hand in the ms notebook. While the print editions of 1818, 1823, and the manuscript agree (yellow reading), the print edition of 1831 introduces new text (blue reading). The pointers are expressed according to the TEI XPointer Schemes defined in Chapter 16 of the TEI Guidelines and are subject to change.

Though we think of XML as a stable sustainable archiving medium, the repeated collapsing and expansion of hierarchies in our collation process makes us consider that for the viability of digital textual scholarship, ordered hierarchies of content objects might best be designed with leveling in mind, and that building with XML may be optimized when it is open to transformation. Preparing diversely encoded documents for collation challenges us to consider inconsistent and overlapping hierarchies as a tractable matter for computational alignment—where alignment becomes an organizing principle that fractures hierarchies, chunking if not atomizing them at the level of the smallest meaningfully sharable semantic features.

Non-normative Data From The Global South And Epistemically Produced Invisibility In Computationally Mediated Inquiry

Sayan Bhattacharyya
sayan@illinois.edu)
Price Lab for Digital Humanities, University of Pennsylvania, United States of America

This paper is an intervention that addresses an epistemological conundrum likely to become increasingly common and acute as the digital humanities both grow more diverse and increasingly encompass knowledge that has been produced outside of the parameters of Euro-American normativity. This contribution is in the spirit of addressing the need for cultural critique in the digital humanities along the lines that Alan Liu has called for (2012). I make use of an instantiated example, a text analysis tool for visualizing properties of a particular digitized text corpus in relation to trends of usage of specific words in the corpus, but I argue that the key insights are generalizable to a large spectrum of digital humanities tools.

Techno-social ensembles, acting as apparatuses of knowledge production through which computationally inferred knowledge is produced, are themselves power-laden. Data that is epistemically heterogeneous can be rendered illegible or less legible within a representational scheme that enforces standardization, creating a situation in which it can be visible only at the cost of relinquishing, in favor of the dominant episteme's normative assumptions, the variability that constitutes its heterogeneity — as these normative assumptions tend to privilege the homogeneity of data. I name and describe this problematic in a way that fosters a dialog between philosophy

and critical theory on the one hand and digital humanities on the other hand, placing it on a theoretical footing in relation to which that dialog can happen.

I describe possible approaches to this problematic, both conceptually and in the form of actionable solutions that follow from the conceptual issues. I also suggest a way to redress the unintended illegibility or invisibility that epistemologically heterogeneous and non-normative knowledge — such as, for example, many knowledge artifacts from the global South — can undergo in computationally mediated knowledge apparatuses. In the first, critical, section of the paper — "critical" in the sense of pertaining to critique — I show, building on insights that I have described elsewhere, how even powerful tools for text analysis and visualization that are state-of-art in the field may tend to produce an undercount in the number of accumulatively retrieved records of occurrence for non-western-language material encountered written in western script within western-language text (Bhattacharyya 2017). Considering such a tool as a knowledge apparatus, I show that the problem arises because the knowledge objects in question — non-western-language words — typically tend to present, when transliterated into morphological expressions in the Latin alphabet, much more representational variation than the extent of heterogeneity that such tools implicitly assume their normative knowledge objects, namely western-language material, to present. I describe the mechanism by means of which the problem arises in this particular knowledge apparatus, and I argue that the problem is homological, and therefore generalizable, beyond the particular constellation of words, scripts and language to a wider set of similar configurations in the humanities, especially when data from the global South is at play.

Computational inquiry into humanistic knowledge regarding non-normative knowledge objects such as knowledge objects from the global south is particularly vulnerable to the general problem: an apparatus for knowledge production tends to render invisible certain kinds of inscriptions that, for one reason or another, do not conform to the epistemic normativity that the apparatus presupposes. Cultural forces, through the sociotechnical ensemble that they are a part of, shape computational, algorithmic inquiry, so that the problem becomes especially acute in the digital humanities at scale. I argue that epistemological problems concerning legibility caused by the logic of scale and accumulation on the one hand, and the complementary logic of networks on the other hand, have a relation to the logics of hierarchical production and nonhierarchical (network-based) production, to whose increasing complementarity in the sociocultural sphere Luc Boltanski and Eve Chiapello, among others, have drawn attention (2005).

I will end by describing possible ways of addressing the issue in the context of undergraduate classes in comparative literature among the likes of which I have used a tool of the above kind. These possibilities point towards one possible kind of a decolonial approach in the digital humanities. I will suggest that the most promising solution has to do with "persistent annotation": a way for users (students for my use case) to annotate the invisibilities/illegibilities as and when they discover them, in the form of a written record that persists (from one term of teaching (one iteration of a course) to another term (another iterarion of the course). A sophisticated implementation of this solution would incorporate such a document, in the form of a user-contributable manifest, into the software tool itself (such as by including a visible pointer to such a manifest from within the GUI for the tool). For my small-sized use case, however, something as simple as a document carried over and renewed from semester to semester across the content-management system for the class can be sufficient as such a manifest. I will argue that this is roughly similar, in principle, to the way that one can make, edits (or, more generally (and more similarly to this situation), editing suggestions in Wikipedia, whether non-anonymously or anonymously as desired (but even in the case of anonymity, with an audit trail of accountability visible to a monitoring party). The epistemological stakes of this kind of approach in the case of Wikipedia have been addressed by Lih (2009) and can provide a useful point of comparison.

While my specific use case pertains to textuality, I will also make points of connection with instances of the illegibility or invisibility of non-normative knowledge in other modalities of computational media in the context of certain specific kinds of data or cultural knowledge. Shannon Mattern, for example, has examined the question of how computationally mediated representations of spatial data can produce illegibility or invisibility (2015), and Irit Rogoff has shown how curatorial practice can do the same for visual artifacts (2005, 2009). Finally, I will conclude by arguing that a connection exists between coloniality, legibility and accountability. Jon Wilson has recently argued that, rather than imperial certainty and confidence, coloniality was often distinguished by administrative anxiety about governance over strangers who are epistemically 'other' (2017) — an anxiety partially redressed by external informants who are tolerated, but only when their participation is underwritten by mechanisms of trust and accountability legible within the imperial episteme. There is an interesting parallel here with digital tools created by well-intentioned tool builders ending up governing the legibility of non-normative cultural artifacts that have their origin in zones of epistemic otherness.

References

Bhattacharyya, Sayan. "Words in a World of Scaling-up: Epistemic Normativity and Text as Data." *Sanglap: Journal of Literary and Cultural Inquiry* 4, no. 1 (2017). http://sanglap-journal.in/index.php/sanglap/article/view/157.

Boltanski, Luc, and Eve Chiapello. *The New Spirit of Capitalism*. Translated by Gregory Elliott. London: Verso, 2005.

Lih, Andrew. *The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia*. New York: Hyperion, 2009.
Liu, Alan. "Where Is Cultural Criticism in the Digital Humanities?"." In *Debates in the Digital Humanities*, edited by Matthew Gold. Minnesota: University of Minnesota Press, 2012.
Mattern, Shannon. "Gaps in the Map: Why We're Mapping Everything, and Why Not Everything Can, or Should, Be Mapped." *Words in Space*, September 18, 2015. http://wordsinspace.net/shannon/2015/09/18/gaps-in-the-map-why-were-mapping-everything-and-why-not-everything-can-or-should-be-mapped/.
Rogoff, Irit. "GeoCultures: Circuits of Art and Globalization." *Open!: Platform for Art, Culture and the Public Domain*, no. 16 (2009). https://www.onlineopen.org/download.php?id=53.
———. "Looking Away: Participations in Visual Culture." In *After Criticism: New Responses to Art and Performance*, edited by Gavin Butt. Malden, MA: Blackwell, 2005.
Wilson, Jon. *India Conquered: Britain's Raj & the Chaos of Empire*. Simon and Schuster, 2017.

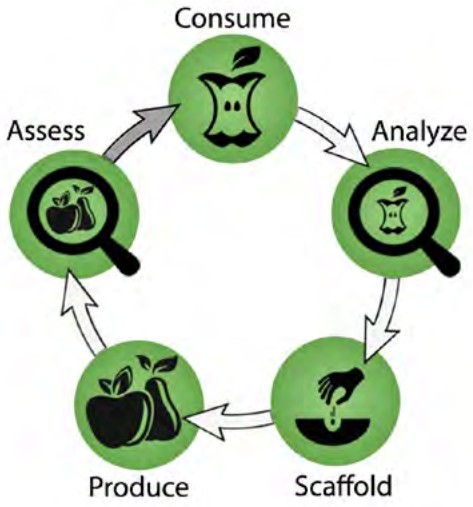
The CASPA Model: An Emerging Approach to Integrating Multimodal Assignments

Michael Blum
mxblum@wm.edu
College of William & Mary, United States of America

Abstract: The digital landscape for teaching and learning increasingly supports the inclusion of multimedia-based instructional strategies. In higher education, a culminating class project or final assignment often requires students to synthesize, analyze, and create content using a variety of communication modalities. Such projects, when closely aligned with course content and desired learning outcomes, allows for relevant and authentic assessment that leverages digital communication strategies. This paper presents an emerging practical model, CASPA, to promote curriculum-based integration of multimodal projects for assessment in higher education courses. Additionally, we report on initial implementation results and recommendations for practice and further research.

The CASPA Model

The model has five components; consume, analyze, scaffold, produce, and assess. The following sections address each component of the model and explain the instructional design processes in each.



C: Consume

The CASPA model starts by having students consume an exemplar of the mode of communication they will later be asked to produce. To be specific, have students consume podcasts that are similar in length and format to the podcasts the instructor anticipates having them create. In the consume phase, students should be asked to talk about the basic message or story and react to it. Start with whole class or small group discussions on evaluative questions. Using both positive and negative models in the Consume stage is useful, since the former is aspirational and the latter is easier to critique and sets an achievable bar for the students' own productions. By having students consume and evaluate on a personal level, the instructor sets the stage for the next phase, the analyze phase, where students will begin looking critically at the medium.

A: Analyze

Once students have consumed and can intelligently communicate the message of the mode and media being consumed, they should be encouraged to analyze the medium. If the instructor is uncomfortable analyzing a medium they may not be expert in, we suggest using this opportunity to discuss the basic concepts of narrative, storytelling, and argument, and how those concepts all depend on the chosen mode of communication. In analyzing the story, one might ask how effective the story is; what arguments, explicit or implicit, are evident; or what aspects of the story are most powerful. In this way, instructors can start developing a rubric, with or without input from their students, from which to analyze and critique a narrative based more on the success of the message than on the expertise of the medium. Certainly, where poor use of the medium is a barrier to understanding or appreciating the message, that should also be taken into

account (i.e., sound quality of a video), but unless the students are being asked to develop professional products, the medium can often take a backseat to the message. As instructors consider this analysis phase of multimodal assessments, other possible questions could include:

- How does this mode of communication affect the message?
- What are the strengths and weaknesses of this form of communication?
- What might be the strengths and weaknesses of another form of communication?

Once the students have a basic understanding of the medium, and they have had the opportunity to analyze narratives created in those media and decide on their own criteria for judging success and failure, they will have a better idea of what makes for a compelling, well-argued, well-researched piece and where such narratives fall short. This is where the right mix of successful and unsuccessful exemplars in the Consume phase pays off. Students can often learn more from failed storytelling than from successful storytelling.

S: Scaffold

The creation of a multimodal product should be completed in phases with the appropriate learning support, or scaffolds, necessary for success in each phase. In scaffolding multimodal assignments, production-based assignments are used to build skills and/or help students communicate in multiple modes. An assignment on podcasting might be built upon teaching good interview skills and, therefore, the first assignment may be to create a series of interview questions and test them out with a partner. The second assignment might be capturing sounds using audio capture, etc. Another type of scaffolding that is worth considering here asks the students to tell a story in various ways at various points in the semester. For example, first, as an elevator pitch, then as a storyboard, then as a PowerPoint, etc. In this case, the final project might be a mashup of all these different media and an analysis of the effectiveness of each in communicating the central thesis. Instructors should consider the multiple pathways available to help students arrive at their final goal. Walking through a scaffolded project will lead to a much less intimidating production phase for students as well as a more transparent assessment process for instructors. Once the pieces have been laid out and assessed in the scaffolding phase, it's time to assemble those pieces in the production phase.

P: Produce

In the production phase of the assignment, there are two basic scenarios with a multitude of variations. In the first scenario, students assemble the discrete pieces of the scaffolding assignments into a final product. Again, a podcast where students have been assessed and received feedback on the various pieces, such as interview questions, music choices, and narrative, is a good example of an assignment that is assembled in this way. In the second scenario, individual assignments are used to help students tell the same story in a variety of ways and then to identify the strengths and weaknesses of each mode. In this sort of assignment, the production phase may look like a curated analysis of the different individual mini-assignments, along with a description of how each mode communicates in different ways.The final product should reflect the recursive phases of analysis and instructor feedback in such a way that students are submitting a polished final creation rather than a simple redraft. From here, students and instructors proceed to the final stage of the CASPA model and benefit from the constructive feedback of others.

A: Assess

Once students are sensitive to the affordances and constraints of the vehicle and how that vehicle affects communication, the production of the assignment is then only the penultimate step in the learning process. An assessment of the assignment, including an assessment of the effectiveness of the vehicle in conveying the story, is highly encouraged as the final step of the CASPA model. For example, students may create a video, not because it is novel, but because the video modality informs the content produced and the story told. Multimodal assignments allow for reflection and assessment of that interaction between vehicle and content. It allows the student to ask why. Why a video over a website? Why a podcast over a PowerPoint presentation? What is gained and what is lost when choosing a vehicle? This line of inquiry might be valued as an area of analysis with a simple multimedia project, but it is paramount in a multimodal project. In this final stage of the assignment, the students return to an analytical mode, critiquing their own work and the work of their peers, ideally using rubrics they have developed or used in the analysis phase. This assessment phase allows students to see the process of consumption and production as an integral whole and an iterative process. Here is where the distinction between multimedia and multimodal really matters. In the assessment phase of the assignment, for the assignment to be truly multimodal, instructors should guide students to identify how the mode of communication alters, enhances, or hinders the telling of the story. Peer review and feedback on the effectiveness of the story will inform this process and empower students to refine their work for future use or retelling.

Various mini-assessments can, and should, come at various stages in the process (e.g., after each scaffolded assignment or before students submit their final pro-

jects). However, a more formal final assessment can be a powerful culminating activity, driving home the importance of self-reflection as part of an ongoing process of self awareness and improvement. Guided questions here may also be helpful. Instructors could select a few key questions, such as:

- What could be removed from the final product and why?
- What should be added and why?
- Is there anything that is in the wrong place within the narrative? Where should it go?

The tenets of the CASPA model support instructors and students alike in the important academic processes of analyzing, synthesizing, and conveying compelling arguments. The following section illustrates these processes in an application scenario derived from real experiences at William and Mary.

References

Barra, E., Aguirre Herrera, S., Pastor Caño, J. Y., & Quemada Vives, J. (2014). Using multimedia and peer assessment to promote collaborative e-learning. *New Review of Hypermedia and Multimedia*, *20*(0), 1–19. http://doi.org/10.1080/13614568.2013.857728

Cox, A. M., Vasconcelos, A. C., & Holdridge, P. (2010). Diversifying assessment through multimedia creation in a non-technical module: the MAIK project. *Assessment and Evaluation in Higher Education*, *35*(7), 831–846. http://doi.org/10.1080/02602930903125249

Hamm, S., & Robertson, I. (2010). Preferences for deep-surface learning: A vocational education case study using a multimedia assessment activity. *Australasian Journal of Educational Technology*, *26*(7), 951–965.

Harper, L., & Ross, J. (2011). An application of Knowles' theories of adult education to an undergraduate interdisciplinary studies degree program. *The Journal of Continuing Higher Education*, *59*, 161–166. http://doi.org/10.1080/07377363.2011.614887

Krippel, G., Mckee, a J., & Moody, J. (2010). Multimedia use in higher education: promises and pitfalls. *Journal of Instructional Pedagogies*, *2*, 1–8.

Morel, G. M., & Keahey, H. L. (2016). Student-Generated Multimedia Projects as a Multidimensional Assessment Method in a Health Information Management Graduate Program. In *Society for Information Technology & Teacher Education International Conference*, *2016*(1), 1120–1125.

Reynolds, C., Stevens, D. D., & West, E. (2013). "I'm in a professional school! Why are you making me do this?" A cross-disciplinary study of the use of creative classroom projects on student learning. *College Teaching*, (61), 51–59. http://doi.org/10.1080/87567555.2012.731660

Tham, J (2015, May 19). Multimedia vs. multimodal: A matter of terms. Retrieved from https://jasontham.com/2015/05/19/multimedia-vs-multimodal-a-matter-of-terms/

Quechua Real Words: An Audiovisual Corpus of Expressive Quechua Ideophones

Jeremy Browne
jeremy_browne@byu.edu
Brigham Young University, United States of America

Janis Nuckolls
cvd6262@gmail.com
Brigham Young University, United States of America

Introduction

Ideophones, sometimes called "mimetics" (Akita, 2009) or "expressives" (Diffloth, 1976) are expressions that communicate sensory aspects of the physical word such as sound (i.e., onomatopoeia), movement, color, etc., or cognitive/emotional states (e.g., "ta-da" in English). Although most linguistics description of and inquiry into ideophones have focused on vocal expressions, gestures are integrated with ideophonic utterances in some languages. The analysis of these gestures and their symbolism may augment scholars' understanding of the target language including how native speakers mentally represent their environment.

In this paper, we describe a web-based tool, *Quechua Real Words*, used by ideophonic linguists at [institution] to catalog and study multimedia representations of gestured ideophones as performed by native speakers of Pastaza Quichua. Research based on this tool is opening new understanding of the target language's aesthetics, especially regarding the non-arbitrariness of gestured signs. We also discuss the relationship between this tool and other digital humanities efforts.

Project Background

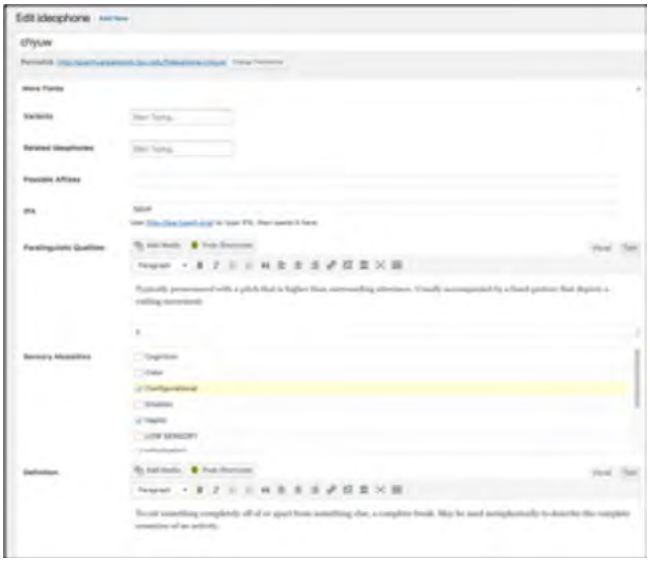
The indigenous people of eastern Ecuador speak Pastaza Quichua (PQ), a dialect of Northern Quechua. Descended from the language of the Incan civilization, Quechua is still spoken by as many as 10 million people in the Andes region stretching from Ecuador in the north to Argentina in the south. In 2015 [second author], a linguistic professor at [institution] led a group of student researchers who spent one semester in Ecuador recording and appreciating the indigenous culture and language. This team videoed over a hundred hours of interviews with PQ speakers, including thousands of examples of PQ ideophonic gestures.

The team returned to [institution] baffled at the scope of archival work that lay between their raw footage and

their research goals. In consultation with their [institution]'s Office of Digital Humanities, they constructed a WordPress-based website that facilitated their archival activities, accelerated their research, and opened their work to a global audience.

The Website

Quechua Real Words uses custom content types within WordPress and simple data entry forms that allow students and professors with little computer experience to record ideophones and link entries to specific segments of recorded videos. The project's footage is hosted on YouTube for simplicity and accessibility, and the data entry form only requests the video segment's URL and start and stop times.



Quechua Real Words ideophone recording form.

The entry form includes two other important features: First, the researcher may classify each ideophone by one or more "sensory modality" (e.g., color, haptic, movement, etc.). Second, each scholar—be they professor or student—may add their name to the list of the entry's contributors.

Once an ideophone is saved, it immediately appears on two indexes: the list of all ideophones, and the list of ideophones by modality. The first index allows researchers to look up specific ideophones, while the second promotes synthetic exploration where relationships between apparently unrelated ideophones can be made clear.

Each ideophone page displays the pronunciation (in IPA format), definition and other information one would expect from a traditional dictionary entry. It also shows a text description of the ideophone's paralinguistic qualities and one or more videos of native speakers expressing the ideophone in candid conversation. These videos are segments of longer YouTube videos, and the segments may be looped, paused, and replayed. (Such functionality is not native to YouTube's standard embedded player, so the site's video player is a custom JavaScript that connects to YouTube's published API.)

An ideophone page from Quechua Real Words.

As insisted on by the supervising professor, each ideophone page displays a "How to Cite" section with a citation in the Linguistics Society of America's preferred format. To recognize the collaborative nature of the website, the credited parties in the citation include everyone who contributed to the entry, even students.

Research Potential

During the first two years of its existence, [first author] used *Quechua Real Words* for research published in a special issue of the *Canadian Journal of Linguistics* ([second author], 2017), in three presentations at international conferences ([second author], 2015a; [second author], 2015b; [second author], 2014), and in two invited book chapters ([second author], in press; [second author], in press). Additionally, the website's content will inform an upcoming monograph ([second author], in preparation).

These publications focus on contextually-rich methods of understanding PQ ideophones, comparing specific gestures and intonations between speakers and contexts, and discovering how the ideophones are integrated with—rather than distinct from—the language's verbal aspects. As Akita and Tsujimura (2016) point out, the goal is to seek typological generalizations for ideophones rather than consider them in isolation. [Second author] seeks to extend these integrative studies and semantic generalizations beyond the vocal utterances into the gestured space.

Quechua Real Words as a Model for DH Collaboration

When [second author] proposed this website to the Office of Digital Humanities, [she/he] had little notion that it would lead to such a level of scholarly productivity. It was only as [she/he] saw how the site could function that [she/he] began to grasp its potential. Similarly, [first author], the digital humanists who crafted the website, overlooked its potential because, quite frankly, the technology behind *Quechua Real Words* is rudimentary for most DH centers.

Perhaps [first author]'s estimation was clouded by the fact that DH as a field has favored text-based literary analysis over multimedia research. Despite the work of the ARTeFACT project (Coartney & Wiesner, 2009) and a few others who have considered digital analysis of performing arts, DH has contributed much less to the analysis of video interactions, such as these ideophones, than it has to the analysis of written text. Garrard, Haigh, and de Jager (2011) demonstrate the status-quo for dealing with nonverbal communication in DH research: "…the recording and representation of various types of paralinguistic feature in transcription is somewhat idiosyncratic, and thus unreliable, suggesting that they should be removed in the interests of consistency."

This lack of emphasis on paralinguistic and nonverbal communication is in spite of those features' apparent value. "The nonverbal channel carries important information about emotional expressions… Systems that combine multiple modalities usually outperform single-modality systems in recognizing emotional" (Truong, Westerhof, Lamers, & de Jong, 2014). Unfortunately, even Truong et al. restricted their valuation of nonverbal channels to prosodic qualities such as timing and rhythm; they did not address issues of body language or gestures.

Regardless of why [first author] overlooked the website's potential, [she/he] has since changed how [she/he] evaluates potential collaborative DH projects. [She/He] now focuses on evaluating the use of the tools, websites, and other resources [she/he] would develop *relative to the target discipline* rather than relative to the state of the art within DH. This new approach has already proven fruitful (first author, 2017).

Future Plans

While [second author] continues to leverage *Quechua Real Words* for [her/his] scholarship, [first author] has combed the DH literature to discover methods of extending the site's capacity. One DH project that could contribute guidance to this project is the work of Paquette-Bigras and Forest (2014) who attempted to build a descriptive vocabulary for dance movements. A similar effort to construct a vocabulary for describing non-vocal expressions may reveal yet-unnoticed relationships between expressive gestures. This would require intense, non-automated markup of the gestures, but the *Quechua Real Words* website and the student-involved structure of [second author]'s courses would be facilitative. Such detailed modeling of the gestures would extend the modality-based clustering currently available on the website to include form-based clustering of the gestures.

Additionally, we are working with [institution's library] to add *Quechua Real Words* to their federated search databases. This will increase the site's discoverability by scholars and students throughout the world.

References

Akita, K. 2009. *A grammar of sound-symbolic words in Japanese: Theoretical approaches to iconic and indexical properties of mimetics*. PhD Dissertation. Kobe University.

Akita, K. & Tsujimura, N. 2016. "Mimetics". In T. Kageyama and H. Kishimoto (eds), *Handbook of Japanese Lexicon and Word Formation,* 133–160. Berlin: Gruyter De Mouton.

Coartney, J. S. & Wiesnet, S. L. (2009). Performance as digital text: Capturing signals and secret messages in a media-rich experience. *Literary and Linguistic Computing, 24*(2), pp. 153–160. https://doi.org/10.1093/llc/fqp012

Diffloth, G. 1976. "Expressives in Semai" *Oceanic Linguistics Special Publications*

No. 13, *Austroasiatic Studies Part I*, pp. 249-264

Garrad, P., Haigh, A., & de Jager, C. (2011). Techniques for transcribers: assessing and improving consistency in transcripts of spoken language. Literary and Linguistic Computing, 26(4), pp. 389–405. https://doi.org/10.1093/llc/fqr018

Paquette-Bigras, E. & Forest, D. (2014). A Vocabulary of the Aesthetic Experience for Modern Dance Archives. Paper presented at DH 2014, Lausanne, Switzerland.

Truong, K. P., Westerhof, G. J., Lamers, S. M. A., de Jong, F. (2014). Towards modeling expressed emotions in oral history interviews: Using verbal and nonverbal signals to track personal narratives. Literary and Linguistic Computing, 29(4), pp. 621–636. https://doi.org/10.1093/llc/fqu041

[The following references will be added following double-blind review:]

[first author], 2017

[second author], 2014

[second author], 2015a

[second author], 2015b

[second author], 2017

[second author], in press

[second author], in press

[second author], in preparation

Negentropic linguistic evolution: A comparison of seven languages

Vincent Buntinx
vincent.buntinx@epfl.ch
EPFL (École polytechnique fédérale de Lausanne), Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
EPFL (École polytechnique fédérale de Lausanne), Switzerland

Introduction

The relationship between the entropy of language and its complexity has been the subject of much speculation – some seeing the increase of linguistic entropy as a sign of linguistic complexification or interpreting entropy drop as a marker of greater regularity (Montemurro and Zanette 2011, Juola 2016, Bentz et al. 2017). Some evolutionary explanations, like the learning bottleneck hypothesis, argues that communication systems having more regular structures tend to have evolutionary advantages over more complex structures (Kirby 2001, Tamariz and Kirby 2016, Ferrer I Cancho 2017). Other structural effects of communication networks, like globalization of exchanges or algorithmic mediation, have been hypothesized to have a regularization effect on language (Kaplan 2014).

Longer-term studies are now possible thanks to the arrival of large-scale diachronic corpora, like newspaper archives or digitized libraries (Westin and Geisler 2002, Fries and Lehmann 2006, Lyse and Andersen 2012, Rochat et al. 2016). However, simple analyses of such datasets are prone to misinterpretations due to significant variations of corpus size over the years and the indirect effect this can have on various measures of language change and linguistic complexity (Buntinx et al. 2017). In particular, it is important not to misinterpret the arrival of new words as an increase in complexity as this variation is intrinsical, as is the variation of corpus size.

This paper is an attempt to conduct an unbiased diachronic study of linguistic complexity over seven different languages using the Google Books corpus (Michel et al. 2011). The paper uses a simple entropy measure on a closed, but nevertheless large, subset of words, called kernels (Buntinx et al. 2016). The kernel contains only the words that are present without interruption for the whole length of the study. This excludes all the words that arrived or disappeared during the period. We argue that this method is robust towards variations of corpus size and permits to study change in complexity despite possible (and in the case of Google Books unknown) change in the composition of the corpus. Indeed, the evolution observed on the seven different languages shows rather different patterns that are not directly correlated with the evolution of the size of the respective corpora. The rest of the paper presents the methods followed, the results obtained and the next steps we envision.

Method and Results

We use the concept of kernel entropy (Buntinx et al. 2017), defined as the Shannon entropy measure applied on word occurrences distribution normalized on the kernel of a given corpus. To calculate this measure, the corpus is subdivided into yearly sub-corpora. Next, we then calculate the word occurrences for the words that are present in each sub-corpus for each year. These words form a set, called a kernel. The word frequencies are normalized on the kernel for each year and the formula of Shannon entropy (using napierian logarithm) is applied on these distributions providing a measure that can be compared diachronically with robustness to corpus size evolution and to noises. The kernel entropy of a kernel for the year is given by the formula:

$$H^{K,y} = -\sum_{i=1}^{N^K} f_i^{K,y} \ln\left(f_i^{K,y}\right)$$

Where is the number of words composing the kernel and the relative occurrence frequency of the word normalized on the kernel in the year . The kernel entropy measure is computed for seven languages of Google Books corpora. *Figure 1* shows the kernel entropy variations normalized with respect to the average value (which change over the languages because kernels of different corpus also have different sizes).

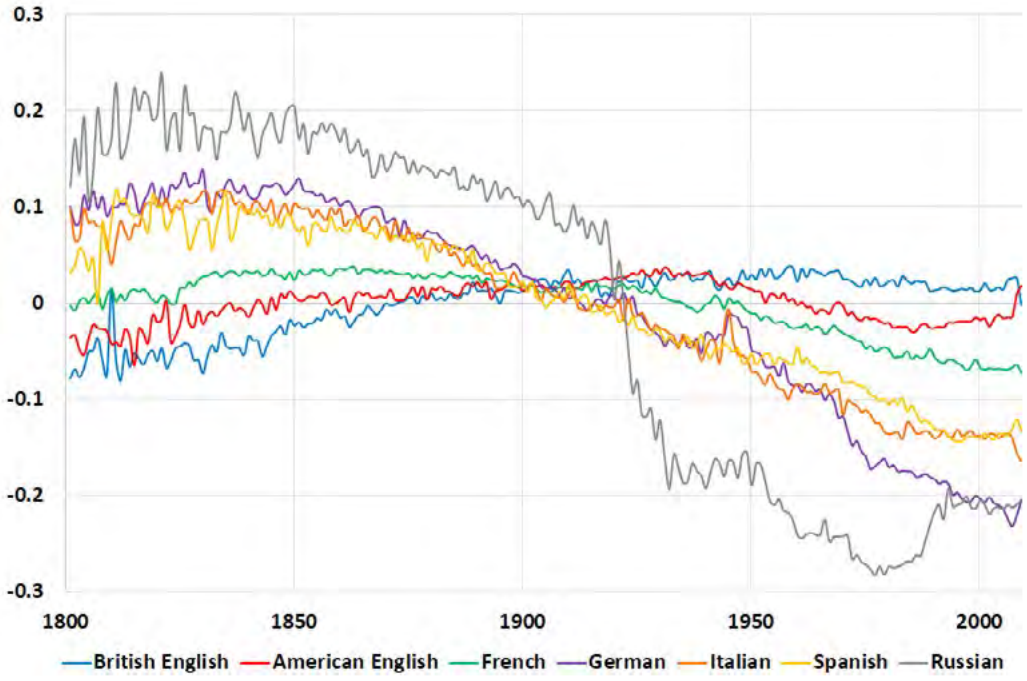
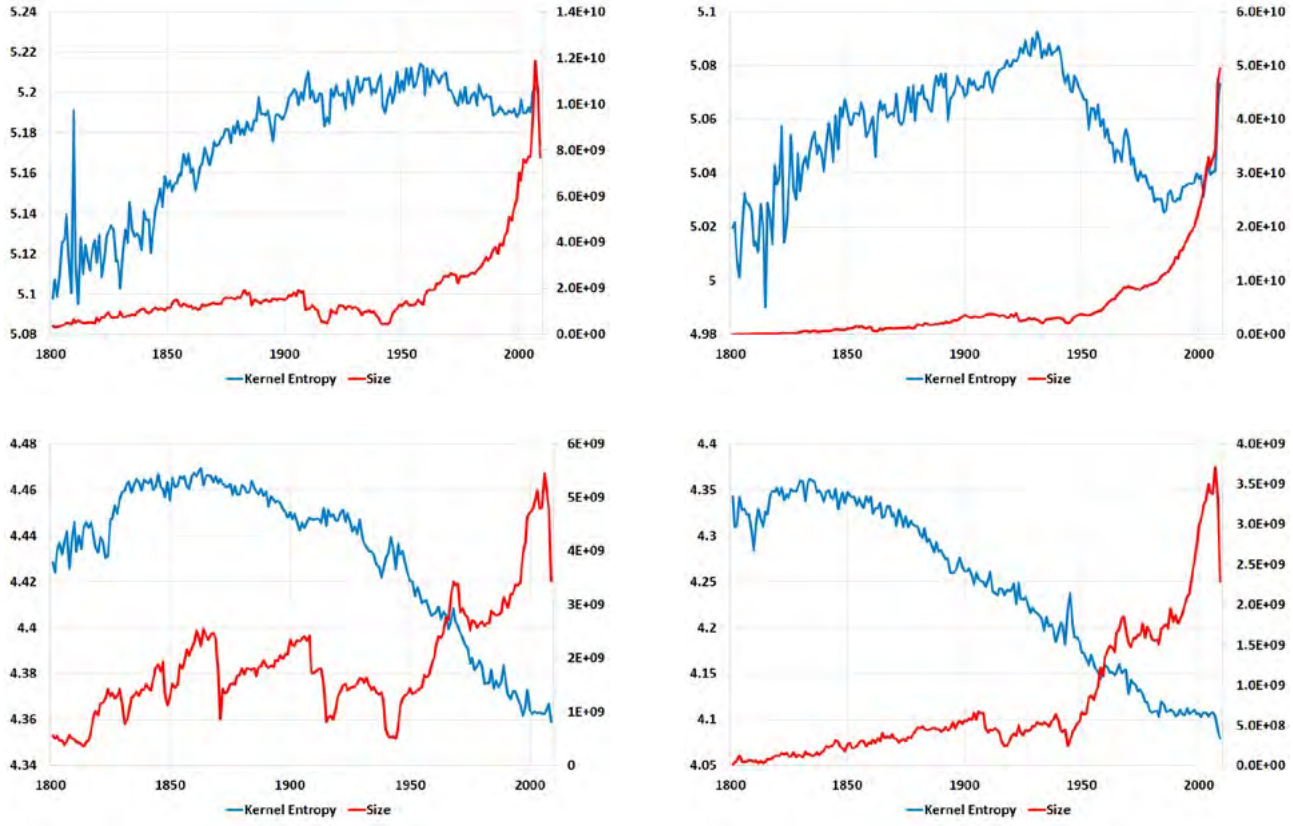


Figure 1: Normalized yearly kernel entropy evolution from 1800 to 2009 of seven Google Books corpora: British English, American English, French, German, Italian, Spanish and Russian.

We observe that even if all the seven language have different patterns and inflection points, they tend generally to show an effect of negentropy with increasing years. We note that most languages have a crosspoint in 1905, except for the Russian language, showing variations particularly from 1920 to 1930. We present in Figure 2 the kernel entropy evolution for each language in comparison to the corpus size.



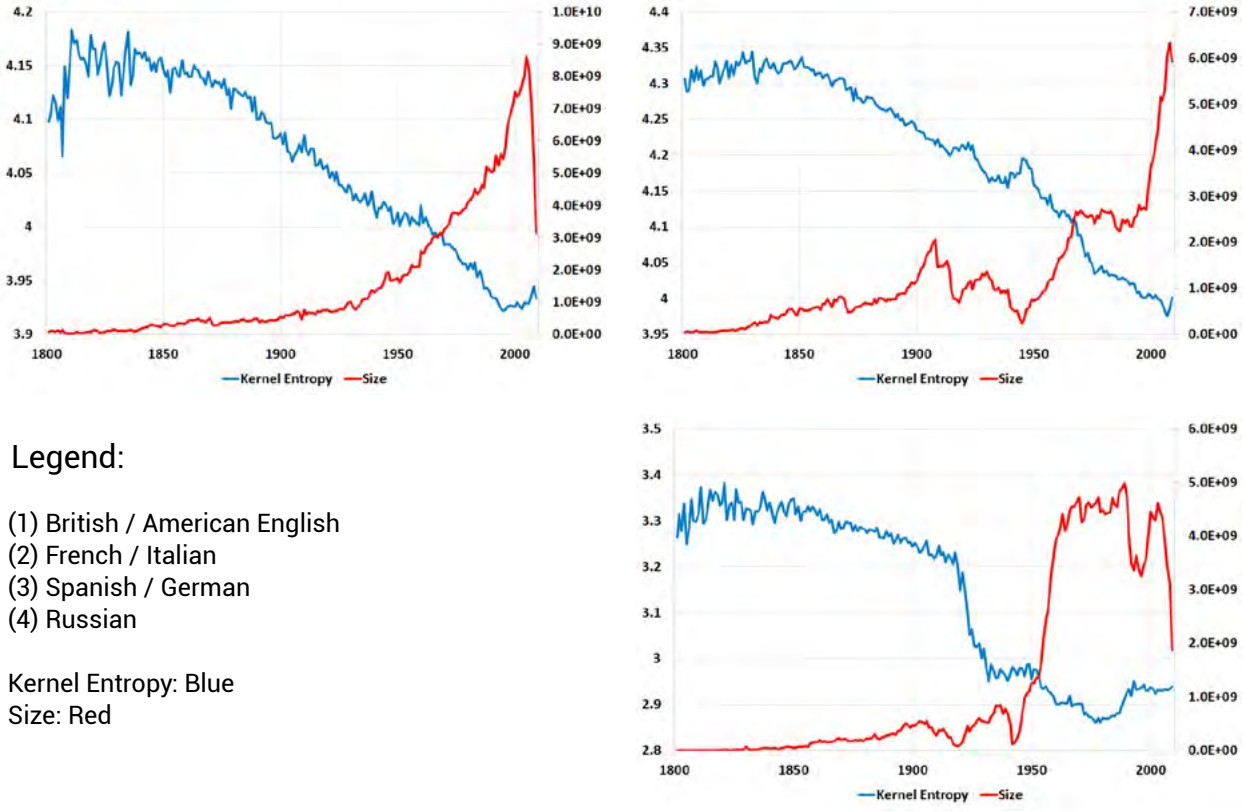


Figure 2: Yearly kernel entropy evolution and size evolution from 1800 to 2009 of seven Google Books corpora: British English, American English, French, German, Italian, Spanish and Russian.

Google Books corpora may experience sudden changes in composition depending on the year. For example, the addition of scientific literature and medical journals (Pechenick et al., 2015). In this case, the words kernel distribution, even if it is robust because composed of the most stable words, can change for a year which is subject to a change of composition of the corpus. However, this effect is still reduced because the words appearing and disappearing during this transition phase are not taken into account. We observe that the entropy of the kernel seems not to be affected by the size variations of corpora and when it appear to be affected, the direction of variation is unpredictable.

The British English and American English are the least affected languages by the negentropic effect. Their kernel entropy increases over time until 1960 (British English) and 1940 (American English). However, American English kernel entropy decrease quickly from 1940 to 1985. We observe that the obtained curve for the French language is similar to the one corresponding to the study of language evolution through 200 years of newspapers written in French despite a different kernel size (Buntinx et al. 2017).

Interesting inflection points are detected and should be poignant to specialists of the targeted language. We present in *Figure 3* the number of words in the kernel and inflections points for the seven languages.

Language	Number of words in the kernel	Inflection point 1	Inflection point 2
British English	82'332	1959	
American English	44'949	1931	1985
French	79'575	1825	1885
German	36'660	1850	1946
Italian	30'996	1983	
Spanish	25'582	1995	
Russian	5'123	1920	1988

Figure 3: Number of words in the kernel and kernel entropy inflection points for the seven Google Books corpora: British English, American English, French, German, Italian, Spanish and Russian.

Furthermore, it is possible to show the languages proximity in terms of kernel entropy evolution behavior through the determination of a distance based on kernel entropy correlations. A projection of the resulting matrix distance using PCA is presented in Figure 4.

We observe that British English and American English are represented together to the left of the plan because they have a relative opposite pattern with respect to other languages. Russian is also particular because of the brutal effect of the negentropy observed between around

1920 and the sudden increase at the end of the 1980s. The last four languages, French, Spanish, German and Italian share a more similar behavior and are represented in the right-bottom part of the plan.

Although much more in-depth investigation must be done, it is reasonable to make the hypothesis of different internal and external factors for explaining these various patterns. The Russian case clearly invites to investigate correlations between linguistic policies during the Soviet-ic period and their actual effects of the Russian language.

The similarity between French, German, Italian and Spanish pushes in the direction for similar processes of standardization, potentially due to linguistic convergence at national levels suppressing some regional particularities. In contrast, American and British English evolution is likely to be explained through the particular histories of the respective English-speaking populations and their relation to the rest of world. The progressive rise of English as a global language, spoken and written by many non-native speakers, is certainly playing a role in the shaping these particular curves.

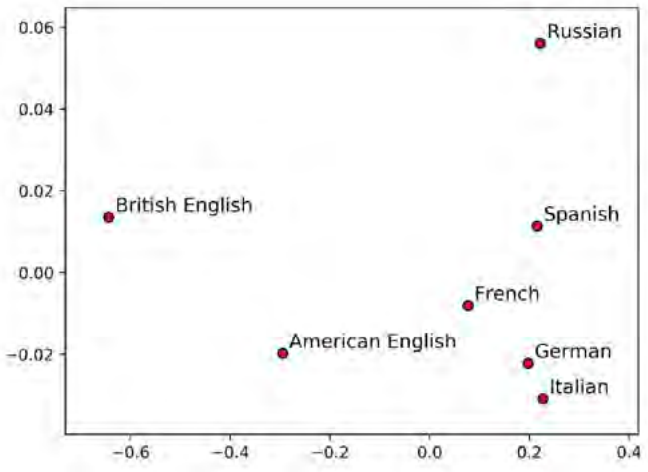


Figure 4: PCA projection of distance matrix using kernel entropy correlation-based distance for Google Books corpora: British English, US English, French, German, Italian, Spanish and Russian.

References

C. Bentz, D. Alikaniotis, M. Cysouw and R. Ferrer-i-Cancho. The Entropy of Words—Learnability and Expressivity across More Than 1000 Languages. *Entropy*, 19(6), 275, 2017.

V. Buntinx, C. Bornet and F. Kaplan. Studying Linguistic Changes on 200 Years of Newspapers. *DH2016*, Kraków, Poland, July 11-16, 2016.

V. Buntinx, F. Kaplan and A. Xanthos (Dirs.). Analyse multi-échelle de n-grammes sur 200 années d'archives de presse. Thèse EPFL, n° 8180, 2017.

R. Ferrer-i-Cancho. Optimization models of natural communication. *Journal of Quantitative Linguistics*, 1-31, 2017.

U. Fries and H. M. Lehmann. *The style of 18th century english newspapers: Lexical diversity. News Discourse in Early Modern Britain*, pages 91–104, 2006.

P. Juola. Using the Google N-Gram corpus to measure cultural complexity. *Literary and linguistic computing*, 28(4), 668-675, 2013.

F. Kaplan. Linguistic capitalism and algorithmic mediation. *Representations*, 127 (1):57–63, 2014.

S. Kirby. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, vol. 5, no 2, p. 102-110, 2001.

G. I. Lyse and G. Andersen. *Collocations and statistical analysis of n-grams. Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian. Studies in Corpus Linguistics*, John Benjamins Publishing, Amsterdam, pages 79–109, 2012.

J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak and E. Lieberman-Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182, 2011.

M. A. Montemurro and D. H. Zanette. Universal entropy of word ordering across linguistic families. *PLoS One*, 6(5), e19875, 2011.

E. A. Pechenick, C. M. Danforth and P. S. Dodds. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS One*, 10(10), e0137041, 2015.

Y. Rochat, M. Ehrmann, V. Buntinx, C. Bornet and F. Kaplan. Navigating through 200 years of historical newspapers. In *iPRES 2016*, numéro EPFL-CONF-218707, 2016.

M. Tamariz and S. Kirby. The cultural evolution of language. *Current Opinion in Psychology* 8: 37-43, 2016.

I. Westin and C. Geisler. A multi-dimensional study of diachronic variation in british newspaper editorials. *International Computer Archive of Modern and Medieval English*, (26):133–152, 2002.

Labeculæ Vivæ. Building a Reference Library of Stains Found on Medieval Manuscripts with Multispectral Imaging

Heather Wacha
wacha2@wisc.edu
University of Wisconsin-Madison, United States of America

Alberto Campagnolo
alberto.campagnolo@gmail.com
Library of Congress, United States of America

Erin Connelly
erincon@upenn.edu
University of Pennsylvania, United States of Americ

The stains found on medieval manuscripts are signs that indicate a past life, more specifically the visible and invisible remains of human interaction over time. Reading these signals - in concert with conventional information such as script, collation, illumination, and size - can add to our understanding of their history and use. While recent work has been done on the uses of multispectral imaging in understanding the degradation and preservation of parchment, there is little pre-existing scholarship on the presence and nature of stains in medieval texts. Indeed, the significance of stains has traditionally been underestimated.[1] This project focuses on those very manuscripts that are often overlooked due to heavy soiling and damage, effects that diminish their perceived quality and value. We are a team of interdisciplinary postdoctoral scholars and collaborators working on constructing a Library of Stains in order 1) to provide an online database that will allow scholars, librarians, and conservators to better analyze materiality, provenance, use and preservation of manuscripts/early-printed books; 2) to document and disseminate a methodological approach for analyzing stains; and 3) to provide a model for public-facing interdisciplinary collaboration. To our knowledge, this is the first interdisciplinary attempt to build a library of medieval and early-modern stains using the tools of medieval literature, medieval history, codicology and bibliography, multispectral imaging, chemical analysis, and data science.

Our presentation will include information about the the Library of Stains project framework and methodology, as well as the dissemination of its data and results. The project timeline, supported by a microgrant from the Council on Library and Information Resources (CLIR), covers a one-year period. Imaging will be complete by December 15, 2017; processing and analyzing the images will follow and results will be documented by April 2018 and codified by August 31, 2018. Our presentation will report on the project findings, their broader implications, public engagement, and best practices for conservators, archivists and librarians who will use the project's database.

This pilot study aims to provide an identified, open-access database of a number of common stains found on parchment, paper, and bindings in medieval manuscripts and early printed books in order to help researchers answer questions such as manuscript provenance, transmission, and material culture. It also highlights how using scientific technologies - in this case, multispectral imaging - aids in answering traditional arts and humanities questions. The database will hold metadata collected from the multispectral imaging (JSON files), as well as information about the processed image data leading to stain identification. The latter will take the form of tiff files representing images taken at different light wavelengths, ultraviolet and fluorescent settings, in conjunction with associated specific spectral curves, and the relevant data collected from XRF/FTIR/FORS scanning. The Schoenberg Institute for Manuscript Studies will preserve the database and all files as part of their open-access Colenda repository.

Identifying the stains present in a book and understanding the relationship between the placement of the stain and its surrounding text brings to light new information about how manuscripts were used, read, and applied in situ. We have identified forty Western European manuscripts held in the University of Pennsylvania Libraries, the Chemical Heritage Foundation, the Library of Congress, the University of Wisconsin Special Collections and the University of Iowa Special Collections, with dates ranging from the twelfth to sixteenth centuries. The type of stains anticipated vary according to genre of manuscripts, and will likely indicate the presence of such elements as blood,[2] heavy metals, candle wax, urine, various oil-based concoctions, wines or spirits, and possibly zoological matter such as crushed spiders or flies. For example, it is our hope that once we have processed the results for a manuscript entitled "On the Colors of Urine," (De Coloribus Urinae, University of Pennsylvania Schoenberg Institute for Medieval Studies, MS Codex 133), it will show that the liquid stains throughout are indeed urine, perhaps stemming from a doctor or patient's accidental spill when consulting the manuscript. This type of analysis builds upon the significance of intellectual and material analysis concerning written culture, and extends beyond current analytical approaches to text, illumination, and bibliographical description.

Once the results are verified and each type of stain has been characterized, other interested parties will be able to access the database and verify their own stains against the fixed dataset. A methodological approach will be documented, disseminated, and openly accessible to those wanting to work with unknown stains so that researchers can model and replicate the workflow and process when faced with an unidentified stain. With data gathered directly from multispectral images, it will be possible to create a graphic representation in the form of spectral curves of each identified stain so that when a user seeks to identify a stain in a particular manuscript, an image can be processed and compared to the graphics held in the Library of Stains database. In this way we

1 The multispectral imaging dataset and physical samples developed by Giacometti *et al.* (2015) will be taken into consideration for comparison and as further reference material on staining substances on parchment. See also Giacometti *et al.* 2012; MacDonald *et al.* 2013; Campagnolo *et al.* 2016; Giacometti *et al.* 2016, 2017.

2 Confirming blood stains, such as those recorded on the *Declaration to the World by Agustin de Iturbide* (see https://www.wdl.org/en/item/2969/#institution=center-for-the-study-of-the-history-of-mexico-carso - accessed 2017/11/26) is particularly interesting for obvious forensic and historical reasons.

engage the scholarly community in an on-going collaboration resulting in the continual growth of the Library and in the open access data it creates. This is a new way for researchers, conservators, librarians, and the public to access important information and gain a greater appreciation of the material makeup of old books, their historical uses, and new approaches for modern studies.

We envision that scholarly audiences will use our data and methodology to advance knowledge into the provenance of manuscripts, their uses within a historical context, their working environment, their transmission, and their circulation. For conservators and librarians in particular new information will help determine proper storage conditions, as well as health and safety issues, in particular the identification of heavy metal contamination, such as mercury residue in alchemical manuscripts or herbaria.[3] For librarians and archivists, the results of this project will also deliver a heightened awareness of the value of interdisciplinary research and model for future collaborations that can create new content and context for rare book and special collections.

Finally, bringing together multispectral imaging experts and humanists offers an opportunity to explore and develop a working model for best practices when engaging in interdisciplinary collaboration that will actively gain the attention of public audiences. There is an enduring interest in medieval themes as a broad concept within the public sphere. Even if these themes are often caricatures or historically inaccurate, this interest in the medieval period in the public imagination offers the perfect opportunity to invite the public in to experience the academic discipline of medieval studies through an engaging and public-facing project. Accessibility to primary sources through an online database like the proposed library of stains juxtaposed with descriptive metadata will contextualize the project, connect with public interest, and provide value in the form of education. Our focus on public engagement is supported through the regular dissemination of information on the project to both public and scholarly communities through a variety of social media platforms, including facebook, twitter (#StainAlive), instagram, flickr, and a blog. With frequent posts across all formats, we hope to engage and excite both academic and public audiences interested in the medieval world and the lived experiences of medieval scribes, scholars, and readers.

References

Campagnolo, A., Giacometti, A., MacDonald, L., Mahony, S., Robson, S., Weyrich, T., Terras, M. and Gibson, A. (2016). Cultural Heritage Destruction: Experiments with parchment and multispectral imaging. In Bodard, G. and Romanello, M. (eds), *Digital Classics Outside the Echo-Chamber*. London: Ubiquity press, pp. 121–46 http://www.ubiquitypress.com/site/books/detail/21/digital-classics-outside-the-echo-chamber/ (accessed 19 May 2016).

Giacometti, A., Campagnolo, A., Macdonald, L., Mahony, S., Robson, S., Weyrich, T., Terras, M. and Gibson, A. (2015). *UCL Multispectral Processed Images of Parchment Damage Dataset*.

Giacometti, A., Campagnolo, A., MacDonald, L., Mahony, S., Robson, S., Weyrich, T., Terras, M. and Gibson, A. (2016). Visualising macroscopic degradation of parchment and writing via multispectral images. Care and Conservation of Manuscripts 15: Proceedings of the *Fifteenth International Seminar Held* at the University of Copenhagen, 2nd-4th April 2014. Copenhagen: Museum Tusculanum Press; University of Copenhagen and the Royal Library of Denmark, pp. 89–102.

Giacometti, A., Campagnolo, A., MacDonald, L., Mahony, S., Robson, S., Weyrich, T., Terras, M. and Gibson, A. (2017). The value of critical destruction: Evaluating multispectral image processing methods for the analysis of primary historical texts. *Digital Scholarship in the Humanities*, 32(1): 101–22 doi:10.1093/llc/fqv036.

Giacometti, A., Campagnolo, A., MacDonald, L., Mahony, S., Terras, M., Robson, S., Weyrich, T. and Gibson, A. (2012). *Cultural Heritage Destruction: Documenting Parchment Degradation via Multispectral Imaging. EVA London 2012: Electronic Visualisation and the Arts*. London: BCS, The Chartered Institute for IT, pp. 301–08 http://ewic.bcs.org/upload/pdf/ewic_ev12_s17paper2.pdf (accessed 4 October 2012).

MacDonald, L., Giacometti, A., Gibson, A., Campagnolo, A., Robson, S. and Terras, M. M. (2013). *Multispectral Imaging of Degraded Parchment*. Chiba, Japan.

Purewal, V. J. (2012). *Novel detection and removal of hazardous biocide residues historically applied to herbaria University of Lincoln Ph.D.* http://eprints.lincoln.ac.uk/13573/ (accessed 8 October 2017).

3 Purewal (2012) has developed a UV-based methodology to identify visually the presence of mercury in herbaria.

Dall'Informatica umanistica alle Digital Humanities. Per una storia concettuale delle DH in Italia

Fabio Ciotti
fabio.ciotti@uniroma2.it
Università di Roma Tor Vergata, Italy

Introduzione

Negli ultimi anni abbiamo assistito a una rinnovata attenzione nei confronti della dimensione storica delle Digital Humanities. Le ragioni di questo interesse si possono rintracciare da una parte nella rilevanza assunta del dibattito sul multiculturalismo e sulla dimensione geopolitica

delle DH; dall'altra nella ricorrente questione della definizione disciplinare e dei suoi confini: risalire alle radici storiche sembra una efficace strategia di analisi e di argomentazione per affrontare entrambe le problematiche. Tuttavia anche questo nuovo fervore storiografico fatica a riconoscere la molteplicità delle tradizioni culturali e nazionali, e il loro ruolo nello sviluppo delle DH globali. In questo paper intendo contribuire alla costruzione di una prospettiva storiografica plurale, delineando il primo abbozzo di una storia concettuale delle DH in Italia.

Il dibattito recente sulla storia delle DH

La riflessione storica sulla propria origine ed evoluzione ha sempre accompagnato il dibattito nel campo delle DH (Hockey, 2004). Tuttavia nell'ultimo lustro il genere auto-storiografico è stato particolarmente frequentato. Per citare solo alcuni dei lavori più rilevanti possiamo ricordare la *lectio* "Getting there from here: Remembering the future of digital humanities", tenuta da Willard McCarty alla conferenza DH2013 in occasione del conferimento del *Busa Award* (McCarty, 2014), dove lo studioso canadese delinea una stimolante *genealogia* delle DH, contrappuntata da una personale biografia intellettuale. Notevole anche il saggio di Edward Vanhoutte nel volume miscellaneo *Defining Digital Humanities*, intitolato "The Gates of Hell. History and Definition of Digital | Humanities | Computing" (Terras et al., 2013: 119–56)"event-place":"Williston","abstract":"Digital Humanities is becoming an increasingly popular focus of academic endeavour. There are now hundreds of Digital Humanities centres worldwide and the subject is taught at both postgraduate and undergraduate level. Yet the term 'Digital Humanities' is much debated. This reader brings together, for the first time, in one core volume the essential readings that have emerged in Digital Humanities. We provide a historical overview of how the term 'Humanities Computing' developed into the term 'Digital Humanities', and highlight core readings which explore the meaning, scope, and implementation of the field. To contextualize and frame each included reading, the editors and authors provide a commentary on the original piece. There is also an annotated bibliography of other material not included in the text to provide an essential list of reading in the discipline. This text will be required reading for scholars and students who want to discover the history of Digital Humanities through its core writings, and for those who wish to understand the many possibilities that exist when trying to define Digital Humanities.","URL":"http://www.ashgate.com/isbn/9781409469636","language":"en","editor":[{"family":"Terras","given":"Melissa"},{"family":"Nyhan","given":"Julianne"},{"family":"Vanhoutte","given":"Edward"}],"issued":{"date-parts":[["2013"]]}},"locator":"119-156"}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} , che ripercorre questa storia dalle origini fino al primo decennio di questo secolo, focalizzando soprattutto l'area degli studi linguistici e letterari e offrendo un dettagliato resoconto della transizione da *Humanities computing* a *Digital Humanities*. Una prospettiva interessante è offerta dal recente libro di Julianne Nyhan e Andrew Flynn *Computation and the Humanities: Towards an Oral History of Digital Humanities* (Nyhan and Flinn, 2016), in cui lo studio dei documenti e lo scavo negli archivi si combina con la storia orale narrata dai protagonisti. La stessa studiosa ha dedicato grande attenzione all'opera di Padre Roberto Busa, grazie all'esame degli archivi personali conservati presso l'Università Cattolica di Milano; al "padre fondatore" ha dedicato un recente volume anche Steven Jones (Jones, 2016). Maggiormente focalizzato sulla storia del rapporto tra studi letterari (anglo-americani) e metodi informatici è il saggio di Amy Earhart *Traces of the old, uses of the new: the emergence of digital literary studies* (Earhart, 2015).

Un elemento che accomuna questi lavori, nonostante i diffusi e convinti richiami alla necessità di adottare una visione plurale, multiculturale e globale delle DH, è che essi sono fondamentalmente centrati sulla tradizione anglo-americana la quale, di fatto, appare da queste ricostruzioni come l'unica ad avere veramente prodotto risultati teorici e operativi degni di nota. Non intendo discutere in questa sede questioni di geopolitica delle DH, peraltro affrontate egregiamente, tra gli altri, da Domenico Fiormonte sul piano della critica teorica e politica (Fiormonte, 2016; Fiormonte, 2017) e da Marin Dacos su quello dell'indagine sociologica empirica (Dacos, 2016).

D'altra parte è innegabile che, almeno per quanto riguarda i risultati pratici, storicamente le DH di origine angloamericana siano state più efficaci, se non altro in virtù degli assai più ingenti finanziamenti di cui hanno potuto godere. Ma resta pur vero che la storia delle DH è stata assai più plurale, se non caleidoscopica, di quanto gli autorevoli studiosi che abbiamo ricordato non riconoscano (con poche rare eccezioni e per pochi nomi eccellenti). Una pluralità che è anche e soprattutto teorica ed epistemologica, e che oggi si manifesta ancora e di nuovo nei diversi e non sempre conciliabili modi in cui si declina il sintagma Digital Humanities.

La storia dell'Informatica Umanistica italiana rappresenta una tessera importante di quel caleidoscopio, che manifesta la sua peculiarità a partire dal nome della cosa stesso, dove spicca la funzione sostantivale del temine "informatica". Questo paper intende fornire un primo contributo per una storia concettuale, prima e più che evenemenziale, dell'Informatica Umanistica.

La preistoria: dopo Busa

Iniziamo tuttavia con una osservazione fattuale: la tradizione italiana nell'informatica umanistica è maturata e si è sviluppata per un lungo periodo di tempo e senza soluzione di continuità. Il riferimento a padre Busa, uni-

versalmente riconosciuto come il fondatore di questo dominio scientifico, e alla sua attività di digitalizzazione e indicizzazione delle opere di Tommaso d'Aquino iniziato addirittura alla fine degli anni 40 dello scorso secolo, è piuttosto ovvio. Ma voglio far notare che l'impresa di Busa non era assolutamente isolata in Italia. È sufficiente ricordare che nel 1961 il prestigioso annuale "Almanacco Letterario Bompiani" (Morando, 1961), pubblicazione il cui ruolo innovativo nel dibattito culturale degli anni 60 italiano è difficilmente sottostimabile, dedicava la sua parte monografica al tema "Le Applicazioni dei Calcolatori Elettronici alle Scienze Morali e alla Letteratura". Nel dossier, arricchito dalla splendida grafica di Sergio Munari e da un lussureggiante apparato iconografico, trovano spazio una serie di interventi originali e di estratti da opere preesistenti, che spaziano dai fondamenti teorici delle macchine computazionali, alle prime pionieristiche ricerche nel campo della traduzione automatica, in quello della linguistica computazionale (con un intervento dello stesso Busa), e della filologia informatica (con la descrizione di un progetto di Aurelio Roncaglia); ma non mancano testi di riflessione teorica e critica, come il bel saggio di Franco Lucentini sul tema dell'automa nella letteratura e il saggio di chiusura di Umberto Eco "La forma del disordine" che allude ai temi del capitale *Opera aperta*; e vi compare uno dei primi esperimenti di letteratura elettronica (probabilmente il primo in assoluto): il poema computazionale *Mark 1* di Nanni Balestrini. Sin da quegli anni lontani, insomma, i più avvertiti e innovativi tra gli intellettuali italiani mostravano una visione a un tempo plurale e teoricamente rigorosa delle prospettive aperte dall'incontro tra informatica e scienze umane.

Sulla scorta di queste esperienze seminali, a cavallo tra la fine del decennio e l'inizio del successivo, vengono fondati i primi centri in cui il rapporto tra scienze umane e informatica trova una collocazione istituzionale. Ci riferiamo in particolare all'Istituto di Linguistica Computazionale del CNR fondato alla fine degli anni 60 dal professor Zampolli a Pisa (già culla dell'informatica italiana), che divenne ben presto un riferimento di eccellenza per l'elaborazione automatica del linguaggio a livello internazionale. Sempre nell'ambito del CNR si colloca l'Istituto per il lessico intellettuale europeo e la storia delle idee (ILIESI), fondato dal Professor Gregory. In stretta connessione con l'esperienza dell'ILC, il centro sin dagli anni 70 si dedicò alla creazione di risorse testuali in formato digitale e all'analisi lessicografica computazionale, con uno specifico interesse per la storia delle idee nell'età moderna.

La fondazione dell'Informatica Umanistica: Tito Orlandi e la scuola romana

Se la genealogia del sapere informatico umanistico italiano affonda le sue radici in epoche lontane, la sua manifestazione teoricamente più rilevante si colloca negli anni '80 dello scorso secolo presso l'Università di Roma La Sapienza: nasce qui, infatti, l'idea dell'Informatica Umanistica come disciplina *autonoma* con uno spiccato orientamento metodologico. La figura trainante di questo percorso intellettuale è Tito Orlandi. Portando a sintesi una serie di esperienze scientifiche e didattiche avviate negli anni precedenti, nel 1984 fonda alla Sapienza il Gruppo di ricerca "Informatica e Discipline Umanistiche", dove raccoglie un gruppo di studiosi, il quali condividevano la "consapevolezza […] che le procedure informatiche rappresentavano un naturale completamento delle proprie ricerche" (Introduzione a Gigliozzi, 1987: IX).

Ciò che caratterizza questa esperienza e che ne definisce la natura fondazionale per la storia concettuale del campo, è il rifiuto di una visione strumentale dell'informatica nelle discipline umanistiche (che era allora già abbastanza diffusa se non predominante nelle pur aurorali sperimentazioni a livello internazionale) e la netta predilezione per un approccio teorico ed epistemologico. L'informatica viene intesa non come ingegneria ma come scienza teorica della rappresentazione ed elaborazione (automatica) dell'informazione, e su questo terreno è evidente la convergenza con le scienze umane (e non solo di quelle basate sul linguaggio, ché sin dalle origini nel gruppo romano grande importanza ebbe l'archeologia, soprattutto in virtù dell'influenza su Orlandi dell'opera di Roger Gardin). Una convergenza che si manifesta fondamentalmente sul piano metodologico (Orlandi, 1992: 17):

il rapporto tra informatica e discipline umanistiche si può esprimere nella questione se vi sia un modo "informatico" di vedere (anche) le discipline umanistiche, che si differenzia a seconda delle discipline (e che dunque, in questo caso, rappresentano l'oggetto di questa disciplina), ma che rimane unitario nel modo di considerarle. Il modo informatico prevede la formalizzazione dei dati […] e la formalizzazione delle procedure per analizzarli e valutarli

Su queste basi non stupisce che il gruppo romano si sia concentrato su aspetti e temi fondativi quali: il problema della codifica intesa come processo semiotico e formale (Gigliozzi, 1987); il concetto di modello e modellizzazione (Gigliozzi, 1992); la ridefinizione del concetto di edizione scientifica (Mordenti, 2001); i fondamenti della critica computazionale e la modellizzazione formale delle strutture narrative (Gigliozzi, 2008). La sintesi di questa stagione di studi viene fornita dallo stesso Orlandi con il suo fondamentale manuale *Informatica Umanistica* (Orlandi, 1990).

Dopo la chiusura di questa prima esperienza sono soprattutto due i membri del gruppo che proseguono il progetto intellettuale originale. Tito Orlandi fonda nel 1991 il CISADU (Centro Interdipartimentale di Servizi per l'Automazione nelle Discipline Umanistiche), il primo centro di informatica umanistica propriamente detto in Italia, e prosegue nella sua esplorazione sui fondamenti teorici e metodologici della disciplina. Giuseppe Gigliozzi – che morirà prematuramente nel 2001– nella seconda metà

degli anni 90 fonda il CRILET (Centro Ricerche Informatica e Letteratura), dove all'aspetto teorico si affianca l'attività applicativa e la creazione di risorse digitali. I temi di ricerca principali sono l'analisi testuale - esemplare il suo studio su *Memoriale* di Volponi (Gigliozzi, 1996) – e la digitalizzazione e codifica dei testi. Questa esperienza ha giocato un ruolo determinante nella diffusione della *Text Encoding Initiative* (e di XML) in Italia e nella sua affermazione come standard di riferimento nella la maggior parte dei programmi di digitalizzazione testuale nel paese (Ciotti, 1994; Ciotti, 1997).

Conclusioni

La storia dell'Informatica Umanistica italiana non si esaurisce ovviamente nella "scuola romana". Già negli anni novanta nel campo digitale si affacciano numerosi altri studiosi, centri, progetti e nuove prospettive e punti di vista emergono: ad esempio la scuola degli studi ipertestuali promossa da Mario Ricciardi (Ricciardi and Bonadonna, 1994); o la realizzazione della Letteratura Italiana Zanichelli da parte di Pasquale Stoppelli (Stoppelli, 2005), che ha sempre avuto una visione strumentalista e ancillare dei metodi informatici. Con il nuovo millennio il panorama si fa sempre più articolato e oggi il movimento italiano è pienamente integrato nelle Digital Humanities globali.

Resta il fatto che diversi decenni di sperimentazioni e di elaborazione teorica hanno una visibilità globale assai scarsa. Senza dubbio la barriera linguistica ha costituito un ostacolo molto arduo da superare per ottenere il dovuto riconoscimento. Ma la questione del multilinguismo e del multiculturalismo nella comunità globale delle Digital Humanities è anche e soprattutto un problema di macro e microfisica dei poteri. Anche per questo occorre raccontare le *nostre* storie.

References

Ciotti, F. (1994). Il testo elettronico: memorizzazione, codifica ed edizione informatica del testo. *Macchine per Leggere. Tradizioni e Nuove Tecnologie per Comprendere i Testi*. Spoleto: Centro Italiano di Studi sull'Alto Medioevo.

Ciotti, F. (1997). Cosa è la codifica informatica dei testi?. *Umanesimo & Informatica: Le Nuove Frontiere Della Ricerca e Della Didattica Nel Campo Degli Studi Letterari*. Fossombrone (PS): Metauro Edizioni, pp. 55–85.

Dacos, M. (2016). La stratégie du sauna finlandais: Les frontières des Digital Humanities. *Digital Studies/ Le Champ Numérique*, 0(0) doi:10.16995/dscn.41. https://www.digitalstudies.org//article/10.16995/dscn.41/ (accessed 24 November 2017).

Earhart, A. E. (2015). Traces of the old, uses of the new : the emergence of digital literary studies.

Fiormonte, D. (2016). Toward a Cultural Critique of Digital Humanities. In Gold, M. K. and Klein, L. F. (eds), *Debates in the Digital Humanities: 2016*. Minneapolis London: University of Minnesota Press, pp. 438–58.

Fiormonte, D. (2017). Digital Humanities and the Geopolitics of Knowledge. *Digital Studies/Le Champ Numérique*, 7(1) doi:10.16995/dscn.274. http://www.digitalstudies.org/articles/10.16995/dscn.274/ (accessed 24 November 2017).

Gigliozzi, G. (1987). *Studi di codifica e trattamento automatico di testi*. Roma: Bulzoni.

Gigliozzi, G. (1992). Modellizzazione delle strutture narrative. *Calcolatori e Scienze Umane: Archeologia e Arte, Storia e Scienze Giuridiche e Sociali, Linguistica, Letteratura*. 1. ed. Milano: Etaslibri, pp. 302–14.

Gigliozzi, G. (1996). Memoriale. In Asor Rosa, A. (ed), *Letteratura italiana. Le opere. Il Novecento: La ricerca letteraria*, vol. 4 2. Torino: Einaudi.

Gigliozzi, G. (2008). *Saggi di informatica umanistica*. Milano: UNICOPLI.

Hockey, S. (2004). The History of Humanities Computing. In Schreibman, S., Siemens, R. and Unsworth, J. (eds), *A Companion to Digital Humanities*. Malden, MA, USA: Blackwell Publishing Ltd, pp. 1–19 doi:10.1002/9780470999875.ch1. http://doi.wiley.com/10.1002/9780470999875.ch1 (accessed 25 November 2017).

Jones, S. E. (2016). *Roberto Busa, S. J., and the Emergence of Humanities Computing. The Priest and the Punched Cards*. London: Routledge (accessed 20 October 2017).

McCarty, W. (2014). Getting there from here. Remembering the future of digital humanitiesRoberto Busa Award lecture 20131. *Literary and Linguistic Computing*, 29(3): 283–306 doi:10.1093/llc/fqu022. http://dx.doi.org/10.1093/llc/fqu022.

Morando, S. (ed). (1961). *Almanacco letterario Bompiani: 1962*. Milano: Bompiani.

Mordenti, R. (2001). *Informatica e Critica Dei Testi*. Roma: Bulzoni.

Nyhan, J. and Flinn, A. (2016). *Computation and the Humanities. Towards an Oral History of Digital Humanities*. (Springer Series on Cultural Computing). Cham: Springer International Publishing doi:10.1007/978-3-319-20170-2. http://link.springer.com/10.1007/978-3-319-20170-2 (accessed 16 November 2017).

Orlandi, T. (1990). *Informatica Umanistica*. Roma: La Nuova Italia Scientifica.

Orlandi, T. (1992). Informatica umanistica: realizzazioni e prospettive. *Calcolatori e Scienze Umane: Archeologia e Arte, Storia e Scienze Giuridiche e Sociali, Linguistica, Letteratura*. 1. ed. Milano: Etaslibri, pp. 1–22.

Ricciardi, M. and Bonadonna, F. (eds). (1994). *Oltre Il Testo: Gli Ipertesti*. (Scienze Umane e Nuove Tecnologie 1). Milano, Italy: F. Angeli.

Stoppelli, P. (2005). Dentro la LIZ, ovvero l'edizione di mille testi. *Ecdotica*(2): 42–59.

Terras, M., Nyhan, J. and Vanhoutte, E. (eds). (2013). *Defining Digital Humanities - A Reader*. Williston: Ashgate http://www.ashgate.com/isbn/9781409469636.

Linked Books: Towards a collaborative citation index for the Arts and Humanities

Giovanni Colavizza
giovanni.colavizza@epfl.ch
EPFL, Digital Humanities Laboratory, France; The Alan Turing Institute, United Kingdom; Odoma Sàrl, Switzerland

Matteo Romanello
matteo.romanello@epfl.ch
EPFL, Digital Humanities Laboratory, France; Odoma Sàrl, Switzerland

Martina Babetto
babetto.martina@gmail.com
Tate Britain, United Kingdom

Vincent Barbay
vincent.barbay@gmail.com
EPFL, Digital Humanities Laboratory, France

Laurent Bolli
laurent.bolli@odoma.ch
EPFL, Digital Humanities Laboratory, France; Odoma Sàrl, Switzerland

Silvia Ferronato
silvia.isotta@gmail.com
EPFL, Digital Humanities Laboratory, France

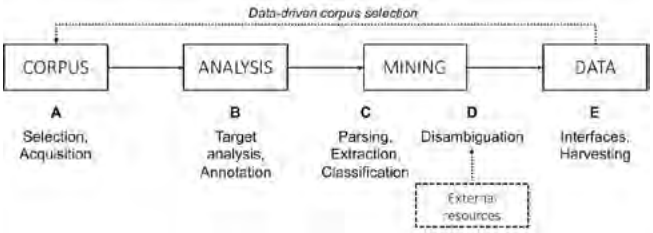
Frédéric Kaplan
frederic.kaplan@epfl.ch
EPFL, Digital Humanities Laboratory, France

We present the Scholar Index: a platform to index the literature and primary sources of the arts and humanities through citations. These resources are becoming increasingly digital, thanks in part to digitization campaigns and a shift towards digital publishing. Nevertheless, the coverage of commercial citation indexes is still poor and mostly limited to publications in English (Mongeon and Paul-Hus, 2015). This situation results in an untapped opportunity, as the literature refers to a wealth of primary sources from institutions such as libraries, archives and museums (Knievel and Kellsey, 2005). As a consequence, a comprehensive indexing of its citations would constitute a unique opportunity to greatly enhance the search capacities of scholars and interconnect collections currently set apart.

The Scholar Index integrates a pipeline to extract citations from scholarly literature in the arts and humanities, along with two interfaces: a digital library (Scholar Library) and a citation index (Scholar Index proper). The prototype Venice Scholar is presented, covering the literature on the history of Venice and currently indexing nearly 3000 volumes of scholarship from the mid 19th century to 2013, from which some 4 million references were extracted. The full citation indexing allowed us for the first time to highlight trends in the large-scale use of archival evidence and scholarly literature made by historians over such a substantial span of time (Colavizza, 2017a, b). We finally argue that a collaborative approach to the indexing of the literature and primary sources of humanists is feasible and would allow to greatly broaden and enrich access to the documentary cultural heritage at large.

Approach

The process of mining citations from digital or digitized scholarly publications entails a set of steps, as sketched below. First of all, a corpus needs to be selected and digitally acquired (including its full-text via OCR). Secondly, the literature needs to be analysed in order to grasp the location of references (typically in footnotes), and the presence of trends in the style of references. These insights can inform a selection of publications to be manually annotated in order to improve the quality of the subsequent automated extraction.



Citation mining can be divided into two tasks: parsing and extraction of references – or the identification of text segments containing a reference to a source – and their disambiguation – or the association of a reference to the unique identifier of the referred source. Having done this, a citation is represented as a relation between a citing publication and a cited source. During parsing, pre-trained text classifiers are used, possibly with adaptation to the domain at hand. During disambiguation, external repositories such as catalogues are optionally queried to establish interlinks. Lastly, citation data can be exposed for a variety of purposes, including search and browsing in a dedicated interface. For more details and evaluations see (Colavizza and Romanello, 2017; Colavizza et al., 2017).

The Venice Scholar

This approach has been applied to create a prototype on the historiography on Venice. There is a sheer amount of literature on Venice, even just considering modern historiography from the 19th century onwards (Dursteler, 2013). We selected the corpus using a variety of means available in research libraries (Colavizza et al., 2017). Once a first seed of literature had been digitized, we proceeded to further expand it with highly cited, usually old sources, as

well as very recent ones. The corpus currently counts over 3000 volumes, circa 20% journal issues and 80% books. This effort has been made possible thanks to the support of several research libraries in Venice.[1] The resulting data has been published in open access: circa 40,000 annotated references used to train reference parsers (Colavizza and Romanello, 2017), while citation data from nearly 4 million extracted references is gradually being ingested into OpenCitations (Peroni et al., 2017), a repository of open citations data.[2]

The platform

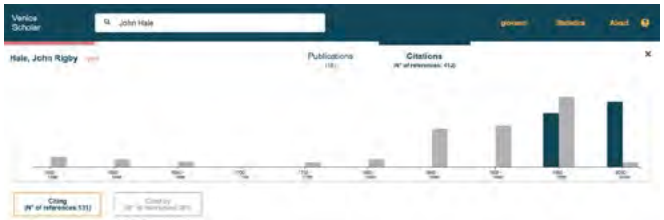
The digital library and the citation index are connected through citations. The interfaces are accessible online.[3] The digital library provides access to the digitized materials, and points to the index through disambiguated references, as shown below.

Viewer (above): allows to read a publication with image and text side by side. This is particularly important in order to appreciate the quality of the OCR.

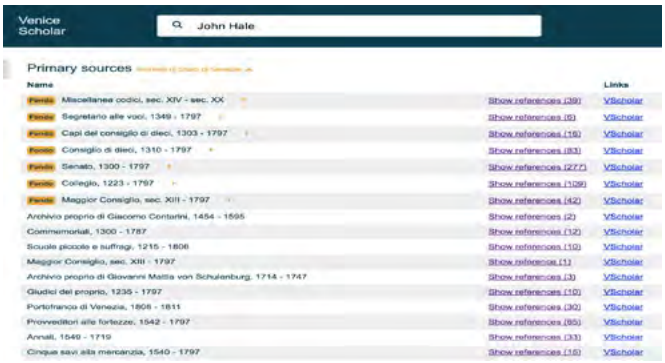
Text view (left): allows the user to search within the full-text of a publication, highlighting all extracted references and links to the relative entries in the index.

The citation index provides instead no access to full-contents – due to reasons of copyright – but allows for the exploration of the network of citations.

Search results (left): citation data is aggregated per author, publication or primary source, with full-text access to the text of extracted references. Search results are conveyed along with their relevant citation information (citations made and received, publications for an author). Authors are linked to the Virtual International Authority File (VIAF) repository, whenever possible.



Citation timeline (left): every aggregated entity has a dedicated page with a timeline of citations (made and received), and a list of relevant sources.



Citations to primary sources (left): the index also links to external collections of primary sources, in this case documentation at the Archive of Venice. Citations to any level of the archival hierarchy are provided, following its structure. The user can easily move from a publication to a document series and see all publications which referred to it, over time.

The platform is thus able to aggregate citation data from many library collections into a unique system, allowing users to not only navigate the resulting network, but also have improved access to collections of primary sources such as archives.

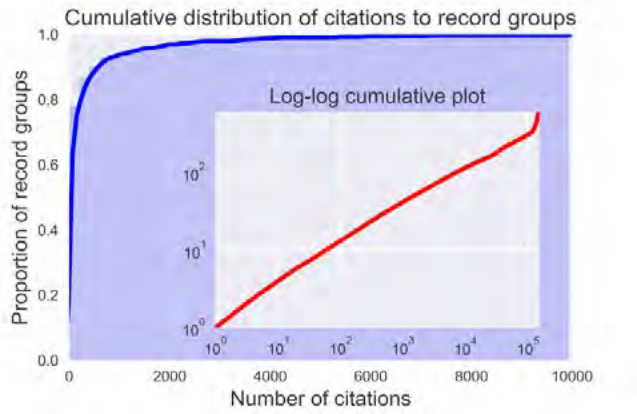
1 For the list of partners see: https://dhlab.epfl.ch/page-127959-en.html.

2 See also https://opencitations.wordpress.com/2018/03/23/early-adopters-of-the-opencitations-data-model/.

3 The (Venice) Scholar Index can be accessed at: www.veniceschol-ar.eu, the (Venice) Scholar Library can be accessed at: www.venice-scholar.eu/library. Try searching for historian "Patricia Fortini Brown", for example. The project's website is at: www.scholarindex.eu.

Citation coverage of the Archive of Venice

The Venice Scholar allowed for the first time to analyse the use made by historians of the vast Archive of Venice over almost 200 years of scholarship. The Archive hosts an estimate 80 linear km of documents and is the main reference for the history of the city. We extracted 157,575 citations to 600 distinct record groups (record groups or smaller series of documents within). Two patterns readily emerge: first, the use of documentary records is highly skewed with few record groups accruing most citations; second, the archival indexation of the records through metadata is key for their discoverability by historians.

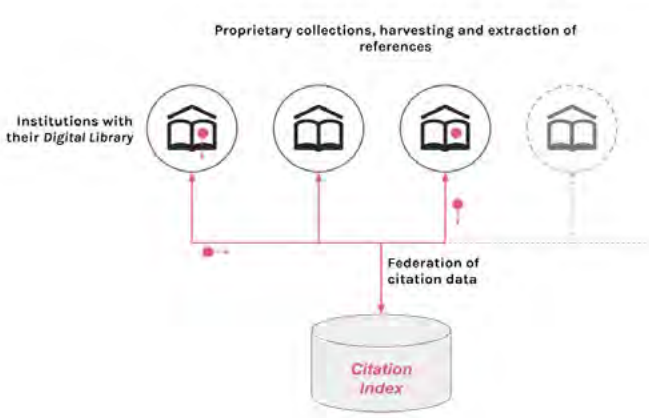


The 600 cited record groups vary from one with almost 10,000 received citations to many having been cited only once. The cumulative distribution of citations suggests the presence of a power law (*above: the inset plot shows the log-log cumulative sum resulting in an almost straight line*).

The Archive possess 14,233 record groups indexed in its information system, meaning only 4,2% of these have been cited from our corpus. Within these record groups, only 25,7% possess information regarding their size in linear meters, which is usually missing for records not yet properly inventoried. Yet 78,5% of cited record groups possess size information: a strong indication of the importance of archival indexation for access. A total of only 18,37 linear km are known to the information system, and of these 64,3% are cited at least once at the aggregated group level. In conclusion, a proportion of ~4% (by archival identifier) to ~15% (by size, (18,37*0.643)/80) of the record groups of the Archive of Venice has been cited from our corpus, and with few aggregates getting most citations: we might conclude that there is still much to explore at the Archive of Venice.

Towards a global citation index for the Arts and Humanities

We believe that the approach used for Venice could be applied to many more collections of scholarly literature, from a variety of libraries. Much in the same way national or international library catalogues are collaboratively created, every library part of the system could take responsibility for an area of scholarship of its interest. This would entail for the library to be in charge for digitization. Once done, the platform would proceed to OCR and mine citations, in view of their federation into a single citation index (*left*). The library could also be responsible for the quality of the so provided citation data, by running regular evaluation and correction campaigns according to its resources. A daunting volume of work would thus be divided into more manageable chunks, and possibly distributed among several players.



The expansion of digital informational ecosystems promises to greatly impact the work of humanists. Besides providing for more rapid access, digital indexing might also make information retrieval a richer experience. Towards this end, we proposed the Scholar Index: an approach to use citations contained in the scholarly literature to index and interlink collections of primary and secondary sources. We believe that our approach has the merit of being able to scale, by catalysing the joint efforts of knowledge institutions towards a common goal, as was shown for the Venice Scholar prototype.

References

Colavizza, G. (2017a). "The Core Literature of the Historians of Venice." *Frontiers in Digital Humanities*.

Colavizza, G. (2017b). "The Structural Role of the Core Literature in History." *Scientometrics*.

Colavizza, G., Romanello, M. and Kaplan, F. (2017). "The references of references: a method to enrich humanities library catalogs with citation data." *International Journal on Digital Libraries*.

Colavizza, G. and Romanello, M. (2017). "Annotated references in the historiography on Venice: 19th-21st centuries." *Journal of Open Humanities Data*.

Dursteler, E. R. (2013). "A brief survey of histories of Venice." In: *A Companion to Venetian History, 1400–1797*, edited by Eric R. Dursteler. Leiden: Brill.

Knievel, J. E. and Kellsey, C. (2005). "Citation analysis for collection development: a comparative study of eight humanities fields." *Library Quarterly*.
Mongeon, P. and Paul-Hus, A. (2015). "The journal coverage of Web of Science and Scopus: a comparative analysis." *Scientometrics*.
Peroni, S., Shotton, D. and Vitali, F. (2016). "Freedom for bibliographic references: OpenCitations arise." In *Proceedings of 2016 International Workshop on Linked Data for Information Extraction*.

Organising the Unknown: A Concept for the Sign Classification of not yet (fully) Deciphered Writing Systems Exemplified by a Digital Sign Catalogue for Maya Hieroglyphs

Franziska Diehr
diehr@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Sven Gronemeyer
sgronemeyer@uni-bonn.de
University of Bonn, Department for the Anthropology of the Americas, Germany; La Trobe University, Department of Archaeology and History, Australia

Christian Prager
cprager@uni-bonn.de
University of Bonn, Department for the Anthropology of the Americas, Germany

Elisabeth Wagner
ewagner@uni-bonn.de
University of Bonn, Department for the Anthropology of the Americas, Germany

Katja Diederichs
katja.diederichs@uni-bonn.de
University of Bonn, Department for the Anthropology of the Americas, Germany

Nikolai Grube
ngrube@uni-bonn.de
University of Bonn, Department for the Anthropology of the Americas, Germany

Maximilian Brodhun
brodhun@sub.uni-goettingen.de
Göttingen State and University Library, Germany

How can the unknown be organised? When working with a script and language that has not been (completely) deciphered yet, primarily an inventory of all signs used must be compiled. What at first seems to be a diligent but routine piece of work, quickly turns out to be a complex classification task, for there is still much unknown. Questions arise about a signs' use-context, the extent of the sign inventory and, above all, how the signs can be classified. Particularly the unambiguous identification of signs has some pitfalls. The difference in meaning of a sign is determined on the one hand by its graphic representation (grapheme) and on the other hand by its phonetic value (phoneme). Since the latter can only be achieved by decipherment work that has already been done, the error-free classification of undeciphered signs is a challenging task. In addition, the investigation of archaic texts creates different contexts that lead to different interpretations. The resulting hypotheses must be included in the classification and decoding tasks of a script if a resilient research basis is to be created.

In our talk we present a concept for the classification and systematisation of characters for writing systems that have not been (completely) deciphered yet. We applied this concept to the Maya hieroglyphs and created a digital sign catalogue that was developed in close interdisciplinary cooperation between epigraphy and information science and technology. The digital sign catalogue can be used to identify, systematise and classify signs and it also offers a starting point for further analyses that can present reliable deciphering.

We took an ontological approach to model the sign catalogue. To use this type of knowledge representation for the classification of signs is a novel approach to digital epigraphy.

Characteristics of Classic Maya Writing

The Maya hieroglyphic script was used between 350 BC and 1550 AD in southern Mesoamerica to record the high level language of Classic Mayan (Wichmann, 2006). The exact number of signs has not yet been determined, it varies from approximately 500 to 1000. Even though Classic Mayan is called a hieroglyphic script due to its iconic character, typologically it is a logo-syllabic writing system in which logograms and syllabograms form the main sign classes. Logograms designate specific terms, such as **PAKAL** (shield). Syllabograms represent open syllables and are also used as phonetic complements of logograms. Words could be written from logograms only (**PAKAL**), syllabograms (**pa-ka-la**), or a combination of both (**PAKAL-la**), see Fig. 1 (Montgomery, 2002). Other sign functions are numerals and diacritic signs. The signs are usually arranged within a hieroglyph block, which

forms a word or a compound of morphemes, similar to Korean Hangul. However, due to its wide range of variants, Maya reveals a much greater degree of calligraphic freedom. Depending on space requirements and aesthetics, graphs can be conflated, infixed or rotated. The formation of graph variants is so complex, that it is a particular challenge to determine the grapheme of a sign.

Maya signs can be polyvalent. A sign can have more than one functional level and thus readings, either as logograms and or syllables. There is a sign labelled T528 according to Eric Thompson's standard sign catalog for Maya writing (1962) that can be read as the logograms **TUN** (stone) and **CHAHUK** (a day name), and also as the syllable **ku**. The graph variants of this sign do not indicate which of the readings is present.

During the investigation of a not yet (fully) decoded script and language, controversial and plausible statements about the deciphering of signs emerge during the research discourse. Each deciphering hypothesis claims to be meaningful in the investigated context. For over 150 years, work has been done on deciphering the Maya script, and the 'birth and death rate' of deciphering-proposals is correspondingly high. The degree of decipherment nowadays ranges between 60 to 80 percent.

Novel Concept of Sign Classification and Modelling of the Digital Sign Catalogue

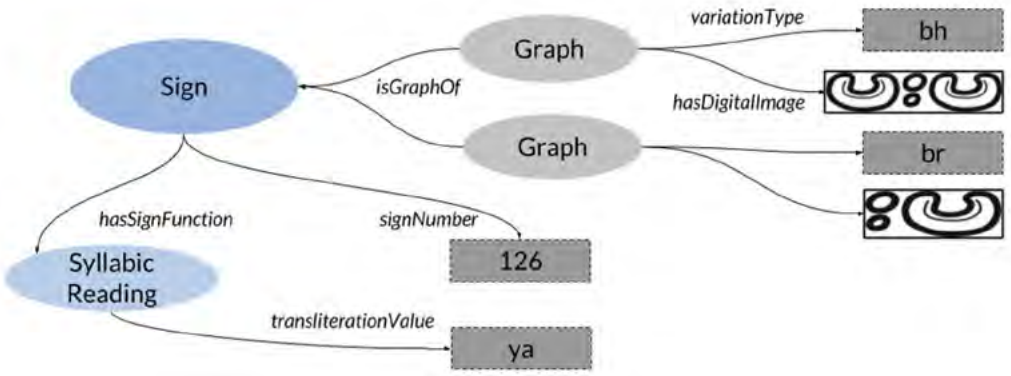
With our digital sign catalogue we want to establish a new concept for the systematisation and classification of signs. We chose a new and unorthodox approach that deliberately differs from previous organisational principles of linguistics and, in particular, from other (Maya) sign inventories. We have specifically questioned traditional practices in order to investigate other ways of systematising and organising signs.

Therefore, we have examined existing classification systems and linguistic terminologies in order to find suitable concepts for describing Maya signs (Department of Linguistics, 2010; Chiarcos and Sukhareva, 2015). We detected that most concepts are not applicable to the development of our catalogue, as they focus too much on the applicability in a specific linguistic context. Thus, they are not applicable to a writing system with a fair deciphering degree, since they are simply not known yet. For this reason, we want to create an organisation system for describing, classifying, and systematising the signs, using linguistic categories only on a meta level and not taking further analysis levels and grammars into account.

The decipherment of Maya signs can only be done by linguistic analyses on the basis of a corpus. To be able to create such a corpus, the signs used in the texts must be identified. In order to allow the processes of sign identification and subsequent text analysis to be interlinked, an organisational system is required that can react flexibly to changes. We achieve the necessary flexibility through ontological-based modelling. To optimally represent the semantic relations between the described entities, the data model of the sign catalogue was implemented in RDF. The documentation of signs in an ontologically based knowledge organisation system has not been done in Maya epigraphy yet and thus represents a new approach in the exploration of this script.

In our catalogue we define the sign as an entity consisting of a functional and phonemic level and a graphical representation. We modelled the class **Graph** which represents all variants of a grapheme (allographs). By the separate recording of discrete graphs we enable an exact method for their identification. Based on preliminary work by the research community (Kelley, 1962; Houston, 2001), for the first time we were able to develop rules and principles on creation of graph variants of Maya signs. 45 variation types in total were defined, which are subdivided into nine classes.[1]

The **Graph**-class is set in relation to the functional and phonemic level of the sign (the class **Sign**, see Fig. 2). This relation is optional, so that even graphs can be recorded that could not have been assigned to any sign yet.

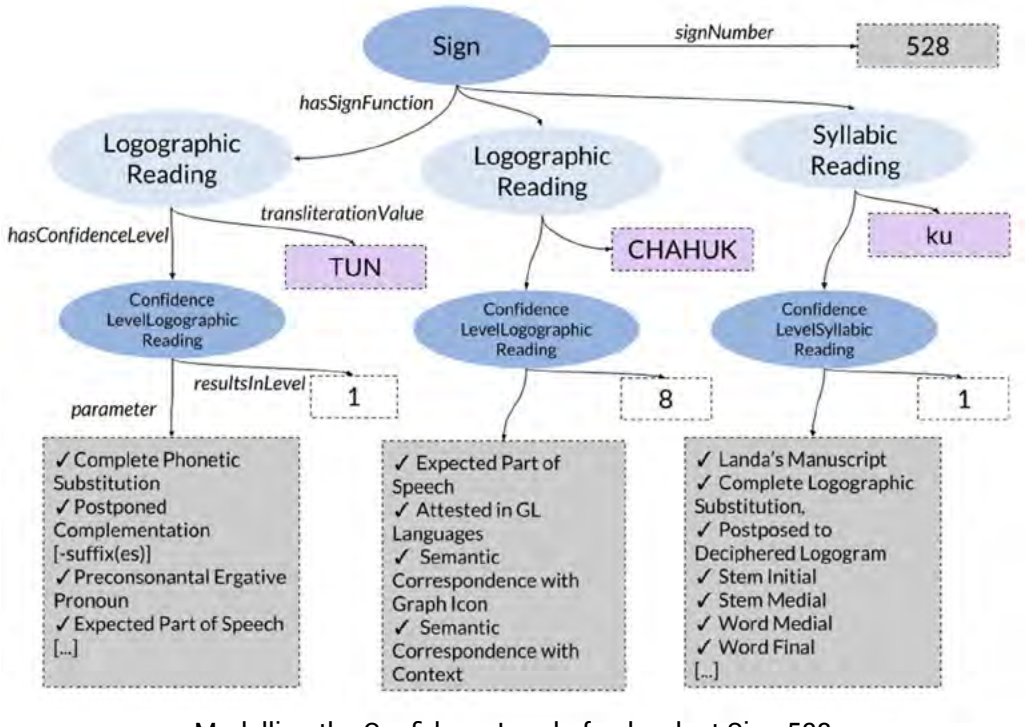


Sign 126 with graph variants bh and br, and phonetic value /ya/

1 mono -, bi-, tri-, and variopartite, division, animation head, animation figure, multiplication and extraction.

The class **Sign** is determined by its function: the use of the sign as a logogram, syllabogram, numeral, or diacritic. The phonemic level of the sign is recorded as the transliteration value with the respective **SignFunction**. Only one value is allowed per function, but one sign can have several sign functions and therefore readings.

In Maya epigraphy, signs that have not yet been deciphered arouse a lively discourse, from which constantly new proposals for their reading emerge. We are faced with the challenge of not only documenting existing as well as our own deciphering hypotheses, but also evaluating and classifying them qualitatively in such a way that they can withstand a critical examination and be used for further (linguistic) analyses. The hypotheses have different quality levels. Some seem more plausible than others. In order to make the quality of the readings formally assessable, we have developed a set of criteria for each sign function, which is oriented, among other things, to the context of use (e. g. plausible text-image-reference) or the proof in modern Mayan languages.[2] The criteria are related by means of propositional logic so that, depending on their combination, a quality level is determined. To represent these in the model, we developed the class **ConfidenceLevel** that is related to the **SignFunction**-class. Therefore, a qualitative evaluation can be made for the transliteration value (see Fig. 3).



Modelling the ConfidenceLevel of polyvalent Sign 528

The qualitative evaluation is particularly relevant to examine the plausibility of reading hypotheses in a corpus. Readings with a high level can be compared with those with a low level. New criteria for the plausibility could always be found in other texts and can then be added to the sign catalogue. This may also increase the quality level of the transliteration value.

Conclusion

Since the concept distinguishes itself specifically from the classification of signs in linguistic categories, it is also transferable to other languages that have not yet been (completely) deciphered or under debate regarding their nature, e.g. Nahuatl, Elamite, Indus, or Rongorongo writing. In particular, the separation of the sign in a graphic and a functional-phonemic level - that can be related to each other depending on the level of knowledge - offers a flexibility that redefines the classification of signs and allows precise identification on the basis of distinguishing characteristics. The ontological modelling approach and the implementation in a RDF data model enables new insights into character classification. By incorporating known and adapting new results, the digital sign catalogue is specifically designed to deal with ambiguousness in research processes.

References

Chiarcos, C. and Sukhareva, M. (2015). OLiA - Ontologies of Linguistic Annotation, *SWJ (Semantic Web Jour-*

2 Notice that these criteria can only be derived from the deciphering work already carried out.

nal), 6(4): 379-386.
Department of Linguistics (2010). General Ontology for Linguistic Description (GOLD), http://linguistics-ontology.org/ (accessed 24 April 2018).
Houston, S. (2001). Introduction. In Houston, S. D., Chinchilla, M. O. F., & Stuart, D. (eds), The Decipherment of Ancient Maya Writing. Norman, OK: University of Oklahoma Press, pp. 3-19.
Kelley, D. H. (1962). Review of A Catalog of Maya Hieroglyphs by J. Eric S. Thompson. *American Journal of Archaeology* 66 (1962): 436-438.
Montgomery, J. (2002). *How to Read Maya Hieroglyphs*. New York, NY: Hippocrene.
Thompson, J. E. S. (1962). *A Catalog of Maya Hieroglyphs*. Norman, OK: University of Oklahoma Press.
Wichmann, S. (2006). Mayan Historical Linguistics and Epigraphy: A New Synthesis. *Annual Review of Anthropology*, 35: 279-294.

Automated Genre and Author Distinction in Comics: Towards a Stylometry for Visual Narrative

Alexander Dunst
dunst@mail.upb.de
University of Paderborn, Germany

Rita Hartel
rst@upb.de
University of Paderborn, Germany

Introduction

Stylometry has a long tradition in linguistics and literary studies and has only gained in popularity with the digitization of text corpora and out-of-the-box tools (Holmes and Calle-Martin & Miranda-García). Stilometric methods for paintings have been advanced in digital art history but remain at an early stage of development (Qu, Taeb & Hughes; Manovich). Stylometric analyses for visual narratives are not yet established. Visual narratives include film and TV, comics and other illustrated print literature, and to an extent computer games, constituting some of the most popular cultural formats of the twentieth and twenty-first centuries. The relative lack of research in this area may be traced to the technical hurdles of image analysis and the absence of suitable corpora. This paper will introduce a method for visual stylometry in comics based on the analysis of a corpus of 209 book-length graphic narratives. In closing, we explore how the method may be applied to other media.

1 Lev Manovich applied stylistic description to manga but his studies remained explorative and did not offer an analysis of categories such as author or genre.

Corpus & Data Analysis

Our analysis is based on the Graphic Narrative Corpus (GNC), the first representative collection of what is commonly called graphic novels (Dunst et al.). The GNC was conceived as a stratified monitor corpus and defines graphic narratives as comics of more than 64 pages in length that tell one continuous or closely-related stories and are aimed at an adult readership. Due to the absence of reliable bibliographies, the total population remains unknown. A random sample is therefore not feasible. To avoid bias, we sampled from a wide array of sources: academic and general audience databases, library collections, international comics prizes, Amazon.com bestseller lists, literary histories, surveys of comics scholars, and media reports. At the time of analysis in November 2017, 209 full-length graphic narratives running to nearly 50.000 pages had been digitized and checked for scanning artefacts.

The focus on image analysis is due to both methodological and practical reasons: stylometric methods for text analysis are more established and are being continuously improved by an existing research community. These methods can be directly applied, or easily adapted, for analyzing text in comics. Automatic text localization and OCR for comics still represent work in progress, and text can not yet be extracted automatically with sufficient quality. This leaves time-consuming manual annotation as the only option, which excludes the analysis of large corpora. Visual style thus represents the most promising avenue for distinguishing between authors and genres. We used five basic measures for analysis, all of which are low-level features that are commonly used in computer vision and information theory. In all these cases, we were interested in finding significant relationships between these measures as indicators of visual style and the critical concepts we are investigating, i.e. genre and authorship.

- **Median Brightness**: the mean value of all brightness values of all pixels of a page. We transformed each page into a grayscale image by computing the Luma of each pixel, i.e., the weighted sums of the gamma-compressed R′G′B′-values of the image.
- **Shannon Entropy**: the expected value of the information in a message. The entropy H(X) of a message $X=(x_1,\ldots,x_n)$ of length n is defined to be $H(X):=S_{i=1..n}P(x_i)*\log_2(P(x_i))$. The message X of the entropy is the list of the brightness values of each pixel, with the x_i range between 0 and 255. In addition, n is the total number of pixels. As $P(x_i)$ denotes the probability or relative frequency of item x_i, we can compute $P(x_i)$ for a given x_i by $P(x_i):=$ (Number of pixels having value x_i)/(n=total number of pixels).
- **Number of Shapes**: describes an image's agitation. To yield normalized values, we scaled each image to a height of 250 pixels. We first split grayscale images into 5 sub-images of different brightness levels and

then measured individual sub-images and filled unconnected areas up to a diameter of four pixels. In a final step, we discounted components that came to less than ten pixels in size.

- **Color Layout**: A color layout descriptor (CLD; MPEG 7) captures the spatial distribution of color using the YCbCr color space. The extraction process consists of image partitioning, representative color selection, discrete cosine transform, and zigzag scanning. The color components Cb and Cr represent the range of blue and red present in an image.
- **Edge Types**: the edge histogram descriptor (MPEG 7) calculates the frequency of different edges in an image: vertical, horizontal, 45° diagonal, 135° diagonal, and non-directional. Each image is divided into 4x4 subframes. Each subframe consists of five bins, each of which represents the different edge types. Subframes are divided into non-overlapping blocks to extract edge types and bin values are normalized by the total number of blocks in the subframe.

After calculating the five basic measures, we derived the median for each of the 209 graphic narratives. To analyze stylistic variation within individual narratives, we calculated standard deviation from each of the five measures. We performed Anova and Tukey's HSD, which are standard statistical methods for testing for significant differences among the means of more than two samples, with $p<0.05$.

Results & Discussion

Genre

The GNC consists of fictional and non-fictional texts, including graphic memoirs and journalism, which are often summarized under the somewhat misleading umbrella term graphic novel. We assigned 23 subgenre categories using plot summaries and information provided by publishers. Their distribution can be seen in figure 1. Subgenres were grouped into six larger categories for analysis: graphic novel, graphic memoir, other non-fiction, humor, graphic fantasy, and miscellaneous.

Figure 1: Larger genre categories are indicated by color ranges: graphic novel (red), graphic memoir (green), other non-fiction (blue), humor (yellow), graphic fantasy (purple), and miscellaneous (gray).

Results show highly significant distinctions for graphic novel, graphic memoir, and graphic fantasy across several measures. Graphic memoirs (including such canonical text as Spiegelman's *Maus* and Bechdel's *Fun Home*) are brighter, show less color variation (cb & cr), and are more regular in their visual style than other genres. Regularity of visual style can be seen in the lowest median scores for entropy and the high frequency of horizontal edges. Graphic fantasy is significantly darker, while showing the highest entropy and lowest number of horizontal edges. Graphic fantasy also distinguishes itself by the highest amount of color variation. Graphic novels are situated between the two extremes of graphic memoirs and fantasy, yet are statistically distinct in their visual style. The measure number of shapes did not return significant results, while the edge histogram only did so for horizontal edges. The boxplots in figures 2-4 show results for entropy, brightness, and horizontal edges.

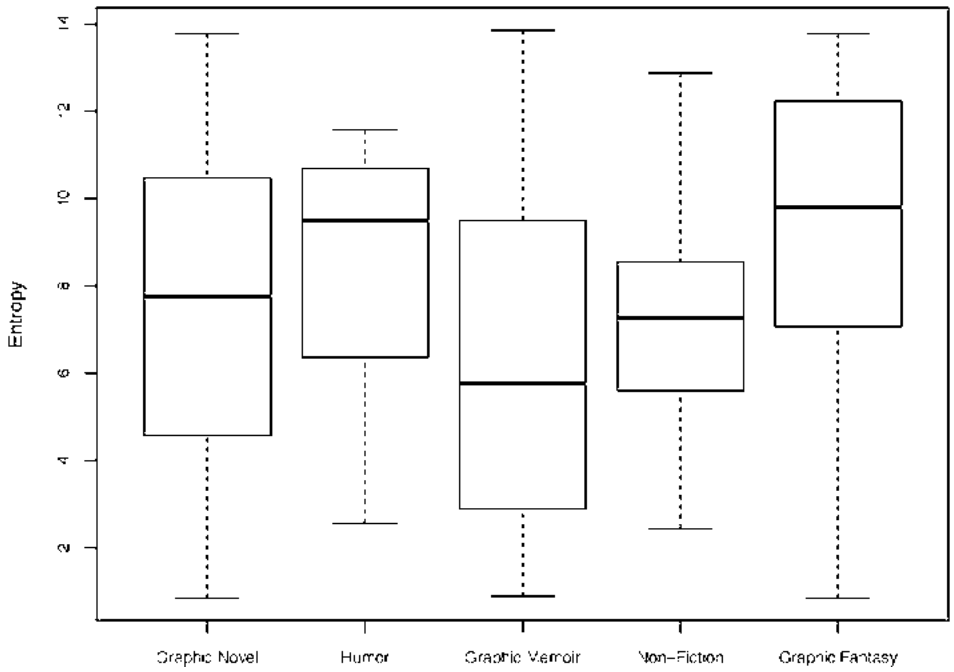


Figure 2: Boxplot Entropy: Graphic Fantasy – Graphic Memoir (p<0.003)

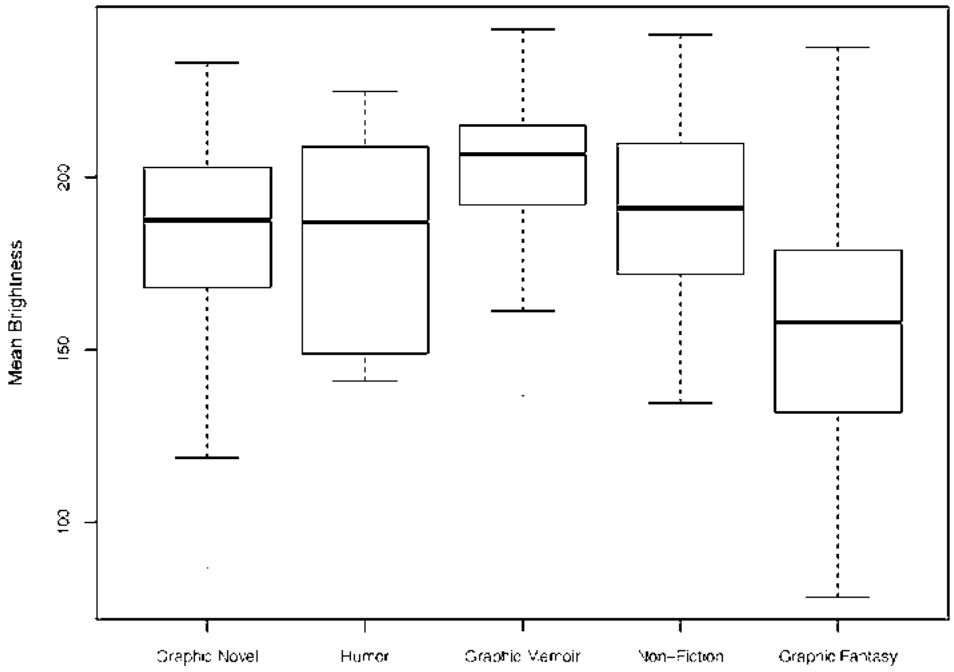


Figure 3: Boxplot Mean Brightness: Graphic Memoir – Graphic Novel (p<0.016); Graphic Fantasy – Graphic Novel (p<0.000)

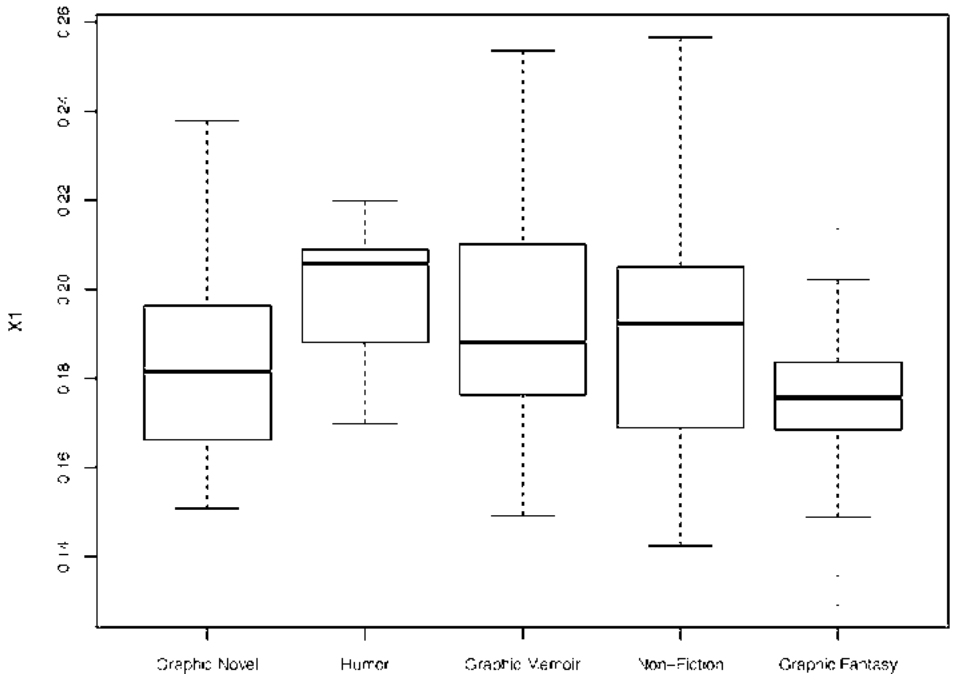


Figure 4: Boxplot Horizontal Edges: Graphic Fantasy – Graphic Memoir ($p<0.001$)

Authorship

The GNC includes several authors that are represented with more than one graphic narrative. The GNC also contains information on single authorship, or collaborations between one writer and one illustrator, or multiple authors. Results returned highly significant distinctions for individual authors and for authorship categories (single, two, and multiple authors). Works by authors such as Neil Gaiman and Frank Miller show consistently higher entropy and a comparatively higher mean brightness than other authors, while the opposite holds for Will Eisner, for instance. Results align with genres in which these authors publish, respectively, graphic fantasy versus graphic novel and memoir. Similarly, the number of shapes and the variation in mean brightness are significantly lower for authors who publish in the latter genres. Individual and multiple authorship also results in distinct visual styles. Graphic narratives written by a single author show lower entropy and number of shapes, are brighter and less colorful, and contain fewer diagonal edges (45° and 135°). Results were highly significant, with $p<0.01$ throughout. Figure 5 and 6 visualize entropy for individual authors and number of shapes for authorship categories.

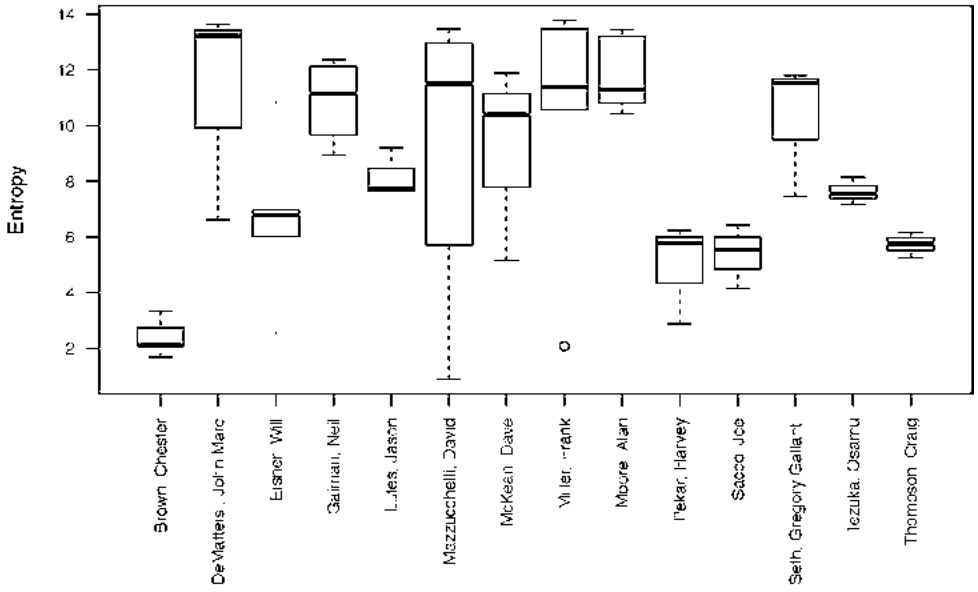


Figure 5: Boxplot Entropy Authors with >3 titles

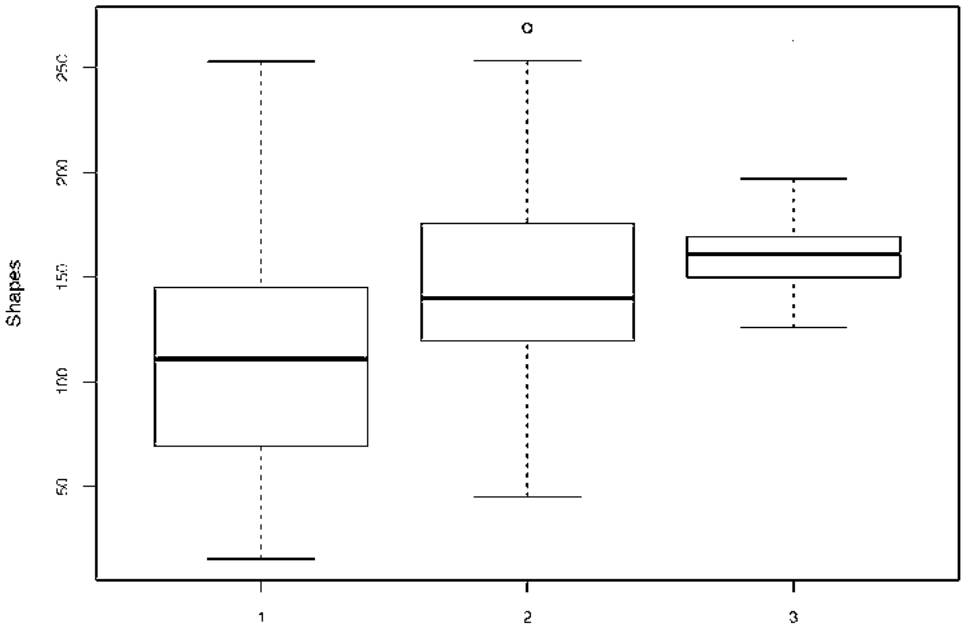


Figure 6: Boxplot Number of Shapes for Authorship Categories: 1 – 2 Authors ($p<0.001$); 1 – >3 Authors ($p<0.001$)

Outlook and Future Work

We introduced image analyses that adapt stylometric distinctions to visual narrative. As our paper shows, comics grouped together under authorship or genre affiliation share numerous visual traits. The correlation between author and genre categories indicates that we need to disentangle these signals. We are working on neutralizing the author signal by penalizing texts from the same writer and will integrate this approach in time for DH 2018 (Tello et al.). As examples of hand-drawn still images, comics have stylistic traits that distinguish them from moving image narratives such as film and TV. Thus, the visual descriptors used here may be adapted most readily to other forms of graphic art, including drawings, woodcuts, and lithographs. Given that the measures we used are highly generic and low-level features, the method also has potential for other media in which the concepts of genre and authorship play a role. Thus, they could be adapted for investigating authorship in film, for instance.

References

Calle-Martin, J. & A. Miranda-García (2012). "Stylometry and Authorship Attribution: Introduction to the Special Issue" *English Studies* 93-3: 251-58.

Dunst, A., R. Hartel, and J. Laubrock (2017). "The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities" in *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)*, 15-20, DOI: 10.1109/ICDAR.2017.286.

Holmes, D (1998). "The Evolution of Stylometry in Humanities Scholarship" *Literary and Linguistic Computing* 13-3: 111-17.

Manovich, L., J. Douglas, T. Zepel (2011). "How to Compare One Million Images", http://manovich.net/index.php/projects/how-to-compare.

Qi, Hanchao, Armeen Taeb & Shannon M. Hughes (2013). "Visual stylometry using background selection and wavelet-HMT-based Fisher information distances for attribution and dating of impressionist paintings" *Signal Processing* 93-3: 541-53.

Tello, José Calvo, et al. (2017)."Neutralising the Author Signal by Penalization: Stylometric Clustering of Genre in Spanish Novels." *DH 2017: Conference Abstracts*, 181-184.

Social Knowledge Creation in Action: Activities in the Electronic Textual Cultures Lab

Alyssa Arbuckle
alyssaa@uvic.ca
University of Victoria, Canada

Randa El Khatib
khatib@uvic.ca
University of Victoria, Canada

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Digital environments now serve as the primary network for academic and non-academic modes of communication, research practices, and knowledge dissemination. This shift has resulted in greater ease of pursuing

collaborative modes of engagement. Social knowledge creation, citizen scholarship, interdisciplinary collaborations, and university-community partnerships have become more common and more visible. Engaging with such transformations in knowledge creation has been a significant research focus for the Electronic Textual Cultures Lab (ETCL) at the University of Victoria. This presentation will detail the intellectual foundations of social knowledge creation, as well as the major initiatives undertaken by the ETCL to pursue and enact this research. The ETCL explores these topics via on-campus activities as well as three substantial environmental scans: Social Knowledge Creation: Three Annotated Bibliographies" (Arbuckle, Belojevic, Hiebert, Siemens, et al., 2014), "An Annotated Bibliography on Social Knowledge Creation" (Arbuckle, Belojevic, El Hajj, El Khatib, Seatter, Siemens, et al., 2018), and "Open Social Scholarship Annotated Bibliography" (El-Hajj, El Khatib, Leibel, Seatter, et al., under development). The annotated bibliographies bring together myriad perspectives on how collaborative knowledge creation and engagement practices have been carried out, historically as well as currently. This work suggests how elements of academia might be reimagined in order to effectively integrate collaborative, interdisciplinary, public-minded praxis. Building on field touchstones like Kathleen Fitzpatrick's *Planned Obsolescence* (2011) and John Willinsky's *The Access Principle* (2006), this work proposes that collaboration-driven academic practices in a new media context can create a more critical work environment that integrates creative options for publishing and disseminating research.

"An Annotated Bibliography on Social Knowledge Creation" updates the previously published "Social Knowledge Creation: Three Annotated Bibliographies." The former version was developed in the ETCL in collaboration with the Implementing New Knowledge Environments (INKE) Research Group, and formulated a snapshot of social knowledge creation scholarship and initiatives up to 2013. The revised document draws on more recent scholarship published in this evolving area of inquiry, and expands the scope to include notable subject additions, including public humanities, crowdsourcing, digital publishing, and open access. In both the 2014 and forthcoming instances, resources are chosen according to their relation to our definition of social knowledge creation: "acts of collaboration in order to engage in or produce shared cultural data and/or knowledge products" (1). Many stress the importance of involving citizen scholars to revitalize research and as a way to respond to a crisis that public humanities draws attention to, namely the ever-expanding gap between the university and the community. The subject additions of the latter document encapsulate the pressing need to create and strengthen community outreach in academic environments.

In 2016, the ETCL team began compiling the "Open Social Scholarship Annotated Bibliography." According to INKE, open social scholarship involves the creation, dissemination, and engagement of research and research technologies that are accessible and significant to a broad audience. The bibliography draws on research that adopts and propagates these knowledge production ideals that have branched out across movements, including open access, open source, public humanities, citizen scholarship, citizen science, and community outreach, among others. The main trends that are explored include: developing and disseminating research in accessible ways; research that draws on university and community interests and needs; active engagement of community members in academic research practices; and the development of research tools that bring these two communities into productive dialogue and serve their needs. Resources range from traditional, foundational forms of open knowledge and resources to highly praxis-oriented projects. Historical publications, starting with *Philosophical Transactions of the Royal Society of London*, exemplify how knowledge was discussed and debated through publication. Advocacy for open access to information is a recurring theme across many of the included works, with a position that publicly funded research should be accessible to the wider public. In addition to the aforementioned discourses, the bibliography addresses the impact of open knowledge on social justice movements through new mediums, and how Internet tools and social networks have been used to mobilize action in activist movements.

These environmental scans lay the foundation for our ETCL-based Open Knowledge Practicum fellowship, launched in January 2017 with two completed rounds, a third one currently running, and more to follow. This initiative puts open social scholarship into action by inviting faculty, staff, students, and members of the community to pursue their own research projects for an academic term in the ETCL. We provide participants with access to resources, library materials, and archives; consultation and guidance from specialists in the field; and other project-specific assistance. The Open Knowledge Practicum is a step toward more publicly engaged scholarship, ranging from discipline-specific foci to research on local public history or the broader community. Practicum findings are published in online, public venues and made discoverable to both general and targeted communities. As a connecting thread, all fellows create, enrich, or revise Wikipedia pages that relate to their topic. This presentation will showcase a number of projects that came out of the Open Knowledge Practicum, available for review at <http://etcl.uvic.ca/?page_id=1919>.

The ETCL also launched Digital Scholarship Fellowships for the 2017-2018 academic year. Digital Scholarship Fellowships support graduate students, postdoctoral fellows, new and visiting scholars, as well as staff, faculty, and librarians making substantial use of digital and/or social knowledge creation methods to carry out humanities or interdisciplinary research. Individuals across dis-

ciplines are able to join the ETCL community in this way, and to work alongside the team in the ETCL on relevant projects within their area of research.

We consider Wikipedia to be a prime example of social knowledge creation, as it is an online encyclopedia comprised, maintained, and expanded by thousands of citizen scholars. In partnership with the U Victoria Libraries, the ETCL appointed two Honorary Resident Wikipedians: Dr. Christian Vandendorpe (2014–16) and Dr. Constance Crompton (2017). So far, Wikipedia edit-a-thons have oriented toward social justice themes.

Moreover, the ETCL runs campus-based digital skills training initiatives. DHSI takes place annually at U Victoria and will run for the 18th consecutive year in June 2018. Participants from different fields and locations attend DHSI for two weeks of workshops, seminars, and other conference activities. In 2017 DHSI launched a course stream that brings the various open knowledge oriented research foundations discussed here into a pedagogical setting. Courses include: "Open Access and Open Social Scholarship," by Arbuckle (U Victoria), "Digital Public Humanities" by Mia Toothill (Cornell U), "Accessibility and Digital Environments," by Erin E. Templeton (Converse C) and George H. Williams (U South Carolina Upstate), "Ethical Collaboration in the Digital Humanities," by Daniel Powell (King's C London), and "Feminist Digital Humanities: Theoretical, Social, and Material Engagements," by Elizabeth Losh (C William and Mary) and Jessica M. Johnson (John Hopkins U). In DHSI 2018, two new courses will join this stream: "Race, Social Justice, and DH: Applied Theories and Methods" by Dorothy Kim (Vassar C) and David Nieves (Hamilton C), and "Queer Digital Humanities: Intersections, Interrogations, Iterations" by Jason Boyd (Ryerson U) and James Howe (Rutgers U). This course stream addresses the theory, methods, and challenges related to open social scholarship in various settings. The ETCL also hosts training throughout the year, the "Digital Humanities Workshop Series," launched in partnership with DHSI and U Victoria Libraries and affiliated with Simon Fraser U (DHIL, SFU Library Research Commons) and U British Columbia (UBC Library, UBC Advanced Research Computing), which provides students, faculty, and staff, and members of the community with a wide range of technical skills and relevant theoretical basis in various digital humanities subfields. The ETCL activities and research directions we outline in our paper share a commitment to address and practice scholarship that is responsive to the evolving needs of the university and the larger community.

The ETCL strives to produce relevant and accessible scholarship, while simultaneously thinking about ways of harnessing the digital medium to benefit all. ETCL initiatives also address the potential for creating and fostering university-community partnerships. We seek to highlight the ever-expanding social nature of knowledge production and how scholarship has expanded beyond the academic context, as evident in the vast amount of research produced by citizen scholars and citizen scientists.

References

Arbuckle, A., Belojevic N., Matthew H., Siemens R., Wong S., Siemens D., Christie A., Saklofske J., Sayers J., INKE Research Group, & ETCL Research Group. (2014). Social knowledge creation: three annotated bibliographies. *Scholarly and Research Communication*, 5(2), http://src-online.ca/index.php/src/article/view/150/299 (accessed 1 May 2018).

Arbuckle, A., El Hajj., El Khatib R., and Seatter L., with Belojevic N., Hiebert M., Siemens D., and Siemens R.G., and with Christie A., Saklofske J., Sayers J., Wong S., and the INKE and ETCL Research Groups. (2018). An annotated bibliography on social knowledge creation. *New Technologies in Medieval and Renaissance Studies*. https://ntmrs-skc.itercommunity.org/.

El-Hajj, T., El Khatib R., Leibel C. and Seatter L., with Arbuckle A., Siemens R., and the Electronic Textual Cultures Lab. (2016). Open social scholarship annotated bibliography.

Fitzpatrick, K. (2011). *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York: New York University Press.

Willinsky, J. (2006). *The Access Principle*. Cambridge: MIT Press.

Network Analysis Shows Previously Unreported Features of Javanese Traditional Theatre

Miguel Escobar Varela
m.escobar@nus.edu.sg
Department of English Language and Literature,
National University of Singapore, Singapore

Andrew Schauf
andrew.schauf@u.nus.edu
Graduate School for Integrative Sciences and Engineering,
National University of Singapore, Singapore

The methods of network analysis are becoming increasingly relevant to the digital humanities, particularly in relation to the study of literary characters (Moretti 2011; Pohl, Reitz, and Birke 2008; Park et al. 2013; Trilcke, Fischer, and Kampkaspar 2015; Elson, Dames, and McKeown 2010; Agarwal et al. 2012; Waumans, Nicodème, and Bersini 2015; Xanthos et al. 2016; Choi and Kim 2007; Bollen 2017; Fischer et al. 2017). Several recent studies and presentations have focused on drama. Most studies deal with European and American drama and this is possibly due in part due to easily available data.

However, we believe that network analysis can be used to interrogate interesting features of Javanese the-

atre as well. Our research focuses on *wayang kulit* (shadow puppetry), one of the oldest and most respected traditions of Southeast Asia. A typical performance lasts all night but usually focuses on a small episode of the Mahabharata, one of the two major Sanskrit epics of Ancient India that provides the narrative material for many traditional theatre forms in Southeast Asia. There are no comprehensive storylines or transcripts available in digital form, so we created our own database by digitizing and annotating the authoritative list of *wayang kulit* storylines compiled by Purwadi (2009).

We used the resulting data to construct a weighted, undirected co-occurence network at the *adegan* (scene) level. Each character is modeled as a node. An edge between two characters means they are present at the same scene, regardless of whether they interact with each other. The weight indicates the number of scenes in which both characters are simultaneously present. While certain favorite characters appear in many stories, many other characters are only present in one storyline. Thus, the network exhibits the ubiquitous characteristics of real-life social networks (Carrington, Scott, and Wasserman 2005; Knoke and Yang 2008). These include a high-level degree of heterogeneity (the number of stories per character in a bipartite projection decreases according to the distribution $P(x) \approx x^{-3.3627}$, see Figure 1) and small world properties:

- A low clustering coefficient (0.863)
- A low average shortest path (0.86)

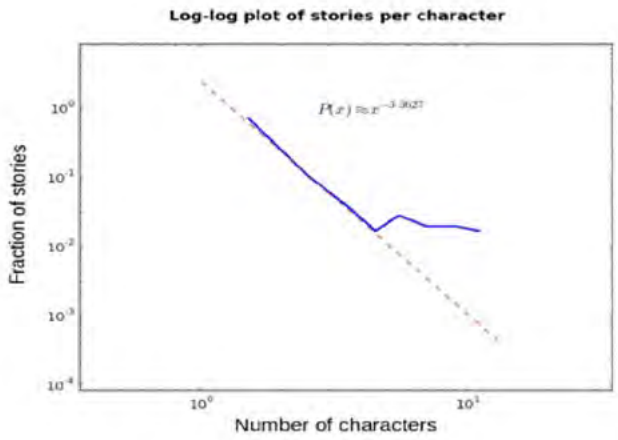


Figure 1. Log-log plot of stories per character in the wayang network (right). The solid lines represent the actual distribution and the dashed line the theoretical power-law distributions.

Characters in Euro-American theatre usually appear in only one play (or at best in a handful of plays) and the kind of structural analysis we conducted here is only possible in a narrative tradition with recurrent characters. We thus feel this contributes an interesting case study to the burgeoning field on network analysis of theatre characters. Beyond merely revealing some interesting quantitative (small-world) properties of this network, further quantitative analysis enabled us to identify previously unreported features of the *wayang* tradition: 1) a network-theoretical perspective reveals some unexpected insights into how various indigenous Javanese elements were integrated into the "structure" of the original Indian epic and 2) there are significant differences in the network properties of characters that can be represented by "interchangeable" puppets and those that can not be changed. To fully appreciate the significance of these findings, a quick overview of the history and performance conventions of *wayang* is needed.

In Java (Indonesia), the recorded history of Mahabharata-derived performances dates back to the 10th century CE, but the performances might have an older history (Escobar Varela 2017). In any case, over this one-thousand year period, Javanese artists have invented a number of new characters and local storylines that they have interwoven with the original Sanskrit epics. However, people still readily acknowledge the stories to be mostly Indian in origin. Much to our surprise, we found out that almost half of the characters used in performance today are local in origin. Our initial estimate (and that of people informally interviewed), put the number of local Javanese characters at only 20% to 30%. Why the discrepancy? By applying network theoretical measures, we found significant differences between the Javanese and the Indian characters. Except for the *punokawan* (the clown-servants), which appear in almost every story, all local Javanese characters have low values for network-theoretical measures such as topological degree and weighted degree (Figure 2). We hypothesize that the reason why people think Javanese characters are less prevalent than they truly are is that, on average, Javanese characters are significantly less "important' in terms of their network-theoretical measurements.

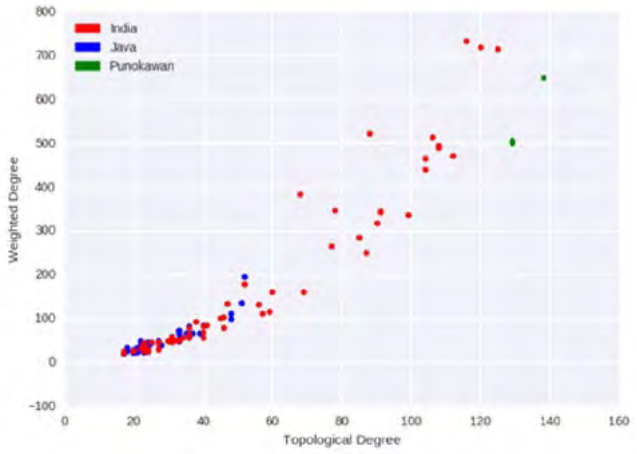


Figure 2. Weighted degree and topological degree of Javanese, Javanese punokawan (clown-servants) and Indian characters.

Our second finding relates to the usage of the puppets in performance. A complete set of puppets can be expensive and often the the *dalang* (puppeteer) would not have all the needed puppets for a given performance. In this case, they have a choice: they can either borrow a puppet from another *dalang*, or they can substitute the required puppet for one that they already posses in their collections. This choice is based on weather the character is considered as "interchangeable" or not. We noted the interchangeability of characters based on interviews with professional *dalang.* When comparing the network-theoretical measurements against this table, we also found that all characters deemed interchangeable have low eigenvector centrality and weighted degree in our network (Figure 3). In other words, the more important a character (in network-theoretical terms), the less likely it is to be exchanged for another character.

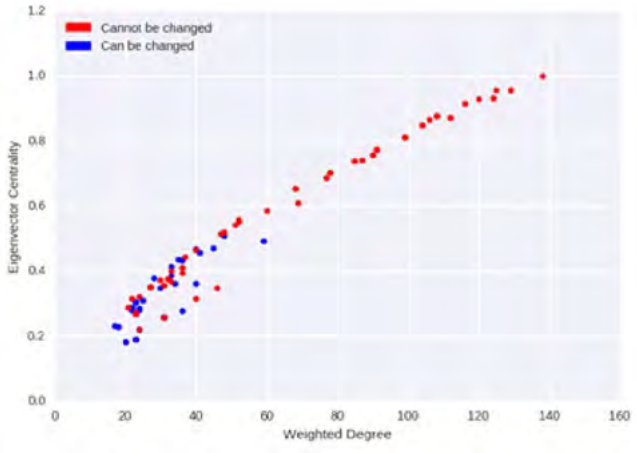


Figure 3. Eigenvector centrality and weighted degree of interchangeable and non-interchangeable characters.

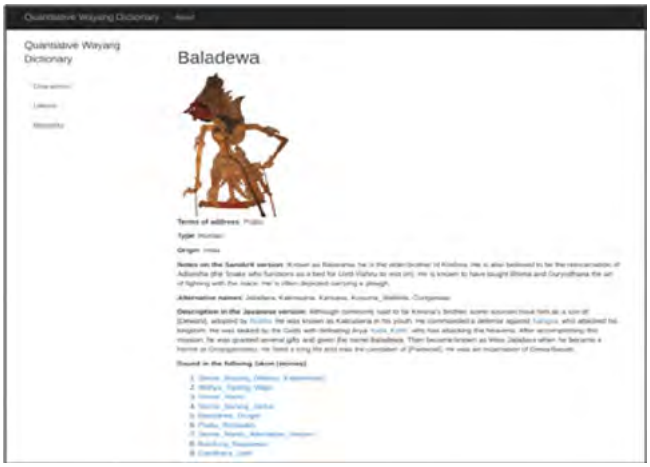


Figure 4. Contextual information for a given character (Baladewa in this example), forthcoming online portal.

Both of this findings seem logical in retrospect, but they were previously unreported. We hope that this application of network analysis can add to the field of network analysis of literary characters and also contribute to the scholarship on Javanese theatre. For this purpose, we are developing an interactive online portal where the contextual information for each character (Figure 4) and the values for network-theoretical measures for each character (Figure 5) and can be consulted in greater detail. All datasets will be openly available for download and we hope this will encourage other research teams to contribute to the quantitative analysis of Javanese theatre.

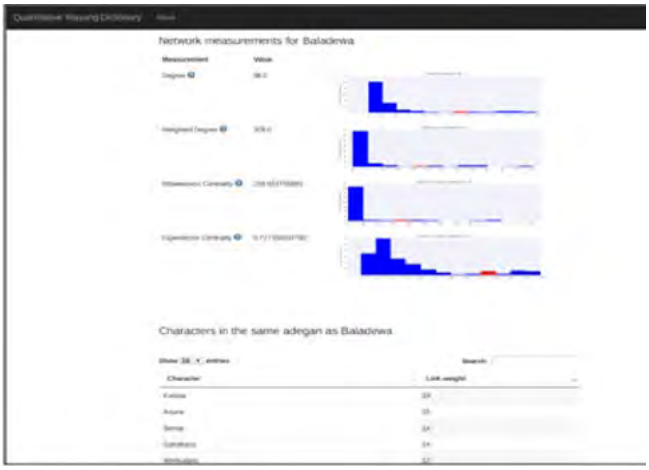


Figure 5. Network-theoretical measures for a given character (Baladewa in this example), forthcoming online portal.

References

Agarwal, Apoorv, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. "Social Network Analysis of Alice in Wonderland." In *CLfL@ NAACL-HLT*, 88–96.

Bollen, Jonathan. 2017. "Data Models for Theatre Research: People, Places, and Performance." *Theatre Journal* 68 (4):615–32. https://doi.org/10.1353/tj.2016.0109.

Carrington, Peter J., John Scott, and Stanley Wasserman. 2005. *Models and Methods in Social Network Analysis*. Vol. no. 28. Structural Analysis in the Social Sciences. Cambridge: Cambridge University Press.

Choi, Yeon-Mu, and Hyun-Joo Kim. 2007. "A Directed Network of Greek and Roman Mythology." *Physica A: Statistical Mechanics and Its Applications* 382 (2):665–71.

Elson, David K, Nicholas Dames, and Kathleen R McKeown. 2010. "Extracting Social Networks from Literary Fiction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 138–47. Association for Computational Linguistics.

Escobar Varela, Miguel. 2017. "From Copper-Plate Inscriptions to Interactive Websites: Documenting Javanese Wayang Theatre." In *Documenting Performance: The Context and Processes of Digital Curation and Archiving*, 203–14. London and New York:

Bloomsbury Methuen Drama.
Fischer, Frank, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, and Peer Trilcke. 2017. "Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts." In *Digital Humanities 2017 (Montréal, 8--11 August 2017). Book of Abstracts*.
Knoke, David, and Song Yang. 2008. *Social Network Analysis*. Vol. 154. Quantitative Applications in the Social Sciences. Los Angeles: SAGE.
Moretti, Franco. 2011. "Network Theory, Plot Analysis." *New Left Review*, no. 68(March):80.
Park, Gyeong-Mi, Sung-Hwan Kim, Hye-Ryeon Hwang, and Hwan-Gue Cho. 2013. "Complex System Analysis of Social Networks Extracted from Literary Fictions." *International Journal of Machine Learning and Computing* 3 (1):107.
Pohl, Mathias, Florian Reitz, and Peter Birke. 2008. "As Time Goes by: Integrated Visualization and Analysis of Dynamic Networks." In *Proceedings of the Working Conference on Advanced Visual Interfaces*, 372–75. ACM.
Purwadi. 2009. *Kempalan Balungan Lakon Wayang Purwa*. Surakarta: Cendrakasih.
Trilcke, Peer, Frank Fischer, and Dario Kampkaspar. 2015. "Digital Network Analysis of Dramatic Texts." In *DH2015 Conference Abstracts. Sydney, Australia*. Vol. 1184.
Waumans, Michaël C, Thibaut Nicodème, and Hugues Bersini. 2015. "Topology Analysis of Social Networks Extracted from Literature." *PloS One* 10 (6):e0126470.
Xanthos, Aris, Isaac Pante, Yannick Rochat, and Martin Grandjean. 2016. "Visualising the Dynamics of Character Networks." In *Digital Humanities 2016: Conference Abstracts*, 417–19.

To Catch a Protagonist: Quantitative Dominance Relations in German-Language Drama (1730–1930)

Frank Fischer
frafis@gmail.com
National Research University Higher School of Economics, Russian Federation

Peer Trilcke
trilcke@uni-potsdam.de
University of Potsdam, Germany

Christopher Kittel
contact@christopherkittel.eu
University of Graz, Austria

Carsten Milling
cmil@hashtable.de
National Research University Higher School of Economics, Russian Federation

Daniil Skorinkin
dskorinkin@hse.ru
National Research University Higher School of Economics, Russian Federation

Related Studies

The idea that "quantitative dominance relations" represent an "important parameter for the central or peripheral position of a character" in a drama has been established by Pfister in his crucial structuralist monograph on the analysis of drama (Pfister 1997, p.227). Digital-empirical studies after Pfister have tested different approaches to provide quantitative descriptions of the dramatis personae. Moretti's suggestion to tie the detection of the protagonist to the network analytical criterion of average distance (Moretti 2011, p.4) was rejected as too simplistic (Trilcke 2013, p.204), although this network-analytical approach was taken up by numerous studies. In this vein, Jannidis et al. (2016) not only calculated quantitative measures for the frequency of a character's appearance, but also the weighted degree to determine the accuracy of the identification of main characters. Moretti himself has adjusted his approach conceptually, insofar as he has shifted the focus from the 'protagonist' to a relationally defined concept of 'centrality'. He also emphasised the tension between different criteria (like *word space* and *character space*) not as a deficit, but as a productive factor of a multidimensional quantitative analysis of the dramatis personae (Moretti 2013, pp.5–9). Moretti's basic ideas – the productive multidimensionality of quantitative analysis and the insight into the relational conceptualisation of quantitative character classification – have been taken up by Algee-Hewitt (2017), who worked with two network-analytical centrality measures (betweenness centrality and eigenvector centrality) and tried to examine the quantitative distribution of the cast of a play.

Goal and Procedure

We will conceptually discuss and complement available approaches to the quantitative description of characters in dramatic texts and test them on the basis of a corpus of 465 German-language dramas. The aim in theoretical and conceptual terms is to gain a better understanding of the dimensions of quantitative character analysis and to present diachronic and typological insights into quantitative dominance relations in German-language drama from 1730 to 1930. The subject of the analyses is the DLINA corpus (Fischer & Trilcke 2015). The data is calculated using the Python tool "dramavis", which has been supplemented for this purpose with new analysis modules (Kittel & Fischer 2017).

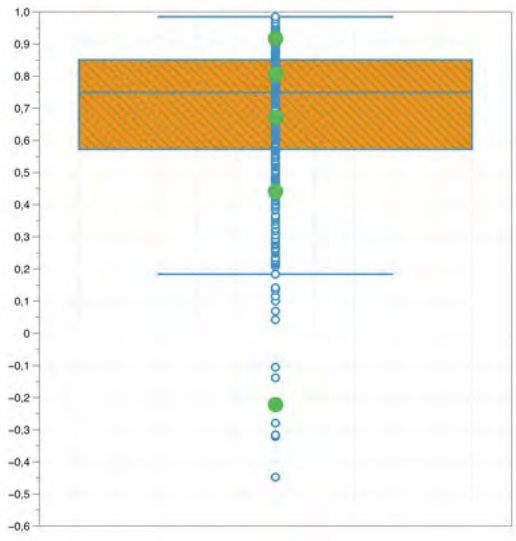
In the first step, we examine the multidimensionality of quantitative descriptions as determined by Moretti (chapter 3). In a second step, we will take up Algee-Hewitt's (2017) proposed approach of working with quartiles and discuss it on the basis of the data from the DLINA corpus (chapter 4). In the third step, we present an approach that describes the quantitative distribution of characters in a play (section 5).

The Correlation of Count-Based and Network-Based Rankings of Characters

The above-mentioned approaches have made use of various measures for the quantitative description of dramatic characters, which can be divided in two groups: count-based measures, such as the number of words expressed by a character, and network-based measures, mostly centrality measures. According to Moretti 2013 (cf. also Jannidis et al. 2016), these two descriptive 'dimensions' can differ considerably.

In order to systematically describe the extent of this deviation, we calculate eight values for each character of the 465 dramas of our corpus, three count-based measures (number of scenes a character appears in, number of speech acts, number of spoken words) and five network-related measures (degree, weighted degree, betweenness centrality, closeness centrality, eigenvector centrality). For each measurement a ranking is created. The rankings are then merged into two meta-rankings: one count-based and one network-based. The two meta-rankings are then combined into an overall ranking.

To determine the deviation between the two meta-rankings, we calculate the ranking correlation coefficient Spearmans Rho and check how strongly the two meta-rankings correlate with each other for all dramas of our corpus (fig.1).

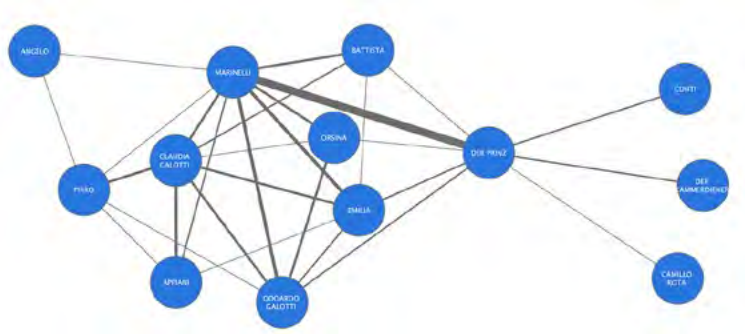


Spearmans Rho for the correlation of count-based and network-based measures.

Complete congruence of the meta-rankings is an exception. In fact, the different measures capture different 'dimensions' of the quantitative character hierarchy. In order to better understand these dimensions, five dramas (marked by the green dots in fig.1) are examined in more detail and discussed in this paper (see figs. 2, 4, 6, 8, 10 for rankings, figs.3, 5, 7, 9, 11 for the network graphs of these dramas).

	Countbasiert			Netzwerkbasiert					Aggregierte Maße		Gesamt-Rank
Figur	Häufigkeit	Sprechakte	Wörter	Degree	Weighted Degree	Betweenness	Closeness	Eigenvector	Count	Netzwerk	
MARINELLI	1	1	1	1	1	2	1	1	1	1	**1**
DER PRINZ	2	2	2	2	2	1	2	2	2	2	**2**
ODOARDO G.	4	3	3	4	4	3	3	3	3	3	**3**
CLAUDIA G.	3	4	6	3	3	5	5	4	4	4	**4**
EMILIA	5	5	5	4	5	3	3	5	5	4	**5**
ORSINA	6	5	4	7	6	7	5	6	5	6	**6**
APPIANI	7	7	7	7	6	9	9	8	7	9	**7**
PIRRO	8	9	10	6	8	6	8	9	8	8	**7**
BATTISTA	8	11	11	7	8	7	5	7	11	7	**9**
ANGELO	10	8	9	10	10	10	10	12	8	10	**9**
CONTI	10	10	8	11	10	10	11	10	10	10	**11**
KAMMERD.	10	12	13	11	10	10	11	10	12	10	**12**
CAMILLO R.	13	12	12	11	13	10	11	13	13	13	**13**

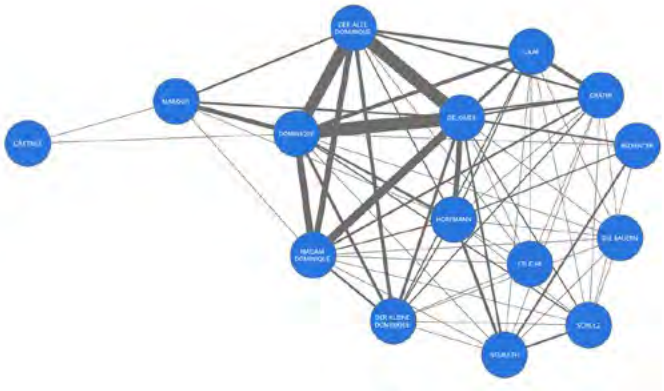
Rankings for Lessing's "Emilia Galotti" (1772) – Spearmans Rho: 0,917.



Network graph for Lessing's "Emilia Galotti" (1772).

	Countbasiert			Netzwerkbasiert					Aggregierte Maße		Gesamt-Rank
Figur	Häufig-keit	Sprech-akte	Wörter	Degree	Weighted Degree	Betweenness	Closeness	Eigenvector	Count	Netz-werk	
DELOMER	1	1	2	2	1	2	2	1	1	2	**1**
DOMINIQUE	2	2	1	1	2	1	1	2	2	1	**1**
HOFFMANN	5	5	5	3	6	3	3	7	4	4	**3**
MME DOMINIQUE	4	7	9	3	4	4	3	4	6	3	**4**
GRAF	6	4	7	5	5	6	5	5	5	5	**5**
ALTE DOMINIQUE	3	3	3	10	3	11	10	3	3	8	**6**
GRÄFIN	8	6	6	5	8	6	5	8	6	7	**7**
KL. DOMINIQUE	8	12	12	5	7	6	5	6	10	6	**8**
MARQUIS	8	8	4	13	10	5	13	9	6	11	**9**
NEURATH	7	9	8	8	9	9	8	10	9	9	**10**
SCHULZ	12	11	10	8	11	9	8	11	12	10	**11**
BEDIENTER	11	10	11	14	12	12	14	12	10	14	**12**
DIE BAUERN	13	13	15	10	13	12	10	14	14	12	**13**
ETLICHE	13	14	14	12	14	12	12	13	14	13	**14**
GÄRTNER	13	14	13	15	15	12	15	15	13	15	**15**

Rankings for Iffland's "Das Erbtheil des Vaters" (1802) – Spearmans Rho: 0.806.



Network graph for Iffland's "Das Erbtheil des Vaters" (1802).

	Countbasiert			Netzwerkbasiert					Aggregierte Maße		Gesamt-Rank
Figur	Häufig-keit	Sprech-akte	Wörter	Degree	Weighted Degree	Betweenness	Closeness	Eigenvector	Count	Netz-werk	
JOHANNA	1	1	1	1	2	1	1	5	1	1	**1**
KARL	3	2	2	3	3	8	5	2	2	3	**2**
DUNOIS	2	3	3	2	1	6	2	1	3	2	**2**
BURGUND	6	4	6	4	6	4	3	6	4	4	**4**
SOREL	4	9	5	6	4	11	10	4	5	6	**5**
LA HIRE	4	6	11	6	5	13	10	3	6	7	**6**
ISABEAU	10	5	7	11	9	5	5	14	7	8	**7**
THIBAUT	12	11	4	8	12	9	5	13	9	10	**8**
RAIMOND	9	8	12	10	11	3	5	16	10	9	**8**
ERZBISCHOF	8	16	13	4	7	7	4	7	14	5	**8**
LIONEL	10	6	8	19	10	17	14	15	8	14	**11**
TALBOT	12	10	10	12	12	2	19	23	11	12	**12**
DU CHATEL	7	12	16	8	8	18	9	8	12	11	**12**
BERTRAND	12	14	9	16	14	12	22	26	12	18	**14**
FASTOLF	15	18	22	13	16	10	12	17	18	12	**14**
LOUISON	15	17	18	27	16	21	28	29	15	23	**16**
MARGOT	15	15	20	27	16	21	28	29	15	23	**16**
CLAUDE MARIE	15	25	28	19	16	16	26	31	22	20	**18**
EDELKNECHT	19	29	34	13	15	19	19	9	29	14	**19**
RATSHERR	23	22	21	27	22	21	33	12	21	22	**19**
RITTER	19	27	27	16	20	20	14	10	26	17	**19**
VOLK	27	27	31	13	21	21	13	11	30	16	**22**
CHATILLON	27	19	23	27	30	21	34	19	23	26	**23**
AGNES	27	32	29	27	30	21	16	20	31	21	**24**
KÖHLER	19	19	19	38	30	14	37	37	19	39	**25**
SOLDAT S.11	23	12	17	42	40	21	24	34	17	42	**26**
A. DER SOLDATEN	23	32	38	32	30	21	17	28	35	25	**27**
HAUPTMANN	27	32	32	32	35	21	23	32	33	29	**28**
RAOUL	27	30	15	32	35	21	35	24	25	37	**28**
M. STIMMEN	27	32	43	16	22	21	19	18	43	19	**28**
KÖHLERWEIB	19	25	25	38	30	14	37	37	23	39	**28**
MONTGOMERY	23	21	14	45	45	21	31	35	20	43	**32**
SCHILDWACHE	27	30	38	32	35	21	26	21	37	27	**33**
HEROLD	27	22	26	38	40	21	24	22	28	36	**33**
ZWEITER	27	32	36	19	24	21	39	40	37	29	**33**
DRITTER	27	32	37	19	24	21	39	40	39	29	**33**
SCHW. RITTER	27	22	24	45	45	21	31	35	26	43	**37**
SOLDAT 2.5	27	32	38	19	24	21	39	40	40	29	**37**
ERSTER	27	32	38	19	24	21	39	40	40	29	**37**
SOLDATEN	27	44	44	32	35	21	17	33	44	28	**40**
EDELMANN	27	32	32	38	40	21	30	27	33	39	**40**
FÜNFTER	27	44	44	19	24	21	39	40	44	29	**42**
VIERTER	27	44	44	19	24	21	39	40	44	29	**42**
VIELE STIMMEN	27	32	38	32	35	21	35	24	40	37	**44**
ETIENNE	27	32	30	42	43	21	45	39	32	45	**44**
KÖHLERBUB	27	32	35	42	43	21	46	46	36	46	**46**

Rankings for Schiller's “Die Jungfrau von Orleans” (1801) – Spearmans Rho: 0.672.

Network graph for Schiller's “Die Jungfrau von Orleans” (1801).

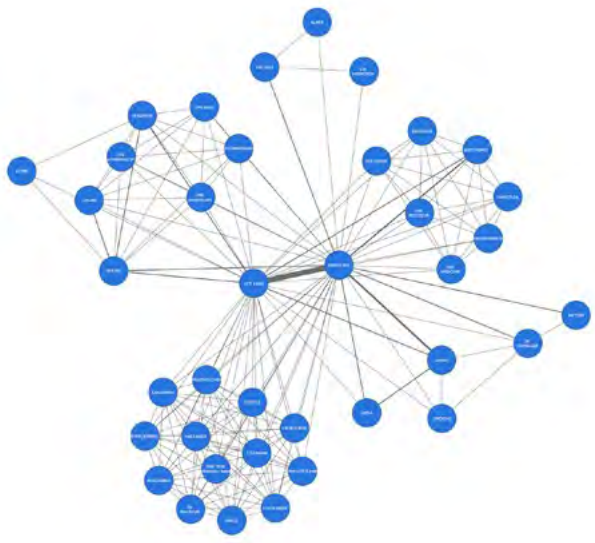
	Countbasiert			Netzwerkbasiert					Aggregierte Maße		Gesamt-Rank
Figur	Häufigkeit	Sprechakte	Wörter	Degree	Weighted Degree	Betweenness	Closeness	Eigenvector	Count	Netzwerk	
VRONI	1	1	1	1	1	1	1	1	1	1	**1**
FERNER	3	2	2	2	2	2	2	3	2	2	**2**
FRANZ	2	2	3	3	2	3	3	2	2	3	**3**
HÖLLERER	5	5	8	11	4	5	4	4	5	4	**4**
LIES	4	4	6	11	6	4	7	6	4	6	**5**
CRESCENZ	6	9	12	11	5	5	4	5	9	5	**6**
MIRZL	13	14	13	4	7	7	8	11	14	7	**7**
TONI	7	7	10	14	7	7	6	7	8	14	**8**
BÜRGEI	13	12	16	4	7	7	8	11	15	7	**8**
MAHM	9	6	5	15	7	7	19	19	7	15	**8**
GROSSKNECHT	7	8	4	21	19	7	15	8	6	19	**11**
MUCKERL	13	14	15	4	7	7	8	14	16	10	**12**
BADER	9	11	11	15	7	7	19	19	11	15	**13**
ROSL	9	12	14	15	7	7	19	19	12	15	**14**
WABERL	13	14	17	4	7	7	8	14	18	10	**15**
ANNERL	13	14	18	4	7	7	8	14	19	10	**16**
CHOR	13	21	23	4	7	7	8	11	23	7	**17**
GRETL	13	20	20	4	7	7	8	14	20	10	**17**
KATHREIN	9	14	19	15	7	7	19	19	16	15	**19**
JAKOB	13	9	7	21	22	7	16	18	10	22	**20**
LEVY	13	14	9	23	23	7	23	23	13	23	**21**
1. SCHWÄRZER	13	21	21	19	19	7	17	9	21	20	**22**
2. SCHWÄRZER	13	21	21	19	19	7	17	9	21	20	**22**

Rankings for Anzengruber's “Der Meineidbauer” (1871) – Spearmans Rho: 0.442.

Network graph for Anzengruber's “Der Meineidbauer” (1871).

	Countbasiert			Netzwerkbasiert					Aggregierte Maße		Gesamt-Rank
Figur	Häufigkeit	Sprechakte	Wörter	Degree	Weighted Degree	Betweenness	Closeness	Eigenvector	Count	Netzwerk	
FRANZISKA	1	1	1	1	1	1	1	1	1	1	**1**
VEIT KUNZ	2	2	2	2	2	2	2	2	2	2	**2**
HERZOG	3	3	3	15	3	3	15	5	3	15	**3**
LAURUS BEIN	12	17	12	3	3	8	3	6	16	3	**4**
BREITENBACH	5	5	4	17	16	8	17	4	5	16	**5**
GISLIND	6	6	7	17	19	8	17	21	6	18	**6**
SPREIZFÜSS.	12	22	22	3	3	8	3	6	23	3	**7**
SOPHIE	3	4	5	29	18	6	29	3	4	23	**8**
GESPENSTERSCH.	12	26	14	3	3	8	3	14	18	11	**9**
MAUSI	12	21	24	3	3	8	3	6	26	3	**9**
HERZOGIN	6	13	19	15	16	3	15	18	13	17	**11**
FAHRSTUHL	12	10	11	17	20	8	17	26	9	24	**12**
CHORUS	12	33	26	3	3	8	3	6	31	3	**13**
DR. MALKOLM	12	33	28	3	3	8	3	6	32	3	**14**
HAGLMEIER	12	33	32	3	3	8	3	6	33	3	**15**
POLIZEIPRÄSI.	12	16	16	17	20	8	17	22	17	19	**15**
HOHENKEMN.	12	13	13	17	20	8	17	26	13	24	**17**
KIESGRÄBER	12	29	37	3	3	8	3	6	34	3	**17**
ROHRDOMMEL	12	22	27	3	3	8	3	14	27	11	**19**
DR. HOFMILLER	6	8	8	29	30	5	29	20	8	30	**19**
HUNDEKOPF	12	22	20	17	20	8	17	22	20	19	**21**
LYDIA	6	10	17	32	31	8	32	19	9	31	**22**
SCHWEINEKOPF	12	22	21	17	20	8	17	22	21	19	**22**
MUTTER	6	7	6	35	34	8	35	33	6	34	**22**
HAGELMEIER	12	33	36	3	3	8	3	6	38	3	**25**
DAS KIND	12	19	25	17	20	8	17	22	23	19	**26**
KULLMANN	12	29	37	3	3	8	3	14	34	11	**27**
DR. HORNSTEIN	12	12	9	35	36	8	35	37	9	37	**28**
DIE MÄDCHEN	12	29	15	17	20	8	17	26	23	24	**29**
KARAMINKA	12	33	35	3	3	8	3	14	37	11	**30**
ALMER	12	13	10	35	36	8	35	37	12	37	**31**
DIRCKENS	12	17	23	31	32	8	31	32	18	32	**32**
PATER	12	9	18	32	34	8	38	36	15	36	**33**
DER DIENER	12	26	31	17	20	8	17	26	28	24	**34**
REGISSEUR	12	29	29	17	20	8	17	26	29	24	**35**
VEITRALF	6	20	29	32	32	7	33	34	21	33	**36**
DER REGISSEUR	12	33	34	17	20	8	17	26	36	24	**37**
TIEFE MÄNNER.	12	26	32	35	36	8	33	35	29	35	**38**

Rankings for Wedekind's “Franziska” (1912) – Spearmans Rho: -0.222.

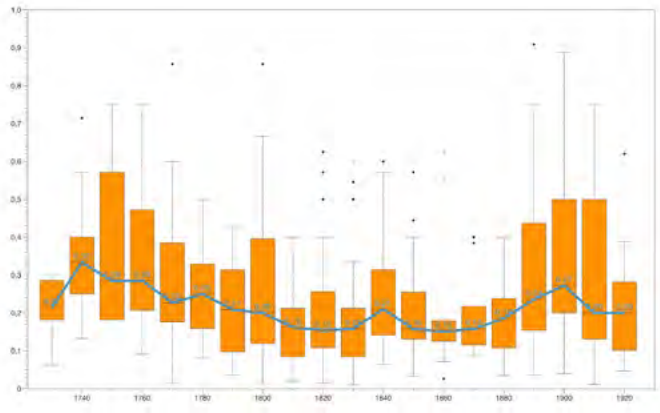


Network graph for Wedekind's “Franziska” (1912).

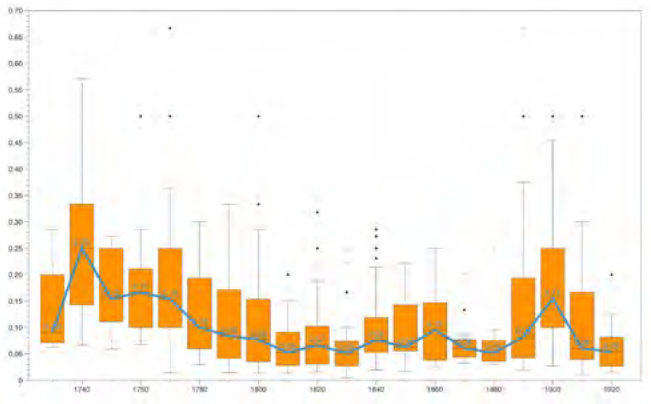
As it turns out, the deviations usually affect characters at the bottom of the hierarchy. Here the network-related measures are particularly sensitive to types of clustering in secondary scenes of a drama (figs. 8 and 9), which has even more severe effects if a drama is quantitatively dominated by very few characters (figs. 10 and 11). On the other hand, both meta-rankings are very similar for the quantitatively dominant characters (top 1 and top 1 or 2). These observations can serve as an argument that the multidimensional description is less relevant to discuss protagonists, but rather for the characterisation of quantitative dominance relations within the cast as a whole.

Percentage of Quantitative Dominant Characters

Algee-Hewitt 2017 made a suggestion for the characterisation of the quantitative distribution of a cast, albeit with a continuing focus on quantitatively dominant characters. Working with an English-language drama corpus of several thousands of plays, he calculated the eigenvector centrality of the characters and then calculated the percentage of characters located in the upper quartile of the distribution. We have reproduced this test with our corpus – for all measures mentioned above. As an example, the box plots show the values for the eigenvector centrality (fig.12) as well as for the count-based measure 'number of words' (fig.13).



Percentage of characters in the upper quartile according to their eigenvector centrality, grouped by decades. Blue line: median.

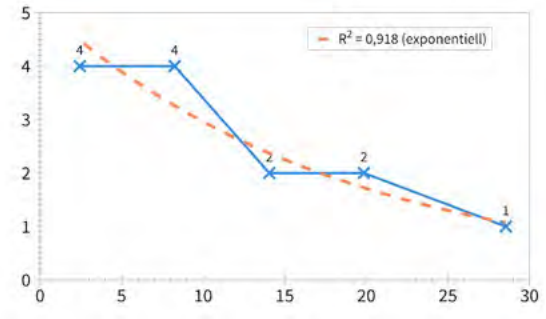


Percentage of upper quartile characters according to number of words, grouped by decades. Blue line: median.

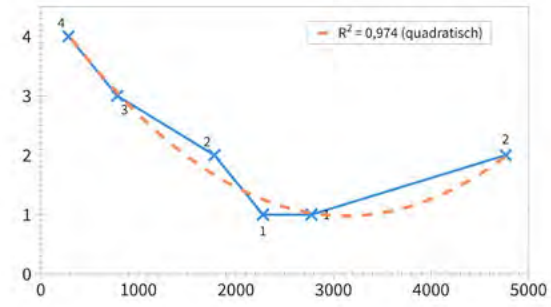
It is interesting to note that the values we calculated for eigenvector centrality (fig.12) are significantly higher than the values presented by Algee-Hewitt (for the period of time covered by our corpus the median is lower than 0.15). A comparison of fig.12 and fig.13 shows that the network-related measures usually locate a larger percentage of characters in the upper quartile. From a network-analytical point of view, a drama tends to be dominated by several characters. The two curves also tend to follow the same course, especially regarding the big flattening curve in the decades after 1740. What we can see there is the reduction of the percentage of quantitatively dominant characters ('main characters') and the emergence of quantitatively less dominant characters ('secondary characters', 'atmospheric characters').

More Detailed Distribution Analysis

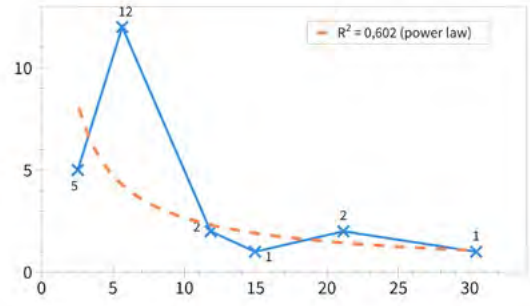
With a focus on quantitatively dominant 'main characters', Algee-Hewitt's approach attempts to describe types of distribution of the dramatis personae and thus to identify 'dominance relations' via distribution analyses. Following analyses of the 'small world' phenomenon in drama (Trilcke et al. 2016), we propose to extend this approach and develop a typology of quantitative distribution of characters in dramatic texts. To this end, we take the above-mentioned eight quantitative measures, calculate the distribution of character values across deciles and subject this distribution to a regression analysis (tests on linear, exponential, quadratic and power-law distribution; typologisation according to the regression line with the highest coefficient of determination). Examples in figs. 14 to 19 show the results for three of the dramas discussed above, for a network-related measure (weighted degree) and a count-based measure (number of words).



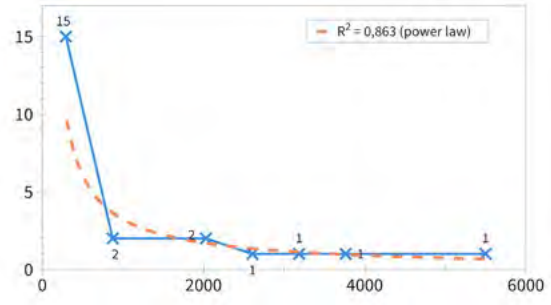
Lessing's "Emilia Galotti" (1772) – decile distribution of weighted degree, y-axis: number of characters.



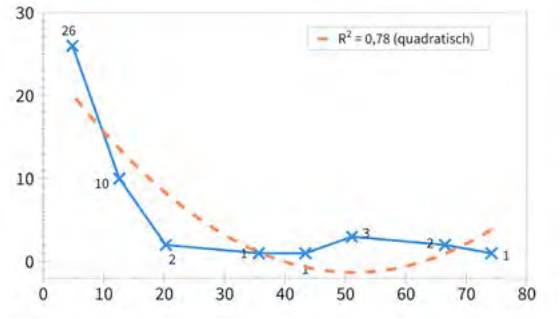
Lessing's "Emilia Galotti" (1772) – decile distribution of number of words, y-axis: number of characters.



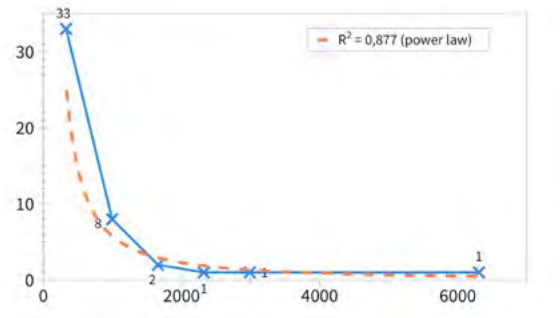
Anzengruber's "Der Meineidbauer" (1871) – decile distribution of weighted degree, y-axis: number of characters.



Anzengruber's "Der Meineidbauer" (1871) – decile distribution of number of words, y-axis: number of characters.



Schiller's "Die Jungfrau von Orleans" (1801) – decile distribution of weighted degree, y-axis: number of characters.



Schiller's "Die Jungfrau von Orleans" (1801) – decile distribution of number of words, y-axis: number of characters.

These typologies are calculated for all eight values and for all 465 dramas – we will present the results for the corpus as a whole at the conference. With our approach, a more precise, multidimensional description of typical quantitative dominance relations in drama will be possible. The increase in the number of less dominant characters observed on the basis of our quartile analysis (figs. 12–13) will be described with more precision and supplemented by a more differentiated examination of types of 'middle characters'.

Summary

This talk brings together several approaches for the quantitative analysis of characters in literary texts, discusses the potential of a multidimensional description beyond top characters (protagonists) and suggests an approach for typologising quantitative dominance relations within the cast of a drama.

References

Algee-Hewitt, Mark (2017): "Distributed Character: Quantitative Models of the English Stage, 1500–1920". *Digital Humanities 2017. Conference Abstracts. Mc-*

Gill University & Université de Montréal, 119–121. URL: <https://dh2017.adho.org/abstracts/103/103.pdf>.
Fischer, Frank / Trilcke, Peer (2015): "Introducing dlina Corpus 15.07 (Codename: Sydney)". *dlina blog,* 20 June 2015. URL: <https://dlina.github.io/Introducing-DLINA-Corpus-15-07-Codename-Sydney/>.
Kittel, Christopher / Fischer, Frank (2017): "dramavis (v0.4)". *GitHub*, September 2017. URL: <https://github.com/lehkost/dramavis>.
Jannidis, Fotis / Reger, Isabella / Krug, Markus / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank (2016): "Comparison of Methods for the Identification of Main Characters in German Novels". *Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków*, 578–582.
Moretti, Franco (2011): *Network Theory, Plot Analysis* (= Literary Lab Pamphlet 2), May 2011. URL: <http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf>.
Moretti, Franco (2013): *"Operationalizing": or, the function of measurement in modern literary theory* (= Literary Lab Pamphlet 6), December 2013. URL: <https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf>.
Pfister, Manfred (1997): *Das Drama. Theorie und Analyse*. 9th ed. Munich: Fink.
Trilcke, Peer (2013): "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft". Ajouri, Philip / Mellmann, Katja / Rauen, Christoph (eds.): *Empirie in der Literaturwissenschaft*. Münster: mentis, 201–247.
Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario (2016): "Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930". *Digital Humanities 2016. Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków*, 385–387.

Visualising The Digital Humanities Community: A Comparison Study Between Citation Network And Social Network

Jin Gao
jin.gao.13@ucl.ac.uk
University College London , UCL Centre for Digital Humanities, Department of Information Studies, United Kingdom

Julianne Nyhan
j.nyhan@ucl.ac.uk
University College London, UCL Centre for Digital Humanities, Department of Information Studies, United Kingdom

Oliver Duke-Williams
o.duke-williams@ucl.ac.uk
University College London, UCL Centre for Digital Humanities, Department of Information Studies, United Kingdom

Simon Mahony
s.mahony@ucl.ac.uk
University College London, UCL Centre for Digital Humanities, Department of Information Studies, United Kingdom

Introduction

An understanding of the intellectual and social structures of Digital Humanities (DH) has been sought by many scholars; some have pointed to the potential usefulness of quantitative methods in such analyses (McCarty, 2003; Terras et al., 2013). A few existing studies have applied quantitative methodologies to analyse publication, conference and social media data (e.g. Nyhan and Duke-Williams, 2014; Weingart, 2016; Grandjean, 2016). This study not only incorporates such approaches but extends them by integrating new analysis and visualisation methods into the wider study of DH's intellectual and social structures.

In this paper, we will introduce research on the citation and social network of DH that is giving rise to new understandings of the field's community structure; scholarly interactions; disciplinary development; and formal/informal communication channels. The citation network of Author Co-Citation Analysis (ACA) comprises 22,321 cited authors across 52,823 cited references from the three core DH journals, while the social network of Twitter Co-Retweet Analysis comprises 3,160 Twitter users and 5,929,609 tweets. To the best of our knowledge, this study is the first to combine bibliometric and social network methods to visualise and compare the DH communities and to uncover their histories. This research contributes to ongoing discussions and debates about the DH knowledge and community structures (Gold and Klein, 2016).

Data Analysis

Citation network

The Author Co-Citation network study was presented at DH2017; this paper extends this earlier analysis. For reasons of clarity, we here give a brief overview of this research. The network has been constructed by collecting the citation data of three core DH journals up to December 2016: *Computers and the Humanities* (*CHum*), *Digital Scholarship in the Humanities* (*DSH*) and *Digital Humanities Quarterly* (*DHQ*). 2,582 articles with 52,823 cited references were collected (see Figure 1).

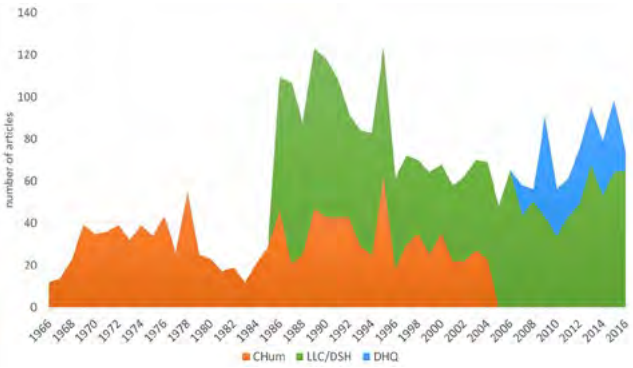


Figure 1. number of articles collected each year (1966-2016)

By using *fractional non-self-citation count* and *exclusive co-citation count* (Zhao and Strotmann, 2008), the weights of nodes and links respectively were calculated for visualisation using the software VOSviewer 1.6.7 (van Eck and Waltman, 2010). An author name disambiguation method (Strotmann et al., 2009) was used, and 22,321 unique cited authors identified. Where possible, other pertinent information was collected (i.e. author full names, gender, country of affiliation, etc.). After counting the occurrences of two authors being cited together, ACA shows the DH intellectual structure and influential topics and scholar groups (Figure 2).

Figure 2. DH Author Co-Citation network

Twitter network

Given DH's early adoption and active use of Twitter (Ross et al., 2011), previous studies have explored the field's scholarly communications and community on Twitter (e.g. Quan-Haase et al., 2015; Grandjean, 2016). This study introduces a new approach (*co-retweet*) to visualise the DH social network.

We have selected all the Twitter users that are followed by the Alliance of Digital Humanities Organisations (ADHO) and its member organisations' (see http://adho.org/) Twitter accounts. As dynamic and interdisciplinary as the DH community, it is often difficult and subjective to select the users by their account descriptions. In contrast, the following relationships by the DH organisations indicate more genuine and representative identification. A total of 3,160 unique users have been collected along with all the 6 million tweets from 21/03/2006 when Twitter was created up to 22:00 (GMT) on 5/11/2017 (see Figure 3).

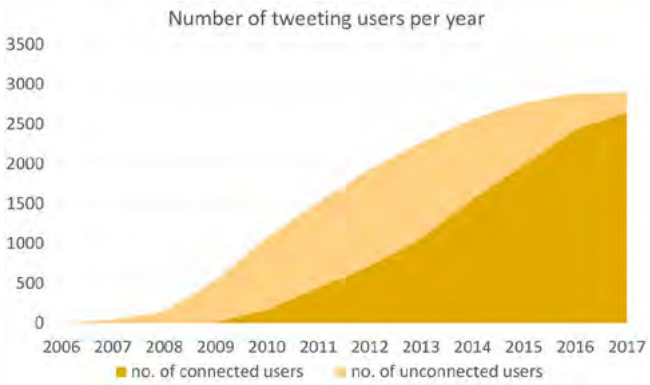


Figure 3. number of tweeting users collected each year

Similar to the citation network method, the Twitter user Co-Retweet network has been constructed by calculating the number of non-self-retweets the user received (*non-self-retweet count*), and the number of same tweets that two users both retweeted (*co-retweet count*) for the weights of the nodes and links respectively. The network resulting visualisations are shown on Figure 4.

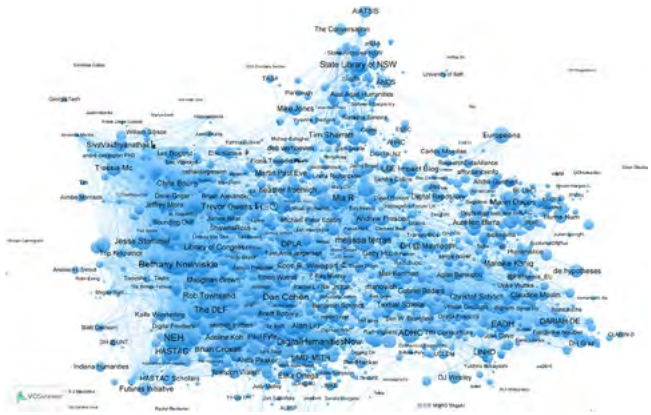


Figure 4. DH Twitter Co-Retweet network

Results and comparison

In the citation network, the authors identified distinct topic-based clusters of researchers with backgrounds in information studies and historical literature; in linguistics; in statistical text analysis; in early concordance projects; and biotech influenced text analysis. In contrast, the co-retweet network exhibits grouping based on language and region, with clusters related to scholars in

North America; in Australia; in the UK; and clusters with Francophonic, Germanophonic and Hispanophonic backgrounds.

The Twitter clusters are connected closely whereas clusters on the citation network are more loosely linked. This makes sense, as topics of study are generally more specific and less likely to change, whereas users on social media probably share a wider range of interests. The citation network is based on formal communications and it would take years to get sufficient citations to form links between two scholars. The Twitter network, however, is constructed by more informal interactions between users, and once two users retweeted the same tweet, they immediately build a link on the network.

By visualising both networks during different time periods, this study will also present the topics, disciplines, countries that are involved, and how the networks have been developed and formed over time.

As shown in Figure 5, we divided the 51 years (1966-2016) of historical citation data into five different periods (Hockey, 2004) for visualisation. The citation clusters experienced isolation (1966-1970); connection (1971-1985); consolidation (1986-1990); sub-fields development (1991-2005); and new specialties expansion (2006-2016). Over time, the most cited topics moved from concordance construction, to computational linguistics, then to information and historical literature studies.

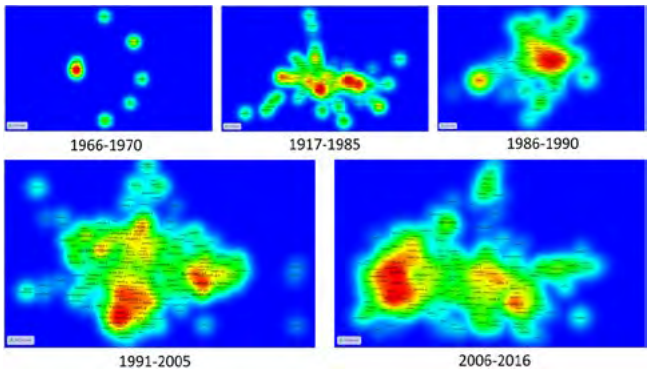


Figure 5. DH Author Co-Citation networks in five periods

As shown in Figure 6, DH Twitter users started to have co-retweet connections in 2009; and then they experienced the beginning of connection (2010); multi-region connection (2010); Anglophonic cluster to centre (2011); Francophonic cluster to develop (2012); North America and UK to separate (2013); Germanophonic to come out (2014); Australian cluster to show (2015); Hispanophonic cluster to emerge (2016); Density continue to move to North America cluster (2017). Over time, the network visualisations show that the density is moving from European clusters towards the North American cluster.

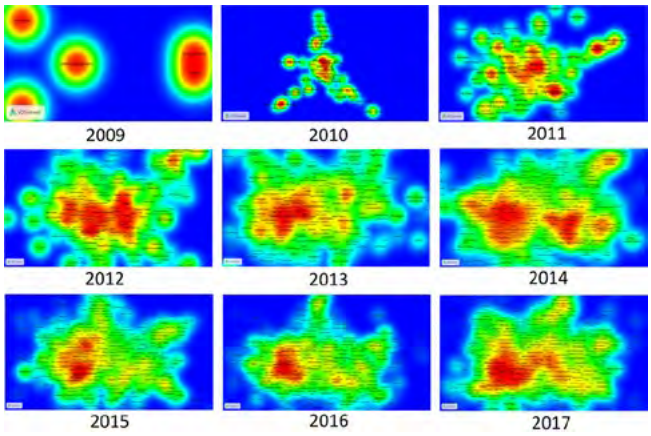


Figure 6. DH Twitter Co-Retweet networks in different years

Discussion and conclusion

This study is not only the first to contribute to the DH history and community studies by visualising and comparing bibliometric and social networks, but also introduces new network approaches (*co-retweet*) to study communications on social media that could support wider social network and data visualisation studies.

As we will discuss, network studies offer powerful but partial ways of studying the aspects of communities that are amenable to quantitative methods. We do not present the visualisations included in this paper as normative representations of the DH "community" or "communities". Nevertheless, when used with caution, network studies can shed new light on important aspects of the historical formation of DH.

There are methodological limitations exist. For example, because the research subjects (cited authors and retweeting users) are not the same group of people (although with much overlap), obvious differences are expected. Besides, the citation lag time has been considered. Other practical methods to identify and study the DH Twitter communities can also be applied.

References

Davis, L.S., Johns, S.A., Aggarwal, J.K. (1979). Texture Analysis Using Generalized Co-Occurrence Matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1*, 251–259. https://doi.org/10.1109/TPAMI.1979.4766921

Gold, M.K. and Klein, L.F. (2016). *Debates in the digital humanities: 2016*. University of Minnesota Press, Minneapolis London.

Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*,3:1171458. https://doi.org/10.1080/23311983.2016.1171458

Hockey, S. (2004). The History of Humanities Computing, in: Schreibman, S., Siemens, R., Unsworth, J. (2004).

A Companion to Digital Humanities. Blackwell Publishing Ltd, Malden, MA, USA, pp. 1–19.
McCarty, W. (2003). Humanities Computing, in: *Encyclopedia of Library and Information Science*. Marcel Dekker, New York.
Nyhan, J. and Duke-Williams, O. (2014). Joint and multi-authored publication patterns in the Digital Humanities. *Literary and Linguistic Computing*,29:387–399. https://doi.org/10.1093/llc/fqu018
Quan-Haase, A., Martin, K., McCay-Peet, L. (2015). Networks of Digital Humanities Scholars: The Informational and Social Uses and Gratifications of Twitter. *Big Data & Society*,2. https://doi.org/10.1177/2053951715589417
Ross, C., Terras, M., Warwick, C., Welsh, A. (2011). Enabled Backchannel: Conference Twitter Use by Digital Humanists. *Journal of Documentation*,67:214–237. https://doi.org/10.1108/00220411111109449
Strotmann, A., Zhao, D., Bubela, T., (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology*,46:1–20. https://doi.org/10.1002/meet.2009.1450460218
Terras, M.M., Nyhan, J., Vanhoutte, E. (2013). *Defining Digital Humanities: A Reader*. Ashgate Publishing Company, Farnham, Surrey, England: Burlington, VT.
van Eck, N.J. and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*,84:523–538. https://doi.org/10.1007/s11192-009-0146-3
Weingart, S.B., (2016). dh quantified: A Review of Quantitative Analyses of the Digital Humanities. *the scottbot irregular: data are everywhen*.
Zhao, D. and Strotmann, A. (2008). All-author vs. first-author co-citation analysis of the Information Science field using Scopus. *Proceedings of the American Society for Information Science and Technology*, 44:1–12. https://doi.org/10.1002/meet.1450440262

SciFiQ and "Twinkle, Twinkle": A Computational Approach to Creating "the Perfect Science Fiction Story"

Adam Hammond
adam.hammond@utoronto.ca
University of Toronto, Canada

Julian Brooke
julian.brooke@gmail.com
Thomson Reuters, Canada

Introduction

In Fall 2015, we were approached by author (and prominent DH-skeptic [2012]) Stephen Marche with a request that we help him use computational textual analysis to allow him to write "the perfect science fiction story." His specific request was for a set of "rules" to follow in composing such a story. On consultation with Marche, we devised an approach that would derive "rules" related to theme (using topic modelling) and style (using a variety of techniques, including our own original work on quantifying style) from a hand-selected corpus of Marche's 50 favourite science fiction stories. In Fall 2016, we sent Marche a list of 14 thematic rules and created a system to provide real-time stylistic feedback on his efforts to meet a set of 24 stylistic targets. In December 2017, the resulting story, "Twinkle Twinkle," was published in the popular technology magazine *Wired* alongside a set of detailed footnotes explaining and reflecting upon the process of its creation. Whereas Marche and his editors viewed the rule-creating process as "objective" and the publishing professionals interviewed in the piece complained that the resulting story lacked "humanity," the process in fact blended computational analysis and human interpretation. We regard it as an instance of computer-assisted "creative deformance" rather than "robotic authorship."

Thematic rules

Below is a selection of the 14 rules we sent to Marche:

1. The story should be set on a planet other than earth.
2. The story should thus NOT be set in space itself.
3. On this planet, there should be an existing, non-human civilization. This civilization should have a hierarchical social structurewith a powerful ruler. Inhabitants of this alien civilization should be given clearly non-human names. The protagonists of the storyshould be humans who are directly observing this civilization from a certain distance and do not consider themselves part of it.
4. The story should be set in a city. The protagonists should be seeing this city for the first time and should be impressed and dazzled by itsscale.
5. Part of the action should unfold at night during an intense storm.
6. Include a pivotal scene in which a group of people escape from a building at night at high speed in a high-tech vehicle made of metal andglass that is large enough to live in for an extended period (it should have a bed in it, for instance).
9. Include a scene set on a traditional earth farm, with apple trees and/or corn fields. In this scene, a mother and father are present. (Given theother rules, this is most likely a flashback to a protagonist's childhood, but I leave the details to you, of course.)
10. Include extended descriptions of intense physical sensations and name the bodily organs that perceive these sensations.
13. DO NOT focus on conventional domestic family life. Marriage should not be a theme. No scenes should depict a conventional bourgeoisfamily (especially a happy bourgeois family) at the dinner table.

These rules were derived as follows. First, we assembled a corpus of some 4,000 texts, of which approximately 3,000 were fiction *other* than science fiction, 1,000 were science fiction, and 50 were the stories hand-selected by Marche. All texts were processed in GutenTag (Hammond and Brooke 2017). Next, we performed topic modelling using Mallet in R (400 topics, 500-word chunks, nouns only). Once this was complete, we examined all the topics that distinguished Marche's corpus from both comparison sets, positively and negatively, with statistical significance values of $p < 0.05$. Some of these topics were easily converted into thematic rules, such as topic 199 (Fig. 1), which became the basis of rule 5. Less legible topics required a more creative, less "objective" approach. This was the case for topic 33 (Fig. 2), the topic which most clearly distinguished Marche's hand-selected corpus from both comparison groups, and which became the basis of rule 6. To preserve the aura of computational objectivity that he craved, Marche was not shown the topic modelling word clouds, and the process by which the rules were devised was not explained to him in detail.

Figure 1: Topic 199, the basis of rule 5.

Figure 2: Topic 33, the basis of rule 6.

Stylistic rules

Marche began composition once he had received the list of thematic rules. We instructed him to insert his drafts into an online system we had devised, SciFiQ, which would provide stylistic feedback based on 24 criteria:

1. Literariness
2. Abstractness
3. Objectivity
4. Colloquialness
5. Concreteness
6. Subjectivity
7. Positivity
8. Text length
9. Average word length
10. Average sentence length
11. Average paragraph length
12. Average variance in sentence length
13. Average variance in paragraph length
14. Average commas per sentence
15. Percentage of Latinate words
16. Nouns per 100 words
17. Verbs per 100 words
18. Adjectives per 100 words
19. Adverbs per 100 words
20. Lexical density
21. Speaking characters
22. Percentage of text which is dialogue
23. Percentage of dialogue by female characters
24. Major named locations

All criteria were calculated using GutenTag, which was configured to identify parts of speech, structure, and named locations (using LitNER [Brooke et al. 2016b]); to distinguish narration from dialogue; to identify individual characters and their gender; and to tag each word for stylistic properties (Brooke et al. 2016a) and sentiment polarity. SciFiQ displayed results for each of these criteria, colour-coded to indicate how close Marche was to his targets. Only once the values were within 0.5 standard deviations of the mean value for all stories in the 50-story corpus would the value read green; otherwise they would read purple (too low) or red (too high). Once Marche had all values in the green range (with the exception of story length, which was at the discretion of his editor), composition was considered complete. The story was not modified at all during the editing process for *Wired*.

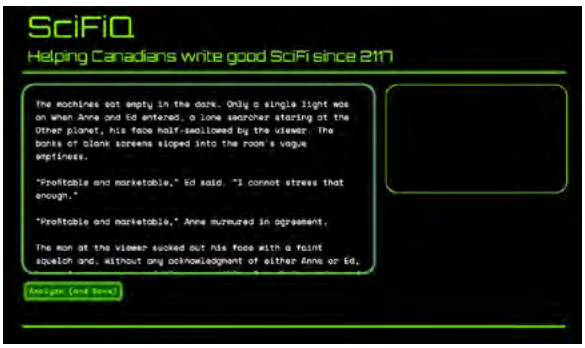


Fig 3: SciFiQ in composition mode.

Fig 4: SciFiQ's analysis report.

Discussion

We approached this project with several aims in mind. First, and most prosaically, we regarded it as a means of validating our work on quantifying style (Brooke et al. 2016a). If our system told Marche to make his style less literary, for example, and he made a series of edits based on his intuitive concept of literariness, would our system respond in a way that he regarded as intuitive? In practice, Marche found that our system's analyses did correspond to his intuitions. We thus consider our approach to style to be further validated. Second, we viewed the project as a way of testing the extent to which a set of tools designed for computational textual *analysis* (some developed by ourselves) could be useful in *composition*. Following critics like Jerome McGann (2001) and Stephen Ramsay (2011) we were interested in computational tools' ability to "deform" the literary work in such a way as to model a helpful and non-traditional form of composition. In this, too, we believe the project was successful. In the notes he supplied in *Wired*, Marche reported, "the algorithm affected the story much more than I would have thought," noting in particular the manner in which conflicting rules (1 and 9) presented welcome imaginative challenges, and in which the SciFiQ system demanded a "Rubik's cube"-like writing process whereby alterations in one part of the story required counterbalancing edits elsewhere.

Of course, we understood that the project would be received in a manner not entirely in keeping with our intentions. Whereas we regarded the project as a somewhat playful creative disruption of the conventional process of composition, Marche presented SciFiQ as "a computational system that [would] optimize [his] prose" and "make [him] better at his job." In a sidebar, *Wired* asked two prominent literary editors (Andy Ward, editor-in-chief of Random House, and Deborah Treisman, fiction editor of *The New Yorker*) to provide general feedback on Marche's story, "without knowing who (or, more specifically, what) wrote it." The implication that the story was *written by* an algorithm is inaccurate (indeed, *Wired* credited it to Stephen Marche, not to SciFiQ). Further, the claim that opinions were solicited "blind" appears disingenuous given that Ward's primary complaint was that the story "doesn't sound human" — an unusual comment unless one has been told that computers were somehow involved in the story's composition. From the perspective of *Wired*, then, the story's appeal as well as its danger lay in the notion of computers taking over the essentially human act of storytelling.

Conclusion

The "Twinkle, Twinkle" project will likely be received by the public as another instance in the longstanding debate about the proper role of machines in storytelling. Yet for us, its significance is rather different. It served as an opportunity to use tools and approaches designed for the analysis of literature in creative production; to validate our 6-dimensional approach to style; and to reach across the divide that continues to separate academic work on literary analysis both from contemporary creative writers and their popular audiences.

References

Brooke, J., Hammond, A., and Hirst, G. (2016a). Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction. *Digital Scholarship in the Humanities*, **2**(2): 1–17.

Brooke, J., Hammond, A., and Baldwin, T. (2016b). Bootstrapped Text-level Named Entity Recognition for Literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*.

Hammond, A., Brooke, J. (2017). GutenTag: A User-Friendly, Open-Access, Open-Source System for Reproducible Large-Scale Computational Literary Analysis. In *Digital Humanities 2017 Abstracts*. Montreal: DH2017, pp. 246-249.

Hammond, A., Brooke, J. (2016). Project Dialogism: Toward a Computational History of Vocal Diversity in English-Language Literature. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 543-544.

Marche, S. (2017). Twinkle, Twinkle. *Wired*, (Dec 2017): 108–115.
Marche, S. (2012). Literature Is Not Data: Against Digital Humanities. *Los Angeles Review of Books.*
McGann, J. (2001) Radiant Textuality: Literature after the World Wide Web. New York: Palgrave.
Ramsay, S. (2011). Reading Machines: Toward an Algorithmic Criticism. Urbana, IL: University of Illinois Press.

Minna de Honkoku: Learning-driven Crowdsourced Transcription of Pre-modern Japanese Earthquake Records

Yuta Hashimoto
yhashimoto1984@gmail.com
National Museum of Japanese History, Japan

Yasuyuki Kano
kano@rcep.dpri.kyoto-u.ac.jp
Kyoto University, Japan

Ichiro Nakasnishi
ichiro@kugi.kyoto-u.ac.jp
Kyoto University, Japan

Junzo Ohmura
ohmura1204@yahoo.co.jp
Bukkyo University, Japan

Yoko Odagi
odagi@kugi.kyoto-u.ac.jp
Kyoto University, Japan

Kentaro Hattori
hattori@kueps.kyoto-u.ac.jp
Kyoto University, Japan

Tama Amano
tama@npo-kikou.com
Kyoto University, Japan

Tomoyo Kuba
tomoyokuba@gmail.com
Kyoto University, Japan

Haruno Sakai
sakai.haruno.36r@gmail.com
Tokyo Metropolitan Library

Introduction

In the last decade, crowdsourcing has become a major technique for transcribing a large volume of historical manuscripts. The volunteers of Transcribe Bentham[1] have transcribed more than 19,000 pages of manuscripts written by Jeremy Bentham (Causer and Wallace 2012). More than 480,000 pages of weather observations from the US Government Arctic logbooks written in the 19th century were transcribed by 4,730 people through the Old Weather[2] project (Eveleigh et al. 2013).

However, managing a crowdsourcing project remains a big challenge for humanities scholars. The following practical difficulties are encountered:

1. The need to draw public attention to the project successfully.
2. The need to encourage participants' long-term involvement.
3. The tasks requiring crowdsourcing in humanities studies (e.g. transcribing ancient handwritten manuscripts) are often difficult for non-trained participants.

In case of Japanese Studies, the last difficulty is particularly crucial; due to the drastic change in the writing system that occurred at the end of 19th century, 99% of modern Japanese people are unable to read *kuzushiji*, classical calligraphic renderings of Japanese characters that were common for both publishing and handwriting. Therefore, the crowdsourcing technique has never been successfully applied to pre-modern Japanese materials.

However, humanities scholars can use education to draw the attention of a large number of people, promote their long-term participation, and train them to tackle difficult tasks. The fundamental idea in this paper is to develop a crowdsourcing system embedded in a collaborative learning environment that enables learners to conduct crowdsourced tasks as a part of their learning with their peers.

Minna de Honkoku[3] (https://honkoku.org/) is a crowdsourced transcription project of pre-modern Japanese earthquake records, developed by the members of the Historical Earthquake Study Group (HESG) at Kyoto University based on this idea. In this paper, we will briefly describe the aim, materials, approach, and results of Minna de Honkoku.

The Background and aim of the project

HESG is a joint group of seismologists and historians including the authors at Kyoto University who have been studying pre-modern earthquake records for seismic research and disaster prevention. Since instrumental observation of earthquakes in Japan began only after the end of 19th century, transcribing written records are required for studying past earthquakes. Therefore, Japanese seismologists have developed an extensive collaboration with historians and archivists.

1 http://www.transcribe-bentham.da.ulcc.ac.uk/.

2 https://www.oldweather.org/.

3 The literal translation of Minna de Honkoku in English is "Transcribe with everyone." Also, the video tutorial of Minna de Honkoku in English is available at: https://www.youtube.com/watch?v=iX5xN4vZeao.

However, the number of records to be transcribed is vast and cannot be handled by a small group of scholars. This prompted the members of HESG to think of using crowdsourcing for transcribing historical earthquake records.

We have set the first goal of our project, Minna de Honkoku, to transcribe all the 114 books from the Ishimoto Collection, which is composed of historical earthquake records collected by a seismologist Mishio Ishimoto (1893-1940) and digitized by Earthquake Research Institute (ERI), Tokyo University. The number of pages in the books ranges from 14 to 268. The total number of pages across the 114 books is 6,386. Each digital image in the collection contains two pages, as presented in Figure. 1.

An example of two digitized pages in a book from the Ishimoto Collection

The challenge and our approach

The biggest challenge of our project is to crowdsource the reading of *kuzushiji*, which is illegible for most modern Japanese people except trained experts. Our approach to this challenge is to design our crowdsourcing system as an online learning environment where participants can learn *kuzushiji* by transcribing the earthquake records in a collaborative manner.

More specifically, Minna de Honkoku integrates crowdsourcing with online learning in the following two ways:

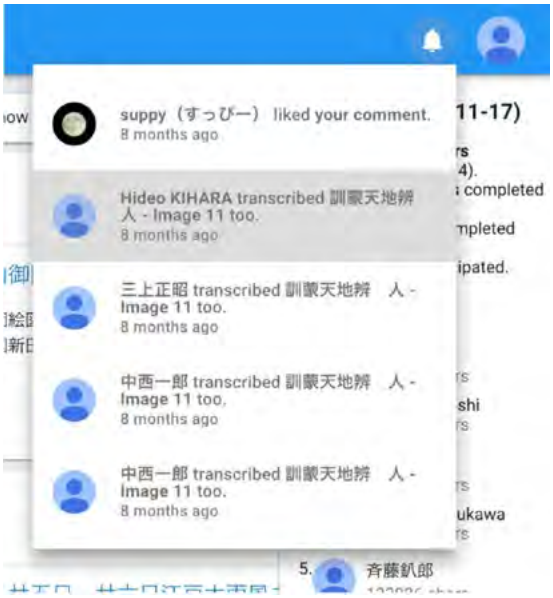
- **Collaboration with a mobile learning app**: Minna de Honkoku collaborates with KuLA[4] (Kuzushiji Learning App), a mobile learning app for reading *kuzushiji* that was developed by one of the authors (Hashimoto 2017) and has been downloaded 85,000 times since its release in 2016 (see Figure. 2). After completing a set of basic lessons for reading *kuzushiji*, the users of KuLA are invited to Minna de Honkoku as an opportunity to acquire more practical training by transcribing actual materials from pre-modern Japan. They can thus begin participating in the project as a continuation of their learning.
- **Collaborative learning through distributed proofreading**: Transcribing *kuzushiji* correctly is quite difficult, and beginners usually make a lot of mistakes. For quality control of transcriptions, Minna de Honkoku uses "distributed proofreading" adopted by Project Gutenberg (Newby 2003) but with an educational purpose; when you finish transcribing an image from a book on the transcription editor of Minna de Honkoku (see Figure. 3), your transcription will be shared and reviewed by other participants on the timeline that shows user activities in real-time (see Figure. 4). When another participant makes corrections on your transcription, you will receive a notification with the feedback, informing you of the mistakes you made and the corrections (see Figure. 5, 6).

Screenshots of KuLA

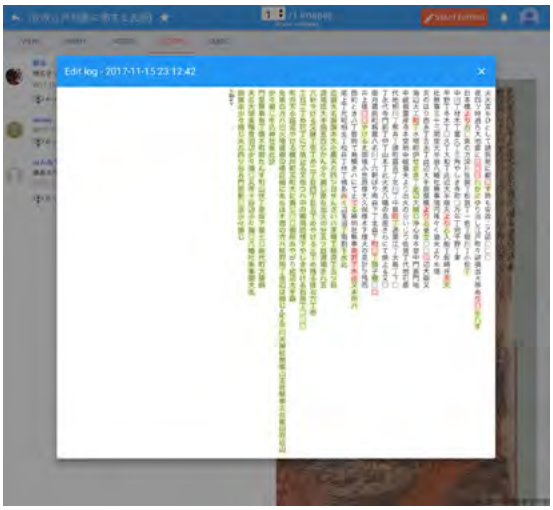
Transcription editor of Minna de Honkoku

4 Android version: https://play.google.com/store/apps/details?id=yuta.hashimoto.kula and
iOS version: https://itunes.apple.com/jp/app/id1076911000.

The timeline view of user activities



The notification panel



Corrections made by another participant (added texts are colored in green and deleted texts in red)

The results

The website of Minna de Honkoku was launched on January 10, 2017. The transcription of 114 books (6,386 pages) from the Ishimoto Collection was completed on May 31, 2017. Thus, our initial goal was completed in less than five months since the project launch. We extended our goal and added another 223 books stored in ERI. As of November 2017, 271 books out of 337 (9,254 pages out of 9,716) including those from the Ishimoto Collection have been transcribed by volunteers. A total number of 3.12 million characters have been transcribed.

A total of 3,457 people have registered an account, and 285 of them have transcribed at least one character on the website. While we were unable to include all registered users in the transcription process, a small number of regular volunteers have eagerly contributed to the project: 35 users have transcribed more than 10,000 characters, and 6 of them more than 100,000.

The background and motivations of the participants

In order to understand the backgrounds and motivations of the participants, we administered an online questionnaire to them via Google Form between March 8 to May 13, 2017. We obtained responses from 64 participants. The following is a brief summary of the questionnaire results:

- 70% of respondents (45 people) are KuLA users.
- We asked the respondents to choose the reasons of their participations from 12 pre-defined choices (multiple choices up to three are allowed). The most selected reasons are as follows:
  1. “Transcribing historical manuscripts is fun” (70%, 45 choices).
  2. “I can learn from other participants’ transcriptions and reviews” (50%, 32 choices).
  2. “I can contribute to seismic research and disaster prevention through the project” (44%, 28 choices).

The results above suggest the following: (1) KuLA works effectively as an “entrance” to Minna de Honkoku, and (2) the possibilities of collaborative learning greatly motivate the participants, although the most powerful motivation is the enjoyment gained from transcribing.

Conclusion

In this paper, we have described the background, aim, approach, and results of Minna de Honkoku, a crowdsourced transcription of historical earthquake records of pre-modern Japan. It had been often said that crowdsourced transcription of pre-modern Japanese materials

is not possible because reading *kuzushiji* is too difficult for non-trained volunteers. However, our learning-centered approach appears to have achieved considerable success. The same approach may also be used in many other countries that are facing difficulties in reading historical manuscripts due to changes in writing systems.

Lastly, desire to learn is one of the most fundamental characteristics of human beings, fulfilling which is one of the important roles of a scholar as a teacher. We therefore believe that considering academic crowdsourcing in the context of education will bring beneficial outcomes.

References

Causer, T., and Wallace V. (2012). Building a volunteer community: results and findings from Transcribe Bentham. *Digital Humanities Quarterly*, 6(2).

Eveleigh, A., et al. (2013). "I want to be a Captain! I want to be a Captain!": Gamification in the Old Weather Citizen Science Project." *Proceedings of the first international conference on gameful design, research, and applications*. ACM.

Hashimoto, Y., et al. (2017). The Kuzushiji Project: Developing a Mobile Learning Application for Reading Early Modern Japanese Texts. *Digital Humanities Quarterly*, 11(1).

Newby, G. B., and Franks, C. (2003). Distributed proofreading. In *Digital Libraries, 2003. Proceedings. 2003 Joint Conference* on (pp. 361-363). IEEE.

Data Scopes: towards Transparent Data Research in Digital Humanities

Rik Hoekstra
rik.hoekstra@huygens.knaw.nl
KNAW, Huygens ING, The Netherlands

Marijn Koolen
marijn.koolen@huygens.knaw.nl
KNAW, Huygens ING, The Netherlands

Marijke van Faassen
marijke.van.faassen@huygens.knaw.nl
KNAW, Huygens ING, The Netherlands

Introduction

For quite some time, humanities scholars have been using digital tools in addition to their established methodology to try and make sense of large and expanding data sources that cannot be handled with traditional methods alone. The digital methods have computer science aspects that may be combined with but do not readily fit into humanities methodology; an issue which is still too implicit in scholarly debate. This gives rise to a need for methodological consolidation to structure the combination of digital and established humanities methods. In this paper, we propose an approach to such consolidation, that we call *data scopes* (also see Graham, Milligan, Weingart 2016).

In principle, the methodology is relevant for many humanities disciplines, especially those that deal with large-scale heterogeneous data sets and sources. We think that digital methods should extend, not replace established methodology. Digital tools are now often employed in a methodological vacuum as if they would yield results all by themselves, but we propose that they should always be embedded in broader research methods.

Digital Data and the Research Process

Data and data sets are often presented as external to research, as the data are the sources upon which the research draws. However, all research considers datasets from the vantage point of research questions. Data and questions shape and transform each other in cycles of searching, selecting, close and distant reading, and extending the data with other data sets and annotations. In this process, the scope of the data and the scope of the research questions are aligned so the latter can be addressed. Preparing data for analysis requires interpretation and is therefore inseparably part of research and should be incorporated into the disciplinary methodology. This calls for an extension of usual source criticism with more specifically digital source criticism. In a typical research project, involving digital data, they are processed with a variety of tools that change them in many ways, making tool results and data at times inseparable. Tool and data criticism are therefore intertwined.

We will illustrate our argument with an example from research on the change in discourse regarding migration in Western Europe from 1913-2013 (van Faassen and Hoekstra, 2017). That study focuses on a 'scientization' of the migration debate and how the scientists and politicians in the debate were connected to each other. The overall research question can be addressed in many ways, but not straightforwardly answered from a single data set as the discourse spans a very long period and a lot of different media. Therefore the question was split into several specific questions that can be more directly operationalized as analyses of a combination of two digital sources, the *Publications of the Research Group for European Migration Problems* (REMP) and the online *International Migration* bulletins of the Intergovernmental Committee for European Migration (ICEM), which merged with the *REMP-bulletin*. For the first dataset, it was possible to identify key actors and their roles and to address specific questions: "who were writing forewords, prefaces or introductions to each other's work; Who ordered the research? Who financed it? etc." (van Faassen and Hoekstra, 2017, p. 7). This requires modelling of actors (persons and organisations), their roles and the relationships

between them, normalizing names of persons and mapping changing roles and names of roles over time, and linking them across publications. This in turn requires interpretations relying on domain knowledge that need to be argued for. For the long term trends in the migration discourse, the frequency in the occurrence of key terms was analysed using the other dataset (and a control set) consisting of series of article titles in *International Migration* and *International Migration Review*, two important long running journals supplemented by topic overviews from WorldCat. This required not only the use of weighted frequency analysis, but also a critical assessment of the value of the various series. Consequently, preparing data sets and analyzing them tends to happen in iterations, where initial analyses inform a next iteration of selecting, modelling, normalizing and linking, and data and research scopes are brought ever closer together.

Conceptual model

Researchers start a research project with a research question, that may be adjusted and expanded in the course of the research process. From the onset, these questions determine the research scope and therefore the scope of which data are relevant. As the research proceeds, partial questions are either answered, or prove to be unanswerable with the available data, because of their form or because of the nature and extent of the data. Researchers then interact with their data, to annotate them in such a way that it enables them to answer questions. They may also pull in other data sets to expand the existing cluster of data, so that the scope of the data will fit the research scope better.

These dynamics of the research process transform the cluster of data into a research instrument, bringing certain aspects of the data into focus, but thereby pushing others to the background or even completely out of view. Making this process transparent allows better reflection on its consequences for interpretation and shaping meaning. Digital sources and data clusters are not just 'raw material' that are the object of study, but points of departure that these iterative research interactions change (Boonstra et al. 2006, 23). It is the research process that turns a data cluster into a data scope:

- We may discern a limited number of separate activities working that are part of this research process that produces the data scope:
- *Modelling* represents the data in such a way that it will fit the research scope
- *Selection* chooses datasets and parts of datasets that are relevant for the research questions
- *Normalizing* structures data and reduces data variation so that they may be queried more easily and they can be used for comparisons, classifications or calculations
- *Linking* data connects previously unconnected data, providing them with context from other data sets. Researchers should be aware that the validity or relevance of links can be context-dependent (person X and Y are linked for a specific question because they played the same role, but the link may be invalid for other questions, see Brenninkmeijer et al 2012).
- *Classifying* data groups them in in order to reduce complexity. This adds a level of abstraction to the data

Because these activities all have the potential to transform the original data, the result is a data scope that is particular to and formed by the research process. This is related to the data life cycle (Boonstra et al. 2006), but the latter focuses on data as the product of research, while data scopes focus on the impact of transformations during the research process. As the research process transforms a cluster of data into a data scope, many levels of annotations are added to the original data sets. For the purpose of our analysis these annotations comprise all types of data enrichment, that range from structuring (for instance by adding markup to a text), to identifying named entities and keywords as structured metadata to adding explanatory notes and everything in between. It is easy to lose track of the changes, as enriching and transforming often goes in small steps and using many different manual and automatic procedures, and because transformations are often cumulative. In light of the incremental transformative effect of the research process upon the data, it is important to keep track of these changes, so that researchers can communicate about and account for their data scopes. Documenting the data changes makes both the research process and the resulting data scope transparent. In this way the research process also becomes reproducible and transferable to other research data clusters. (Groth et al. 2012; Ockeloen et al. 2013)

Discussion

The data scope concept is not a strictly theoretical model, but it is rooted in an experience of many years of empirical research with a lot of different research projects. In the presentation we will illustrate the concept with an example of a research project that combines a number of data sources to analyze the a long-term perspectives on discursive cycles relating to migration and migration policy.

Data scopes are not just a plea to work interdisciplinary and collaboratively, but in a number of research processes they are necessary. While our examples are mostly drawn from the historical sciences, the value of the concept is by no means confined to that disciplinary field.

In our view, such a methodological approach is always indispensable when there are large amounts of data

available for research and when the data are of heterogeneous in nature. Transparency and transferability is also important when researchers from different disciplines collaborate or when there is a collaboration between humanities researchers and more technologically oriented partners. In those cases, researchers, who have often mostly separate tasks, have to make sure that they understand each other to prevent misunderstandings and waste of resources.

Many humanities researchers have already adopted a part of the methodology, but unfortunately they often just use a number of tools, without acknowledging the cumulative transformative effects these have. The value of the proposed model is in the emphasis on a coherent methodological approach to doing research with (large scale) data.

References

Boonstra, Onno, Leen Breure en Peter Doorn, 2006 *Past, present and future of historical information science,* Amsterdam 2006

Brenninkmeijer, C., et al. 2012. Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision. http://linkedscience.org/wp-content/uploads/2012/05/lisc2012_submission_8.pdf

van Faassen, Marijke, Rik Hoekstra. 2017. *Modelling Society through Migration Management. Exploring the role of (Dutch) experts in 20th century international migration policy*. Conference paper. Government by Expertise: Technocrats and Technocracy in Western Europe, 1914-1973. Panel 3. Global Expertise.

Graham, S., I. Milligan, and S. Weingart. 2016. *The Historian's Macroscope: Big Digital History* http://www.themacroscope.org/2.0/

Groth, P., Y. Gil, J. Cheney, and S. Miles. 2012. "Requirements for provenance on the web." *International Journal of Digital Curation* 7(1).

Ockeloen, N., A. Fokkens, S. ter Braake, P. Vossen, V. de Boer, G. Schreiber and S. Legêne. 2013. BiographyNet: Managing Provenance at multiple levels and from different perspectives. In: *Proceedings of the Workshop on Linked Science (LiSC) at ISWC 2013*, Sydney, Australia, October 2013. http://linkedscience.org/wp-content/uploads/2013/04/paper7.pdf

Authorship Attribution Variables and Victorian Drama: Words, Word-Ngrams, and Character-Ngrams

David L. Hoover
david.hoover@nyu.edu
New York University, United States of America

Introduction

For nearly a decade, Brian Vickers has been championing a method of authorship attribution for Early Modern Drama based on the number of rare 3-6-word Ngram matches between an authorial corpus and an anonymous play that are found nowhere else in a reference corpus (Vickers, 2008, 2009, 2010, 2011, 2012). He has consistently argued that word-Ngrams are more appropriate than simple words because of the idiomatic nature of language, but he has recently intensified his attack in "The Misuse of Function Words in Shakespeare Authorship Studies" (2016). Although Vickers claims conclusive results on plays and even parts of plays, his method has been challenged (Burrows, 2012, Craig and Kinney, 2009, Hoover, 2011, 2012, Jackson, 2008, 2010). Three more recent challenges are especially significant.

Antonia, Craig, and Elliott assess "the quality and quantity of authorial markers, rather than success in classification" (2014: 152). They show that, in spite of the inherent sequentiality of language, single words are sometimes the most powerful variables, including in a test on Early Modern Drama. They test only words and word-Ngrams, but their study confirms the effectiveness of frequent words for authorship attribution of drama and offers no significant support for the effectiveness of rare Ngrams.

Jackson accepts the potential usefulness of rare Ngrams while criticizing Vickers's method (2014). He shows that the presence of many rare word-Ngram matches between Kyd's corpus and *Arden of Faversham* that are found nowhere in other plays of the period does not provide conclusive evidence of Kyd's authorship because Vickers has not compared the numbers of such matches in plays by other authors. He reports that two Shakespeare plays each "afford considerably more unique matches than any of the three canonical Kyd plays" (2014: 54). Jackson's work is impressive, but the Early Modern Drama corpus is so problematic that it cannot provide clarity on the question of Vickers's method. Early Modern spelling variability makes fully automated methods impossible, many of the plays are anonymous, many are not reliably dated, and the majority have been lost entirely (Vickers, 2008: 13).

Finally, Hoover's tests of the Vickers method on Victorian drama show that some tests fail and that strong false positives can occur (2015). These results are similar to his earlier results on narrative fiction and Modern American poetry (2010, 2011). Unfortunately, he does not include testing with frequent words, frequent character-Ngrams, or frequent word-Ngrams, and crucially does not test the methods head-to-head on exactly the same texts.

The Vickers attack on frequent function words

Vickers's recent attack on function words adds somewhat contradictory arguments to his usual argument based on

the idiomatic nature of language. First, contrary to the widely-shared view that function words are appropriate variables because their relatively unconscious makes intentional manipulation unlikely, he argues that they are not necessarily used unconsciously (2016, 6-9). Second, he argues that "they are minutiae of usage, of no appreciable significance; and . . . , they exist below the threshold of our unaided perception" (18). He ignores the fact that many analysts use longer word lists including many lexical words. He also argues that authors of drama create distinct idioms for their characters, so that word frequencies cannot capture the author's own idiom (16). This criticism applies to most genres, and seems to apply to his own method as well, as he accepts rare Ngram matches with any character in a play as evidence of authorial identity. Its chief weakness, however, is that the effectiveness of (function) words as variables for the attribution of drama can be tested empirically, so that this argument, like his other a priori arguments, seems largely irrelevant. It is to such testing that I now turn.

Frequent words, frequent word-ngrams, and frequent character-ngrams and Victorian drama

I began with a corpus of 125 Victorian plays and 2,600-word sections of plays that mirrors Vickers's corpus—with Hoover's extensions (2015)—as closely as possible. I selected 6 authors with corpora of 7 or more plays and treated 4 plays and a 2600-word section from each as if they were anonymous. Using the remaining plays as knowns, I tested these 30 test texts using words, character-2grams, -3grams, and -4grams, and word-2grams, -3grams, and -4grams, using 6 of the methods implemented in JGAAP (Juola, 2009): Burrows's Delta, Linear Discriminant Analysis, Nearest Neighbor Driver with metric Kendall Correlation TauB, WEKA Linear Regression, WEKA Multilayer Perceptron, and WEKA SMO r: Polynomial. JGAAP was chosen because of its wide variety of methods and variables; testing with Stylo (Eder et. al., 2016) and other methods gave similar results. The best results were 93.3% correct for the 300 most frequent words and the 300 most frequent character-4grams, though words performed better overall. Character-2grams and 3grams were weaker, and word-Ngrams were weaker still (4grams weaker than 3grams, which were weaker than 2grams).

The best results for 7 authors with smaller corpora were 95% correct, but this was achieved just once, based on the 300 most frequent words, and results were weaker overall, presumably because of the smaller corpora (2-4 plays) and because a higher proportion of the test texts were 2600-word sections. Words again gave the best results, then character-2grams, character-4grams, and word-2grams. Longer word-Ngrams gave much weaker results.

These results clearly refute Vickers's arguments that frequent words are inappropriate for the attribution of drama. Although character-Ngrams are effective for these texts, words alone are even more effective, and increasingly longer word-Ngrams are increasingly ineffective. These results are quite strong, especially considering that 6 of the 30 test texts in the first test and 9 of the 20 in the second are 2,600-word sections.

Rare ngrams and Victorian drama

How well does Vickers's rare Ngram method perform on these same texts? I collected all the 3- to 6-word Ngrams that occur at least twice in the 125-text corpus. I established a reference corpus of 86 plays by 14 authors (minimum 3 plays each), and a set of 14 plays and 8 sections of 2,600 words by authors outside the reference corpus, and 2 additional plays and 15 sections of 2,600 words by reference set authors. These 39 additional texts allow a comparison of the frequencies of Ngram matches between the authorial corpus and the author's test texts and between the authorial corpus and texts by other authors. If Vickers is right, the test texts should contain many more rare Ngram matches with the authorial corpus than texts by other authors do.

To test an author from the reference corpus, I remove that author's plays and delete all the 3- to 6-grams that occur in the remaining reference corpus. I then create an authorial corpus that matches the known texts tested in JGAAP and delete all the 3- to 6-grams that are absent from it. Because the texts differ greatly in length, I divide the total number of occurrences of each Ngram by the length of the text in words to give a measure of frequency relative to text-length, and then sort the test texts and sections on this frequency. (Vickers gives raw numbers, ignoring the effect of the lengths of the plays; Jackson uses matches per 1,000 lines as a measure of relative frequency [2008: 123-25].) The most favorable measure of success is to count just one error for each of the author's test texts that is outscored by any text by another author. By this measure, the overall success of Vickers's method for the 30 test texts 63.3%. All 5 texts by Byron and Phillips are correctly attributed, showing that the test is sometimes effective. Unfortunately, the four other authors show errors for 1, 2, 3, and all 5 test texts. For the 7 authors with smaller corpora, the results are much worse: no author's texts outscore all texts by other authors, and the overall success rate is just 45%. (Many of these small corpora are the same size as Vickers's Kyd corpus.) Yet, even these poor results underestimate just how badly the method fails: 20 or more plays or sections by other authors (more than half) outscore 3 of the test texts in the 2 sets, and 10 or more outscore another 5 of the test texts.

Conclusion

As attractive as Vickers's rare Ngram method initially seems, and in spite of its apparent effectiveness for some authors, it cannot offer the conclusive proof of authorship

that Vickers claims. Frequent words, contrary to Vickers's a priori arguments, are quite effective in attributing plays and even short sections of plays to their authors, and very much more effective than rare Ngrams (as are frequent character-Ngrams). It is presumably possible that Early Modern Drama is different enough from Victorian drama that the method works better there. However, the results presented here, combined with those for narrative fiction and modern American poetry (Hoover, 2011, 2012), strongly suggest that rare Ngram matching is not a sound method of authorship attribution.

References

Antonia, A., Craig, H. and Elliott, J. (2014). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. *Literary and Linguistic Computing*, **29**(2): 147–63.

Burrows, J. (2012). A second opinion on 'Shakespeare and authorship studies in the twenty-first century'. *Shakespeare Quarterly*, **63**(3): 355–92.

Craig, H., and Kinney, A., eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1): 107-121

Hoover, D. L. (2011). Delta, Zeta, and Iota: An ngrammatical investigation. *Language Individuation: A Symposium in Honour of John Burrows*, University of Newcastle, Australia, July 4-8.

----. (2012). The rarer they are, the more there are, the less they matter. *Digital Humanities 2012.* Hamburg: Hamburg University Press: 218-21.

----. (2015). Rare n-Grams, Victorian drama, and authorship attribution. *Digital Humanities 2015: Global Digital Humanities*, Sydney: University of Western Sydney, n.p.

Jackson, MacD. P. (2008). New research on the dramatic canon of Thomas Kyd. *Research Opportunities in Medieval & Renaissance Drama*, 47: 107-127.

----. (2010). Parallels and poetry: Shakespeare, Kyd, and *Arden of Faversham*. *Medieval & Renaissance Drama in England,* 23: 17-33.

----. (2014). *Determining the Shakespeare Canon: Arden of Faversham and A Lover's Complaint.* Oxford: Oxford University Press.

Juola, P. (2009). JGAAP: A system for comparative evaluation of authorship attribution. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(1): 1-5.

Literature Online (LION)

Vickers, B. (2008). Thomas Kyd, secret sharer. *Times Literary Supplement,* 18 April: 13-15.

----. (2009). The marriage of philology and informatics. *British Academy Review,* 14: 41-44.

----. (2010). Disintegrated: Did Thomas Middleton really adapt Macbeth? *Times Literary Supplement*, 28 May: 13-14.

----. (2011). Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly,* **62**(1): 106-42.

----. (2012). Identifying Shakespeare's additions to *The Spanish Tragedy* (1602): A new(er) approach. *Shakespeare*, 8: 13-43.

----. (2012) The misuse of function words in Shakespeare authorship studies, full paper version of what is the appropriate authorship attribution method for Elizabethan drama? *Göttingen Dialog in Digital Humanities*, November 30.

Digital Humanities in Latin American Studies: Cybercultures Initiative

Angelica J. Huizar
ahuizar@odu.edu
Old Dominion University, United States of America

My research is multi- and interdisciplinary focusing on electronic literature and cybercultures in/of Latin America. My latest articles and book manuscript explore the divide and convergence in literature and technology. This project lends itself well to the application of those theories and the evaluation of how they can best be implemented in classroom practices and complemented with co-curricular modules. I will therefore present my research findings on the use of Digital Humanities components specifically for the teaching of Latin American Studies. The presentation would thus serve as a report of: 1) initial research findings and best practices found at other institutions; 2) work accomplished at the DHSI 2018 Workshop (Victoria, Canada) "Critical Pedagogy and Digital Praxis in the Humanities"; 3) feedback gained from presentation at the DHSI 2018 Conference & Colloquium; and, 4) samples of syllabi to foster a lively discussion on the application of such a course with co-curricular components for Latin American Studies programs.

The goal of this project is to do a detailed study of program and curriculum design at other institutions on the use of DH modules specifically for Latin America/US Latino culture with a focus on pedagogical methodologies that engage critically about the problems that DH platforms do and do not resolve in Latin American Studies. The course design and the co-curricular components complement and intersect each other. This project will facilitate the assessment of various curriculums and specialized courses for the digital humanities and would ultimately lead to develop a course for all students interested

in DH Latin American Studies.[1] Course components will include developing language proficiency, learning and using DH tools, and analyzing the effectiveness and drawbacks of such technologies specifically to Latin American Studies.

The interactive, systematic, and innovative features of digital humanities have been proven to advance language learning both in and outside of the classroom. Through exploring different forms of digital humanities, including multi-media, online archives, as well as existing web tools like Google Earth and Twitter, instructors and scholars of foreign languages not only facilitate collective and immersive language learning, but also broaden and deepen students' exposure and knowledge of foreign culture. These projects break the traditional geographical and cultural boundaries in learning a foreign culture and/or language. Therefore, it is essential for instructors to reflect on how best to incorporate digital humanities in language/culture learning, and to determine to what extent digital learning complements and even replaces traditional ways of teaching and learning.

Students will be encouraged to adapt these new tools of analysis to their own future career objectives. The field of Digital Humanities is collaborative and very interdisciplinary as it produces new scales of analysis with varying modules (texts, maps, audio-mapping and networks) which may include experiments across modalities with: distant reading alongside close reading techniques, programming language, audio creation, geotagging, speech recognition encoding documents in TEI (Text Encoding Initiative[2]), learning the basics of computational text analysis, programming chatbots using the Python programming language, etc. The course will also note the drawbacks or pitfalls of the use of technology.

However, the skills needed in DH have less to do with a particular hardware or commercial software and more about engaging in digital literacy (train interpretative methods necessary for critical analysis), and showcase how digital humanities is valuable to better understand Latin America's transformations in the production, circulation and reception as well as its impact on culture, politics, history, literature, music, etc. The course will encourage students to develop more analytical projects from the use of such modalities. The focus will also be to analyze and address *why* this method of learning is complementary or even superior to traditional methods, specifically addressing the impact and implications that technology involved on ideologies, ethics and ideas. For example, a more involved topic would approach the idea of "mapping" as interpretation of geospatial data in GIS, georectify historical maps in Map Warper, manage digital archival objects in Omeka, and use Neatlineto build "deep maps" of particular neighborhoods or landmarks in a city, layering historical photographs, maps, geospatial data, literary texts, and other elements to build analysis about their city.

Additionally, the course will attempt to link to public libraries (Slover in Norfolk), museums (Chrysler, Mariners, Living Museum), research centers, community groups (Norfolk Chamber of Commerce, Hispanic Chamber of Commerce, Hispanic Community Dialogue) or other campus-level initiatives (ODU's Institute of Humanities "Mapping Lambert's Point Project," for instance). The goal is to build projects that make use of the University and community's collections. These public projects can energize students to work that much harder, as they can create materials with a chance of life beyond the classroom itself. The course will draw on resources from, participate in and continue their learning with the Regional, National and International Network[3] aimed to promote digital humanities initiatives to Old Dominion University faculty and to learn from and collaborate with external groups.[4] This network would be dedicated to exploring, analyzing, and sharing the cultural and visual modalities of digital humanities in the research and teaching of Latin America. The network would engage in these discussions through symposia for faculty and students with guest speakers or virtual conferences, virtual exhibitions, and online or hybrid workshops.[5] The network and initiatives that I foresee fostering and/or facilitating may include:

K-12 Service Education: Working with the College of Education and the Licensure Students in the World Languages and Cultures Department to: Expand on its longstanding educational outreach commitments with K-12 educators and students at the local and state level; and, serve as a resource to K-12 educators working to meet Virginia Performance Standards as they relate to Latin American content in the social, natural, and life sciences by

Language Without Borders Initiative: Create the next generation of global professionals through innovative language education, with Superior level proficiency in Spanish and overseas internship experience.

DH and Latin/o American Cybercultures Initiative: Exposure to the digital culture of Latin America through seminars, symposia, courses, exhibitions, and workshops.

---

1 Students in this course would include (but not limited to) those in the Latin American Studies Minor Program, International Business, International Studies (BAIS and GPIS), Humanities, Political Science, Spanish majors and Minors, World Cultural Studies majors and minors.

2 Text Encoding Initiative Markup Language at the University of Virginia, https://dh.virginia.edu/tool/text-encoding-initiative-markup-language-tei (for my future reference)

3 To be featured in the Latin American Studies Program website

4 I already have established contacts and am in current collaborations with: Centro de Cultura Electrónica in Mexico City; the project Cultura Digital Chile (Universidad Diego Portales, Chile); the Latin American and Digital Humanities/Cybercultures at University of Georgia; the Digital Latin American Cultures Network: Researching the Cultural Dimensions of New Media in the United Kingdom; I am also a board member of the organization Lit-e-Lat: Red de Literatura Electrónica.

5 For example, "Tecnoestética y sensorium contemporáneo: arte, literatura, diseño y tecnología" in September 2017 in Córdoba Argentina;

A machine learning methodology to analyze 3D digital models of cultural heritage objects

Diego Jimenez-Badillo
diego_jimenez@inah.gob.mx
Instituto Nacional de Antropología e Historia, Mexico

Salvador Ruiz-Correa
salvador.ruiz@ipicyt.edu.mx
Instituto Potosino de Investigación en Ciencia y Tecnología, Mexico

Mario Canul-Ku
mariocanul@cimat.mx
Centro de Investigación en Matemáticas, A.C., Mexico

Rogelio Hasimoto
hasimoto@cimat.mx
Centro de Investigación en Matemáticas, A.C., Mexico

Thanks to recent advances in scanning technologies there has been an increase in the number of methods developed for digitizing cultural heritage objects. Many of the resulting 3D models are used for visualization or archiving purposes. Unfortunately, there are still few projects oriented to gain archaeological knowledge from point clouds and triangular meshes.

In this paper we present some results of an ongoing project that applies machine learning and computer vision techniques for recognizing, retrieving and classifying cultural heritage objects in an automatic way (Jiménez-Badillo, et al. 2010, 2013; Jiménez-Badillo and Román-Rangel, 2016, 2017; Roman Rangel and Jiménez-Badillo, 2015, Román-Rangel et al., 2014, 2016a, 2016b; Jiménez-Badillo and Ruiz-Correa, 2017). The presentation focuses specifically on a method to analyze style variations of archaeological artefacts with minimal human intervention. This is based on a 3D morphing algorithm proposed by Shelton (2000). Our implementation allows analyzing pairs of objects whose shapes represent the canonical extremes of a continuum, that is, objects that belong to two different "styles" within a cultural tradition. The purpose of the algorithm is taking two extreme shapes (i.e. 3D point-clouds, surface meshes or 3D digital models) as input in order to extract several 3D virtual models whose shape or "style" lies "in-between" the two extremes. This is useful in situations where archaeologists need to decide to which extreme a real artefact is more similar. Archaeologists can also apply the algorithm to compare each object of a collection against all the other members of the set. This would produce an "atlas" of the shape variations expected for such collection, which in turns would facilitate the application of a classification method based on machine learning.

The formal mathematical details of this approach can be found in the original paper by Shelton (2000). During the presentation, we plan to offer an intuitive explanation of the algorithm for the benefit of those Humanists who are not experts in mathematics. This can be summarized as follows:

- Mathematically speaking, the problem consists in finding correspondences between two 3D point-clouds or surface meshes. This means finding points in surface "A" that match corresponding points in surface "B" with minimal user intervention.
- The challenge is how to make that the algorithm recognizes meaningful geometric correspondences between models "A" and "B". In other words, how to find structural geometric correspondences between the points that define, for example, the nose of model "A" with the nose of model "B", and the same for all other features of the masks.
- The solution proposed by Shelton is a mathematical equation that fits three criteria:

  Similarity. For each point a on surface A, the function *C(a)* must find a point close or on the surface B.

  Structure. Function C must produce the least possible distortion in the transition from A to B. In other words, the result of function C(a) must have a geometric structure similar to A.

  Plausibility. Function C must represent a realistic model derived from surface A.

The first property establishes that C must find real points in A that match points in B. The second condition establishes that the correspondences found between A and B must not be arbitrary. On the contrary, there must be matching substructures of A present in B (e.g. the matching of the nose in A must have some correspondence with the nose in B), so that the deformation makes sense (figure 1). The last condition guarantees that the deformation includes the previous knowledge of the user in terms of which forms are acceptable for the deformation, because it makes no sense to transform a face mask into an airplane, for example.

The idea for this project came from the need to rank shape similarities in a collection of archaeological stone masks from Mexico. This includes masks belonging to several well-defined styles, but it also includes many others that cannot be clearly positioned within a specific class because they share features of two or more canonic styles (figure 2).

These masks were found in the Sacred Precinct of Tenochtitlan, the main ceremonial Aztec complex, located in Mexico City. The schematic features of these objects set them apart from other artifacts with more naturalistic style. This has attracted the attention of many specialists and during the last three decades these items have been the subject of intense debate for two main reasons:

First, the 162 masks were located in 14 Aztec offerings dating from 1390 to 1469 A.C., yet they do not show

typical “Aztec” features. Indeed, their appearance resembles artifacts from the southern State of Guerrero, particularly from the Mezcala region, which is hundreds of kilometers away from the ancient Tenochtitlan.

Such origin would not be rare, as it was common for the Aztecs to import goods from other regions either by trade or by extracting tribute from conquered towns. The style of the masks and figurines, however, is more difficult to explain. It is similar, if not identical, to the style of objects produced in Mezcala and other places of Guerrero during much earlier times, probably during Classic (200 to 1000 A.C.) or even Preclassic times (2000 B.C. to 200 A.C.), while the offerings are Late Postclassic contexts. This leads to the question: Did the Aztecs collected “antique” objects to re-use them in their own offerings?, or the Guerrero/Mezcala styles survived till the late Postclassic period and therefore the offering objects were produced during Aztecs times? It is worth noticing that before the Aztec offering findings very few Mezcala style artifacts had been found in Postclassic contexts. Unfortunately, not enough stratigraphc information is available for collections from Guerrero, so specialists rely purely on stylistic considerations to explain the chronology of these artifacts.

Second, it is not clear how many Guerrero/Mezcala styles exist. Some specialists believe there are at least five different traditions (Covarrubias, 1948, 1961; Olmedo and González, 1986; González and Olmedo, 1990), while others recognize only four (Gay, 1967) or two (Serra Puche, 1975). The diversity of views is due in part to a lack of contextual information available for the majority of artifacts found in Guerrero, but it also reflects the subjective criteria used to classify such artifacts.

Clearly, more objective methods are needed to answer questions such as: how many styles were developed in the Guerrero/Mezcala regions?; which specific styles are represented among the offering objects found in the Sacred Precinct of Tenochtitlan?; and more importantly for the purposes of this paper: To which style each mask belongs?

Previous studies have tried to solve some of these questions by analyzing object shapes with clustering methods (Olmedo and González, 1986 , González and Olmedo, 1990, Jiménez-Badillo and Ruiz-Correa, 2017), but we believe that the application of morphing algorithms could produce a more objective assessment to solve the problem of style attribution in this and other archaeological collections.

Our application takes examples of two canonical styles and applies the deformation algorithm in order to produce a hundred virtual 3D models whose shapes go from one to the other extreme (figure 3). The virtual models produced in this way represent intermediate steps from style “A” to “B”. Each virtual model has associated a number that indicates its degree of similarity to style “A” or “B”. We can then examine a real archaeological object “c” to determine if its shape is closest to “A” or “B” and by how much. During the presentation we demonstrate a piece of software that implements the morphing algorithm and show, in a visual way, which parts of a 3D model suffer more deformation while transitioning from style “A” to “B” (figure 4).

As this is a work in progress, we are interested in receiving feedback from the audience about the relevance of our tools to resolve similar or new archaeological questions and welcome collaboration with other research projects willing to try this generic software for new applications.

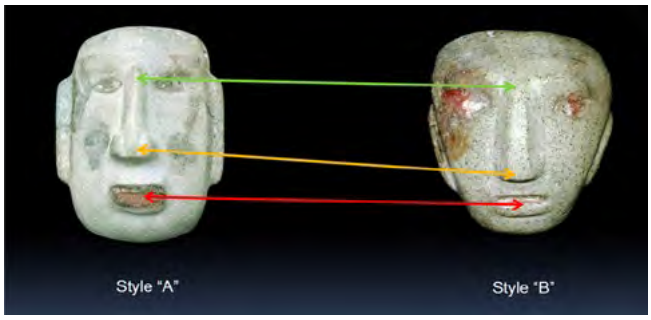


Figure 1. The morphing algorithm proposed by Shelton (2000) seeks to generate a sequence of intermediate virtual models from “A” to “B”. To do that, it needs to identify correspondences in geometric substructures (i.e. noses, mouths, etc.) in both models

Figure 2. The first three columns from left to right show nine masks belonging to three different styles from Guerrero, Mexico. The fourth column on the extreme right shows three masks that cannot easily be attributed to the Sultepec, Chonatl or Mezcala styles

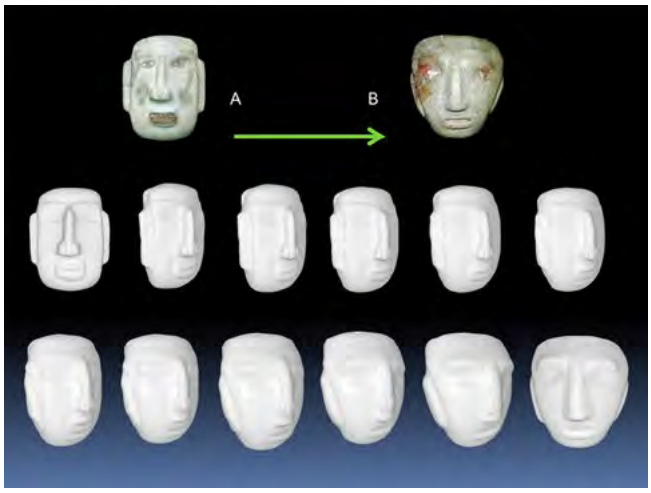


Figure 3. A sequence of 3D virtual models produced with Shelton´s algorithm (Shelton, 2000). Notice that each model represents a transition between the shape of mask "A" and the shape of masks "B"

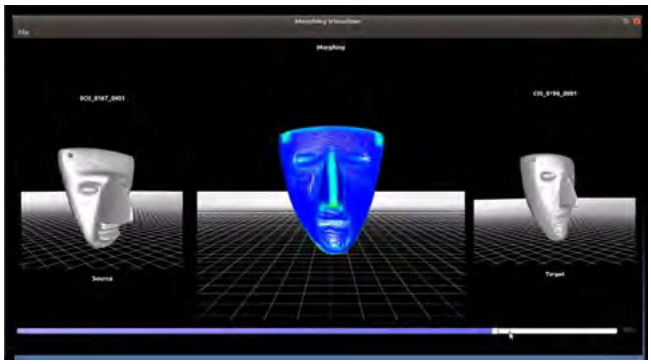


Figure. 4. A snapshot of the morphing software implemented as part of the research project

References

Covarrubias, M. (1948). Tipología de la industria de piedra tallada y pulida de la cuenca del Río Mezcala. In *El Occidente de México*. México: Sociedad Mexicana de Antropología, pp. 86-90.

Covarrubias, M. (1961). *Arte indígena de México y Centroamérica.* México: Universidad Nacional Autónoma de México.

Gay, C. T. (1967). *Mezcala Stone Sculpture: The Human Figure*. New York: The Museum of Primitive Art, Studies Number Five.

González, C., y Olmedo, B. (1990). *Esculturas Mezcala en el Templo Mayor.* México: Instituto Nacional de Antropología e Historia.

Jiménez-Badillo, D. and Román-Rangel, E. (2016). Application of the ´Bags-of-Words Model´ (BoW) for Analyzing Archaeological Potsherds. In Campana, S., Scopigno, R., Carpentiero, G., and Cirillo, M. (eds.), *CAA 2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology,* Vol. 2. Oxford: Archaeopress, pp. 847-856..

Jiménez-Badillo, D. and Román-Rangel, E. (2017). Clasificación automática de fragmentos de vasijas arqueológicas mediante el modelo Bolsa de Palabras. In Jiménez-Badillo (ed.), *Arqueología Computacional. Nuevos enfoques para la documentación, análisis y difusión del patrimonio cultural*. México: Instituto Nacional de Antropología e Historia, pp. 111-126.

Jiménez-Badillo, D. and Ruiz-Correa, S. (2017). Análisis tridimensional de objetos arqueológicos con técnicas de visión por computadora. In Matos Moctezuma, E. and Ledesma Bouchan, P. (eds.), T*emplo Mayor. Revolución y estabilidad.* México: Instituto Nacional de Antropología e Historia, pp. 199-214.

Jiménez-Badillo, D., Ruíz-Correa, S. and García Alfaro, W. (2010). 3D Shape Matching and Retrieval for Archaeological Analysis. In Melero. F. J., Cano, P. and Revelles, J. (eds.), *Fusion of Cultures. Abstracts of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, Granada, Spain, April 6-9, 2010, pp. 583-586.

Jiménez-Badillo, D., Ruiz-Correa, S. and García Alfaro, W. (2013). Developing a Recognition System for the Retrieval of Archaeological 3D Models. In Contreras, F., Farjas, M. and Javier Melero, F. J. (eds.), *CAA Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 6-9, 2010*. Oxford: Archaeopress, BAR International Series 2494, pp. 325-332.

Olmedo, B., y González, C. (1986). *Presencia del estilo Mezcala en el Templo Mayor: Una clasificación de piezas antropomorfas.* Thesis presented as requirement for the degree of Bachelor in Archaeology. México: Escuela Nacional de Antropología e Historia.

Román-Rangel, E. and Jiménez-Badillo, D. (2015). Similarity Analysis of Archaeological Potsherds Using 3D Surfaces. In Carrasco-Ochoa, J. A. et al. (eds.), *Lecture Notes in Computer Science*, Vol. 9116, *Proceedings of the 7th Mexican Conference on Pattern Recognition,* Mexico City, June 24-27, 2015. Switzerland: Springer International Publishing, pp. 125-134. DOI: 10.1007/978-3-319-19264-2 13

Roman-Rangel, E., Jimenez-Badillo, D. and Aguayo-Ortiz, E. (2014). Categorization of Aztec Potsherds Using 3D Local Descriptors. In Jawahar, C. V. and Shan, S. (eds.), *Lecture Notes in Computer Science,* Vol. 9009: *Computer Vision - ACCV 2014 Workshops, Part II,* Singapore, November 1-2, 2014. Switzerland: Springer International Publishing, pp. 567-582. DOI: 10.1007/978-3-319-16631-5 42

Román-Rangel, E., Jiménez-Badillo, D., and March-and-Maillet, S. (2016a). Classification and Retrieval of Archaeological Potsherds Using Histograms of Spherical Orientation. *ACM Journal of Computing and Cultural Heritage 9*(3): 17:1-17:23.

Román-Rangel, E., Jiménez-Badillo, D., and March-and-Maillet, S. (2016b)**.** Rotation Invariant Local

Shape Descriptors for Classification of Archaeological 3D Models. In Martínez Trinidad, J. F. et al. (eds.), *Lecture Notes in Computer Science*, Vol. 9703: *Proceedings of the 8th Mexican Conference on Pattern Recognition,* Guanajuato, Mexico, June 22-23 2016. Switzerland: Springer International Publishing, pp. 13–22. DOI: 10.1007/978-3-319-39393-3 2

Serra Puche, M. C. (1975). Intento de seriación en esculturas de Guerrero. Cronología del estilo Mezcala. In *XIII Mesa Redonda de la Sociedad Mexicana de Antropología*, Jalapa, México: Sociedad Mexicana de Antropología.

Shelton, C. R. (2000). Morphable Surface Models. *International Journal of Computer Vision*, 38(1): 75-91.

Women's Books versus Books by Women

Corina Koolen
corina.koolen@huygens.knaw.nl
KNAW, Huygens ING, The Netherlands

Introduction

Books written by and marketed towards women have been analyzed mostly in the context of popular culture (Radway, 1987; Hollows, 2000; Modleski, 2008). In literary criticism however, fictional work by women is regularly held up to such 'women's novels' to measure the quality (van Boven, 1992; Vogel, 2001; Groos, 2011). This connection made between female author gender and popular feminine novels is likely based on bias, but it is not yet well-researched in computational stylistics. In this paper we present a pilot study for examining this potential bias, through the combination of a reader survey and text analysis.

Related work

Although computational stylistics is now quite common in analysis of fiction (i.e. Semino and Short, 2004), 'women's' genres are not researched often in relation to literature. Jautze et al. (2013) focuses on differences between the syntactic make-up of sentences in literary novels and so-called 'chick lit' (cf. Ferriss and Young, 2013); Montoro (2012) performs computational-linguistic analysis on chick lit as opposed to a BNC sampler corpus – but not to literary fiction specifically.

Women's books

What is the relationship between books by women and 'women's books' according to readers? We examine this through results of the National Reader Survey (2013). Respondents were supplied with a list of 401 recent Dutch-language novels (translated and originally Dutch, published between 2007-2012) that were most often loaned from libraries and bought from bookstores between 2009-2012 (Koolen et al., in preparation).[1],[2] Respondents supplied ratings of literary quality on books they had read (on a scale of 1-7) and were allowed to motivate one of their scores.

Overall, works by female authors are judged to have lower literary quality (M=3.92, SD=0.81) than those by male authors (M=4.73, SD=1.04); t(344)=-8.34, p < 0.01. This is partially caused by romantic novels, which are mainly written by women (M=3.02, SD=0.60).[3] More surprisingly, within general fiction female authors' works scores' (M=4.55, SD=0.84) are significantly lower than for male's (M=5.53, SD=0.73); t(120)=-7.60, p<0.01.

An analysis of the motivations shows that the concept of the 'women's book' ('vrouwenboek') and similar gendered terms are used dozens of times to explain what literary quality is **not**; a male equivalent is mentioned twice ('men's book', 'boy's book'). Examples of novels referred to as 'women's' book' are translations of *Eat, Pray, Love* by Gilbert (general fiction), *Remember Me?* by Kinsella (romantic fiction) and *The Ice Princess* by Läckberg (suspense). Thus, works by female authors are equated with 'women's books' regardless of the novel's own genre. Perceived connections that respondents provide are: bad story (about love), a simple style, no deeper layers, etc.. But how much do 'women's books' differ from novels that are perceived as literary? And are they more strongly connected to other female-authored novels than to male-authored ones?

Text analysis

We perform two experiments as a first exploration. We compare present-day romantic novels by female authors (R), predominantly chick lit, to general fiction by women (GF) and general fiction by men (GM). We select the lowest scoring novels in the romantic genre and the highest in the general fiction genre (i.e. the most 'literary' ones according to our respondents), to find the clearest contrast (cf. Table 1). We use only one novel per author, unless the author uses a different pen name (Kinsella/Wickham).

Genre / gender author (av. rating literariness)	Transl. from English	Originally Dutch
Romantic / female (2.8)	10	2
General fiction / female (5.2)	10	2
General fiction / male (5.9)	10	2

Table 1. Division of books in the sub-corpus

1 Note that the Riddle corpus' novels show the one-sidedness of the market: it consists of few genres, there are very few novels by people of color, it contains mostly European and North-American novels.
2 The factor of translation will be taken into account in further development of this pilot, for information on effects within the larger project, see van Dalen-Oskam, 2016.
3 To distinguish genres, we roughly base ourselves on Dutch publishers' assignments of genre, which is done through a uniform classification system in the Netherlands.

Experiment 1: style

As we have shown, the style of 'women's books' is seen as inferior. We use stylometric analysis to explore this notion, adding Gilbert's *Eat, Pray, Love* to this experiment (cf. Section 3); a hybrid of general fiction and romance. Stylometric analysis is most often used to perform authorship recognition, but has been successfully applied to identify gender (Rybicki, 2015)"container-title":"Digital Scholarship in the Humanities","page":"746-761","volume":"31","issue":"4","ISSN":"2055-7671","journalAbbreviation":"Digital Scholarship in the Humanities","author":[{"family":"Rybicki","given":"Jan"}],"issued":{"date-parts":[["2015"]]}}},"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} and fictional genres (Allison et al., 2011). We apply the method detailed in Eder (2017). First, with R-package Stylo (Eder et al., 2016), we construct a bootstrap consensus tree based on the 100 through 1,000 most frequent words with 100-word intervals, using Classic Delta to calculate stylistic similarity (cf. Eder, 2017). Second, we use network analysis and visualization tool Gephi to visualize the novels' connectedness (Bastian et al., 2009). Color-codes are based on modularity, which visualizes groupings of greater inner coherence (Blondel et al., 2018). Finally, we apply the ForceAtlas2 algorithm to make groupings more visually distinct.

Network visualization of the novels' stylistic proximity (R = romantic, GF = general fiction/female author, GM = general fiction/male author). Colors indicate groupings based on modularity

Fig. 1 shows six clusters. Part of the romantic novels (blue, soft pink) are indeed separated from the general fiction (other colors); Stockett's *The Help* is stylistically connected strongest to romantic novels. General fiction by female and male authors hardly form clusters of their own. Except for one 'male' cluster which contains a Barnes' novel and an outlier: Gilbert's novel – which is seen as a 'women's novel' by our respondents. Weiner, known for opposing the 'chick lit' label to her work (Mead, 2014) has a stronger connection to general fiction. In other words, stylistically seen, part of the romantic novels appear to have a specific signature, but most novels by female authors are not obviously stylistically connected to them.

Experiment 2: sentiment

We now use Linguistic Inquiry and Word Count (LIWC), a word list analysis tool, which has a dictionary for Dutch (Boot et al., 2017) and has been applied to literary fiction in genre analysis (Nichols et al., 2014). LIWC contains a number of content and sentiment-related categories that are of interest. Attention to physical appearance, a (heterosexual) love story, work and friendship and have been identified as themes of chick lit novels (Gill and Herdieckerhoff, 2006), which are the main component of the romantic genre in this corpus. We report significant differences on salient categories in an independent *t*-test between averages of groups ($p < 0.01$).

LIWC category	Romantic-Gen. Female	Romantic-Gen. Male	Gen. Female-Gen. Male
Articles		X	
Prepositions		X	
Affect	X	X	
Posemo	X	X	
Negemo			
Social		X	
Communication	X	X	
Friends	X	X	
Job	X		
Swearwords	X		

Table 2. Significant differences ($p < 0.01$) between groups

Table 2 shows that romantic novels differ from general fiction in some ways: more positive emotions, but no significant difference in negative emotions, more words pertaining to friendship. The romantic novels differ in other ways from either the female or the male-authored literary novels: there are more job-related words in the romantic novels than in female-authored general fiction; less articles and prepositions than male-authored general fiction. Female-authored literary novels and male-authored ones do not significantly differ on any category. This might indicate that when comparing literary fiction to romantic novels, readers choose to focus on commonalities with female authors and differences with male authors, whereas differences between female authors and commonalities with male authors are overlooked. However, we need to be careful with interpretations of *t*-tests in LIWC (cf. Koolen and van Cranenburgh, 2017). Additional analysis will need

to be performed to identify within-group differences. Finally, physicality and the body do not appear to be specific to romantic novels. This finding corroborates earlier research, see Montoro (2012) and Koolen (2018).

Conclusion

Romantic novels appear to be more different from all general fiction than the general fiction differs among authors of female and male gender. They contain signature elements, albeit not all the expected ones (positive emotions and friends, not attention to appearance). Part of the romantic novels are clearly different from general fiction stylistically, but a number of them cluster with male-authored general fiction; most notably work by Gilbert and Weiner. Although further testing is needed, they show that computational stylistic analysis might be used to paint a more objective picture of the actual style of contemporary novels by female authors and the relationships between them. We offer a speculation: if we consider the romantic novels in this corpus to be 'women's novels', there are a several indications that commonalties between female-authored general fiction and romantic novels are stressed heavily and this might be a reason female authors' novels are judged to have less literary quality. Nevertheless, we do not aim to assert 'low' literary quality of the romantic novels, either. To examine gendered quality perceptions further, we will include other fictional genres in future research.

References

Allison, S. D., Heuser, R., Jockers, M. L., Moretti, F. and Witmore, M. (2011). Quantitative formalism: an experiment. *Stanford Literary Lab* https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf (accessed 27 November 2017).

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 8: 361–62.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008.

Boot, P., Zijlstra, H. and Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1): 65–76.

Eder, M. (2017). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, 32(1): 50–64.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1): 107–21.

Ferriss, S. and Young, M. (2013). *Chick Lit: The New Woman's Fiction*. New York: Routledge.

Gill, R. and Herdieckerhoff, E. (2006). Rewriting the romance: new femininities in chick lit?. *Feminist Media Studies*, 6(4): 487–504.

Groos, M. (2011). Wie schrijft die blijft? Schrijfsters in de literaire kritiek van nu (Who writes remains? Female writers in today's literary criticism). *Tijdschrift Voor Genderstudies*, 3(3): 31–36.

Hollows, J. (2000). *Feminism, Femininity and Popular Culture*. Oxford: Manchester University Press.

Jautze, K., Koolen, C., van Cranenburgh, A. and de Jong, H. (2013). From high heels to weed attics: a syntactic investigation of chick lit and literature. *Proceedings of the Workshop on Computational Linguistics for Literature*. Atlanta, GA, USA: Association for Computational Linguistics, pp. 72–81.

Koolen, C. (2018). *Reading Beyond the Female: the Relationship between Perception of Author Gender and Literary Quality*. Amsterdam: University of Amsterdam.

Koolen, C. and van Cranenburgh, A. (2017). These are not the stereotypes you are looking for: bias and fairness in authorial gender attribution. *Proceedings of the First Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, pp. 19–29.

Koolen, C., van Dalen-Oskam, K., van Cranenburgh, A., Nagelhout, E. and de Jong, H. (in preparation). Literary quality in the eye of the Dutch reader: the National Reader Survey and its results.

Mead, R. (2014). Written off: Jennifer Weiner's quest for literary respect. *The New Yorker* https://www.newyorker.com/magazine/2014/01/13/written-off (accessed 27 November 2017).

Modleski, T. (2008). *Loving with a Vengeance: Mass Produced Fantasies for Women*. New York: Routledge.

Montoro, R. (2012). *Chick Lit: The Stylistics of Cappuccino Fiction*. London, New York: Bloomsbury Publishing.

National Reader Survey (2013). *Het Nationale Lezersonderzoek*. https://www.hetnationalelezersonderzoek.nl/ (accessed 26 April 2018).

Nichols, R., Lynn, J. and Purzycki, B. G. (2014). Toward a science of science fiction: applying quantitative methods to genre individuation. *Scientific Study of Literature*, 4(1): 25–45.

Radway, J. A. (1987). *Reading the Romance: Women, Patriarchy and Popular Literature*. London: Verso.

Rybicki, J. (2015). Vive la différence: tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities*, 31(4): 746–61.

Semino, E. and Short, M. (2004). *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. New York: Routledge.

van Boven, E. (1992). *Een Hoofdstuk Apart: 'Vrouwenromans' in de Literaire Kritiek 1898-1930 (A Separate Chapter: 'Women's Novels' in Literary Critique 1898-1930)*. Amsterdam: Sara/Van Gennep.

van Dalen-Oskam, K. (2016). 'Could be the translation, of course'. Analysing the perception of literary fiction and literary translations. *Digitalität in Den Geisteswissenschaften*. Loveno di Menaggio, Italy.

Vogel, M. (2001). *'Baard Boven Baard': Over Het Nederlandse Literaire En Maatschappelijke Leven 1945-1960 ('Beard over Beard': On Dutch Literary and Societal Life 1945-1960)*. Maastricht: Maastricht University.

Digital Modelling of Knowledge Innovations In Sacrobosco's Sphere: A Practical Application Of CIDOC-CRM And Linked Open Data With CorpusTracer

Florian Kräutli
fkraeutli@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Matteo Valleriani
valleriani@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Esther Chen
echen@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Christoph Sander
csander@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Dirk Wintergrün
dwinter@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Sabine Bertram
bertram@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Gesa Funke
gfunke@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Chantal Wahbi
cwahbi@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Manon Gumpert
mgumpert@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Victoria Beyer
vbeyer@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Nana Citron
ncitron@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Guillaume Ducoffe
guillaume.ducoffe@ici.ro
ICI Bucharest, Romania

Introduction

In the frame of the research project *The Sphere. Knowledge System Evolution and the Shared Scientific Identity of Europe* we investigate the knowledge tradition that is interwoven with the history of one text: the *Tractatus De Sphaera* by Johannes de Sacrobosco. This 13th century treatise on cosmology has been published as part of university textbooks up until the 17th century. We have identified a corpus of more than 300 printed books related to Sacrobosco's text and obtained digital copies – a process that took three years to complete. These textbooks, which were part of the mandatory curriculum in most European universities at that time, contain Sacrobosco's text in its original version, as well as in translated, annotated or commented form. In addition, publishers included other texts that were seen as relevant for the study of cosmology from fields such as medicine, astronomy or mathematics (Valleriani, 2017).

Based on this corpus we seek to study how knowledge innovations have proliferated through the dissemination of texts, and identify the structural and social factors that contribute to or hinder the spread of certain kinds of knowledge. We do so by making use of methods from the area of network analysis which we apply on a dataset that we derived from our literary corpus.

This paper presents the foundational work that enables this kind of research with immediate application for similar projects concerned with editorial histories and structural analyses of corpora. We demonstrate the practical application of linked semantic data and the CIDOC-CRM model for shaping and addressing research questions in the humanities (Crofts et al., 2011).

Challenges

Our main challenge for this part of the project is the digital representation of the structure of the books and relevant contextual data.

The data model needs to be detailed. Individual texts can be derived from and include other texts. This genealogy of a text needs to be represented. We also require a suitable way of inputting complex data in a user friendly way. We need to be able to query and extend the data in a flexible manner. The data needs to support not only our initial research questions, but also future ones and those by other researchers. We need to be able to maintain an audit trail and trace occurrences that appear as a result of a network analysis to the original source. Last but not least we want to be able to publish our data in an understandable and reusable format.

We meet those challenges by modelling our data in adherence to the formal ontology CIDOC-CRM (Crofts et al., 2011) and the FRBRoo extension for bibliographic records (Bekiari et al., 2015), by storing our data in RDF and according to the 5-star deployment scheme for Linked Open Data (Berners-Lee, 2006), and by making use of

the Metaphactory (Metaphacts, n.d.) and ResearchSpace platform for semantic data creation (Oldman, 2016).

The next challenge is the development of a mathematical model that allows us to analyse the evolution of knowledge innovations – initially based on the textual sources and social structures, and later including other kinds of evidence such as book illustrations, family and business relationships, etc.

Related work

Our project builds on previous work in the area of semantic data, specifically CIDOC-CRM, and network analysis for research in the humanities.

Historical research that makes use of network modelling and analysis is increasingly relevant (Renn et al., 2016). A recent example is the establishment of the Journal for Historical Network Analysis (Rollinger et al., 2017). The evolution of scientific ideas in particular lends itself to be studied through networks (Lalli and Wintergrün, 2016) as well as how academic funding structures are of influence (Bellotti, 2012).

CIDOC-CRM (Crofts et al., 2011) has been developed and successfully used as a way of reconciling and connecting sources coming from different cultural and technical contexts. Examples include CLAROS (Kurtz et al., 2009), which brings together classical art research databases, PHAROS (Reist et al., 2015), which provides consolidated access to photo archives, or the reconciliation of the Arachne database of the German Archaeological Institute (Krummer, 2006). A RDF implementation of CIDOC-CRM and FRBRoo has been developed at the University of Erlangen (Goerz et al., 2008). The team is also involved in Wiss-Ki (Goerz et al., 2009), along with ResearchSpace (Oldman, 2016) one of few tools that support data creation in CIDOC-CRM compatible RDF (CIDOC/RDF).

Our approach

CorpusTracer

To address the outlined challenges we developed CorpusTracer. CorpusTracer is our front-end for creating and querying the dataset (Figure 1). It is a custom configuration of the Metaphactory semantic data platform and relies on modules developed as part of the ResearchSpace initiative. ResearchSpace is a cultural heritage research platform that builds on Metaphactory as a middleware and introduces modules for CIDOC-CRM compatible data creation and access. It allows to write data directly in CIDOC/RDF to a Blazegraph triple store. Crucially, it is possible to harvest the expressivity of CIDOC-CRM while not having to expose users of the tool to its complexity. We will demonstrate the tool, which is available open-source for download and use.

Figure 1. The home screen of CorpusTracer, featuring a search field and recently edited book and person records, with images and biographical data of persons drawn from Wikidata

Data model

Our data model (Figure 2) relies on generic concepts defined in CIDOC-CRM and FRBRoo, making it understandable and reusable outside the scope of our project. We have earlier described the model in more detail (Kräutli and Valleriani, 2017). Since then, we have slightly expanded the model to account for more complex derivations of texts, and for illustrations. The FRBRoo approach, which separates the concept of a book into several layers of physical and conceptual abstractions, fits well to the research framework.[1] It allows us to accurately capture the composition of each book: the texts it contains and, for each text, whether it is an original text or how it derives from existing texts.

We employ a strict separation between the data that is based on our corpus and data that provides context, such as biographical details or location data. We achieve this by linking relevant entities to external sources from Wikidata and the CERL thesaurus. Researchers are able to search for and link to resources on Wikidata directly within the CorpusTracer user interface.

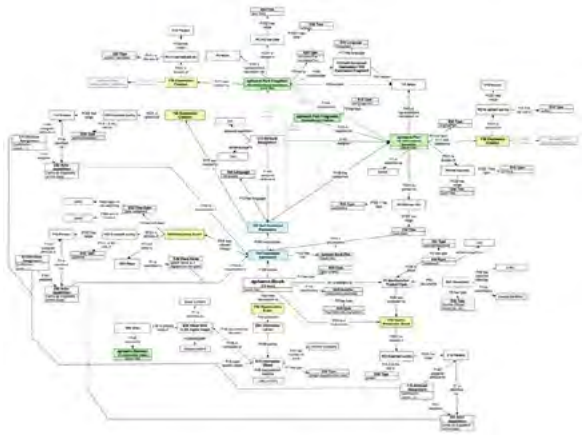


Figure 2. A graphical representation of our CIDOC-CRM/FRBRoo data model

1 FRBR introduces the concepts of Item, the material book, Manifestation, the prototypical book, Expression, the text of a book, and Work, the overall work conveyed by the book.

Discussion

The technical foundation provided through Metaphactory and ResearchSpace allowed us to develop the data model and implement a version of CorpusTracer ready for inputting data within a few months. Our team could then start with inputting the bibliographic data while scholars simultaneously performed the structural analysis of the publications. Changes on both the model and the interface were implemented as we gained a better understanding of the material at hand.

Although we use the platform primarily for data creation, we designed it in a way that will also allow the general public to access and navigate the dataset – which ultimately also benefits expert users. The structured search component of the Metaphacts platform is implemented to allow querying the graph database without having to know the underlying data model (Figure 3). Queries can be made for different entities (books, texts, persons, etc.) and the relationships between them.

Data can be downloaded in CSV format on different pages of the interface as well as by using the structured search. In order to extract the network data required for our analysis we however rely on custom SPARQL queries.

To construct the queries a good knowledge of the data model, the SPARQL syntax and the architecture of graph databases is required. While we found the data created through the platform to be reliable, one has to be careful not to introduce errors when querying the data manually. Unlike in relational databases, where one row in a table corresponds to one item of data, the boundaries of individual entities are not strictly defined in the Blazegraph triple store. We often found errors in our own custom queries that produced a higher number of results than we would have expected.

Despite the above reservations we find it preferable to not to rely on the graphical interface and CSV downloads to access the data, but to use custom SPARQL queries: for reasons of transparency, for maintaining an audit trail between original and extracted data, and for better reproducibility when the dataset changes.

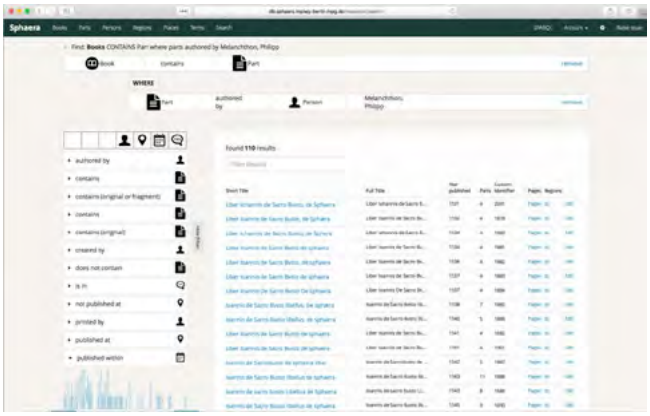


Figure 3. The Structured Search interface of the Metaphacts platform allows also non-expert users to formulate complex query on the graph database

Future

We have now completed the work on the dataset for the structural analysis of the corpus. The dataset can be accessed and downloaded, along with CorpusTracer, via our website (sphaera.mpiwg-berlin.mpg.de).

We continue to extend the dataset, particularly with regards to other forms of evidence to study exchange of knowledge. CorpusTracer implements an annotation tool which we use to mark illustrations in the digitised pages of the books (Figure 4). By employing an image hashing algorithm we identify shared illustrations across books that indicate relationships between printers.

Currently we are working on a mathematical model that enables us to identify the contributing structural and social factors that lead to the successful proliferation of particular knowledge innovation.

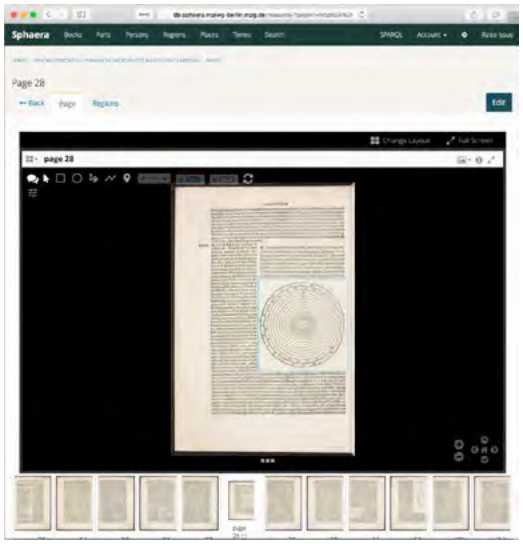


Figure 4. ResearchSpace provides a Mirador IIIF Viewer with annotation functionality, which we use to mark illustrations within pages of the books

References

BekiariC.,DoerrM.,La BoeufP. andRivaP.(eds.) (2015). Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism. Retrieved 27 April 2017 from https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf.

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved 23 November 2017, from https://www.w3.org/DesignIssues/LinkedData.html

Bellotti, E. (2012). Getting funded. Multi-level network of physicists in Italy. *Social Networks*, 34(2): 215–229. http://doi.org/10.1016/j.socnet.2011.12.002

Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M. (eds.). (2011). Definition of the CIDOC Conceptual Reference Model. *ICOM/CIDOC CRM Special Interest Group*.

Goerz, G., Scholz, M., Merz, D., Krause, S., Fichter, M. and Reinfandt, K. (2009). About WissKI. Retrieved November 22, 2017, from http://wiss-ki.eu/about

Goerz, G., Schiemann, B. and Oischinger, M. (2008). An implementation of the CIDOC conceptual reference model (4.2.4) in OWL-DL. *2008 Annual Conference of CIDOC, Athens, September 15-18*.
Kräutli, F. and Valleriani, M. (2017). CorpusTracer: A CIDOC database for tracing knowledge networks. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx047
Krummer, R. (2006). Integrating data from The Perseus Project and Arachne using the CIDOC CRM An Examination from a Software Developer's Perspective. Retrieved 22 November 2017, from http://www.perseus.tufts.edu/~rokummer/KummerCIDOC2006.pdf
Kurtz, D., Parker, G., Shotton, D., Klyne, G., Schroff, F., Zisserman, A. and Wilks, Y. (2009). CLAROS - Bringing Classical Art to a Global Public. *Fifth International Conference on e-Science*, pp. 20–27. IEEE. http://doi.org/10.1109/e-Science.2009.11
Lalli, R. and Wintergrün, D. (2016). Building a scientific field in the Post-WWII Era: A network analysis of the renaissance of general relativity. *Invited talk at the Forschungskolloquium zur Wissenschaftsgeschichte, Technische Universität, Berlin, 15 June 2016*.
Metaphacts (n.d.). Metaphactory. https://metaphacts.com/product
Oldman, D. (2016). ResearchSpace. https://public.researchspace.org
Renn, J., Wintergrün, D., Lalli, R., Laubichler, M. and Valleriani, M. (2016). Netzwerke als Wissensspeicher. In J. Mittelstraß and U. Rüdiger (eds,), *Die Zukunft der Wissensspeicher: Forschen, Sammeln und Vermitteln im 21. Jahrhundert*. Konstanz: UVK Verlagsgesellschaft mbH, pp. 35–79.
Reist, I., Farneth, D., Stein, R. S. and Weda, R. (2015). An Introduction to PHAROS: Aggregating Free Access to 31 Million Digitized Images and Counting... Retrieved 22 November 2017, from http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/BoardMeetings/CIDOC_PHAROS_Farneth-Stein-Weda_1.pdf
Rollinger, C., Düring, M., Gramsch-Stehfest, R. and Stark, M. (eds.). (2017). Journal of Historical Network Research 1. *Luxembourg Centre for Conte.mporary and Digital History*. https://doi.org/10.25517/jhnr.v1i1.
ValleianiM.(2017). The tracts of the sphere. Knowledge restructured over a network. InValleianiM. (ed.),*The Structures of Practical Knowledge*.Dordrecht:Springer, pp.421–73.

Quantitative microanalysis? Different methods of digital drama analysis in comparison

Benjamin Krautter
benjamin.Krautter@ilw.uni-stuttgart.de
QuaDramA / University of Stuttgart, Germany

Introduction

Recent results of computer-aided research suggest that characters in novels – measured by their character speech – can be laid out stylistically distinct from other characters of the same novel (Hoover, 2017; Fields, Bassist, Roper, 2017). Thus, experienced authors are able to create characters with 'distinctive voices' which can be identified by word frequencies. Unlike stylometrically determined signals in respect to author, genre or period, it is then an intratextual criterion for similarity and disparity. The study's subject is therefore not a large text corpus of different authors and periods, but a single literary text that comes into analytical focus. This approach to text selection is oftentimes called 'microanalysis' (Hoover, 2017). The term does not only differ from buzzwords such as 'big data', it also emphasizes the differences to concepts such as 'macroanalysis' (Jockers, 2013) and 'distant reading' (Moretti, 2000; 2005) despite their comparable quantitative techniques.

Surprisingly, studies on the stylistic differentiation of character speech are mostly limited to novels even though the structure of dramatic texts makes a quantitative examination of dramatic character speech easier: The speech is neither sorted nor commented nor framed by a narrator. By consequence and in contrast to narrative texts, the character speech can be isolated automatically. Initial approaches are already available: E.g., John Burrows and Hugh Craig show that individual drama characters can indeed be successfully assigned to an author's signal (Burrows, Craig, 2012). Both argue against critics who question a successful attribution of dramatic texts to an author, as Masten (1997) does who claims that the lack of narrators would lead to many indistinguishable voices.

Distinctive Character Speech in Dramatic Texts?

Figure 1[1] is based on David Hoover's approach in *The Microanalysis of Style Variation* (2017) but is applied to the genre of drama. The hierarchical cluster analysis in *Figure 1* illustrates the various characters of Gotthold Ephraim Lessing's *Minna von Barnhelm, oder das Soldatenglück (1767)* in regard to their similarity. As one of the plays of "Lessing's maturity" (Worvill, 2005: 177) *Minna von Barnhelm* seems to be an appropriate drama to discuss its characters and their speech. Michael Metzger, e.g., argues that Lessing created "a characteristic pattern of language for each of the various roles he has written" (Metzger, 1966: 196; see also Worvill, 2005; Asmuth, 2009).

1 Figure 1, 2 and 6 were generated using the 'stylo' package for R. Figure 3, 4 and 5 were created using the 'DramaAnalysis' package for R (Nils Reiter, Marcus Willand). https://github.com/quadrama/DramaAnalysis. The visualization of Figure 2 was done in Gephi.

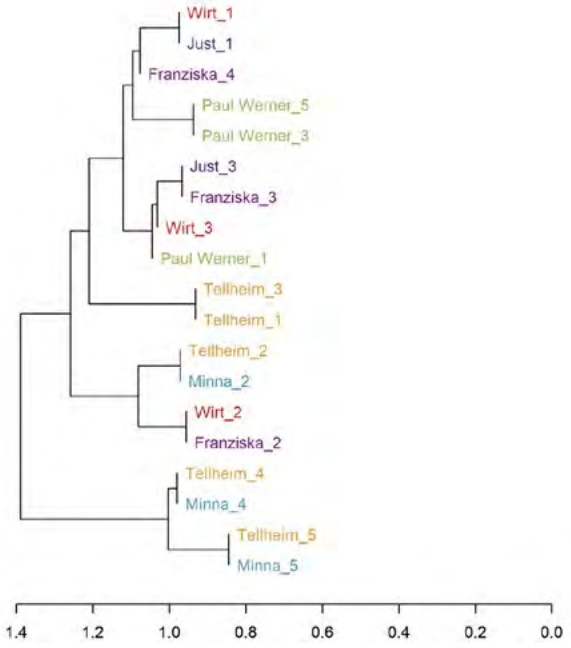


Figure 1: Dendrogram of *Minna von Barnhelm*, 1000 MFW, no culling, Cosine Delta, Ward Clustering.[2]

The stylometric analysis is based on word frequency lists which are extracted from the individual characters' utterances. With the help of 'Cosine Delta', that is claimed to achieve more reliable results than 'Burrows's' or Argamon's Delta' (Evert et al., 2017), the speeches' relative stylistic similarity is calculated by means of word frequencies. Contrary to Hoover's approach, the character speech is not divided into artificial segments of 1500 words each but by its 'naturally given' act boundaries.[3] This is helpful for the interpretation of the stylometric results based on the conditions of their emergence, such as the co-presence of characters. The procedure's disadvantages are the speech segments' inconsistencies: Some segments fall below a length of 700 words and must be excluded.[4] It also eliminates the so-called possibility of 'randomization', as it is practiced by Hoover: the individual character speeches' word distribution is randomly assigned to the segments in order to 'normalize' outliers. However, one should be cautious regarding the random distribution of words since potentially better results can only be measured by the underlying hypothesis.

Minna von Barnhelm's stylometric analysis indicates certain signs of stylistically distinctive character speech: E.g., Tellheim's speech – he is the male protagonist of the play – from Act 1 and 3 is grouped in immediate vicinity. The same holds true for the speech in Act 3 and 5 taken from Paul Werner. However, most of the speech segments seem to follow a different criterion. This is particularly evident for the uppermost section of the chart: The speeches by Major Tellheim, Minna von Barnhelm, Franziska (Minna's chambermaid) and the landlord (Wirt) are grouped on a contiguous branch, i.e. they resemble the other segments stylistically. Those four segments of speech belong to the drama's second act. There are other examples that seem to confirm act boundaries as an important factor for the analysis' results. The most striking ones are those of Tellheim and Minna in both Act 4 and 5. The analysis shows that the results by Hoover, Fields, Bassist and Roper cannot be transferred to Lessing's dramatic text directly.

A single dendrogram, however, must not be more than a first indication for the assessment of the hypothesis. To avoid a potential 'cherry picking' problem at this point, further stylometric analyses on an expanded corpus were conducted.[5] Both, the author's signal (175 of 175 segments matching) and the text unity (171 of 175 segments matching) can be clearly identified. Thus, the cluster analysis does not seem to be influenced negatively by the relatively small sizes of the speech segments. *Figure 2*, a network plot that uses the same corpus, consolidates this finding.[6]

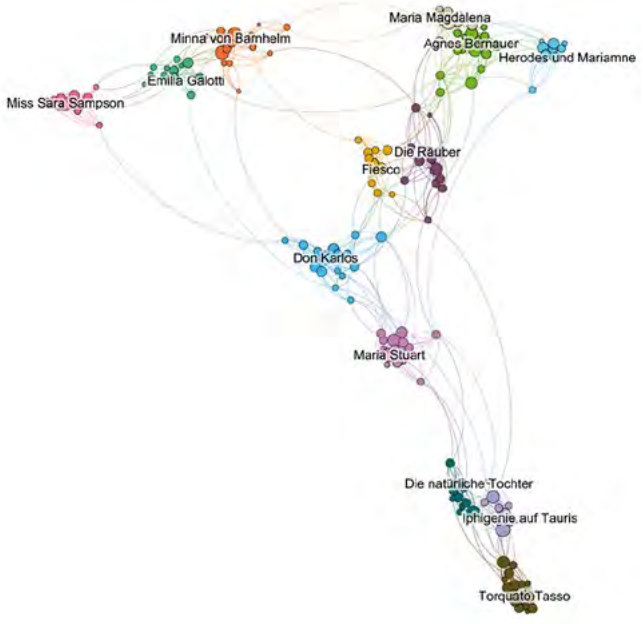


Figure 2: Stylometric network of 13 dramas. 500–1500 MFW, no culling, Cosine Delta, three nearest neighbors. Node sizes represent average degree, node colors represent modularity rank.

2 Although some speech segments fall below a length of 1000 words, it should still be feasible to use a vector length of 1000 MFW (Eder, 2017b). The results of Figure 2 support this hypothesis, but a larger scale study on this topic is a future task.
3 The act boundaries are marked with underscores in the illustration.
4 To compare: Fields, Bassist and Roper use segments of only 200 words each.
5 I analyzed 13 texts – three by Lessing, four by Friedrich Schiller, three by Johann Wolfgang Goethe and three by Friedrich Hebbel – with a total of 175 speech segments. Parameters used: 1000 MFW, no culling, Cosine Delta, Ward Clustering. The visualization is not shown in the paper.
6 See Eder (2017a) for advantages of stylometrics visualized by network plots.

Co-presence and Character Semantics

Stylometrics are not the only method to determine relative similarities within a text corpus. The extent to which they are suitable to discuss open questions – in contrast to, e.g., author attribution – remains to be examined. If parameters such as distance measures, word size or culling must be redefined with respect to the text corpus, 'cherry picking' would then become inherent to the method (Schöch, 2014; Jannidis 2014; Eder, 2013). It is therefore necessary to compare the established observations to other quantitative methods. This is done by means of analyzing co-presence and semantics of character speech.

Figure 3 illustrates the speech parts of the six most important characters in Lessing's *Minna von Barnhelm*. The following investigation focuses on the protagonists Tellheim and Minna. In the second, but especially in the fourth and fifth act, Tellheim and Minna are mainly co-present. This structural data correlates with the observations in *Figure 1*. The speech segments of those acts are grouped closely together, while Tellheim's speech in Act 1 and 3 is clearly separated. In these two acts Tellheim and Minna are not co-present.

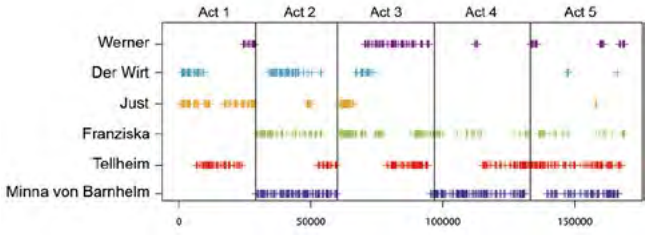


Figure 3: Co-presence in *Minna von Barnhelm*.

The observation that stylistic similarities of the character speech is related to structural characteristics challenges earlier research and demands further investigation: Is it possible to expand or specify this finding? A semantic word field analysis, as used by Willand and Reiter (2017), serves to operationalize the thematic conception of character speeches.[7] *Figure 4* illustrates two diagrams that compare different segments of Tellheim's and Minna's speech. The figure on the left compares Tellheim's speech in Act 1 and 5. It indicates significant semantic differences in those segments that also showed little similarity in terms of style. The themes 'love' and 'ratio' are given greater weight in Act 5, while the context of 'family' is invoked less frequently. All in all, one can clearly detect a discrepancy in the semantic fields' word frequencies.

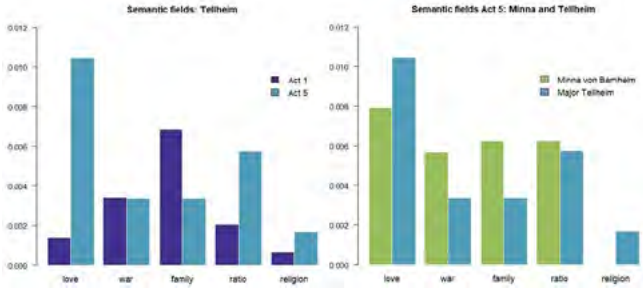


Figure 4: Semantic fields in *Minna von Barnhelm*.

The diagram on the right shows the semantic fields of Minna and Tellheim in the fifth act. Compared to the diagram on the left, the two speeches of Tellheim and Minna seem to correlate better with each other, especially considering the word fields 'love' and 'ratio'. None of the word fields is conspicuous due to extreme differences. Whether this observation can actually be used as a marker for similar topics or not has to be proofed within a larger text corpus. By consequence, this would be useful to determine a threshold value to mark similarity and disparity. I started this task using the Euclidean distance to measure the similarity between different segments of character speech in *Minna von Barnhelm*. It results in the following values of similarity:

	Tellheim 1	Tellheim 2	Tellheim 3	Tellheim 4	Tellheim 5
Tellheim 2	0,01723				
Tellheim 3	0,00781	0,01344			
Tellheim 4	0,01370	0,01279	0,01334		
Tellheim 5	0,00773	0,01167	0,00438	0,01396	
Minna 2		0,01070*			
Minna 4				0,0114*	
Minna 5					0,0049*

Figure 5: *denotes co-presence, yellow colored values are nearest neighbor segments as taken from the stylometric analysis (*Figure 1*), lower values display a higher similarity.

Average of the four nearest neighbor segments: 0,008703
Average of Tellheim's segments (without nearest neighbors): 0,012026667

The difference of the two groups' average margin is a value of 0,00332, or 38,2 percent. Although the sample size is still small, one dramatic text only, this seems to be quite a significant result. Thus, the word field semantics do at least provide an indication that style, theme and presence of characters are related to some extent.

Conclusion

A closer examination of the character speech in *Minna von Barnhelm* has shown that it is plausible to combine

7 For this purpose, five dictionaries on the topics of family, war, love, ratio and religion were created, enlisting 65 to 110 words each. The words were used in dramas between 1770 and 1830 (Willand, Reiter 2017).

different analytical methods. Thus, the investigation benefits from their respective strengths. Herein, results can be validated and opened for broader questions. In the chosen dramas, co-presence seems to have an impact not only on style but also on the semantics of character speech. The segments spoken by the two protagonists in Act 5 of *Minna von Barnhelm* exemplify this thesis. These results differ from Hoover's and suggest having a closer look on co-presence and its influence on the distinctiveness of character speeches in dramas as well as in novels. The absence of a narrator in dramatic texts is one possible starting point to explain the differences outlined in this paper.

References

Asmuth, B. (2009). *Einführung in die Dramenanalyse*. 7th ed. Stuttgart, Weimar: J. B. Metzler.

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Web and Social Media*, North America.

Burrows, J. and Craig, H. (2012). Authors and characters. *English Studies*, 93(3): 292–309.

Eder, M. (2013). Computational Stylistics and Biblical Translation: How Reliable Can a Dendrogram Be?. In Piotrowski T. and Grabowski Ł. (eds.), *The Translator and the Computer*. Breslau: WSF Press, pp. 155–170.

Eder, M. (2017a). Visualization in Stylometry: Cluster Analysis Using Networks. *Digital Scholarship in the Humanities*, 32 (1): 50–64.

Eder, M. (2017b). Short Samples in Authorship Attribution: a New Approach. *Digital Humanities 2017. Conference Abstracts*. Montréal: McGill University and Université de Montréal, pp. 221–224.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a Suite of Tools. In: *Digital Humanities 2013. Conference Abstracts*. Lincoln: University of Nebraska, pp. 487–489.

Evert, S., Proisl, Th., Jannidis, F., Reger, I., Pielström, S., Schöch, Ch. and Vitt, Th. (2017). Understanding and Explaining Delta Measures for Authorship Attribution. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx023.

Fields, P. J., Bassist, L. and Roper, M. (2017). Characters in 19th Century Novels Display Distinctive Voices as Seen by Stylometric Analysis. In *Digital Humanities 2017. Conference Abstracts*. Montréal: McGill University and Université de Montréal. https://dh2017.adho.org/abstracts/494/494.pdf.

Hoover, D. (2017). The Microanalysis of Style Variation. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx022.

Jannidis, F. (2014). Der Autor ganz nah. Autorstil in Stilistik und Stilometrie. In Schaffrick M. and Willand M. (eds.), *Theorien und Praktiken der Autorschaft*. Berlin, Boston: De Gruyter, pp. 169–195.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago, Springfield: University of Illinois Press.

Masten, J. (1997). *Textual Intercourse: Collaboration, Authorship and Sexualities in Renaissance Drama*. Cambridge: UP.

Metzger, Michael M. (1966). *Lessing and the Language of Comedy*. The Hague, Paris: Mouton.

Moretti, F. (2000). Conjectures of World Literature. *New Left Review*, 1: 54–68.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Schöch, Ch. (2014). Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. *Literaturwissenschaft im digitalen Medienwandel. Beihefte zu Philologie im Netz*, 7: 130–157.

Willand, M. and Reiter, N. (2017). Geschlecht und Gattung: Digitale Analysen von Kleists 'Familie Schroffenstein'. *Kleist Jahrbuch*, 2017: 177–195.

Worvill, R. M. (2005). *'Seeing' Speech: Illusion and the Transformation of Dramatic Writing in Diderot and Lessing*. Oxford: Voltaire Foundation.

Computational Analysis and Visual Stylometry of Comics using Convolutional Neural Networks

Jochen Laubrock
laubrock@uni-potsdam.de
University of Potsdam, Germany

David Dubray
ddubray@uni-potsdam.de
University of Potsdam, Germany

Introduction

Stylometry is a very successful application area of the digital humanities (e.g., Juola, 2006). However, to date it is mainly confined to the study of linguistic style, perhaps reflecting a general focus of the digital humanities on text. Stylometric tools for *visual* material are not yet as well-established, despite recent advances in digital art history (Saleh and Elgammal, 2015; Manovich, 2015). Part of this deficit may originate in the challenging aspects of traditional computational image analysis, which requires deep expert knowledge for hand-crafting engineered features applicable to a problem domain. Recent advances in artificial intelligence together with the availability of large corpora of annotated images have partly ameliorated the situation. Now we can delegate to the machine the task of discovering the features relevant for classification. In particular, deep convolutional neural networks (CNNs; LeCun et al., 2015) have been very successful in many image classification tasks, using a feature hierarchy akin to levels of processing in the human visual system. Here we

propose a method for a visual stylometry of comics based on CNN features. We test transfer to comics by using a large corpus of graphic narratives. We further show how CNN features can be used to predict readers' attention. In closing we explore how the approach might be used for tasks such as locating text or finding characters, as well as in other domains such as art history.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of neural networks specialized in analyzing data with an implicit spatial layout, such as the stack of three 2D matrices commonly used to represent RGB images in the computer. CNNs are characterized by local connections, shared weights, pooling, and the use of many layers. Within each convolutional layer, a stack of different filters (feature maps) is trained. Each unit is connected to local patches in the feature maps of the previous layer through a set of learned weights. These are called a filter kernel, and learned by backpropagation. A local weighted sum computed by applying the filter kernel to the image is passed through a non-linearity[1], often a rectified linear unit (ReLU). All units in a feature map share the same filter kernel; feature maps in a layer differ by using different kernels. The receptive size of each filter (i.e. the region of the image it responds to) is small at the lower layers, and becomes progressively larger at higher layers. Conversely, the higher the layer, the more complex the features encoded by the filters. Pooling layers typically replacing a local patch by its maximum value are added to further reduce the number of parameters and to provide a more coarse-grained and robust description.

Lower-level filters often respond well to edges and boundaries and thus resemble simple cells in human visual cortex. Higher-level features, in contrast, can code for complex stimuli like textures or facial parts. Just like the visual system, CNNs compose objects out of simple features by using compositional feature hierarchies. Edges combine into motifs, motifs into parts, and parts into objects. CNNs pre-trained on large-scale image classification tasks like ImageNet (14 million images with over 1,000 classes) can be adapted to specific material by re-training just a few layers, assuming that basic features at the lower level are more or less generic. Therefore, we expected transfer to comics drawings, even for networks that had been pre-trained on photographic images. Note that all of the networks we use have been trained on photographs, i.e., they have never seen graphic novels. However, since they have learned filters and filter combinations that are useful for the interpretation of our environment, we hypothesized that they might also be useful for the analysis of drawings. Drawings are abstractions, but as such they do have a relationship to our environmental reality.

Method

The material we used is the Graphic Narrative Corpus (GNC; Dunst et al., 2017). The GNC is a representative collection of graphic novels, i.e., book-length comics that tell continuous stories and are aimed at an adult readership. The stratified monitor corpus currently includes 209 graphic narratives amounting to nearly 50,000 digitized pages. A subset of the first chapter of these works is annotated by human annotators with respect to the location and identity of panels, main characters, character relations, captions, speech bubbles, onomatopoeia, and the respective text. Furthermore, eye movement data is collected for these pages to measure readers' attention. At the time of writing we had available the first 10 pages of 95 works by 87 authors.

In order to test generalization of the features and their transfer to graphic illustrations, we describe material from the GNC using a specific CNN, Inception V3 (Szegedy et al., 2015) using pre-trained weights from ImageNet. We chose Inception V3 for stylometry and artist attribution due to its state-of-the art performance, economic parameterization, and relative independence of input sizes. Because of the small amount of training data, we trained a support vector machine (SVN) rather than a neural network to classify drawing style, based on a description of 9 pages of each of the works in terms of the visual features coded in each of the main (mixed) layers of Inception V3. One randomly determined page per comic was held out to evaluate performance.

Second, we were interested in which features of the material were guiding visual attention of the reader. For the prediction of the distribution of attention we used DeepGaze II (Kümmerer et al., 2016), currently the leading entry in the MIT saliency benchmark. DeepGaze II uses features from several layers of VGG-19 (Simonyan and Zisserman, 2014) to predict "empirical saliency", i.e., where people look or where they move the mouse to unblur an image[2]. We were again interested in determining how good the transfer from photographs to graphic illustrations is. We compared DeepGaze II predictions to empirical gaze locations of 100 readers reading a subset of 105 pages from 6 graphic novels using the metric of information gain explained (Kümmerer et al., 2015).

Analysis and Results

Overall, the top-1 classification accuracy in the artist attribution analysis, based on the highest vote, was 93%. That is, the artist of 93% of the hold-out pages was correctly

1 This nonlinearity is needed because neural networks with just linear activation functions are effectively only one layer deep, and therefore cannot be used to model the full range of real world problems, many of which are nonlinear.

2 We are convinced that in principle, the DeepGaze architecture could also use features of a different CNN such as Inception as predictors. However, it is currently only available based on VGG features.

classified based on Inception features. It is instructive to further inspect the few misclassifications. For example, a page of "Batman: The Long Halloween" drawn by Tim Sale was mistakenly attributed to David Mazzuchelli, the artist responsible for "Batman: Year One", which was also part of the training corpus. Tim Sale got only the second highest vote for this particular page. Our analysis illustrates that stylistic classification is generally possible using out-of-the box features of a pre-trained neural network and given only very limited amount of training material. We expect the results to yet improve given more training material, and possibly using a neural network classifier rather than an SVM. An in-depth analysis and the corresponding visualizations of which features are the most discriminative and most strongly associated with a given artist are underway and will be presented at DH.

The empirical fixation distribution is very convincingly reproduced using DeepGaze II, even without additional training (Figure 1). Overall, for all 105 pages the match of empirical fixation distribution by the model predictions was quite high (Figure 2). CNN features can thus be used to predict which image regions will attract attention. Most likely this is due to their encoding of image properties that combine into objects such as text boxes, faces, and characters, on which most of the empirical fixations are concentrated.

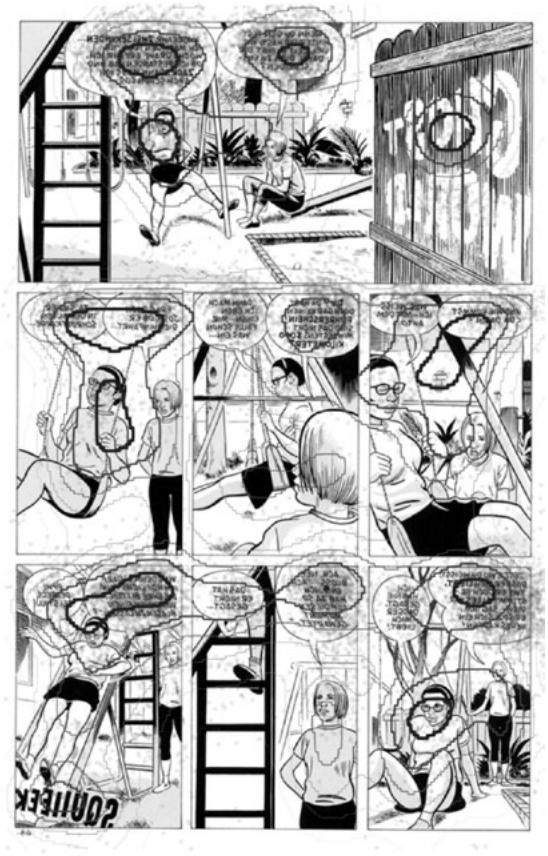


Figure 1. Empirical fixation distribution of 100 readers (dots) and DeepGaze II predictions (contour lines) on a page from the German translation of Daniel Clowes' (1997/2000) graphic novel Ghost World

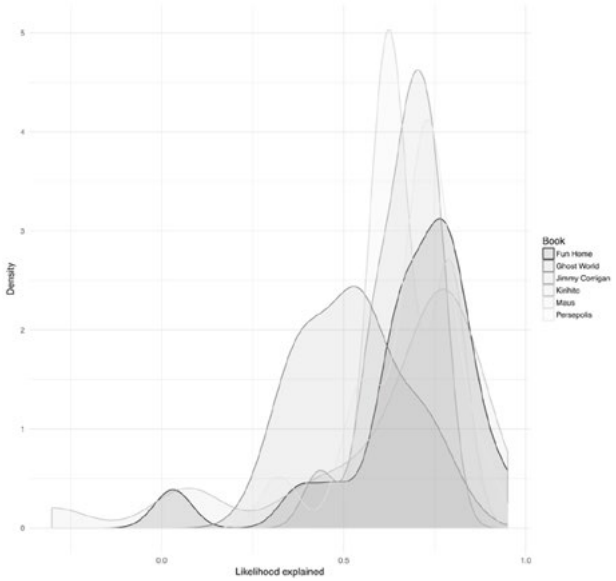


Figure 2. Distribution of the measure *information gain explained*, comparing theoretical Deep Gaze II predictions with empirical fixation locations obtained by measuring eye movements of 100 readers on 105 pages from six graphic novels

Outlook

We obtained very promising results of a stylometric analyses of comics artist based on CNN features trained on photographs. Given this successful transfer, we suspect that such features are general enough to be applied in a wide variety of other domains in which a visual stylometry may be useful. For example, arts historians may be interested in combining the method with nearest neighbor search to describe how close different artist are in feature space (cf. Saleh and Elgammal for a similar approach based on classic features). Historians may find visual feature based analyses useful for annotation of documents containing images along with text.

If stylometry works so well, CNN features can probably be used for detecting visual elements such as speech bubbles or characters within panels. We currently experiment with YOLO 9000 (Redmond and Farhadi, 2017), which does a very good job at locating objects and persons in panels, and is likely to also function well as a speech bubble locator with additional training. If such object classes can be located automatically, implementation into an annotation tool might make the tedious task of annotation significantly easier, so that annotators have more time to concentrate on work at the narratological level.

References

Clowes, D. (1997/2000). *Ghost World. Translated by Heinrich Anders*. Berlin: Reprodukt.

Dunst, A., Hartel, R. and Laubrock, J. (2017). The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities. *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 03, 15–20.*

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv*:1311.2524.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Kümmerer, M., Wallis, T. S. A. and Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences of the United States of America*, 112(52), 16054–59.

Kümmerer, M., Wallis, T. S. A. and Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *CoRR*, abs/1610.01563.

Juola, P. (2006). Authorship Attribution. *Foundations and Trends in Information Retrieval, 1*, 233-334.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature, 521, 436–444.

Manovich, L. (2015). Data Science and Digital Art History. *International Journal for Digital Art History, 1*, 13-35. http://dx.doi.org/10.11588/dah.2015.1.21631.

Redmon, J. and Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. CVPR 2017.

Saleh, B. and Elgammal, A. M. (2015). Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *CoRR*, abs/1505.00855.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Classical Chinese Sentence Segmentation for Tomb Biographies of Tang Dynasty

Chao-Lin Liu
chaolinliu@gmail.com
National Chengchi University, Taiwan

Yi Chang
black.heptagram@gmail.com
National Chengchi University, Taiwan

Introduction

Figure 1 shows a slab of tomb biography of the Tang dynasty.[1] Researchers can copy the words on such slabs to produce a collection of tomb biographies for research. A typical tomb biography contains various types of information about the deceased and their families and, sometimes, a rhyming passage of admiration. Employing software tools to analyze the texts, one can extract useful information from the collections of tomb biographies to enrich databases like the China Biographical Database (CBDB) to support further Chinese studies.[2]

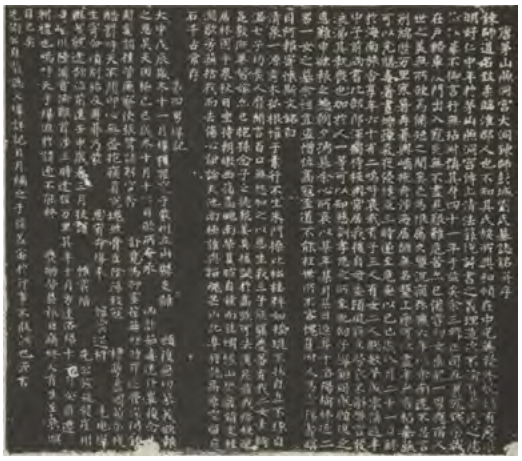


Figure 1. A slab of tomb biography of the Tang dynasty

It is well known that modern Chinese texts do not include delimiters like spaces to separate words. Hence, researchers design algorithms for segmenting Chinese character strings into words (Sun et al., 2004; Shao et al., 2017).

In contrast, it is not as well known that, in classical Chinese, there were no markers for the separation of sentences. The characters in Figure 1 simply connect to each other. In modern Chinese, texts are punctuated for pauses in sentences and ends of sentences. The research about algorithmically inserting these syntactic markers into classical Chinese is receiving more attention along with the growth of digital humanities in recent years. The needs of segmenting ancient texts for humanities studies are not unique to Chinese studies, interested readers can find examples for German texts (Petran, 2012) and Swedish texts (Bouma and Adesam, 2013).

Huang et al. (2010) employed the techniques of conditional random fields (CRFs) to segment texts of literature and history. They achieved 0.7899 and 0.9179 in F_1,[3] respectively, for segmenting the texts in Shiji and Zuozhuan.[4] Wang et al. (2016; 2017) applied recurrent neural networks to segment texts in a diverse collection

1 This image was downloaded from <http://www.lyqtzz.com/uploadfile/20110817165325665.jpg>. The Tang dynasty existed between 688CE and 907CE. More images of tomb biographies are available at <http://goo.gl/XHCL9P>.

2 The China Biographical Database (https://projects.iq.harvard.edu/cbdb/home) is a free and open database for Chinese studies.

3 The **precision** rate, **recall** rate, and **F measure** are designed for evaluating the effectiveness of information retrieval and extraction. F_1 is a popular choice of the F measure.

4 Shiji (史記) and Zuozhuan (左傳) are two very important sources about Chinese history.

of classical Chinese sources. They achieved F_1 measures that are close to 0.75, and item accuracies that are near 0.91.[5] The researchers achieved different segmentation results for different corpora even when they applied the same techniques and procedures. It is thus inappropriate to just compare the numbers for ranking because the nature of the corpora varies widely.

In this proposal, we report our attempts to segment texts in tomb biographies with CRF models (Lafferty, 2001). We studied the effects of considering different types of lexical information in the models, and achieved 0.853 in precision, 0.807 in recall, 0.829 in F_1,[3] and 0.940 in item accuracy.[6] Better results were accomplished when we employed deep learning techniques, including applicaitons of long short-term memory networks and sequence-to-sequence networks, for segmenting our tomb biographies.

Data Sources

We obtained digitized texts for three books of tomb biographies of the Tang dynasty (Zhou and Zhao, 1992; 2001). The collection consists of 5119 biographies which contain 423,922 periods, commas, and semicolons. There are 5505 distinct types of characters and a total of more than 2461 thousand of characters in the collection.[7] When counting these statistics, we ignored a very small portion of characters that cannot be shown without special fonts. Hence, these statistics are not perfectly precise, but they are accurate within a reasonable range. On average, a biography has about 480 characters. Some of them are very short and have many missing characters. Hence, we exclude biographies that have no more than 30 characters in our experiments.[8]

Training and Testing CRF Models

We consider the segmentation task as a classification problem. Let C_i denote an individual character in the texts. We categorize each character as either **M** (for "followed by a punctuation mark") and **O** (for "an ordinary character"). Assume that we should add only a punctuation mark between C_3 and C_4 for a string "C_1 C_2 C_3 C_4 C_5". A correct labeling for this string will be "O O M O O".[9]

We can convert each character in the texts into an **instance**, which may be used for training or testing. We provide with each instance a group of contextual **features** that may be relevant to the judgment of whether or not a punctuation mark is needed. For instance, we may use one character surrounding a character X and itself as the group of features. The following are two instances that we create for C_3 and C_4. The instance for C_3 is (1), and the leftmost item is the correct label for C_3, and the rest are the features for C_3.[10]

M w[0]=C_3,w[-1]=C_2,w[1]=C_4 (1)

O w[0]=C_4,w[-1]=C_3,w[1]=C_5 (2)

We can train a CRF model with a selected portion of the instances (called **training data**), and test the resulting model with the remaining instances (called **test data**). The instances in the training and the test data are mutually exclusive.

We employ a machine-learning tool that learns from the training data to build a CRF model.[11] We then apply the learned model to predict the classes of the instances in the test data. The labels of the instances in the test data are temporarily concealed when the learned model makes predictions.[12] The precision rate and recall rate of the learned model are calculated with the correct and the predicted labels.

We report four sets of basic experiments next, each investigating an important aspect for analyzing Chinese texts. The 5119 biographies were resampled and randomly assigned to the training (70%) and test (30%) sets for every experiment.[13] We repeated every experiment three times, and report the averages of the precision and recall rates.

Changing the Size of the Context

We certainly can and should consider more than one cha-

5 The **item accuracy** evaluates the labeling judgments including both punctuated and non-punctuated items. In a typical sentence segmentation task, there are many more non-punctuated items than punctuated items, so it is relatively easier to achieve attractive figures for the item accuracy than for the F measure

6 We interviewed Hongsu Wang (王宏甦), the project manager of the China Biographical Database Project at Harvard University, about his preferences in post-checking the segmentation results that are produced by software. He suggests that higher precision rates are preferred. When seeking higher recall rates (often sacrificing the precision rates), the false-positive recommendations for punctuation are annoying to the researchers.

7 In terms of Linguistics, we have 5505 character types and 2,461,000 character tokens.

8 30 is an arbitrary choice, and can be changed easily.

9 Due to the constraint on the word count in DH 2018 proposals, we can only briefly outline the steps for training and testing CRF models. More details can be provided in the presentation and in an extended report.

10 Here, we adopt typical notations for CRF-based applications. w[0] is the current word, w[-1] is the neighbor word to the left of the current word, w[1] is the neighbor word to the right of the current word. Two actual instances that are produced from "孝敬夭啟，動必以禮" for character-based segmentations will look like the following.

O w[-1]=敬,w[0]=夭,w[1]=啟

M w[-1]=夭,w[0]=啟,w[1]=動

Two instances that are produced from "母子 忠孝 ， 天下 榮之" for the word-based segmentations will look like the following.

M w[-1]=母子,w[0]=忠孝,w[1]=天下

O w[-1]=忠孝,w[0]=天下,w[1]=

11 CRFSuite: <http://www.chokkan.org/software/crfsuite/>

12 Thus, the instances for testing CRFs look like (1) and (2) that do not carry the correct labels, M and O, respectively..

13 Recall that we used only those biographies that have no less than 30 characters.

racter around the current character as the context. Figure 2 shows the test results of using different sizes of contexts for the instances. The horizontal axis shows the sizes, e.g., when *k*=2, the feature set includes information about two characters on both sides of the current character. P1 and R1 are the average precision and recall rates, respectively.

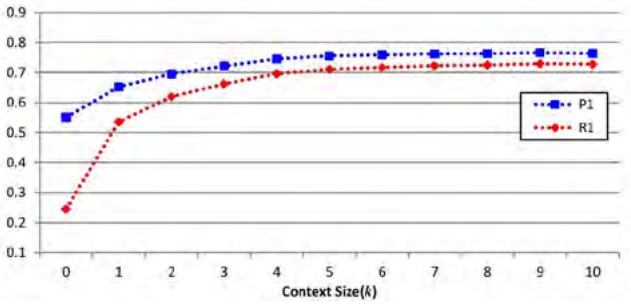


Figure 2. Effects of varying context sizes

We expected to improve the precision and recall rates by expanding the width of the context. The margin of improvements gradually decreased, and the curves level off after the window sizes reached six. The recall rises sharply when we add the immediate neighbor word into the features, emphasizing the predicting power of the immediate neighbor character. When *k*=10, the precision and recall are 0.765 and 0.729, respectively, and the item accuracy exceeds 0.91.

Adding Bigrams

We added bigrams that were formed by consecutive characters into the features. The following instance shows the result of adding bigrams to the features in (1).[14]

$$M \quad w[0]=C_3, w[-1]=C_2, w[1]=C_4, \mathbf{w[-1\_0]=C_2C_3, w[0\_1]=C_3C_4} \qquad (3)$$

Figure 3 shows the test results of adding bigrams while we also tried different sizes of context. The curves named P1 and R1 are from Figure 2, and P2 and R2 are results achieved by adding bigrams to the features. Both rates are improved, and the gains are remarkable.

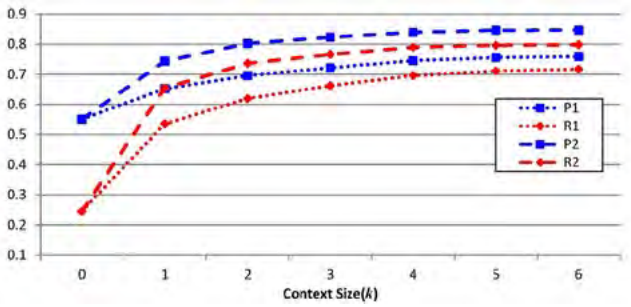


Figure 3. Adding bigrams improves the results.

14 Here, w[-1_0] is the bigram on the left side of the current word, and w[0_1] is the bigram to the right of the current word. When we consider bigrams for a wider context, we may consider bigrams like w[-2_-1] and w[1_2].

Effects of Pronunciation Information

Using the characters and their bigrams in the features is an obvious requirement. Since the tomb biographies may contain rhyming parts, it is also intriguing to investigate whether adding pronunciation information may improve the overall quality of the segmentation task.

We considered two major sources of the pronunciation information for Chinese characters in the Tang dynasty: *Guangyun* and *Pingshuiyun*.[15] The statistics in Table 1 show that adding pronunciation information into the features did not improve the overall performance for the segmentation task significantly.[16] The results suggest that, given the characters and their bigrams, adding pronunciation did not contribute much more information. Huang et al. (2010) reported similar observations when they used *Guangyun* in their work. Relatively, *Guangyun* is more informative than *Pingshuiyun* for the segmentation tasks.

Adding Word-Level Information

We can obtain information about the reign periods, location names, and office names in the Tang dynasty from CBDB. By segmenting characters for these special words and adding appropriate type information, we added word-level information into the features. The statistics in Table 2 show that the word-level information did not raise the performance very much.[17]

We examined the training and test data, and found that, although we gathered the special terms for the Tang dynasty, those words were not used in the biographies often. As a consequence, we did not add a lot of word-level information in the features in reality.

We have also adopted pointwise mutual information (PMI) of bigrams as features, but the net contributions are not significant.

Discussions

We have consulted historians,[6,18] and learned that our current results are useful in practice. The best precision rates and F measures are better than 0.8 in Figure 3 and Table 2. The best item accuracy is better than 0.94.

15 *Guangyun* and *Pingshuiyun* are 廣韻》and 平水韻》, respectively

16 This does not suggest that using the pronunciation information alone was not useful. We have conducted more experiments to evaluate the effectiveness of using the pronunciation information for the segmentation tasks, and will provide more details in the presentation and in an extended report.

17 In Table 2, WOC stands for "Width of Context", "P" stands for precision, "R" stands for recall, "C+B" stand for "Characters and Bigrams" and "C+B+W" stands for "Characters, Bigrams, and Words".

18 In addition to Hongsu Wang of Harvard University, we also consulted Professor Zhaoquan He (何兆泉) of the China Jiliang University. They use tomb biographies of the Tang and the Song dynasties in their research.

Features	Width of Context = 1			Width of Context = 2		
	Precision	Recall	F_1	Precision	Recall	F_1
Characters	0.652	0.535	0.588	0.695	0.620	0.655
Characters+Bigrams	0.743	0.654	0.696	0.802	0.736	0.768
Characters+Bigrams+Guangyun	0.748	0.671	0.707	0.781	0.707	0.742
Characters+Bigrams+Pingshuiyun	0.737	0.659	0.696	0.763	0.698	0.729

Table 1. Contributions of pronunciation information

Features	WOC = 1		WOC = 2		WOC =3		WOC = 4		
	P	R	P	R	P	R	P	R	F_1
C+B	0.743	0.654	0.802	0.736	0.823	0.766	0.839	0.790	0.814
C+B+W	0.747	0.671	0.800	0.741	0.818	0.767	0.832	0.787	0.809
C+B+PMI	0.748	0.661	0.804	0.740	0.824	0.769	0.839	0.791	0.814

Table 2. Adding word-level information

In fact, we have designed an advanced mechanism to further improve our results.[19] The new approach employs a second level learning step that learns from the errors of the current classifiers.

One may plan to consider more linguistic information in the segmentation tasks. If appropriate corpora or sources are available, it is worthwhile to explore the effects of adding part-of-speech information in the task (Chiu, 2015; Lee, 2012). We have applied deep learning techniques for the segmentation tasks, and achieved better results.

Although we look for methods to reproduce the segmentations in the given texts, we understand that not all experts will agree upon "the" segmentations for a corpus. Different segmentations may correspond to different interpretations of the texts, especially for the classical Chinese. The results of asking two persons to segment Chinese texts may not match perfectly either (Huang and Chen, 2011).

References

Bouma, G. and Adesam, Y. (2013). Experiments on sentence segmentation in Old Swedish editions. *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA* 2013. NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87:11–26.

Chiu, T.-s., Lu, Q., Xu, J., Xiong, D. and Lo, F. (2015). PoS tagging for classical Chinese text. *Chinese Lexical Semantics* (*Lecture Notes in Artificial Intelligence* 9332), pp. 448–456.

Huang, H.-H. and Chen, H.-H. (2011). Pause and stop labeling for Chinese sentence boundary detection. *Proceedings of the 2011 Conference on Recent Advances in Natural Language Processing*, pp. 146 153.

Huang, H.-H., Sun, C.-T., and Chen, H.-H. (2010). Classical Chinese sentence segmentation. *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 15 22.

Lafferty, J., McCallum, A., and Pereira, F. C.N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289.

Lee, J. (2012). A classical Chinese corpus with nested part-of-speech tags. *Proceedings of the Sixth EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 75–84.

Petran, F. (2012). Studies for segmentation of historical texts: Sentences or chunks? *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities*, pp. 75–86.

Shao, Y., Hardmeier, C., Tiedemann, J., and Nivre, J. (2017). Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. *Proceedings of the 2017 International Joint Conference on Natural Language Processing*, pp. 173 183.

Sun, M.-S., Xiao, M. and Tsou, B. K. (2004). Chinese word segmentation without using dictionary based on unsupervised learning strategy. *Chinese Journal of Computers*, 27(6):736 742. (in Chinese)

Wang, B., Shi, X. and Su, J. (2017). A sentence segmentation method for ancient Chinese texts based on recurrent neural network. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 53(2):255 261. (in Chinese)

Wang, B., Shi, X., Tan, Z., Chen, Y. and Wang, W. (2016). A sentence segmentation method for ancient Chinese texts based on NNLM. *Proceedings of the Chinese Lexical Semantics Workshop* 2016, *Lecture Notes in Computer Science* 10085, pp. 387–396.

19 Again, we could not provide details about more experiments because of the word limit for DH 2018 submissions.

Zhou, S. (周紹良) and Zhao, C. (趙超). (1992). *A Collection of Tomb Biographies of Tang Dynasty* (唐代墓誌彙編). Shanghai Ancient Books Publishing House (上海古籍出版社). (in Chinese)
Zhou, S. (周紹良) and Zhao, C. (趙超). (2001). *A Collection of Tomb Biographies of Tang Dynasty: An Extension* (唐代墓誌彙編續集). Shanghai Ancient Books Publishing House. (in Chinese).

Epistemic Infrastructures: Digital Humanities in/as Instrumentalist Context

James W. Malazita
malazj@rpi.edu
Rensselaer Polytechnic Institute, United States of America

In his essay "How Not to Teach Digital Humanities," Ryan Cordell outlines some of the pedagogical and institutional challenges of integrating DH into larger humanities curricula. Importantly, Cordell argues that successful Digital Humanities pedagogy must always take into account local institutional and infrastructural contexts, and notes his structuring of previous classes in order to afford students' leveraging of campus archival collections.

Cordell's focus on material and institutional infrastructure as the "context" of Digital Humanities work dovetails with other scholars' calls to productive engage with the wider "structures" that enable DH work,[1] and, more recently, invocations of "critical infrastructure studies.[2] While the highlighting of the local and distributed material systems and institutions that underpin digital technologies (and therefore DH practices) can provide crucial insights into the hidden labor and material translations that shape DH institutions, this highlighting can serve to flatten and neuter the ideological-epistemological structures that *also* undergird digital practices.

In the context of the STEM educational apparatus, where I find myself enmeshed as a member of the Science and Technology Studies (STS) Department at an Engineering-Centered Institution, these epistemological frameworks can be especially influential, and are often cast as emphasis of "technical expertise" at the cost of the kinds of critical knowledge work that humanities faculty claim to encourage in our students. At face value, this may not be particularly surprising to other humanities scholars. In advocating for the need for DH faculty to resist overplaying "the digital" card in our classrooms, Cordell describes the orientation of the kinds of students we find enrolled in Humanities majors:

> Many of our students honestly, truly, really choose literature or history or art history or religious studies because they wanted to read and think deeply rather than follow what they perceive as a more instrumentalist education in business or technical fields. To do so they often resist substantial pressure from family and friends pushing them toward "more practical" majors, which are often perceived to be more technical majors.[3]

Cordell's characterization fits the standard understanding of where DH takes place—in English departments[4] and classrooms where computational methods are being used to augment "traditional" humanities education. These students are of a different sort than students in more "instrumentalist" programs and majors—usually stereotyped in DH scholarship as STEM students interested in quantification, technology, and the ability to get a job.[5] However, if DH is to truly operate not as an "interdisciplinary bridge,"[6] but rather as a force to resolve and heal the divides between computational/technical practices and interpretive/critical scholarship, we must begin to take seriously the kinds of epistemic-infrastructural contexts STEM disciplines are embedded in, as well as the understand the ideological histories that have shaped those contexts. We must attend to students and scholars in educational contexts *the opposite* of which Cordell outlines above: in Engineering-Centered Institutions, Polytechnics, and other *instrumentalist*[7] educational contexts.

In this essay, I want to talk about instrumentalism not as pragmatic practice, but as ideological-epistemological apparatus. Instrumentalism not only resists the kinds of non-deterministic scholarship practiced in many humanities spaces, but it is also explicitly designed to account for, consume, and subvert the impacts of critical perspectives on technological systems. I thus want to inflect the concept of "infrastructure" differently than Alan Liu, who defines infrastructure as "the social-cum-technological milieu that at once enables the fulfillment of human experience and enforces constraints on that experience."[8] Rather than enabling and constraining the activities of users, I argue that infrastructures operate epistemically, as "machineries of knowledge,"[9] to *produce* those users

1 Kirschenbaum, Matthew. (2012). "Digital Humanities Is/As a Tactical Term." Debates in the Digital Humanities, ed. Matthew Gold, University of Minnesota Press
2 Parks, Lisa and Starosielski, Nicole. (2015). Signal Traffic. University of Illinois Press
3 Cordell, 2016
4 Kirschenbaum, Matthew. (2012). "What is Digital Humanities and What's it Doing in English Departments?" Debates in the Digital Humanities, ed. Matthew Gold, University of Minnesota Press
5 Cordell 2016
6 Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler. (2014). "Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities", Dagstuhl Reports, Vol. 4, Issue 7, pp. 80–111
7 Nieusma, Dean. (2015). "Conducting the Instrumentalists: a Framework for Engineering Liberal Education." Engineering Studies. Vol. 7, 2-3, pp. 159-163
8 Liu 2016
9 Knorr Cetina, Karin. (1999). Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press

themselves.[10] I borrow from STS scholar Karin Knorr Cetina in arguing that infrastructures of scientific and technical production—including those relevant to the digital humanities—should be understood less as "knowledge infrastructures" and more as "epistemic infrastructures." For Knorr Cetina, the term "knowledge structures" implies that material-social systems work to produce *what* we know. The term "epistemic structures," in contrast, highlights how those systems work instead to produce *how* we know, by producing the practices, tools, spaces, and boundaries of "knowing" and of knowable objects.[11] Machineries of knowledge thus produce "epistemic subjects" and "epistemic objects:" practitioners and their always-in-negotiation objects of study.[12]

If we take seriously the epistemic infrastructures of STEM education, it would be wrong to think of engineering students as instrumentalist persons who enter STEM in order to "be filled" with narrow technical expertise, or of engineering instructors as conspiratorial anti-political agents. Rather, the instrumentalist epistemic structures of engineering education *produce* students and teachers as technical practitioners; experts who, through their mastery of the fundamentals of math and physics, practice the production of "non-political" material systems. Simultaneously, though engineering students generally understand that technology "in the world" has social dimensions, engineering's epistemic infrastructures produce technology as an *epistemic object*—"Technology" as abstract and ideal, methodological and *apolitical*—and define the boundaries of STEM's knowledge domain as the exploration of that epistemic object of Technology.

Instrumentalist epistemic infrastructure is frighteningly effective at producing anti-political practices. Erin Cech's longitudinal study of engineering students at four different universities shows that engineering students' interest in public welfare, social concerns, and the political impacts of technological systems steadily *declines* over the course of their education.[13] This is despite the fact that, in most engineering programs, what little hands-on design, making, and human-interaction work that students do engage in almost always occurs towards the end of their coursework. This heavy declination of interest in social and political good should be especially concerning given that early outreach programs, particularly at the grade school level, combine building activities with "use technology to change the world" rhetoric to recruit students into STEM career paths. These programs, which include activities like *Lego Mindstorms* workshops and hands-on hackathons—and are not altogether unlike celebrated "making" pedagogies in the digital humanities--even consciously recruit women and underrepresented minorities, ostensibly in an effort to diversify the STEM workforce. Upon entrance into STEM higher education, however, students are subjected to a double "bait-and-switch:"[14] as making and building activities are immediately sidelined in favor of math and science foundations courses, so too are political and ideological concerns systematically excised from the epistemic object of engineering. This double bait-and-switch is coupled with a systemic administrative devaluing of interpretive humanities and social science courses. While engineering students in the U.S. are required (for now) to take "broad educational" courses, in my experience engineering students are often encouraged by their academic advisors to take "easy" humanities courses that they can mostly ignore in order to concentrate on their core educational work and simultaneously boost their GPA. Instrumentalist infrastructures thus practice the double-move of simultaneously *accounting for* and *defanging* the political ramifications of humanities scholarship.

Unlike Cordell's students who, for various reasons, approach technologically-centered humanities classes with reticence and suspicion, engineering and STEM students interested in taking seriously their humanities classes are often attracted to elective classes that are viewed as fitting in with or dovetailing with their technical education, such as economics or philosophy of technology classes, or that allow them to apply their technical skills in the hands-on, self-directed ways that they are unable to pursue in their core coursework, such as digital arts classes. The technological inflection of the digital humanities thus offers a unique incentive for STEM students, as well as pathway for critical humanities and social sciences faculty to productively engage with those students. Ideally, the digital humanities can even begin subverting the instrumentalist epistemic infrastructures of Engineering-Centered Institutions (and the neoliberal university in general).

However, digital humanities pedagogy is also in a unique position to *reinforce* instrumentalist epistemological infrastructure, as well. Partly, this comes from the difficulty of teaching technical skills and critical thought to undergraduates at the same time, due in no small part to the epistemic infrastructures erected in the university post-instrumental turn. Ian Bogost has opined that humanists have to bracket criticality in order to get our grounding in technical skills.[15] I certainly sympathize with the pragmatic difficulties of teaching undergraduates code and close reading at the same time, particularly in our contemporary instrumental episteme. But bracketing technological practice as apolitical skills with potential social impacts—even in the context of a humanities cour-

10 Knorr Cetina, 1999
11 Knorr Cetina, Karin (2007). "Culture in global knowledge societies: knowledge cultures and epistemic cultures" Interdisciplinary Science Reviews. Vol. 32, 4, pp. 361-375
12 Knorr Cetina, 199
13 Cech, Erin. (2014). "Culture of Disengagement in Engineering Education?" Science, Technology, and Human Values. Vol. 39, 1, pp. 42-7
14 Lachney, Michael and Nieusma, Dean. (2015). "Engineering Bait-and-Switch: K-12 Recruitment Strategies Meet University Curricula and Culture." Proceedings of the American Society for Engineering Education.
15 McPherson, 2014

se—only continues to produce Technology as apolitical epistemic object, as something that can be learned apart from the social and political world. As Tara McPherson suggests, the ontology of brackets is particularly pervasive in digital culture, and can actively undermine critical perspectives on technology and ontologies of difference that emerge from feminist, queer, and postcolonial positions.[16] Thus, DH's relative lack of attention to the epistemic practices of Technology can *encourage* students to assume the instrumentalist stance, and, worse, to pre-tune them to the rejection of politics of difference.

Digital humanities *can* provide a model of transformational resistance to technocratic culture. Rita Raley argues that "the digital humanities should not, and cannot, bear the burden of transforming the technocratic knowledge economy."[17] But if not us, then whom? And who better to build material-epistemic infrastructures that subvert the bracketing of critical thought and technical practice, that challenge the very ideological tenets of instrumentalism, than digital humanists? By entangling ourselves in the apparatuses of STEM education, and by building frameworks for STEM students—especially those in engineering and computer science—to ideologically contextualize their own educational experiences in their technical majors, digital humanities pedagogy can make inroads into dismantling technocratic culture by allying with the very persons in the best position to reproduce it.

Acknowledgements:

Activities described in this paper were made possible in part by the National Endowment for the Humanities.

References

Kirschenbaum, Matthew. (2012). "Digital Humanities Is/As a Tactical Term." *Debates in the Digital Humanities*, ed. Matthew Gold, University of Minnesota Press

Parks, Lisa and Starosielski, Nicole. (2015). *Signal Traffic*. University of Illinois Press

Cordell, 2016

Kirschenbaum, Matthew. (2012). "What is Digital Humanities and What's it Doing in English Departments?" *Debates in the Digital Humanities*, ed. Matthew Gold, University of Minnesota Press

Cordell 2016

Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler. (2014). Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities", *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 80–111

Nieusma, Dean. (2015). "Conducting the Instrumentalists: a Framework for Engineering Liberal Education." *Engineering Studies*. Vol. 7, 2-3, pp. 159-163

Liu 2016

Knorr Cetina, Karin. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press

Knorr Cetina, 1999

Knorr Cetina, Karin (2007). "Culture in global knowledge societies: knowledge cultures and epistemic cultures" *Interdisciplinary Science Reviews*. Vol. 32, 4, pp. 361-375

Knorr Cetina, 1999

Cech, Erin. (2014). "Culture of Disengagement in Engineering Education?" *Science, Technology, and Human Values*. Vol. 39, 1, pp. 42-72

Lachney, Michael and Nieusma, Dean. (2015). "Engineering Bait-and-Switch: K-12 Recruitment Strategies Meet University Curricula and Culture." *Proceedings of the American Society for Engineering Education.*

McPherson, 2014

McPherson, 2014

Raley, Rita. 2014. "Digital Humanities for the Next Five Minutes." *differences: a Journal of Feminist Cultural Studies*. Vol 25, No. 1, pp. 26-45

16 McPherson, 2014

17 Raley, Rita. 2014. "Digital Humanities for the Next Five Minutes." differences: a Journal of Feminist Cultural Studies. Vol 25, No. 1, pp. 26-45

Visualizing the Feminist Controversy in England, 1788-1810

Laura C Mandell
mandell@tamu.edu
Texas A&M University, United States of America

Megan Pearson
mpearson42@tamu.edu
Texas A&M University, United States of America

Rebecca Kempe
rkempe08@tamu.edu
Texas A&M University, United States of America

Steve Dezort
sdezort@tamu.edu
Texas A&M University, United States of America

Recently, text miners have analyzed gendered discourse based on a binary opposition, male/female (M/F), trying to determine distinctively 'female writing style,' 'female keywords,' or 'female themes' (Rybicki 2015; Jockers 2011, 2013). The terms 'male' and 'female' suggest biology and hence were abandoned by literary critics during the big feminist recovery projects of 'women writers': "women' was used in preference to 'female' despite the fact that the former is a noun, not an adjective, indicating that gender was a cultural formation rather than a biological one. More theoretically enlightened text miners have used the tools of data analytics to trace changes through time in those attributes assigned to women and

those assigned to men, examining how notions of gender change over time (Garg, Scheibinger, Jurafksy, Zou 2017; Underwood, Bamman, and Lee 2018; Olsen 1992). There is another way to analyze gender historically using digital tools without assuming a biological basis for differences between men and women that involves searching for gender categories beyond the binary opposition M/F. In her book *Gender Trouble*, Judith Butler encourages undermining the M/F binary opposition by 'proliferating' identity categories (Butler 1990, pp. 17, 146). Butler says that 'the very notion of the subject, intelligible only through its appearance as gendered, admits of possibilities that have been forcibly foreclosed by . . . various reifications of gender' into the M/F binary (Butler 1990, p. 33).

The result of Butler's call to multiply gender categories has been the creation of what Bowker and Star call "boundary objects" (1999): "cisgender" and "transgender" have expanded the binary while still relying on the underlying classification of m/f. But to apply these categories on historical documents is anachronistic: there are historically accurate gender categories that have been identified by others for eighteenth-century such as "molly" (Alan Bray 1988; Randolph Trumbach 1991) and "sapphist" (Lisa Lynn Moore 1997; Yopie Prins 1999). But what about others that have not yet been identified by readers? The Feminist Controversy in England project tries to find foreclosed identity categories, to uncover historically specific gender designations in novels, pamphlets, and essays written by women between 1788 and 1810 in England.

In 1974, Garland Publishing (now no longer in existence) published a collection of 44 treatises by women authors published on topics related to emerging feminism, edited by Gina Luria Walker (https://books.google.com/books/about/The_Feminist_Controversy_in_England_1788.html?id=j1pqMwEACAAJ). Mary Wollstonecraft's ground-breaking *Vindication of the Rights of Women* (1792) was among them. They were facsimile editions. We have used Optical Character Recognition software (Tesseract 3, trained for 18th-century typefaces by the Early Modern OCR Project, http://emop.tamu.edu), corrected the OCR using TypeWright (http://www.18thconnect.org/typewright/documents), run through Named Entity Recognition software to identify character names, and uploaded into the Catma.de interface where they have been tagged by three different teams: undergraduate students, graduate students, and the Professor who is the Principal Investigator on the project. Each team used its own taxonomy explicitly defined in Catma.de except the PI who derives a set of tags from the texts themselves.

The first, most basic taxonomy according to which the texts were tagged by undergraduat students identifies personality traits and activities of characters. The second more interpretive set of tags, encoded by graduate students, involves formal features of novels and essays--protagonists, narrators, and other character types. After these two procedures, each text's tags are clustered by character in a graph, an interactive d3.js interface that allows a third round of tagging by the PI: the personality traits and activities (character attributes) are tagged either as 'gender-normative' or 'different,' and the different categories are given what Johnny Saldaña calls 'In Vivo' codes, short phrases that come from the language of the text itself (Saldaña 2009, 2016). Afterwards, these In Vivo codes are regularized across the whole set of documents. A second visualization interface provides a network view of all the characters grouped by their connections to tagged attributes, both gender normative and different. The goal has been to discover characters clustering around a set of non-normative character attributes--that is, to find personality traits and activities that are both different and shared, which is to say not merely a matter of any specific character's personality. We argue that such clusters present alternative gender categories, based of course upon m/f norms (as are 'cisgender' and 'transgender') but contesting those norms nonetheless.

At DH2018, we present preliminary findings using our prototype visualization interfaces. As we have discovered so far, many characters share attributes in common with Harriet Freke, a character in Maria Edgeworth's *Belinda*. Thus we argue that 'freke' represents a specific gender category found in many of the transcribed texts. The goal is to postulate non-binary gender terms that have been derived from the texts themselves, and to demonstrate how this procedure offers an alternative method for historicizing gender.

References

Bowker, Geoffrey C. and Star, Susan Leigh. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Alan Bray. (1988). *Homosexuality in Renaissance England*. 2d ed. London: Gay Men's Press.

Butler, Judith. (1990) *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Edgeworth, Maria. *Belinda*. 1801.

Garg, Nikhil, Scheibinger, Londa, Jurafksy, Dan, and Zou, James. (2017). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Computation and Language*. arXiv:1711.08412v1

Jockers, Matthew. (2011). The LDA Buffet is Now Open: or, Latent Dirichlet Analysis for English Majors. http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/. Accessed 19 April 2016.

Jockers, Matthew. (2013). *Macroanalysis: Digital Methods and Literary History*. Chicago: University of Illinois Press.

Moore, Lisa Lynn. (1997). *Dangerous Intimacies: Toward a Sapphic History of the British Novel.* Durham, NC: Duke University Press.

Olsen, Mark. (1992). Qualitative Linguistics and *Histoire des Mentalités*: Gender Representation in the *Trésor de la Langue Française*. *QALICO*.
Prins, Yopie. (1999). *Victorian Sappho*. Princeton, NJ: Princeton University Press.
Rybicki, Jan. (2015). *Vive la différence*: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies. *Digital Scholarship in the Humanities*, pp. 1-16. doi: 10.1093/llc/fqv023.
Saldaña, Johnny. (2009, 2016). *The Coding Manual for Qualitative Researchers*. Thousand Oaks, CA: Sage Publications.
Trumbach, Randolph. (1991). Sex, Gender, and Sexual Identity in Modern Culture: Male Sodomy and Female Prostitution in Enlightenment London," *Journal of the History of Sexuality* 2(2): 187- 88.
Underwood, Ted, Bamman, David, and Lee, Sabrina. (2018). The Transformation of Gender in English Language Fiction. *Journal of Cultural Analytics* (February 13, 2018) http://culturalanalytics.org/2018/02/the-transformation-of-gender-in-english-language-fiction/
Wollstonecraft, Mary. *Vindication of the Rights of Woman*. 1792.

ZX Spectrum, or Decentering Digital Media Platform Studies approach as a tool to investigate the cultural differences through computing systems in their interactions with creativity and expression

Piotr Marecki
piotr.marecki@ha.art.pl
Jagiellonian University, Poland

Michał Bukowski
yerzmyey@poczta.onet.pl
Jagiellonian University, Poland

Robert Straky
hellboj@centrum.cz
GH University of Science and Technology, Poland

The point of departure for our paper is the statement that "The computer is not a tool to help us do whatever we do, it *is* what we do, it is the medium on which we work" (Dene Grigar, Electronic Literature Organization), and that the Platform Studies approach is essential in the Digital Humanities field to better understand rules of the contemporary digital world. Our goal is to present an output of our 2-years research project devoted to the 8-bit computer ZX Spectrum (especially the ZX Spectrum 16/48K and 128K models). According to Nick Montfort and Ian Bogost: "Platform Studies investigates the relationships between the hardware and software design of computing systems and the creative works produced on those systems. Particular platform studies may emphasize different technical or cultural aspects and draw on different critical and theoretical approaches, but they will be united in being technically rigorous and in deeply investigating computing systems in their interactions with creativity, expression, and culture." (http://platformstudies.com/)

One could imagine a narration about the ZX Spectrum platform as the official history of the British company Sinclair Research Ltd., in which the official and copyrighted market products of the company would be presented (both the hardware and the software, like games, word processors, graphics processing programs). Sir Clive Marles Sinclair created the ZX Spectrum at the beginning of the 80s as a machine that first and foremost was meant to serve educational purposes. As is the case with the inventions of many creators, Sinclair's broke away and began a life of its own. This unofficial grassroots and human story is the one we wish to tell.

The starting point for our narrative is the belief that the ZX Spectrum platform is unique as compared to other 8-bit machines. Its uniqueness lies in the reception of the platform by users on a scale which is incomparable to that of any other platform. The traditional way of using platforms (not only the 8-bit) is based on their consumption, or the use of the official equipment, as well as programing, delivered by the manufacturer. And although the stories about platforms such as the C-64 or Atari are no strangers to creative and bottom-up approaches, these are based on the creation of independent programs. Besides the ZX Spectrum, none of these platforms generated the same hardware systems or clones on such scale and creative level. This is related to the simplicity of the computer's construction and the cheap cost of the accessories as well as the geopolitical conditions in the world in the period of the platform's popularity, the 80s and 90s. It should also be added that the UK platform was popular mainly in Europe (despite attempts, the platform was never popularized in the United States).

One of the novel aspects of the output of our project is the attempt to compare the East and the West - two worlds with different approaches to the same platform. The story of the Spectrum is used as a focal point that will enable us to describe the differences between the two sides of the Iron Curtain, also after its fall, in the period of political transformation in the countries of the former Eastern Bloc. Briefly, in one of these worlds software and platforms were easily accessible commercial goods, while in the other they were coveted symbols of a different reality in a situation where legal software was almost inaccessible and the platforms were sold only in special stores with foreign goods, or distributed illegally. Both financial and political matters, and the aforementioned simplicity, decided that the platform was cloned en masse. Creativity in both naming the clones (ZX Evolu-

tion, Didaktik, Scorpion – just to mention a few), as well as ways of tuning the equipment, is the subject of our study and description. It is also worth noting that no other platform inspired as many equipment parties, bringing together fans of platforms that to this day create clones of the hardware, or magazines devoted to it. Currently, clones of this platform can still be purchased.

By describing the ZX Spectrum platform, we try to tackle trends that are relevant to contemporary studies on digital media, taking into account and affirming the local perspective, different from the dominant one. We are interested in the aspect of creation in the field of digital media, as well as the use of computers for artistic purposes or programing for fun. During the several decades of the existence of digital media, a number of creative fields and worlds bringing together users of different platforms, used for their creative purposes, have flourished. Alongside the fields of electronic music, video games, new media art and electronic literature, there is the demoscene, separate from and not having many links with the rest of the digital world.

What is the demoscene? This phenomenon is apparent to those with advanced understanding of digital media. In the book *Freax. The Brief History of Computer Demoscene* it is stated that "almost all modern art genres have an underground stream that can not be found anywhere, or bought in shops, and only insiders know of its existence." (Polgár, 2005: 6) Adjectives such as illegal, grassroots, independent are often related with this field and practice. The term itself is derived from the word "demonstration" and refers to the demonstration of the capabilities of a platform and the skills of a programmer. A basic understanding of the demoscene will treat it as "a subculture in the computer underground culture universe, dealing with the creative and constructive side of technology" (Demoscene FAQ).

The reasons, however, for telling the story of this particular platform through this perspective are several. Among the platforms, the 8-bit Spectrum is widely considered to be the cheapest (this aspect is important to taking up the issues of accessibility, universality and democratic nature of the platform). This is a very important factor considering the fact that the prices of other platforms could be added up to a number of monthly salaries. One of the objectives of the demosceners is to circumvent the technical limitations of the platform. Demosceners fell in love with the ZX Spectrum, more or less because it is recognized as a platform characterized by the simplest technical solutions, so it was natural to perform impossible operations on it.

Our proposal can be compared above all to recent approaches from the book series on platforms studies. To indicate the most recently published, among them are *Now the Chips Are Down: The BBC Micro Alison Gazzard* (2016, MIT Press, due to the British local context) and *The Future Was Here: The Commodore Amiga* (2012, MIT Press) by Jimmy Maher. Another work that addresses the issues of a local approach to personal computer is an *Electronic Dreams: How 1980s Britain Learned to Love the Computer* by Tom Lean (Bloomsbury Sigma, 2016). The book by Tom Boellstorff, Bonnie Nardi, Celia Pearce & T. L. Taylor, *Ethnography and Virtual Worlds: A Handbook of Method* (Princeton University Press, 2012), mainly the ethnographic research by T. L. Taylor, is a reference point for our ways of working with the community that uses digital media.

Our paper will present the findings of a two year research project devoted to the platform. The research work will involve semi-structured interviews (with 20 demosceners from Russia, the Czech Republic, Poland, Slovakia), centered around the creative possibilities of the platform.

References

Polgár T. (2005). *Freax. The Brief History of the Computer Demoscene.* Winnenden: CSW Verlag.

Ciências Sociais Computacionais no Brasil

Juliana Marques
juliana.marques@fgv.br
FGV CPDOC, Brazil

Celso Castro
celso.castro@fgv.br
FGV CPDOC, Brazil

Introdução

O rótulo das Humanidades Digitais (HDs) ainda não é amplamente conhecido, ou deveríamos dizer reivindicado, pela academia no Brasil. Isto não quer dizer que não haja reflexões e produções sobre tecnologia digital e com tecnologia digital sendo realizadas em departamentos de Letras, de Artes, de Ciências Humanas e Sociais Aplicadas.

As HDs podem designar tanto um conjunto de práticas de pesquisa que têm em comum a utilização de tecnologias digitais - sejam seus temas de interesse e objetos de pesquisa do mundo virtual ou não-, quanto um novo campo de conhecimento, muitas vezes com pretensão transdisciplinar e de fato, em grande parte, interdisciplinar (Alves, 2016; Ortega and Gutiérrez, 2014; Schreibman et al., 2004).

A história humana é também a história da criação de tecnologias, que ao longo do tempo, contribuíram para formular e solucionar problemas da vida em sociedade, criando variadas formas de conexão entre pessoas, desde a coleta, o estoque e a preparação de alimentos até, milhares de anos depois, as máquinas à vapor, a eletricidade e a eletrônica (Derry and Williams, 1993). Para além

de ferramentas e máquinas, no início do século 20, o termo passa a abranger uma gama crescente de meios, processos e idéias utilizados pelas pessoas para mudar ou manipular seu ambiente (Behrent, 2013; De Landa, 1997).

O recorte aqui empregado delimita este conceito ao empregar a categoria de **tecnologia digital**, que se caracteriza pela transformação de qualquer linguagem ou informação, incluindo textos, sons, imagens fixas ou em movimento, entre outras, em registros numéricos binários, isto é, em zeros e uns (0 e 1). Essa transformação depende de sistemas computacionais criados na primeira metade do século XX a partir de conceitos matemáticos do século XVII e é o que conhecemos, hoje, como informação digitalizada, ou seja, gravada neste código binário, também chamado de bits. Ela permite que grandes quantidades de informações sejam compactadas em pequenos dispositivos de armazenamento que podem ser facilmente preservados e transportados. Ou seja, a digitalização acelera a velocidade e aumenta a capacidade de transmissão e, posteriormente, de troca de informação. Sem dúvida, a tecnologia digital transformou e continua a transformar radicalmente a maneira como nos comunicamos e como atuamos no mundo (Mansell, 2002).

A fim de produzir novos conhecimentos e desenvolver habilidades de trabalho adequados à revolução causada pelas tecnologias digitais de informação e comunicação, passaram a atuar, juntos, profissionais das Humanidades e das Ciências da Informação e Tecnológicas. O trabalho conjunto tem visado, sobretudo, a constituição e a difusão de acervos, repositórios e bibliotecas digitais, o uso de sistemas de informação geográfica, o tratamento computacional de linguagens, tanto as visuais, em seus variados formatos, como as verbais, em formatos de texto ou de áudio, e a simulação de realidades virtuais.

Segundo o Digital Humanities Manifesto 2.0, a primeira onda de trabalho das HDs (no contexto internacional) foi quantitativa, mobilizando os poderes de pesquisa e recuperação do banco de dados, automatizando a chamada linguística de corpus etc. A segunda onda seria de caráter mais qualitativo e interpretativo, mobilizando a experiência e a emoção criadora. A partir de então, a riqueza hermenêutica das humanidades é resgatada. O Estado da Arte deste campo em formação apontaria agora para a necessidade de novas conexões e mudanças, que são facilitadas tanto por novos modelos de prática de pesquisa quanto pela disponibilidade de novas ferramentas digitais.

Em 2013, a Alliance of Digital Humanities Organizations (ADHO) estimava que o termo Humanidades Digitais, empregado de forma polissêmica, era empregado em mais de 29 países, nomeava mais de cem centros de pesquisa e movimentava pelo menos 40 milhões de dólares por ano, em um movimento de expansão que se iniciou por volta de 1989. Foi também em 2013 que foi organizado, no Brasil, o primeiro congresso internacional dedicado ao assunto. O que sabemos sobre o histórico de práticas e do campo das HD no Brasil?

Objetivo

Este trabalho se propõe a construir um panorama, ainda que incompleto, de uma parte circunscrita da cena nas Humanidades Digitais no Brasil, que permanece em grande parte misteriosa e inexplorada.

O objetivo é mapear, nas Ciências Sociais brasileiras desde a década de 1980 até o presente, práticas, publicações, instituições e pessoas que aproximam tecnologias digitais e questões, teorias e métodos de pesquisa considerados tradicionais em seu campo. Nossa amostra inclui pesquisadoras e pesquisadores doutores que usam tecnologias digitais como ferramentas para análise e para construção do conhecimento (42%) e também aquelas que as encaram como tópico para reflexão (58%), na medida em que procuram dar conta de como essas tecnologias foram e vêm sendo concebidas, criadas e usadas.

Por Ciências Sociais, queremos dizer as subáreas da Sociologia, Antropologia, Ciência Política e Relações Internacionais. De acordo com a avaliação quadrienal da CAPES, publicada em 2017, são 126 Programas de Pós-Graduação a abrigar 197 cursos. Estima-se que são mais de mil professores e mais de 2.000 pesquisadores em formação.

Um estudo como este contribui para reunir informações sobre esse campo em formação, mesmo quando seus atores não rotulam o que fazem como Humanidades Digitais. É notório que muitos dos temas abordados pelas Hds dão título a pesquisas, cursos, publicações e eventos ocorridos no país. Justamente pela diversidade de estudos e aplicações que conjuga, a área vem ganhando cada vez mais espaço, mesmo que ainda não forme uma comunidade de práticas bem definida.

Proposta

A partir do mapeamento realizado, criar-se-á uma base de dados sobre a relação entre Ciências Sociais e Humanidades Digitais no Brasil que permitirá a posterior realização de análise de redes e georreferenciada, além de avaliações qualitativas das práticas e temas privilegiados por cientistas sociais em contexto nacional.

A pesquisa foi realizada digitalmente em um processo que chamamos, sem perder o humor, de raspagem artesanal de dados, tendo como guia uma lista de 23 termos-chave, usados principalmente no banco de dados bibliográfico SciELO Brasil, no catálogo de teses e dissertações da CAPES e na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). Os termos humanidades digitais, humanidades computacionais e tecnologias digitais guiaram as buscas no diretório de grupos de pesquisa do CNPq. A busca pelo tema de interesse nas páginas dos programas de pós-graduação e, a partir daí, na Plataforma Lattes foi realizada através da leitura de todos os perfis docentes. No portal Periódicos CAPES, buscamos por

revistas do campo das Ciências Sociais voltadas para o tema das tecnologias digitais, de forma ampla, e encontramos apenas duas revistas, avaliadas nos estratos B4 e B5.

Os resultados apontam para maior inserção da subárea da Sociologia (42%) na cena das HDs, seguida pela Antropologia (36%) e então pela Ciência Política e Relações Internacionais (22%). No total, 70% dos 67 pesquisadores da amostra é composta por homens e apenas 30% de mulheres, o que registra uma acentuada desigualdade de gênero, com destaque para maior equilíbrio entre antropólogos (54% homens, 46% mulheres). A maior parte dos pesquisadores e das iniciativas está concentrada em instituições públicas e nas regiões Sudeste e Sul do país, que é onde está a maior parte dos programas e dos cursos de pós-graduação. Mais uma vez, na Antropologia, o interesse está mais bem distribuído geograficamente, com 29% de participação das regiões Sul, Sudeste e Nordeste. Chamou atenção como o Sul concentrou o movimento de vanguarda na criação de linhas e centros de pesquisa voltados para o mundo digital, ainda na década de 1980. O ponto de inflexão, com expressivo crescimento no número de pessoas, projetos e publicações que poderiam fazer parte da comunidade das HDs aconteceu por volta dos anos 2000 e, surpreendentemente, a tendência recente (dos últimos 5 anos) não parece ser de intensificação do engajamento com tecnologias digitais por parte dos cientistas sociais.

Nesse contexto, é imperante a formação de um novo profissional com habilidades e competências conectadas com as transformações tecnológicas de nossos tempos. Esta é a realidade tanto para aqueles que já se encontram inseridos no mercado como para as gerações de profissionais ainda em formação. A interdisciplinaridade implica mudanças na linguagem, nas práticas, nos métodos, resultados e produtos de pesquisa, tendo cada vez mais como diretriz normativa a concepção de uma Ciência Aberta.

O artigo publicado em maio deste ano no periódico da FAPESP, intitulado A realidade que emerge da avalanche de dados, diz com muita propriedade que "Humanidades digitais se disseminam por várias disciplinas, influenciam formação de pesquisadores e inspiram políticas públicas" (Marques, 2017: 19). De fato, a apropriação das tecnologias permite que façamos mais rápido e mais além do que fazíamos anteriormente - principalmente com relação à análise de grandes volumes de dados -, motivando pesquisadores a buscarem formações inovadoras que os capacitem.

Logo, seja como novo campo seja como comunidade de práticas, as HD se colocam o desafio de incorporar novas perguntas, epistemologias e métodos à tradicional forma de trabalhar das Ciências Sociais. Emergem, portanto, questões teórico-metodológicas, éticas e políticas que se impõem a pesquisadores, educadores e estudantes.

References

Alves, D. (2016). As Humanidades Digitais como uma comunidade de prática dentro do formalismo académico: dos exemplos internacionais ao caso português. *Ler História*, 69, 2016, pp. 91-103.

Behrent, M. C. (2013). Foucault and Technology. *History and Technology*, 29:1, pp. 54-104.

De Landa, M. (1997). *A Thousand Years of Nonlinear History*. New York: Zone Books.

Derry, T. K. and Williams, T. I. (1993). *A Short History of Technology: From the Earliest Times to A.D. 1900*. Dover Publications.

Mansell, R. Ed. (2002). *Inside the Communication Revolution: Evolving Patterns of Social and Technical Interaction*. Oxford and New York: Oxford University Press.

Marques, F. (2017). A realidade que emerge da avalanche de dados. *Revista Pesquisa Fapesp*, 255, maio de 2017.

Ortega, E. and Gutiérrez, S. E. (2014) MapaHD. Una exploración de las Humanidades Digitales en español y portugués. In Frías, E. R. and González, M. S. Eds. *Ciencias Sociales y Humanidades Digitales Técnicas, herramientas y experiencias de e-Research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social, pp. 103-104.

Schreibman, S., Siemens, R. and Unsworth, J. (2004) *A Companion to Digital Humanities*. Oxford: Blackwell.

Distributions of Function Words Across Narrative Time in 50,000 Novels

David William McClure
dclure@mit.edu
Massachusetts Institute of Technology, United States of America

Scott Enderle
scott.enderle@gmail.com
University of Pennsylvania, United States of America

What can be said, at an empirical level, about the internal structure of literary narratives? Can we model the "shape" of a plot? In recent years, there has been a surge of interest in what might be thought of as a computational form of narratology. Instead of flattening out the text into an unordered bag of words, a series of studies have looked at the fluctuation of different types of literary signals across "novel time," the linear space between the beginning and end of a text. Most well-known is probably Matt Jockers' work with Syuzhet, an R package that calculates the dispersion of positive and negative sentiment across novels. Ben Schmidt, working with a corpus of movie and TV scripts, tracked the distribution of topics across the screenplays,

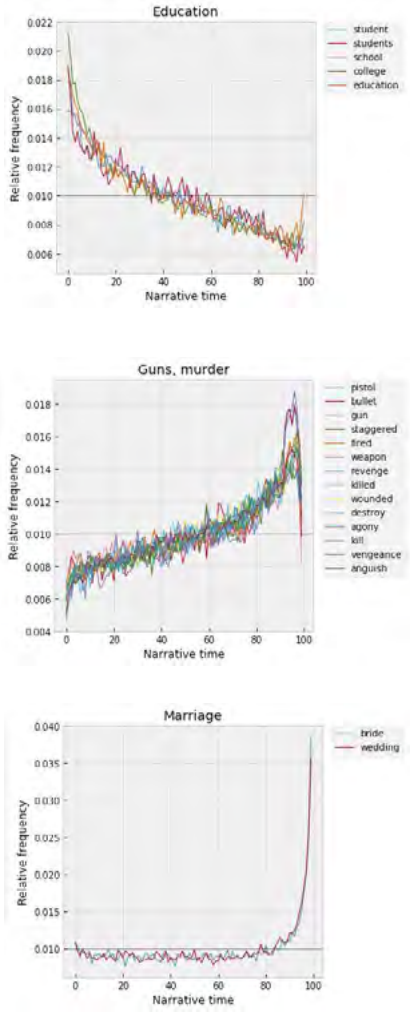
finding a footprint of the prototypical cop drama, with a crime at the beginning and a trial at the end. Andrew Piper, writing in New Literary History, identifies a signature for the "conversional" narrative, based on Augustine's *Confessions*, and then traces this signal forward in literary history. At the Stanford Literary Lab, Holst Katsma tracked the "loudness" of speech utterances across chapters in Dostoevsky and Austen. And Mark Algee-Hewitt, working with collaborators on the Suspense project, has trained a classifier that can score the "suspensefulness" of passages of text at different regions across narrative time.

Each of these signals -- sentiment, topic, suspense, loudness -- is a fascinating object of study in its own right. But what are they signals of? How do they interact with one another? Do they track distinct literary phenomena, or are they moved by the same underlying forces? Are they hierarchically related to one another -- is "conversion" a subset of "topic"? Does "sentiment" encompass "suspense"? From among the infinite proliferation of threads that weave through the text, why should we select these? Are they the most explanatory, the most cross-cutting, the most fundamental? Or are there any fundamental, cross-cutting aspects to narrative at all?

Far from trying to offer definitive answers to these questions, we explore a minimalist, bottom-up approach, with the goal of providing a basic corpus-linguistic survey of the internal structure of English language novels -- we are interested in a very simple treatment of the question, with the aim of providing framing and context for higher-level studies. Instead of starting with a relatively complex signal like sentiment or suspense, we just look at the distributions of individual words across narrative time in a corpus of ~50,000 novels, and then systematically identify the words that have the most non-uniform distributions across narrative time when averaged across the entire corpus -- words that are most skewed, the most distinctive of beginnings, middles, ends, and anything in between.

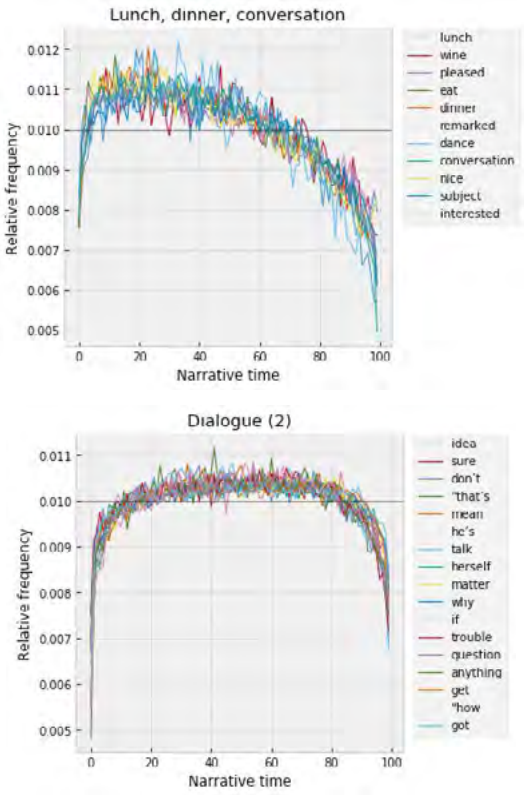
This is extremely simple, essentially just a particular way of counting words. Working with Gale's American Fiction corpus (~18k novels from 1820-1940), a subset of cleaned texts from the Chicago Novel Corpus (~7k American novels from 1880-2000), and a subset of HathiTrust (a sample of 20k works identified as fiction by Underwood et al.), we split each text into a set of N equally-sized chunks and then count the total number of times that each word appears in each of these chunks across all novels. For example -- the word "love" appears 9,418 times in the first 1/100th of novels in the corpus, compared to 25,132 in the last percentile. With this, we can represent each word as a distribution across narrative time and compare the variance to what would be expected, given the frequency of the word, under a uniform distribution -- the baseline variation that we would expect if the frequency of the word had no significant relationship with the position inside the narrative. This gives a simple way to score each word, to quantify the degree to which it tends to cluster in some regions of the narrative at the expense of others.

With content words, many of these results confirm basic intuitions about genre conventions and the pragmatic requirements of storytelling. For example, beginnings are filled with descriptions of people, places, and things; birth, youth, education; and enumerations of family relationships. Guns, death, war, and criminal justice peak right around 95%, the moment of climax and peak action. Endings are marked by marriage, death, and expressions of emotion, both happy and sad.



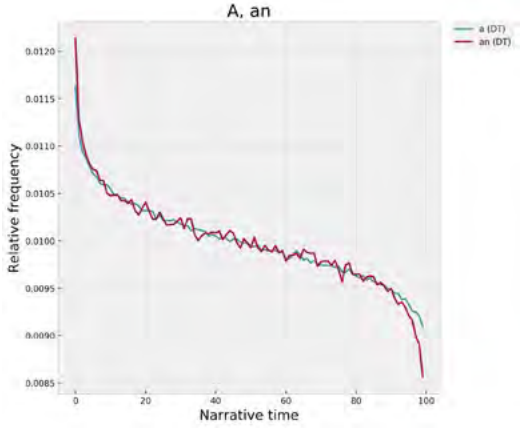
Education, guns, and marriage across ~30k novels. (Gale American Fiction, Chicago Novel Corpus)

Other words have patterns that are somewhat less self-evident. For example -- words related to food, eating, talking, and female characters peak strongly around the 10-20% marker in the novel, as if -- once the cast of characters is introduced at the start, it's common novelistic practice to sit them down and put them in conversation together over a meal, as an early set-piece. Or, at the 50% marker -- novelistic middles seem to be dominated by speaking and thought, words related to dialogue and psychological experience.



Food and conversation at ~20%; dialogue at ~50%. (Gale American Fiction, Chicago Novel Corpus)

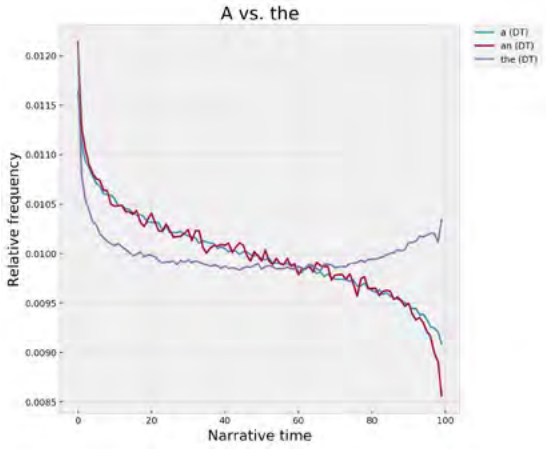
More interesting, though, it turns out that function words -- including many of the most frequent words in English -- also have very irregular distributions across narratives when averaged across tens of thousands of texts, often in ways that seem to suggest the presence of low-level (and highly consistent) narratological tendencies that sit well below the conventions of genre or plot. For example, the indefinite articles "a" and "an" fall off dramatically across narrative time -- they are overrepresented at the start, fall off quickly in the first 20%, decline more slowly across the middle, and then fall quickly again at the end.



"A" and "an" across ~30k novels. (Gale American Fiction, Chicago Novel Corpus)

Since "a" carries very little semantic content, the interpretation of this is more complex. Perhaps this is related to the fact that "a" is used when an entity is referred to for the first time, when it is "unfamiliar" in the narrative frame? For instance, to use an example from Abbot (2006), we might first say -- "Mary saw a movie last week" -- before then switching to the definite "the," once the entity has been introduced -- "The movie was not very interesting." It seems plausible, then, that "a" would be in higher demand at the beginning, when everything is unfamiliar and the fictional world needs to be described for the first time. If this is the narratological role of "a" -- can we treat it as a marker for a general notion of "speed" or "motion" in narrative, the degree to which the text is moving into new fictive contexts that need to be described for the first time?

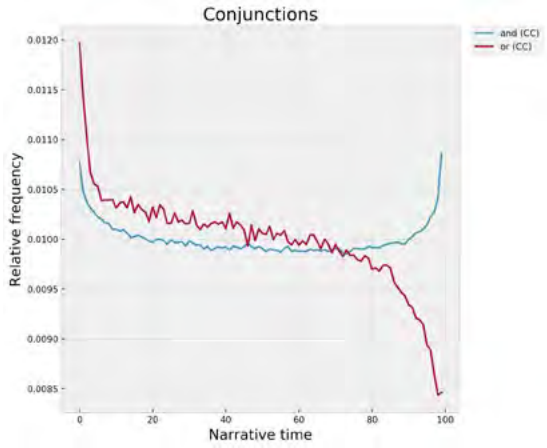
The distribution of "the," though, is neither the same nor precisely the opposite of "a," and seems to be marking some different configuration of narrative pressures. "The" is high at the start, like "a," but with a much faster falloff; then flat across the middle, and with a significant rise at the end.



Indefinite vs. definite articles. (Gale American Fiction, Chicago Novel Corpus)

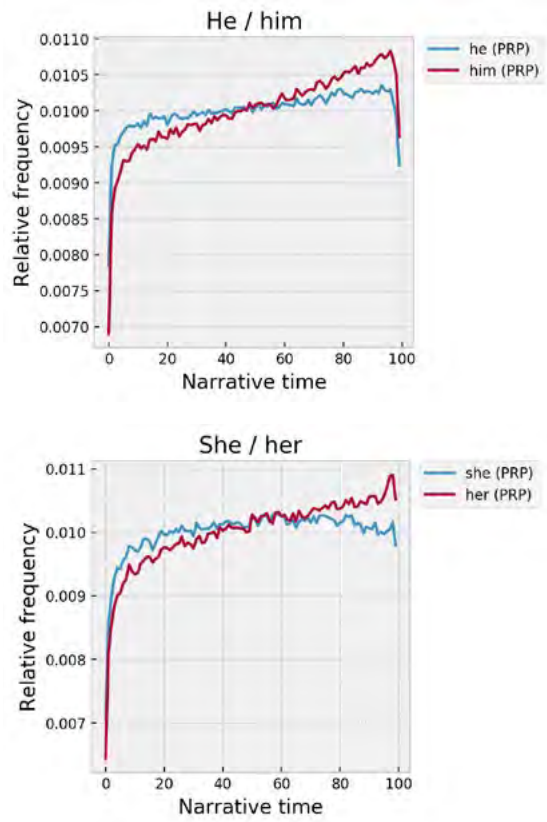
This seems to muddy the picture. "A" seems to be marking something about how beginnings and ends are different, whereas "the" is marking something about how they are similar. But how precisely, and why? Why do "a" and "the" diverge at the end?

These trends are highly variable across the ~50 most frequent words in English, often in ways that seem to suggest a kind of basic taxonomy of narrative variation, a set of lenses for thinking about the ways in which narratives can change across the axis of the text. For example, "and" and "or" also separate cleanly at the end:



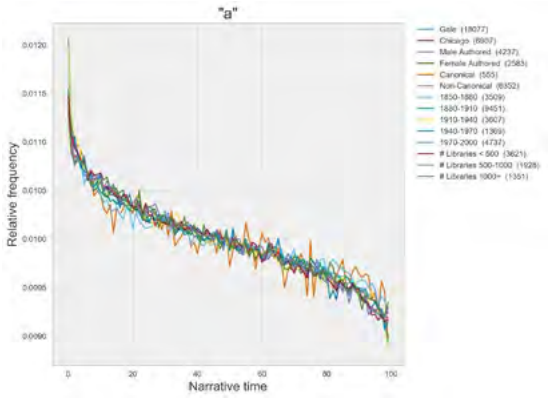
"And" vs. "or." (Gale American Fiction, Chicago Novel Corpus)

Where, perhaps, "or" tends to introduce a state of indeterminacy, a potential fork, and thus falls off as the text approaches the end, as the "circle" of the plot comes to a close, as James would say? Or, less intuitive -- personal pronouns tend to increase across the narrative. But, the object pronouns "him" and "her" rise more steeply than the subject pronouns "he" and "she" -- so, as the narrative progresses, people increasingly become grammatical *objects*? People do things at the beginning of stories, and increasingly have things done to them at the end?

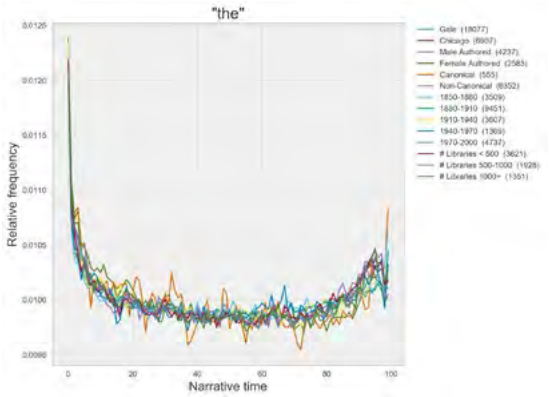


Object pronouns rise more steeply across the narrative than subject pronouns. (Gale American Fiction, Chicago Novel Corpus)

This paper will investigate these patterns and attempt to provide linguistic and literary explanations for why they look the way they do, with a focus on the differences between "a" and "the." Overall, we find that the distributions are highly consistent across corpora (Gale, Chicago, HathiTrust), date of publication (1850-2000), and available metadata for canonicity and author identity:



The distribution of "a," sliced by corpus, publication date, author gender, and canonicity. (Gale American Fiction, Chicago Novel Corpus)



The distribution of "the," sliced by corpus, publication date, author gender, and canonicity. (Gale American Fiction, Chicago Novel Corpus)

We also find that the patterns are significantly different than trends observed in a corpus of ~20k nonfiction volumes from HathiTrust (Underwood et al., 2015), which suggests that they mark something specific about the structure of (fictional) narratives, and not just something that arises generally in long documents. Beyond the aggregate trends -- we explore the degree to which individual texts do and don't conform to the corpus-level averages, with a focus on what can be learned from the extreme examples that most strongly exemplify and resist the overall trends -- for example, the ~1% of novels for which "a" increases consistently across the entire text at a statistically significant level.

References

Abbott, B., 2006. Definite and indefinite. *Encyclopedia of language and linguistics*, 3, pp.392-399.

Chu, E. and Roy, D., 2017. Audio-Visual Sentiment Analysis for Learning Emotional Arcs in Movies. arXiv preprint *arXiv*:1712.02896.

Froehlich, H., 2012. Independent women? Representations of gender-specific possession in two Shakespeare plays. *7th Lancaster University Postgraduate Conference in Linguistics & Language Teaching*.

Jockers, M., 2015. *Revealing Sentiment and Plot Arcs with the Syuzhet Package*. Matthew L. Jockers [blog], http://www.matthewjockers.net/2015/02/02/syuzhet/.

Katsma, H., 2014. Loudness in the Novel. Stanford Literary Lab, *Pamphlet* 7. https://litlab.stanford.edu/LiteraryLabPamphlet7.pdf

Piper, A., 2015. Novel devotions: Conversional reading, computational modeling, and the modern novel. *New Literary History*, 46(1), pp.63-98.

Reagan, A.J., Mitchell, L., Kiley, D., Danforth, C.M. and Dodds, P.S., 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1), p.31.

Schmidt, B.M., 2015, October. Plot arceology: A vector-space model of narrative structure. In *Big Data (Big Data), 2015 IEEE International Conference* on (pp. 1667-1672). IEEE.

Underwood, T., Black, M.L., Auvil, L. and Capitanu, B., 2013, October. Mapping mutable genres in structurally complex volumes. In*Big Data, 2013 IEEE International Conference* on (pp. 95-103). IEEE.

Underwood, T., Capitanu, B.,Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C., Downie, J.S., 2015. *Word Frequencies in English-Language Literature, 1700-1922 (0.2)* [Dataset]. HathiTrust Research Center. doi:10.13012/J8JW8BSJ.

Challenges in Enabling Mixed Media Scholarly Research with Multi-media Data in a Sustainable Infrastructure

Roeland Ordelman
rordelman@beeldengeluid.nl
University of Twente, Netherlands Institute for Sound and Vision, The Netherlands

Carlos Martínez Ortíz
c.martinez@esciencecenter.nl
Netherlands Escience Center, The Netherlands

Liliana Melgar Estrada
melgar@uva.nl
University of Amsterdam, Netherlands Institute for Sound and Vision, The Netherlands

Marijn Koolen
marijn.koolen@huygens.knaw.nl
KNAW, Huygens ING, The Netherlands

Jaap Blom
jblom@beeldengeluid.nl
Netherlands Institute for Sound and Vision, The Netherlands

Willem Melder
wmelder@beeldengeluid.nl
Netherlands Institute for Sound and Vision, The Netherlands

Jasmijn Van Gorp
j.vangorp@uu.nl
Utrecht University, The Netherlands

Victor De Boer
v.de.boer@vu.nl
Vrije Universiteit Amsterdam, The Netherlands

Themistoklis Karavellas
tkaravellas@beeldengeluid.nl
Netherlands Institute for Sound and Vision, The Netherlands

Lora Aroyo
lora.aroyo@vu.nl Vrije
Vrije Universiteit Amsterdam, The Netherlands

Thomas Poell
t.poell@uva.nl
Vrije, University of Amsterdam, The Netherlands

Norah Karrouche
karrouche@eshcc.eur.nl
Erasmus University Rotterdam, The Netherlands

Eva Baaren
ebaaren@beeldengeluid.nl
Netherlands Institute for Sound and Vision, The Netherlands

Johannes Wassenaar
jwassenaar@beeldengeluid.nl
Netherlands Institute for Sound and Vision, The Netherlands

Julia Noordegraaf
j.j.noordegraaf@uva.nl
Vrije, University of Amsterdam, The Netherlands

Oana Inel
oana.inel@vu.nl
Vrije Universiteit Amsterdam, The Netherlands

Providing scholarly access to large collections that are distributed across various content providers, and to create user-friendly applications to work with that data in a diversity of scholarly projects is the underlying goal of the development of the Common Lab Research Infrastructure for the Arts and Humanities (CLARIAH)[1].

Big-scale infrastructure projects in the humanities and social sciences such as the Digital Research Infras-

1 https://clariah.nl/

tructure for the Arts and Humanities (DARIAH) (Edmond et al., 2017), or the Common Language Resources and Technology Infrastructure (CLARIN) (Hinrichs and Krauwer, 2014) aim to provide solutions for both preservation and access to collections and data necessary for scholarly research (Zundert, 2012). Some infrastructure projects build decentralized "atomic" software services, e.g., as in the LLS infrastructure project (Buchler et al., 2016), while others prefer to build more centralized virtual research environments, as in the European Holocaust Research Infrastructure (EHRI) (Lauer, 2014). Also, even within a single infrastructure project, these two models can coexist. This is the case of the CLARIAH infrastructure, where different approaches have been taken to date for serving different user groups, i.e., several specialized tools for linguists (Odijk, Broeder & Barbiers, 2015), or a research environment (the *Media Suite*) that serves the scholarly needs for working with audiovisual data collections and related mixed-media contextual sources that are maintained at cultural heritage and knowledge institutions. This paper discusses the rationale and challenges behind the development of the *Media Suite*.

The CLARIAH Media Suite

Whereas in some domains of scholarly research the focus is on the creation of private data collections, in other domains scholarly research focuses on already *established* data collections maintained at heritage and knowledge institutions. Access to and use of these latter collections is often restricted, especially when they concern audiovisual media, due to intellectual property rights (IPR) or privacy issues (e.g., with respect to recorded interviews). Therefore, many scholars end up using collections that are openly available. Or, they spend considerable amounts of time in doing on-site visits to archives where data are available for consultation. Data collections at these institutes are "locked," scattered, or at least hard to use for scholarly research.

To open up these collections for research and let scholars take advantage of the sheer quantity and richness of these data sets, we developed the *Media Suite* (Figure 1),[2] a research environment or high-level tool that works as a data aggregator where the metadata and the media content can be explored, browsed, analyzed, stored in personal collections, annotated manually, enriched automatically, and visualized, and where the user annotations can be exported.

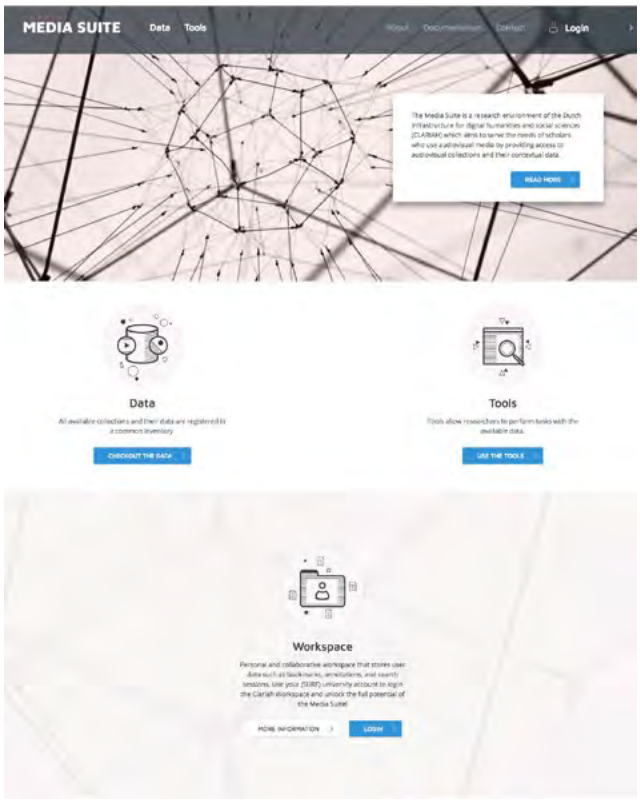


Figure 1. The CLARIAH Media Suite's homepage (http://mediasuite.clariah.nl/, version 2)

The ultimate goal of the *Media Suite* is to: (i) enable distant reading (Schulz, 2011), that is, identifying patterns or new research questions in all aggregated collections; (ii) facilitate close reading: the detailed examination of individual items (e.g., videos) in a collection or parts of these items (e.g., video segments) during search and scholarly interpretation, and (iii) make sure that the "scholarly primitives" (Unsworth, 2000, also described as an infrastructure framework in Blanke and Hedges, 2013) are well supported.

Even though these are accepted scholarly approaches that should be taken into account by infrastructure projects in the humanities nowadays, the question is: How to facilitate "close reading" when the media objects cannot be accessed because of copyright issues? How to enable "distant reading" when the content cannot be fully automatically processed or when their metadata is diverse and incomplete? How to cater to the needs of scholars with specific research questions and methods in the context of an infrastructure that has to be generic enough to be feasible?

Challenges and solutions

The approach of the CLARIAH Media Suite to tackle these challenges is: (i) to organise and implement a federated authentication mechanism to overcome access barriers

2 The first of a four release versions was introduced in April 4, 2017.

(Figure 2, number 5), and (ii) to provide mechanisms that enable researchers to work with tools and aggregated data *within* the infrastructure. We refer to this approach as to "bringing the tools to the data", as opposed to "bringing the data to the tools".

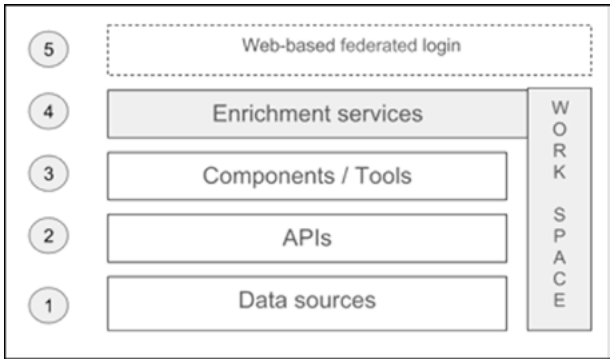


Figure 2. The building blocks of the CLARIAH Media Suite

Figure 2 shows the main elements that constitute the *Media Suite* research environment. We explain below to which challenge is each element an answer to:

Data Sources -- Data Governance

Institutional collection maintainers have internal data governance processes to ensure that data assets are formally managed. One important aspect covered by governance processes is licensing: who has permission to access the data. However, data governance with respect to external processes --loosely defined as being part of an 'infrastructure'-- is typically not accounted for. This means that key data governance areas such as availability (e.g., metadata can be harvested), usability (e.g., source data can be viewed), integrity (e.g., protocols are in place to handle duplication and enrichment) and security (e.g., provenance information is maintained), need to be (re)organized or (re)considered, formalized and supported by the *Media Suite* and the emerging infrastructure in which it is embedded.

APIs -- Sustainable development

A digital infrastructure should use existing protocols, conventions, and standards. Besides obtaining data by harvesting using the OAI-PMH protocol, or using APIs, the functionalities have been organised in a modular approach, which includes (Matínez et al., 2017):

- Components: which use the API's to perform specific tasks.
- Tools: which incorporate a number of components in a tool.

Moreover, all components and tools developed in the project are open source. In addition, the Media Suite offers public APIs, which provide mechanisms for software programmers to create functionalities using API building blocks or components. We offer a Collection API, a Search API, and an Annotation API, which provides functionality for adding data annotating existing data, using the W3C Web Annotation data model (Sanderson et al., 2017).

Components/Tools -- User friendly interaction design

Developing new tools "from scratch" would be a very inefficient (and costly!) endeavour. The digital infrastructure should provide tools that are suitable both for common scholarly tasks, and for specific tasks required by each discipline. However, the digital humanities community incorporates a wide diversity of scholars with different research questions, methods, and levels of expertise in working with information processing techniques and technologies. As every infrastructure, we also have to tackle "the generalization paradox" (Zundert, 2012). We address this challenge by (i) focussing on the similarities in research methods from different disciplines (e.g., De Jong, Ordelman, Scagliola, 2011; Melgar et al., 2017), (ii) analyzing tools that support qualitative methods (Melgar & Koolen, 2018), and (ii) working with scholars as co-developers in the process[3]. The resulting functionalities are built in a modular (lego) approach that supports both flexible software development of components and user friendly interaction with assembled tools. A current challenge is to provide entity-based browsing (Verhoeven and Burrows, 2017) of both linked open data collections (RDF) and tabular data via an exploratory browser (see De Boer et al., 2017).

Enrichment services and Work Space -- Working with audio-visual content and private data

In addition to IPR and privacy restrictions, access to the audiovisual content in the *Media Suite* is also limited due to its nature; consisting of pixels (video) and samples (audio) and some manually generated metadata or subtitles (text). Typically, scholars want to search audiovisual data using (key)words that may be 'hidden' (encoded) in the pixels or the samples. This is called the semantic gap (Smeulders, 2000) that needs to be "bridged" by decoding the information in the pixels and the samples to semantic representations, e.g., a verbatim transcription of the speech or labels of visual concepts in the video (a car, a face, the Eiffel Tower), that can be matched with the keywords from the scholars. These semantic represen-

3 Indeed, an adopted strategy at the CLARIAH project level, has been to offer grants to scholars to conduct small scale research pilot projects using the CLARIAH infrastructure. In the media studies focus that we describe in this paper, almost ten scholars participate as co-developers. We follow an Agile methodology for implementation, which despite criticisms has proved to be useful for this type of projects (van Zundert, 2012)

tations can be generated manually or, especially when data collections are large, automatically using automatic speech recognition (ASR) or computer vision technology.

The generation of semantic representations is addressed in different ways. On the one hand, we are currently developing an ASR service that resides within the CLARIAH infrastructure that can handle requests from the infrastructure itself (e.g., to process a data set that exists within the infrastructure), but also request from individual scholars that want to process their private collections. On the other hand, supporting manual annotation is key for interpretation in scholarly contexts (Melgar et al., 2017). The *Media Suite* aims to support the generation of both ways of semantic representations in complementary ways via information workflows centered around a "Work Space" (see Figure 3) which stores private session data and enables collaboration.

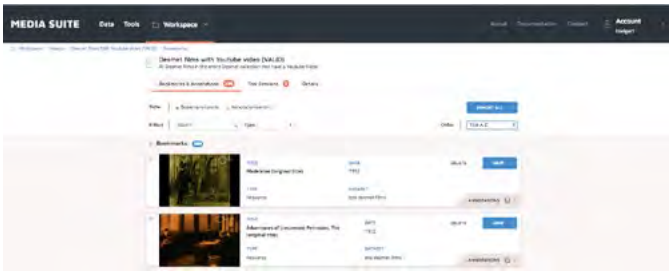


Figure 3. The CLARIAH Media Suite's Workspace

Conclusions and Future work

The paper describes the challenges found in building a sustainable, dynamic, multi-institutional infrastructure that can properly serve media scholars and digital humanists in general. We choose the approach of building a research environment that adheres to infrastructural requirements while at the same time being flexible and user-friendly. In order to ensure its used and further development after the project's lifetime, we need to carefully align the requirements of scholars with the context of the ecosystem the *Media Suite* needs to live in: an ICT infrastructure hosted and maintained by multiple institutions that in turn, adheres to a diverse set of institutional requirements with respect to, for instance, data access permissions and software development and maintenance. The *Media Suite* is currently functional and used by scholars doing actual research projects and will be developed further, e.g., by incorporating additional data sources (e.g., social media data), increasing metadata granularity (e.g., adding computer vision or emotion recognition), adding advanced annotation tools, and supporting missing data visualization (data critique) for heterogeneous datasets.

References

Blanke, T., & Hedges, M. (2013). Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems*, 29(2), 654–661.

Buchler, M., Franzini, G., Franzini, E., & Eckart, T. (2016). Mining and analysing one billion requests to linguistic services (pp. 3230–3239). Presented at the *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*.

De Boer, V., Melgar Estrada, L., Inel, O., Martínez Ortiz, C., Aroyo, L., & Oomen, J. (2017). Enriching Media Collections for Event-based Exploration. Presented at the *11th International Conference on Metadata and Semantics Research*, Tallinn, Estonia.

De Jong, F., Ordelman, R., & Scagliola, S. (2011). Audio-visual Collections and the User Needs of Scholars in the Humanities: a Case for Co-Development. Presented at the *2nd Conference on Supporting Digital Humanities (SDH 2011)*, Copenhagen, Denmark.

Edmond, J., Fischer, F., Mertens, M., & Romary, L. (2017). The DARIAH ERIC: Redefining Research Infrastructure for the Arts and Humanities in the Digital Age. *ERCIM News*, (111). Retrieved from https://hal.inria.fr/hal-01588665/document

Krauwer, S., & Hinrichs, E. (2014, May). The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *LREC-2014*, pp. 1525 - 1531.

Lauer, G. (2014). Challenges for the Humanities: Digital Infrastructures. In A. Duşa, D. Nelle, G. Stock, & G. G. Wagner (Eds.), *Facing the Future : European Research Infrastructures for the Humanities and Social Sciences*. Berlin: Scivero Verlag.

Martínez Ortiz, C., Ordelman, R., Koolen, M., Noordegraaf, J., Melgar Estrada, L., Aroyo, L., … Poell, T. (2017). From Tools to "Recipes": Building a Media Suite within the Dutch Digital Humanities Infrastructure CLARIAH. Presented at the *Digital Humanities Benelux*, Utrecht.

Melgar Estrada, L., & Koolen, M. (2017). Audiovisual media annotation using Qualitative Data Analysis Software: a comparative analysis. The Qualitative Report.

Melgar Estrada, L., Koolen, M., Huurdeman, H., & Blom, J. (2017). A process model of time-based media annotation in a scholarly context. In *CHIIR 2017: ACM SIGIR Conference on Human Information Interaction and Retrieval*. Oslo.

Odijk, J., Broeder, D., & Barbiers, S. (2015). *CLARIAH Linguistics Plan (v0.95)*. Retrieved from https://clariah.nl/werkpakketten/focusgebieden/taalkunde

Sanderson, R., Ciccarese, P., & Young, B. (Eds.). (2017). *Web Annotation Data Model: W3C Recommendation 23 February 2017*. W3C.

Smeulders, A. W. M., Worring, M. , Santini, S. ,Gupta A., and Jain, R. (2000), Content-based image retrieval at the end of the early years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, Dec 2000.

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common,

and how might our tools reflect this? Presented at the *Symposium on "Humanities Computing: formal methods, experimental practice*," London: King's College.
Van Zundert, J. (2012). If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities. *Historical Social Research / Historische Sozialforschung*, 37(3 (141)), 165–186.
Verhoeven, D., & Burrows, T. (2015). Aggregating Cultural Heritage Data for Research Use: the Humanities Networked Infrastructure (HuNI). In *Metadata and Semantics Research* (pp. 417–423). Springer, Cham.

El campo del arte en San Luis Potosí, México: 1950-2017. Análisis de Redes Sociales y Capital Social

José Antonio Motilla
jamotilla@gmail.com
Universidad Autónoma de San Luis Potosí, Mexico

La presente ponencia tiene como objetivo analizar la estructura del campo del arte en San Luis Potosí, México, de 1950 a 2017, desde el estudio del capital social de sus actores, mediante la metodología del Análisis de Redes Sociales.

Por medio de las herramientas propias de la historia oral, sociología del arte, investigación bibliográfica y documental, se ha logrado generar una compleja base de datos que nos permite contar con los elementos necesarios para comprender la dinámica que el campo del arte, con énfasis en las artes visuales, ha tenido en los últimos 67 años. La información recopilada y generada con las metodologías anteriormente citadas, ha sido examinada y sistematizada mediante herramientas propias de las humanidades digitales, y estudiada por medio del Análisis de Redes Sociales. Así, se han generado una serie de indicadores que han sido procesados para ingresarlos en un sistema de información que nos permite conocer las características del campo del arte, y el peso por centralidad que cada uno de los actores tienen dentro de la red.

Los resultados obtenidos hasta el momento nos permiten comprender las diferentes corrientes, los intereses de los artistas, la manera en que están conectados entre sí, el papel de las instituciones, el vínculo que han tenido con el exterior, el estado del mercado del arte, y fundamentalmente, para los intereses de esta ponencia, el capital social acumulado por los artistas.

La pertinencia de este trabajo reside en lo planteado por Azam y de Federico (2014), quienes sostienen que si bien las investigaciones pioneras en el campo de la sociología del arte dan un lugar central a las interacciones y relaciones entre los integrantes del campo en cuestión, las investigaciones que recurren al Análisis de Redes Sociales son prácticamente inexistentes.

Ante este panorama, la presente ponencia busca estudiar el impacto que esta perspectiva puede tener para el estudio del campo del arte y la manera en que el uso de una metodología que es resultado de la interacción de herramientas y perspectivas teóricas de diversos campos, con enfoques contemporáneos como el Análisis de Redes Sociales, nos pueden permitir acceder a explicaciones más completas y contextualizadas de lo que los estudios tradicionales, generalmente disciplinares, pueden ofrecer.

The Search for Entropy: Latin America's Contribution to Digital Art Practice

Tirtha Prasad Mukhopadhyay
tirthamukhopadhyay@gmail.com
Universidad de Guanajuato, Mexico

Reynaldo Thompson
thompson@ugto.mx
Universidad de Guanajuato, Mexico

Introduction

What, we may ask, is Latin America's contribution to global art? The answer assumes special importance in the context of the twentieth century when pioneers in the intersection of art and technology were creating new precepts, like in the works of the 'Fluxus' group in USA. If abstraction and penetration through formal stasis were acknowledged as the basic style in the art world, then pioneers in Latin America were also in pursuit of a set of most innovative possibilities for their art. We may say that Latin American artists through the sixties and seventies created such kinetic artworks that no group of artists, joined together either by contiguity or ideology, had yet achieved anywhere else in the world. What Latin American artists did were to reinvent kinetic possibilities in their most entropic and unprognosticated formats, denying subjectivity and creating ooportunities of looking at movement in art, not movement as a vector but as a function with an unknown trajectory. Our ongoing archival project on Digital Art in Latin America is especially oriented toward a perception of a digital art prototypes that evolved as a response to infiltrations of technology into Latin America.

Thematic Content

We should first emphasize on the rise of pre-electronic art in the middle of the twentieth century, which was initially represented in structural invariances of optical and sometimes pre-digital templates. Julio Le Parc's metallic illuminations, and Martinoya and Joel's Abstratoscopio

Cromático are exemplars of this new beginning with the experimental media, a new search for entropy in kinetic objects (Fariat, 2015). Leading artists like LeParc were experimenting with projection of light on reflective material. In Brazil, Abraham Palatnik tried trans-positioning colors through mechanical movements. Waldemar Cordeiro was a similar innovator with punch cards. Chileans Carlos Martinoya and Naum Joel created the Abstratoscopio Cromático, an installation which anticipates an entirely new artistic usage of polarized light effects, something the world never witnessed before (Martinoya and Joël, N. 1968). In Mexico, Manuel Felguérez, produced innovative pictorial compositions using paleo-computational programming in an age when the PC was nonexistent, and the Mexican artist Pola Weiss embraced video art. The archival project on digital heritage preservation could be an attempt to save the history of this transformation in the arts and to restore the place of Latin-American artists in the trajectory.

History of Digital Art in Latin America

The beginning may be marked from the late 1940s and early 1950s, especially in countries such as Argentina and Brazil where some of the most innovative artists began experimenting with new tools and technologies. Unfortunately, these pioneering Latin-American artists have neither been recognized nor absorbed in mainstream literature or the history of art. The initiative is specially oriented to create an appropriate perception of a digital art prototype that evolved as a response to infiltrations of technology in the Ibero-American world. Our objective is to show that Latin American artists in the new media evolved a peculiar style which manifests itself in the intelligent use of kinetic actions in an art work, to create effects which supersede categorization. They were able to explore movement and its entropic combinations. The moment could no longer be predicted, and the teleological design would lie outside subjective and interventionist approach. Perhaps Julio Le Parc or Abraham Palatnik, the agents of creative deconstruction, saw the possibility of having an art independent of intentions, one couldn't predict outcomes, an in an interation of lights and angles, or as in an interactive program with random inputs. In Le Parc's kinetic sculptures and in Palatnik's optical moments we see this first evidence of play and disruption of linear structure, later so enormously amplified in the kinetic sculptures of Mexican artists like Rafael Lozano Hemmer or Gilberto Esparza (Thompson and Mukhopadhyay, 2015). The parallels to such art installations in the north are in the K 456 of Nam Jun Paik, and the robots of Norman White, but the peculiarity of Latin American artists lie in their minimal use of technology and the potentially rich suggestivity with more formless, thermodynamic movements.

And also from Brazil, Waldemar Cordeiro was a similar innovator in the 60s who used punch card applications to manipulate images within a prehistoric computer: some of his works were recently restored by the ITAU Cultural in Brazil. In Argentina, back in the 50s Julio LeParc started experimenting with projection of light and only years later his works were shown at the Venice Biennial (1966). Also in Argentina Marta Minujin together with Wolf Vostell in Cologne, Germany and Allan Kaprow in New York, were transmitting the actions of events to artists in Cologne and New York by means of a satellite: this artwork was known as Simultaneidad en Simultaneidad (1966). In Chile, Carlos Martinoya and Naum Joel created the Abstractoscopio Cromático (1960), to anticipate an entirely new artistic object with polarized light effects, something the world had never witnessed before. In Mexico, Lorraine Pinto combined light with the music of Stockhausen in La Quinta Dimensión (1968), an artwork shown in the same year as that of the Olympic Games in Mexico City. Years later Manuel Felguérez, produced innovative pictorial compositions using paleo-computational programming in an age when the PC was nonexistent. He named his project La Máquina Estética (1975-1977). Women like Pola Weiss embraced video-art: her video Flor Cósmica was produced in 1977 and few years earlier in 1973 in Brazil, Analivia Cordeiro created the video-dance installation performance M3x3.

Geo-cultural Markers

We see that many of the artists that contributed to the development of new pathways in art and technology were ones living in other countries, especially Europe or the United States; some others came to Latin America from other regions of the planet and settled and gradually established themselves there. In Argentina for instance a fertile ground for innovation in the arts was created by the Instituto Di Tella that was bringing some of the most important and revolutionary creative minds of the world to work and exchange ideas with Argentinian artists in Argentina (Plotkin and Neiburg 2014). The Instituto Di Tella as such became during the sixties a deservedly international reference for the arts. Some of those important figures showed their work for first time at the Di Tella museum in Buenos Aires. Julio Le Parc and Martha Minujín were among them.

Archival Tools

Any archive could be created with adequate programming for a template that creates a space for storing information on these important art works. The survey for this project would have to be long drawn, with attempts of collecting information on individual works displayed in museums and galleries, and often during important events like biennials and electronic festivals of art. Simultaneously, it is necessary to obtain, wherever available, existing videos or photographs of electronics ins-

tallations and designs (Gumbrecht and Marrinan, 2003). There is no comprehensive digital catalogue of digital art. Archival efforts would have to be directed with the aim of creating a virtual space in which visitors to the archive have an opportunity to share a video or visual image of the art works that have been created as part of a tradition.

Additional Material

Links

http://www.digitalmeetsculture.net/article/digital-lat-in-america-aims-to-show-latin-potential-for-dig-ital-artistic-creation/

Websites

Archive of Digital Art
The Google Art Project / Digital meets Culture
An introduction to the booming world of Latin American digital arts

Articles

Davis, D. (1995). The work of art in the age of digital re-production (An evolving thesis: 1991-1995). *Leon-ardo*, 381-386.

Bertacchini, E., & Morando, F. (2013). The future of mu-seums in the digital age: New models for access to and use of digital collections. *International Journal of Arts Management*, 15(2), 60.

Turnbull, D., & Connell, M. (2014). Curating digital public art. In *Interactive Experience in the Digital Age* (pp. 221-241). Springer, Cham.

Books

Hilbert, M. R. (2001). *Latin America on its path into the digital age: where are we?*. United Nations Publica-tions.

References

Fariat, A. (2015). *Julio Le Parc. Critique d'art. Actualité internationale de la littérature critique sur l'art con-temporain*.

Gumbrecht, H., & Marrinan, M. (2003). *Mapping Benja-min: The work of art in the digital age*.

Martinoya, C., & Joël, N. (1968). The'Chromatic Abstrac-toscope': An Application of Polarized Light. *Leonar-do*, 1(2), 171-173.

Paul, C. (2008). New media in the white cube and be-yond: Curatorial models for digital art. *Leonardo Re-views Quarterly*, 1(2010), 33.

Plotkin, M., & Neiburg, F. (2014). Elites intelectuales y ciencias sociales en la Argentina de los años 60. El Instituto Torcuato Di Tella y la Nueva Economía. *Estudios Interdisciplinarios de America Latina y el Caribe*, 14(1).

Thompson, R., Mukhopadhyay, T. P., & Dufour, F. (2016). *The Latin American digital heritage: methods of digital art archive construction and the retrieval of immateriality. Archiving and Questioning Immateri-ality*, 204.

Ego-Networks: Building Data for Feminist Archival Recovery

Emily Christina Murphy
emurphy@uvic.ca
University of Victoria, Canada

Can data-capture be a tool for feminist historiography? Can contemporary frameworks for understanding networ-ks—actor-network theory, linked open data standards—help to shift our understanding of cultural production and literary history? This paper argues that data capture mo-delled on the principles of the "ego-network" is rich in its potential to address persistent problems in the recovery of non-canonical literary history. An ego-network is a data representation of the way that individuals are "embedded in local social structures" (Hanneman). Women's cultural production has frequently been that of secondary cultural labours like editorship. Literary scholarship has struggled with appraising this secondary labour since at least 1986, with the publication of Shari Benstock's foundational *Wo-men of the Left Bank*, in which she points out that women were frequently the "midwives of modernism," editors and caregivers performing who supported the construction singular male author. Likewise, scholars like Jack Stillin-ger (1991) have tried to break apart the "myth of solitary genius" which perpetuate the burial of these secondary labours. Despite the efforts of traditional scholarship to unsettle these myths and valorize the literary labours per-formed largely by women, the field has not yet succeeded in turning critical attention away from canonical, singular authors and towards what I call a practice of *distributed authorship*. I contend that this is a problem of methodo-logy, and that the careful creation of network graphs to represent distributed authorship may assist in correcting this persistent literary historical sticking point.

This paper emerges from the early stages of "Moder-nism, Feminism, and the Ego-Network," a postdoctoral research project with the major linked open data mode-lling project, Linked Modernisms. It concentrates on the archival collections of British activist, author, anthologist, and editor, Nancy Cunard. Cunard's archives at the Harry Ransom Center (University of Texas at Austin) represent a who's-who of literary and historical figures from the modernist period, paper remnants of a professional and personal network. Multiplicity and polyvocality are the ha-llmarks of her oeuvre, and her texts demonstrate networ-

ked connections amongst modernist writers, ideas, and events. In an echo of the scholarly trend at large, periodic attempts to recover Cunard's work and legacy (Chisholm, 1979; Marcus, 1995; Moynagh, 2002; Gordon, 2007) have not taken root, despite her deep connectedness. In DH, feminist digital scholarship has revealed the way that histories of literature and histories of DH have been obscured in the wake of canonical digital archival projects (Mandell, Earhart), and so the problems of archival recovery affect print and digital scholarship alike. This paper will present visualizations and theoretical concerns that emerge from the on-going building and modelling of a prototype of an ego-network for feminist archival recovery.

This project takes up a relatively simple example as a prototype for feminist data collection: Cunard's conventional anthology *Poems for France* (1944), published in the last years of the Second World War as a tribute to newly fallen and occupied France. The process of its publication is well represented by the archival collection, in which Cunard has meticulously preserved received correspondence in response to calls for contributions in the *Times Literary Supplement* and individual solicitations. Cunard clearly drew upon the breadth of her literary network, as letters from well-known figures like T.S. Eliot, Cecil Day-Lewis, and Vita Sackville-West show up, whether or not they ultimately contributed poems to be anthologized. However, from the point of view of conventional literary studies, this collection offers little in the way of telling examples or golden anecdotes. Cunard has little to no editorial hand over the text of the submitted poems, many of which had been published elsewhere. She rejects few contributions, and those who decline her invitation are brief and polite. Studies of editorship in the early-twentieth century have by and large looked to the depth of involvement of individual editors like Ezra Pound in the writing of authors like Hilda Doolittle or T.S. Eliot or concentrated on authors' own processes of revision (Sullivan 2013). Even in these studies of editorship, an impulse towards individual authorship persists. The work of the cultural contributor is contained in the perceivable strong hand of the individual, and collaboration is reduced to direct editorial intervention amongst canonical figures. The version of cultural contribution that Cunard undertakes in this anthology must be read differently to be read as an instance of cultural contribution.

My methodological approach to modelling this anthology has been to build a small dataset of the poems and letters relating to the collection. The dataset is currently in the form of a relational database. I begin from the position that Cunard represents the central node in an ego-network, and that the anthology can represent the immediate social structure in which she is embedded. I made a few key decisions in modelling my data. First, I have expanded the dataset's definition of *publication*. Inclusion in the published anthology and publication elsewhere are recorded as an instance of a poem's publication. In addition, a poem's inclusion in manuscript form in the Collection, whether that manuscript is received from a contributor or transcribed by Cunard is also recorded as an instance of publication. This decision aims to give similar weight to the work of solicitation and curation as it does to the instance of publication. Second, I created unique rows for each mention of a poem or of the work as a whole in the letters held in the Collection. This additional data has allowed me to sketch the shape of individual relationships across a social network that emerges in the anthology tethered to discussions of the anthology. The current dataset contains over 600 data points relating to one small anthology. This dataset is an ego-network in the sense that it take a single node in a social network—Cunard—as its tether. Rather than expanding to the whole network of modernist literary production, it is interested in the relationships between people and publications (in the expanded sense) in context of the event of the anthology.

In the current phase of the project, I am cleaning and refining my data model. I am incorporating name authority files provided to me by the HRC and maintained by standards like VIAF. I am also incorporating node type descriptions drawn from Linked Open Data ontology developed by Linked Modernisms. This phase of the project aligns with what Laura Mandell calls "guerilla coding"—in which projects that attempt to make a cultural critical interventions into technologies and standards must also make themselves legible to those same existing, problematic technologies and standards. As "Modernism, Feminism, and the Ego-Network" emerges from Linked Modernisms, it is already in conversation with the dynamics of canon creation in digital space. LiMo, already more comprehensive in its scope than many canonical DH archival projects, has made admirable attempts to redress a persistent scholarly bias against women's cultural contributions by partnering with the Orlando project to address the equitable representation of women's histories. But equitable representation is only one part of the on-going project of re-evaluating and re-narrating women's historical experiences. A feminist digital humanities approach also requires that we examine the data structures in which we perform this work. In the "guerilla coding" phase of this project, I hope that my ego-network may shift the way that major digital projects construct whole networks.

The lessons in feminist data capture and modelling that emerge from this prototype dataset are laying the groundwork for modelling data in relation to the work and communities of women's cultural production. The immediate next step in this project will be to expand the data model refined in this prototype to the study of Cunard's other works. As the data model argues that inclusion in an archives has equal weight to traditional publication, it leaves open the possibility of treating private documents like scrapbooks as a work of cultural production. This is particularly fitting for Cunard's private work: Cunard owned a printing press, and frequently mocked up her scrapbooks to look like volumes for publication, blurring the line between private and public work. Cunard, of course, is not the

only cultural producer whose study might benefit from the creation of an ego-network, and later stages of this project will experiment with building overlapping ego-networks in line with the models developed in the prototype.

References

Benstock, Shari. *Women of the Left Bank: Paris, 1900-1940*. U Texas P, 1986. Chisholm, Ann. *Nancy Cunard.* Sidgwick & Jackson, 1979.

Cunard, Nancy. Nancy Cunard Collection. Harry Ransom Center, University of Texas at Austin.

1895-1965.

---. *Poems for France*. France Libre, 1944.

Gordon, Lois. *Nancy Cunard: Heiress, Muse, Political Idealist.* Columbia UP, 2007. Hanneman, Robert A. and Mark Riddle. *Introduction to social network methods*. University of

California, Riverside, 2005. [Textbook published in digital form] faculty.ucr.edu/~hanneman/.

Mandell, Laura. "Feminist Critique vs. Feminist Production in Digital Humanities." Women's History in the Digital World Conference. Keynote. Bryn Mawr College, 2013.

Marcus, Jane. "Bonding and Bondage: Nancy Cunard and the Making of the *Negro* Anthology."

Borders, Boundaries and Frames. Ed. M.G. Henderson. Routledge, 1995.

Moynagh, Maureen, ed. *Nancy Cunard's Essays on Race and Empire*. Broadview, 2002.

Orlando: Women's Writing in the British Isles from the Beginnings to the Present. Eds. Susan Brown, Patricia Clements and Isobel Grundy. Cambridge: Cambridge University Press. 2006-2016. [Textbase. Updated semi-annually.] orlando.cambridge.org/.

Ross, Stephen, ed. *Linked Modernisms*. 2013-2016. linkedmods.uvic.ca/.

Stillinger, Jack. *Multiple Authorship and the Myth of Solitary Genius*. Oxford UP, 1991. Sullivan, Hannah. *The Work of Revision*. Harvard UP, 2013.

Searching for Concepts in Large Text Corpora: The Case of Principles in the Enlightenment

Stephen Osadetz
osadetz@fas.harvard.edu
Harvard University, United States of America

Kyle Courtney
kyle_courtney@harvard.edu
Harvard University, United States of America

Claire DeMarco
claire_demarco@harvard.edu
Harvard University, United States of America

Cole Crawford
cole_crawford@fas.harvard.edu
Harvard University, United States of America

Christine Fernsebner Eslao
eslao@fas.harvard.edu
Harvard University, United States of America

At the beginning of a research project, every scholar in the humanities asks a question: what should I read? This paper presents a new search engine that sifts through large corpora of unstructured text in order to find particular passages that deal with a concept of interest. Its underlying algorithm is based on the practice of concept search, which was originally developed in legal practice to efficiently automate the process of document review (Blair, 1985; Bai, 2005; Algee-Hewitt, 2008; Zhu, 2009; King, 2017). Our search engine builds upon that technique by applying it to large corpora relevant to humanistic scholarship and, crucially, dividing each text into passages of a standard size, improving the specificity of results. In order to demonstrate the fullest potential of this technique, our paper will focus upon its application to a specific problem in eighteenth-century intellectual history, while also discussing its most significant theoretical implications, including a reevaluation of the tight connection that has developed between the computational evaluation of the great unread through distant reading (Cohen, 1999; Moretti, 2013).

Concept search for large text corpora

Our team has expanded the paradigm of text discovery by developing a search engine that sifts through large digitized corpora to identify passages that deal with particular concepts. The algorithm itself is simple. After reading deeply in a subject, a researcher gathers together a number of passages from various sources that focus on a particular concept of interest. The researcher then uses a word-frequency analyzer to judge which of the remaining terms are most important to describing his or her concept. That keyword set is used to perform a search. The search uses a number of statistical measures to determine which items are likely to be most relevant, and produces results either as a downloadable csv spreadsheet or a GUI. The researcher can then read through the results in a preliminary fashion to judge whether they are satisfactory. If they are not, the process can be iterated.

Concept search is a statistically-based method of information retrieval that has been adopted widely in legal practice and the business world, but that is only rarely used in the humanities. Concept search and keyword search are not the same. Classical Boolean keyword search queries are predicated on single terms and phrases. Concept search, on the other hand, searches for a cluster of terms drawn from a loading set of passages identified by a researcher, which po-

tentially encode the underlying semantic quality of the passages in that set. (It should be said, this sense of "concept" is statistical and highly artefactual, quite different from the sense of the term in psychology, linguistics, or philosophy.) Keyword search simply looks for matches between query terms and the documents in a corpus. Concept search, on the other hand, does not require the occurrence of any one of its search terms, instead relying on a variety of statistical measures to judge which passages in the search corpus are most likely to be useful. Keyword search often returns non-relevant items because of the problems of synonymy and polysemy. Concept search attempts to overcome these problems, as well as that of OCR error, by representing a concept as the statistical likelihood of the occurrence of the cluster of terms in its query set.

What makes this search especially useful is that results are ordered, not by **volume**, but by which **passage** is most likely to be relevant to a particular concept. We originally developed our search engine to search through Cengage-Gale's Eighteenth Century Collections Online database. In doing so, we divided its volumes into 16 million passages, each of 1,000 words. Compare this technique to the traditional method of identifying which volumes to consult. It is akin to asking a librarian for material relevant to a research topic, and having that librarian not only identify which books are likely to be of interest, but also opening each volume to the particular page that most clearly deals with that topic.

Each set of results is essentially an index of hundreds of thousands of passages, sequentially ordered by which are most likely to be relevant. Instead of only displaying page after page of search results that must be consulted one by one, it offers researchers the option of downloading a single spreadsheet that is easy to filter by author, work, and date, and that allows for a quick, global view of which texts are relevant. In order to make sense of this data, one has to sort and filter judiciously. We have developed a number of standard search filters that can be used repeatedly to select for literary works, canonical works across disciplines, and an author's gender. Or, should a researcher prefer, she can quickly select for one, two, or fifty authors that she would like to examine. Sorting is equally important. The most basic statistical measure is term frequency, which counts each time a keyword appears in a particular passage. The search engine also allows for sorting by other statistical measures, including the proportion of keywords (useful if there is a great deal of error in a passage due to the process of digitization, or if there is a relatively high number of stopwords), tf–idf.

Searching for eighteenth-century principles and theoretical implications

In order to demonstrate the fullest potential of this technique, we present a use case that concerns a specific problem in intellectual history. In the eighteenth century, many of the most famous authors obsessed over the possibility of encapsulating a whole book, or even an entire discipline, in a single proposition called a principle. Among these are the best-known statements of the Enlightenment, including Newton's inverse-square law of gravity and Kant's categorical imperative. David Hume put it succinctly. A principle, he said, offers "a whole science in a single theorem" (Hume, 1987). By promising to encapsulate and disseminate an author's most fundamental ideas, the principle became the preeminent intellectual device of the Enlightenment. In some cases, however, less famous principles might lie buried in books, some obscure. The most common research tools and methods – long reading, critical bibliographies, and Library of Congress subject headings – would be of only limited help in discovering these. Our search engine provides an efficient means of discovering passages in which authors frame principles and reflect upon the consequences of this rhetorical obsession.

In accounting for the eighteenth-century enthusiasm for principles, one important question that has been difficult to answer using traditional research methods is what women thought of this rhetorical habit. The principle is undeniably masculinist (etymologically, the word itself is tied to seeds and semen), and the prevailing assumption in the period was that men framed principles, so much so that Rousseau claimed in his pedagogical treatise Émile that women lacked the ability to generalize their ideas. By applying standard filters to our search results, we are able to efficiently identify important passages in works by female authors, in which they chafed against the pervasive insistence that women were shut out from the culture of the principle.

The final section of our paper concerns the theoretical implications of the tool we have developed. We interrogate the common denigration of search that many digital humanists have voiced (Moretti, 2007; Berry, 2012; Jockers, 2013), while also questioning the tight connection that has developed between the critical concepts of "the great unread" and "distant reading" (Moretti, 2000; Sculley and Pasanek, 2008; Trumpener, 2009). Our position is that many of the strongest examples of digital scholarship treat the great unread as a textual noumenon that can only be approached obliquely and in its totality, through the analysis of minimal textual features. The search engine we have developed allows computational methods to invigorate more traditional modes of reading by helping researchers to quickly draw together a project-specific list of works that comprise canonical and non-canoni-

cal material. As such, it promises to open new avenues for research, both for those scholars committed to historicist methods, those exploring alternatives to critical modes of reading (Best and Marcus, 2009; Shore, 2010; Pasanek, 2015), and those who wish to rethink scholar's pervasive recourse to context, in an effort to trace how ideas move across history (Felski, 2015)

References

Algee-Hewitt, M. (2017).*The Afterlife of the Sublime: Toward a New History of Aesthetics in the Long Eighteenth Century*. ProQuest Dissertations and Theses.

Bai, J., Song, D., Bruza, P., Nie, J-Y., and Cao, G. (2005). "Query Expansion Using Term Relationships in Language Models for Information Retrieval."*Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 688-95.

Berry, D. M. (2012).*Understanding Digital Humanities*. Palgrave Macmillan.

Best, S. and Marcus, S. (2009). "Surface Reading: An Introduction." *Representations*, 1-21.

Blair, D., and Maron, M. (1985). "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System" *Communications of the ACM*: 289-99.

Cohen, M. (1999).*The Sentimental Education of the Novel*. Princeton University Press.

De Bolla, P. (2013).*The Architecture of Concepts: The Historical Formation of Human Rights*. Fordham University Press.

Felski, R. (2015).*The Limits of Critique*. University of Chicago Press.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009).*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hume, D. (1987). Miller, Eugene F., ed.*Essays, Moral, Political, and Literary*. Rev. ed. Liberty Fund, 1987.

Jockers, M. L. (2013).*Macroanalysis :Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.

King, G., Lam, P., and Roberts, M. E. (2017). "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science*: 289-99.

King, G., Pan, J., and Roberts, M. E. (2017). "Reverse-engineering Censorship in China: Randomized Experimentation and Participant Observation."*Science*. 345.6199: 1251722.

Moretti, F. (2000). "Conjectures on World Literature."*New Left Review*1.54.

Moretti, F.(2013). *Distant Reading*. Verso.

Moretti, F. (2009). "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)."*Critical Inquiry*.36.1: 134-58.

Pasanek, B. (2015).*Metaphors of Mind: An Eighteenth-Century Dictionary*. Johns Hopkins University Press.

Rousseau, J.-J. Bloom, A., ed. (1979).*Emile : Or, On Education*. Basic Books.

Sculley, D., and Pasanek. B. M. (2008). "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities."*Literary and Linguistic Computing*.23.4: 409-24.

Shore, D. (2010). "WWJD? The Genealogy of a Syntactic Form."*Critical Inquiry*.37.1: 1-25.

Trumpener, K. (2009). "Critical Response I. Paratext and Genre System: A Response to Franco Moretti."*Critical Inquiry*.36.1: 159-71.

Zhu, X., and Goldberg, A. B. (2009). "Introduction to Semi-Supervised Learning."*Synthesis Lectures on Artificial Intelligence and Machine Learning*.3:1: 1-130.

Achieving Machine-Readable Mayan Text via Unicode: Blending "Old World" script-encoding with novel digital approaches

Carlos Pallan Gayol
pallan.carlos@gmail.com
University of Bonn, Germany

Deborah Anderson
dwanders@sonic.net
University of California at Berkeley, United States of America

Introduction

In 2015, our team began work to get the Mayan hieroglyphs into the international standard Unicode, so Mayan text can be reliably interchanged on computers and other devices. Thanks to collaboration with standards experts and recent advances in computer technology and Mayan decipherment, the work to encode Mayan in Unicode has progressed significantly from the state we reported at DH2016 [1]. This paper will describe the challenges that prevented scholars from encoding Mayan in the past, and the strategies we used to overcome these hurdles. We will also give examples of the rapidly expanding repository of digitally encoded, machine-readable Mayan texts, and report on the implications for future research.

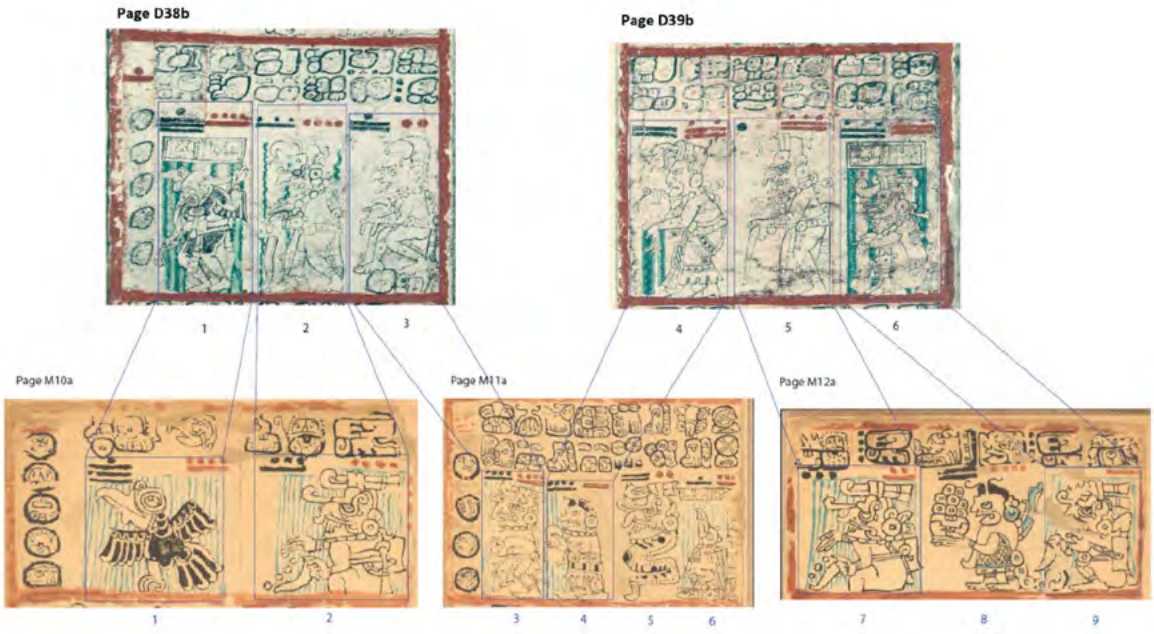


Figure 1. The first two Mayan texts selected for encoding in the Unicode format: parallel "cognate" almanacs on the Dresden (D38b-D41b) and Madrid Codices (M10b-M13a). Cf. Bricker and Bricker 1988; Aveni et al. 1996: Fig. 10; Aveni 2004:158-159

Past Problems and Current Approaches

Early attempts to apply computational approaches to Mayan decipherment from the late 1960s and 1970s proved premature, given the technological limitations of the time [4; 8]. In addition, the stage of decipherment then allowed only a fraction of Mayan syllabic and logographic signs to be read with any certainty (probably less than 30%), compared to what is possible today. Subsequently, a number of Mayan database projects attempted to cover the full corpus of Mayan texts [2; 11; 15], encompassing thousands of inscriptions and spanning almost 2,000 years across multiple regions. Given the wide range of scribal practices across such a broad spectrum of space and time, it has proven difficult to identify a core set of characters.

Our project, on the other hand, decided to focus on the extant Mayan codices, three of which are preserved in libraries in Dresden, Madrid, and Paris (see figure 1). These Codices are widely accepted to originate in Pre-Columbian Yucatan, Mexico, during the late Postclassic period (ca. AD 1250-1519). This strategy proved fruitful, as these ancient documents attest to much greater consistency in the range of scribal practices, and make use of a relatively limited, time-specific, common repertoire of signs. Ultimately, inscriptions and earlier Classic-period texts can be added, building upon the base repertoire of the Codices.

Format

Because most database projects operated under largely non-standardized formats, Mayan textual data could not be widely shared, but was limited to those institutions that shared the same (non-standard) formats. In contrast, we centered our efforts on getting the script into the international character encoding standard Unicode. The advantages of Unicode include:

- wider accessibility, since Unicode is an open-source standard that is freely available to all users and developers. This would facilitate scholars and humanities students to contribute to improving and expanding the ongoing textual repository of encoded, machine-readable Mayan hieroglyphic texts.
- reliable communication, ensuring the recipient of a text will receive the text as originally sent.
- enhanced searching and querying capabilities, as well as advanced text mining, in ways that are not possible with annotated collections of pictures, drawings, or scans of ancient inscriptions.
- greater compatibility and discoverability with/through existing online resources, such as graphemica.com, and other aids.
- Improved long-term archiving and preservation of textual records, given the stability of the standard and the ease of depositing multiple copies.

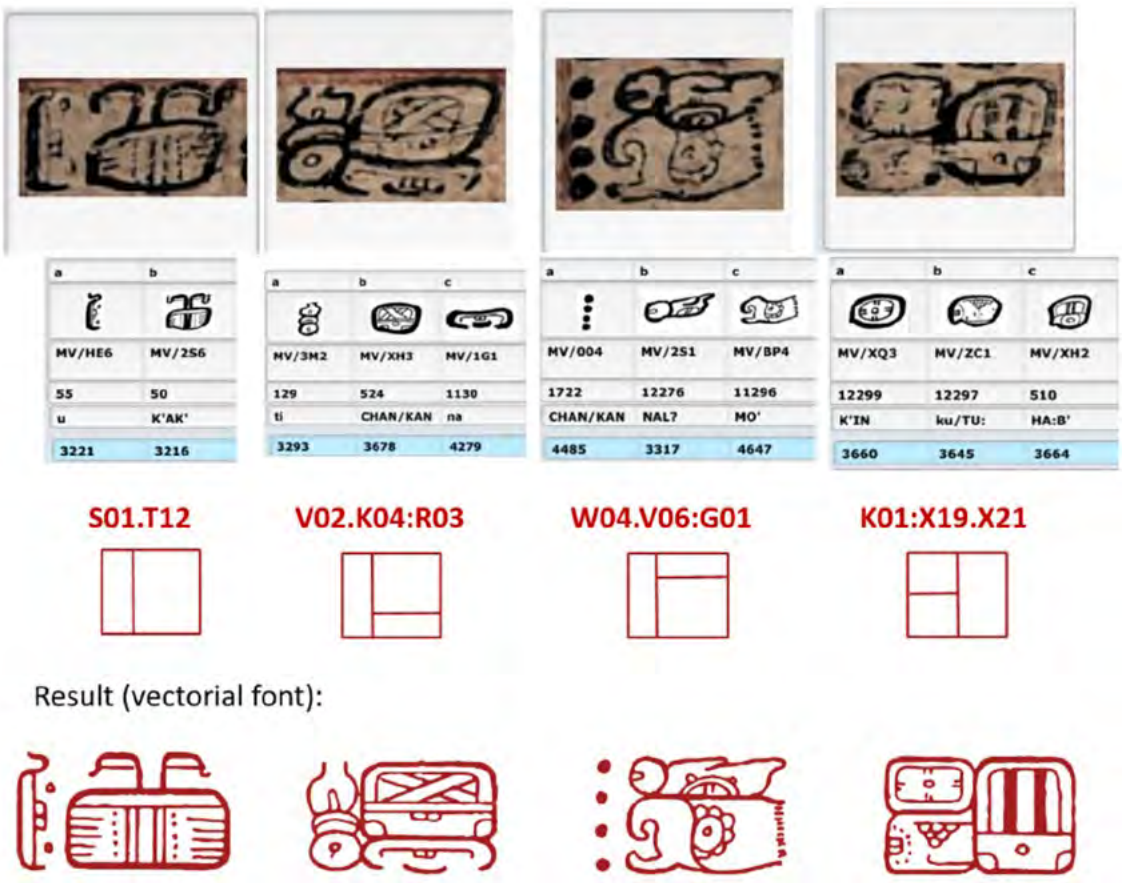


Figure 2. Examples of our workflow relying on glyph-block cluster arrangements or "quadrats" for rendering complex Mayan signs from the Dresden Codex.

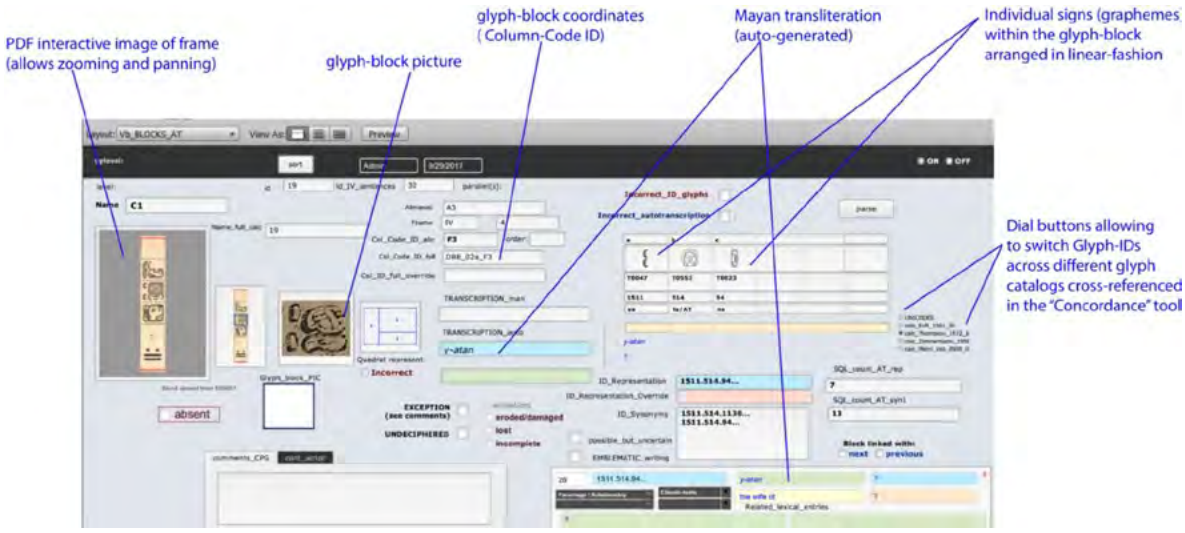
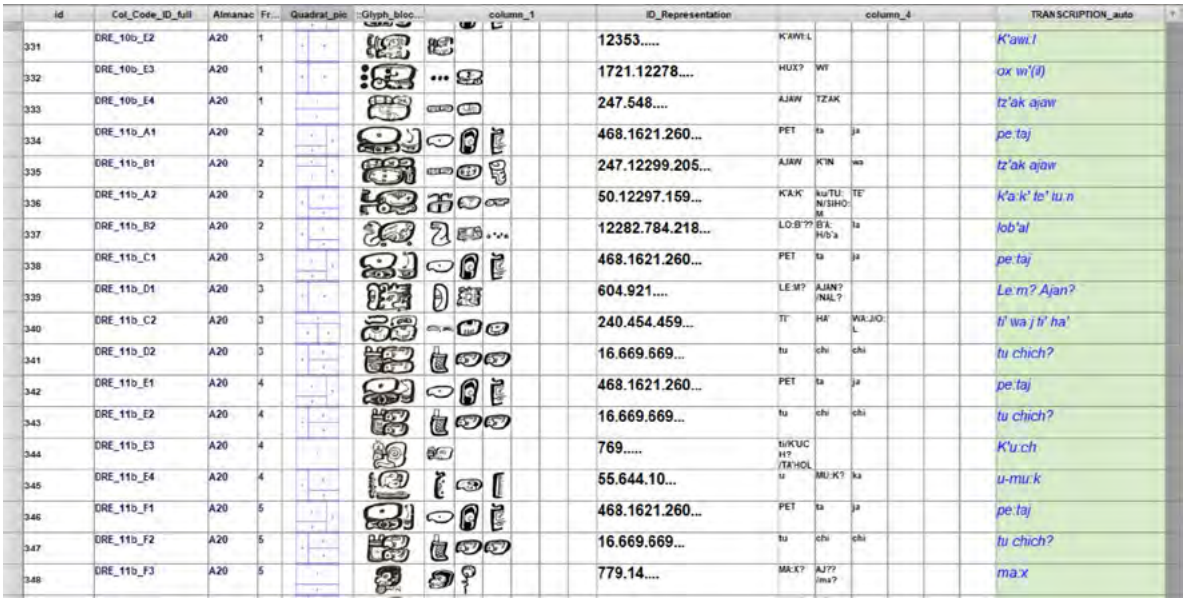


Figure 3. Database screenshot from our new semi-automated Mayan transliteration and translation functionality

Technical Issues and Solutions

The degree of visual complexity prevalent in the Mayan script has posed unique challenges, both in describing the data in a standardized way, and being able to accurately render signs with standard font protocols. This complexity includes ligation of signs, horizontal and vertical joining, truncation, and infixation (figure 2), as well as identification of cluster arrangements ("quadrats") and the placement of signs in the clusters. To our knowledge, no other projects beside ours have focused on identifying all attested cluster arrangements in the codices consistently under the Unicode Standard [12] while describing sign-positioning in ways supported by new font and software upgrades.

To overcome the technical challenges involved, we are working directly with Unicode specialists and implementers, whose experience in encoding other writing systems has informed our methodology, specifically in describing the cluster arrangements and the database structure. Thus, our description of the cluster arrangements drew from work on Egyptian hieroglyph quadrats by Andrew Glass [3]. Based on input from collaborators, the database has been set up so it can generate real-time semi-automated transliterations and translations of Mayan hieroglyphs into English and Spanish (figs. 3-4). This system is also capable of breaking down visually complex glyph-block cluster arrangements ("quadrats") into their constituent individual signs and displaying them in linear fashion.



id	Col_Code_ID_full	Almanac	Fr...	Quadrat_pic	Glyph_bloc...	column_1	ID_Representation	column_4	TRANSCRIPTION_auto
331	DRE_10b_E2	A20	1				12353.....	K'AWI:L	K'awi:l
332	DRE_10b_E3	A20	1				1721.12278....	HUX? WI	ox wi'(il)
333	DRE_10b_E4	A20	1				247.548....	AJAW TZ'AK	tz'ak ajaw
334	DRE_11b_A1	A20	2				468.1621.260...	PET ta ja	pe:taj
335	DRE_11b_B1	A20	2				247.12299.205...	AJAW K'IN wa	tz'ak ajaw
336	DRE_11b_A2	A20	2				50.12297.159...	K'A:K' ku/TU:N/SIHO:M TE'	k'a:k' te' tu.n
337	DRE_11b_B2	A20	2				12282.784.218...	LO:B'?? B'A:Hb'a la	lob'al
338	DRE_11b_C1	A20	3				468.1621.260...	PET ta ja	pe:taj
339	DRE_11b_D1	A20	3				604.921....	LE:M? AJAN? /NAL?	Le:m? Ajan?
340	DRE_11b_C2	A20	3				240.454.459...	TI' HA' WA:JO:L	ti' wa j ti' ha'
341	DRE_11b_D2	A20	3				16.669.669...	tu chi chi	tu chich?
342	DRE_11b_E1	A20	4				468.1621.260...	PET ta ja	pe:taj
343	DRE_11b_E2	A20	4				16.669.669...	tu chi chi	tu chich?
344	DRE_11b_E3	A20	4				769.....	ti/K'UCH? /TA'HOL	K'u:ch
345	DRE_11b_E4	A20	4				55.644.10...	u MU:K'? ka	u-mu:k
346	DRE_11b_F1	A20	5				468.1621.260...	PET ta ja	pe:taj
347	DRE_11b_F2	A20	5				16.669.669...	tu chi chi	tu chich?
348	DRE_11b_F3	A20	5				779.14....	MA:X? AJ?? /ma?	ma:x

Figure 4. Database screenshot from our new semi-automated Mayan transliteration and translation functionality (multi-record table view).

Research Results

Our analysis of the main textual contents of the Mayan Codices has enabled us to attain a full analysis of all the extant hieroglyphic inscriptions in the Mayan Codices (Table 1). It also resulted in the identification of the full range of permutations by which individual signs (graphemes) conform into glyph-block arrangements or specific cluster-configurations (i.e. "quadrats). These quadrats unfold into 167 different types [12], which we ordered into classes, ranging from one up to six signs per glyph-block (Figs. 3-4). We have also developed a "mapping engine" able to segment the Mayan Codices into a meaningful, hierarchical arrangement of their constitutent levels and segments. This tool can help to identify the underlying thematic composition and structure of the Codices and other complex texts at different levels (i.e. glyph-block, phrases, frames, almanacs, pages, sections/chapters, volume), in much the same way as scholars of Western literary tradition have been able to identify medieval text structures [14]. For instance, this engine allows for laying out the structure of the Dresden Codex as a document composed of 74 pages, plus four blank pages, arranged into 22 sections, 96 almanacs and tables, and 575 frames. A key component of our efforts has been the creation of a new Unicode-based "glyphary" tool, a comprehensive digital catalogue of characters (graphemes) exclusively occurring in the Codices, which substantially departs from previous efforts [4;11;16;19] by its underlying methodology and novel taxonomy. In this methodology alphanumeric codes are tied together to code points assigned by the Unicode Standard. In developing this glyphary tool, we relied on previous collaborations between one of us (Pallan) and teams at IDIAP and UniGe (Switzerland) for developing digital multimedia resources and machine vision algorithms suitable for Mayan [6;7].

Implications of Research

Based on the above results, our paper provides a critical look at the implications for Mayan scholarship and the humanities, including the degree to which the codical sign-set compares to the earlier sign-sets from the Classic-period monumental inscriptions. We also explore specific idiosyncrasies and the global patterns that can be identified within codical texts and datasets, partly by programming highly specific SQL (Standard Query Language) queries for addressing these and other culturally significant questions. For example, which lexical terms occur with greater frequency, to which semantic and grammatical categories do these terms belong? Which major languages are represented and what is the affiliation of the lexical terms attested [9; 17;18]? We also develop indicators that permit approaching complex scribal features and practices within the codices, such as the degree of phoneticism and the ratio of individual signs per glyph-block (see Table 1 below). Our system also allows precise mapping of undeciphered and problematic signs and identifies the contexts in which they occur.

Codex:	DRESDEN	MADRID	PARIS
Number of extant pages	74 pages	112 pages	24 pages
Number of almanacs	96 (75 almanacs + 21 tables)	237 (almanacs & tables)	18 (almanacs & tables)
Number of frames	575 / 1659 total	889 / 1659 total	192 / 1659 total
Number of glyph-blocks	2951 / 7122 41.43%	3340 / 7122 total 46.89%	831 / 7122 total 11.66%
Blocks per frame ratio	5.13 blocks per frame	3.75 blocks per frame	4.32 blocks per frame
Number of graphemes (main-text signs, not counting calendric grid)	7208 / 17147 total 42.03%	7913 / 17147 total 46.14%	2026 / 17147 total 11.81%
Signs per glyph-block ratio:	2.442 signs per glyph-block	2.369 signs per glyph-block	2.438 signs per glyph-block

Table 1. Comparative statistics derived from analysis of three extant Mayan Codices

Future Work and Goals

Plans for future work include further development of advanced OpenType Mayan font —in close collaboration with Andrew Glass— providing more accurate rendering of linear signs into glyph-blocks. We are also planning to expand our resources into the realm of the monumental stone inscriptions, in collaboration with Dr. Gabrielle Vail and other researchers, with the aim of generating robust, representative new datasets of texts from all the major Classic and Terminal Classic scribal traditions. thereby substantially increasing the range of chronological/regional variability of our textual repositories.

To facilitate collaborative team data editing and analysis, we are currently migrating some of our core database functionalities into a MySQL-based server that offers greater compatibility with widely used open source solutions such as SQLite and MariaDB. On the longer term, we are collaborating with the READ (*Research Environment for Ancient Documents*) co-creators Andew Glass and Stephen White [13] to adapt and expand this powerful engine into a "Mayan-READ". This tool would provide scholars and humanities students with the full-range of our open-access online resources, allowing them greater access and interactivity with our datasets, plus the ability to contribute in expanding a Unicode-based repository of digitally encoded, machine-readable Mayan hieroglyphic texts. In so doing, we are also seeking to establish innovative collaborations with cultural institutions and research groups (such as INAH in Mexico). Part of this effort involves organizing workshops, where our workflow and methodologies can be learned by other teams in Mexico and other locations, and ultimately put into practice in ways that can have greater impact to benefit the humanities research community as a whole.

References

[1] Anderson, Deborah and Carlos Pallan (2016). "Unlocking the Mayan Hieroglyphic Script with Unicode." *Presentation at DH2016*, Krakow, Poland.

[2] CODICES Project: IDIAP Research Institute (Switzerland) and INAH (Mexico). Home page at URL: https://www.idiap.ch/project/codices

[3] Glass, Andrew. (2016). "Preliminary analysis of Egyptian Hieroglyph quadrat types." URL: http://www.unicode.org/L2/L2016/16232-quadrat-types.pdf

[4] E.B. Evreinov, Y. Kosarev, and B.A. Ustinov (1961). *The Application of Electronic Computers in Research of the Ancient Maya Writing*. USSR, Novosibirsk.

[5] Gates, William E. (1931) *An Outline Dictionary of Maya Glyphs: With a Concordance and Analysis of Their Relationships.* Baltimore: Johns Hopkins Press

[6] D. Gatica-Perez, G. Can, R. Hu, S. Marchand-Maillet, J.-M. Odobez, C. Pallan Gayol, and E. Roman-Rangel MAAYA

(2017) Multimedia Methods to Support Maya Epigraphic Analysis. In Diego Jimenez-Badillo (Ed.) Arqueologia computacional Nuevos enfoques para el analisis y la difusion del patrimonio cultural. INAH-RedTD-PC, in press. Available online at IDIAP: http://publications.idiap.ch/downloads/papers/2017/Gatica-Perez_INAH-REDTDPC_2017.pdf

[7] R. Hu, G. Can, C. Pallán Gayol, G. Krempel, J. Spotak, G. Vail, S. Marchand-Maillet, J.-M. Odobez and D. Gatica-Perez (2015) Multimedia Analysis and Access of Ancient Maya Epigraphy. In *IEEE Signal Processing Magazine*, Special Issue on Signal Processing for Art Investigation, Vol. 32, No. 4, pp. 75-84, Jul. 2015, Available online at IDIAP: http://publications.idiap.ch/index.php/publications/show/3629

[8] Dell H. Hymes [ed], Wenner-Gren(1965) Section about Morris Swadesch In: *The use of computers in anthropology*. Foundation for Anthropological Research, (Studies in General Anthropology, 11.) London, The Hague, Paris: Mouton & Co., pages 524-525

[9] Law, Daniel A. (2014) *Language Contact, Inherited Similarity and Social Difference: The Story of Linguistic Interaction in the Maya Lowlands*, *Current Issues in Linguistic Theory,* Vol. 328. John Benjamins Publishing Company, Amsterdam. 206 pages.

[10] M. J. Macri and G. Vail (2009) *The* New Catalog of Maya Hieroglyphs, vol II, the codical texts. University of Oklahoma Press.

[11] Maya Hieroglyphic Database Project (MHD) home page at URL: http://mayadatabase.faculty.ucdavis.edu/database/

[12] Pallan Gayol, Carlos (2018) "A Preliminary Proposal for Encoding Mayan Hieroglyphic Text in Unicode" (v2), proposal submitted to the Unicode Technical Committe meeting, Google Inc. Mountain View, CA, January 22, 2018. URL: http://www.unicode.org/L2/L2018/18038-mayan.pdf

[13] Research Environment for Ancient Documents (READ) enabled for Gāndhārī Language and Literature. By Stefan Baums and Andrew Glass. URL: https://gandhari.org/blog/?p=251

[14] C. M. Sperberg-Mcqueen (1991). Text in the Electronic Age: Texual Study and Textual Study and Text Encoding, with Examples from Medieval Texts. *Literary and Linguistic Computing*, Volume 6, Issue 1, 1 January 1991, Pages 34–46,https://doi.org/10.1093/llc/6.1.34

[15] Textdatenbank und Wörterbuch des Klassischen Maya (TWKM) http://mayawoerterbuch.de/

[16] J. E. S. Thompson (1962) *A catalog of Maya Hieroglyphs*. University of Oklahoma Press,

[17] Wald, Robert 2004 The Languages of the Dresden Codex: Legacy of the Classic Maya. In The Linguistics of Maya Writing, edited by Søren Wichmann, pp. 27-58. University of Utah Press, Salt Lake City.

[18] Wichmann, Søren and Albert Davletshin. 2006. Writing with an accent: phonology as a marker of ethnic identity. In Sachse, Frauke (ed.), Maya Ethnicity: The Construction of Ethnic Identity from the Preclassic to Modern Times, pp. 99-106. Markt Schwaben: Verlag Anton Saurwein.

[19] G. Zimmermann (1956) Die *Hieroglyphen der Maya Handschriften. Abhandlungenaus dem Gebiet der Auslandskunde*. Band 62- Reihe B, Universit¨at Hamburg. Cram, De Gruyter & Co.,

Whose Signal Is It Anyway? A Case Study on Musil for Short Texts in Authorship Attribution

Simone Rebora
simone.rebora@univr.it
University of Verona, Italy

J. Berenike Herrmann
berenike.herrmann@unibas.ch
University of Basel, Switzerland

Gerhard Lauer
gerhard.lauer@unibas.ch
University of Basel, Switzerland

Massimo Salgaro
massimo.salgaro@univr.it
University of Verona, Italy

State of the art and experimental design

Robert Musil, one of the most important authors of twentieth-century German-written literature, fought in the Austrian army at the Italian front. During WWI, between 1916 and 1917, Musil was chief editor of the *Tiroler Soldaten-Zeitung* (TSZ) in Bozen. This activity has posed a philological problem to Musil scholars, who have not been able to attribute with certainty a range of texts to the author. However, solving the riddle of authorship for this particular set of texts promises a great advancement in the study of Musil's political thinking. With this paper, we present a new approach that combines historical and philological research with stylometric methods.

The determination of possible authorship starts with reviewing the literature for previous attempts. There are 38 articles in the TSZ for which Musil's authorship has been proposed at least once (see Table 1).

Text #	Title	Date of publication	Attributed by
Excl_1	Der Weg zu den Sternen	08.07.1916	C, FL
Excl_2	Aus der Geschichte eines Regiments	26.07.1916	C, FL
1	Kameraden arbeitet mit!	06.08.1916	A, FL
2	Bin ich ein Österreicher?	20.08.1916	A, FL
3	Herr Tüchtig und Herr Wichtig	27.08.1916	C, FL
4	Das Schlagwort	27.08.1916	A, FL
5	Die Erziehung zum Staat	03.09.1916	A, FL

6	Bauernleben	01.10.1916	C
Excl_3	Kunst hinter der Front	08.10.1916	C
7	Sonderbare Patrioten	15.10.1916	A, FL
8	Noch einmal Bauernleben	29.10.1916	C
9	Opportunität	12.11.1916	FL
Excl_4	Kannst Du deutsch [III]	12.11.1916	A, FL
10	Eine gute persönliche Beziehung	26.11.1916	A, FL
11	Eine österreichische Kultur	10.12.1916	R, A, FL
12	Der Nörgler und der neue Österreicher	17.12.1916	A, FL
13	Das Kompromiß	24.12.1916	A, FL
Excl_5	Der Augenzeuge	24.12.1916	C
14	Heilige Zeit	31.12.1916	A, FL
15	Zentralismus und Föderalismus	07.01.1917	FL
16	Föderalismus oder Zentralismus	14.01.1917	FL
Excl_6	Kannst Du Deutsch [V]	21.01.1917	A, FL
Excl_7	Vorpolitische Reinigung	04.02.1917	A, FL
Excl_8	Kannst Du Deutsch [VI]	04.02.1917	A, FL
17	Zu Milde und zu Wilde	11.02.1917	A, FL
Excl_9	Aus einer öffentlichen Schwulstfabrik	18.02.1917	A, FL
Excl_10	Schnucki in der „großen Zeit“	18.02.1917	A, FL
18	Neu-Altösterreichisches	25.02.1917	A, FL
19	Ist die »österreichische Frage« schwierig?	04.03.1917	FL
20	Seiner Hochwohlgeboren!	04.03.1917	D, A, FL
21	Luxussteuern	04.03.1917	A, FL
22	Positive Ziele	11.03.1917	FL
23	Der Frieden versprochen!	18.03.1917	FL
24	Das Staatsprogramm der Deutschen	18.03.1917	A, FL
25	Wehe dem Staatsmann!	25.03.1917	FL
26	Der Frieden und die Zukunft	01.04.1917	FL
27	Presse und Krieg	08.04.1917	FL
28	Vermächtnis	15.04.1917	D,R,C,A,FL

Table 1. TSZ articles attributed to Musil, derived from (Schaunig, 2014). D = (Dinklage, 1960); R = (Roth, 1972); C = (Corino, 1973, 2003, and 2010); A = (Arntzen, 1980); FL = (Fontanari and Libardi, 1987).

The major problem for carrying out a stylometric analysis on the texts published in the TSZ is their shortness. As demonstrated by (Eder, 2015), the minimum length for a reliable authorship attribution is around 5,000 words. However, the average length of the 38 disputed TSZ articles is slightly below 1,000 words (see Figure 1).

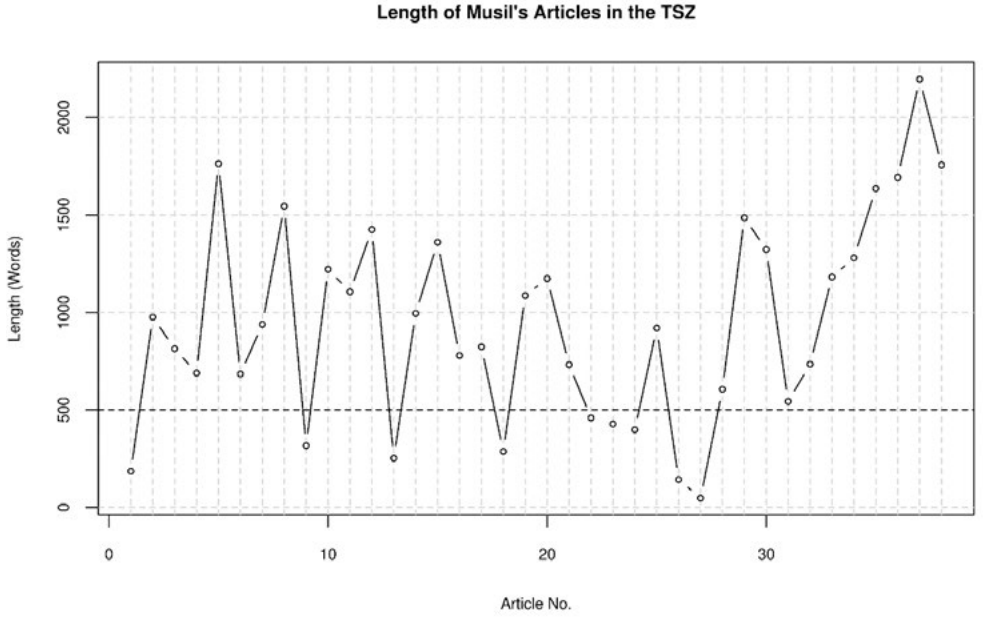


Figure 1. TSZ articles' lengths

As a possible solution for this issue, we developed a combinatory design that analyzes longer chunks composed by the juxtaposition of single texts. To reduce the number of combinations, we excluded the 9 shortest texts (below 500 words), together with the only text (Excl_2 in Table 1) that has been solidly attributed to Musil on philological grounds (Corino, 1973). This leaves us with a corpus of 28 texts, already digitized by (Amann et al., 2009). The optimal configuration was obtained by combining groups of 6 texts. This permutation generated 376,740 text chunks with an average length of N=6,963 words and a standard deviation of 909 words.

As for the composition of the training set, we combined the stylometric "impostors method" (Koppel and Winter, 2014) with historiographical research. Following (Juola, 2015), we thus fixed the number of "impostors" to a minimum of three, identifying as likely candidates Franz Blei, Franz Kafka, and Stefan Zweig. In addition, we selected three possible TSZ collaborators: Marie delle Grazie, Hugo Salus, and Albert Ritter (cf. Urbaner, 2001). While all others were digitally available, we manually retrodigitized Ritter's texts. The training set was completed by a selection of articles authored by Musil in various journals between 1911 and 1919. For each author, the retrieved material was subdivided in three text chunks (length ranging 6,000–8,000 words): the training set was thus composed of 21 text chunks.

The Experiment

Validation and experimentation were carried out using the R package *Stylo* (see Eder et al., 2016). A 20-fold stratified validation had the following results: (1) distance measures (with the exception of Cosine) work slightly better than machine learning algorithms; (2) word-based analysis outperforms 10-character n-grams (cf. Halvani et al., 2016: 39); (3) Fig. 2 shows that accuracy levels fluctuate substantially between 10 and 400 MFWs.

	Mean accuracy	(with 10-char. n-grams)
Delta	99.16	98.96
Eder	98.58	98.57
Canberra	99.37	99.24
Cosine	95.03	95.40
Cosine Delta	98.90	98.79
SVM	98.56	98.46
k-NN	95.28	94.95
NSC	95.55	95.34

Table 2. Stratified validation results

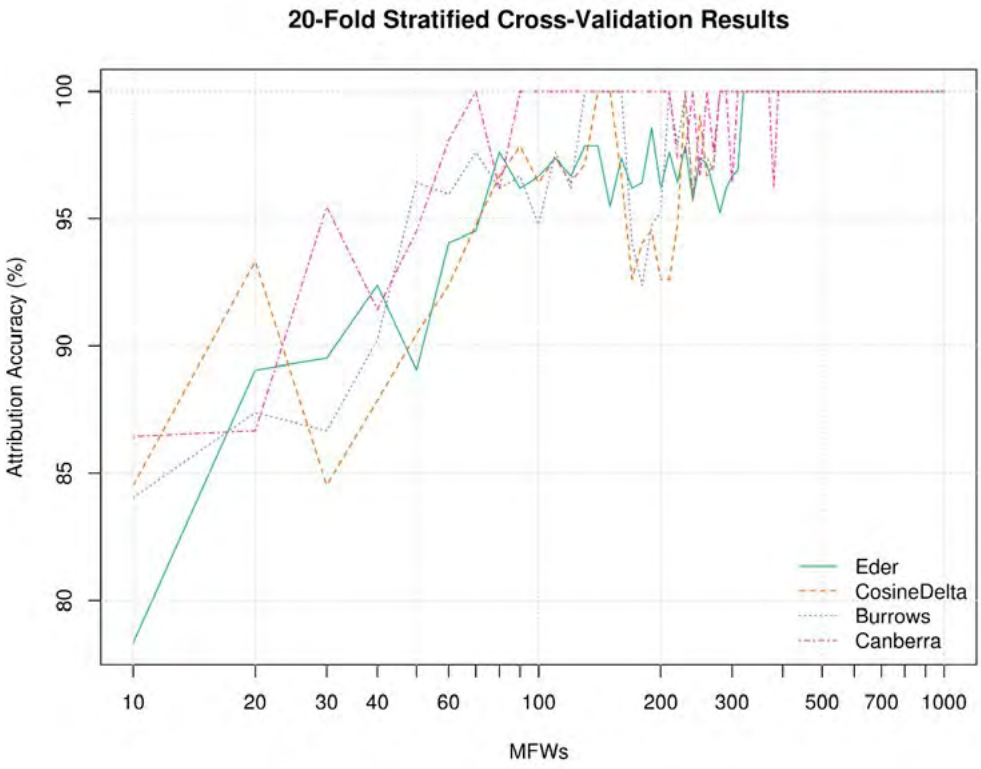
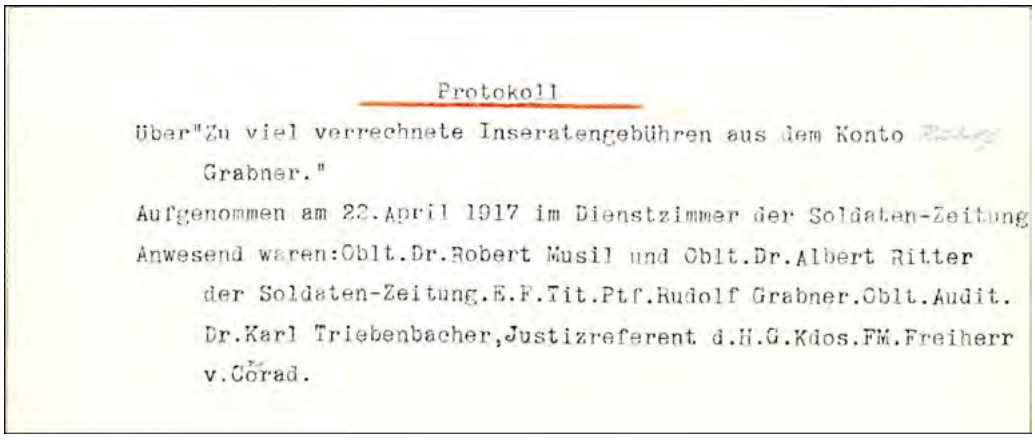


Figure 2. Stratified validation results

For these reasons, we limited feature selection to altogether 16 combinations of: (1) the distance measures Burrows's Delta, Eder's Delta, Cosine Delta, and Canberra; (2) the frequency strata 10–100 MFWs, 20–200 MFWs, 50–500 MFWs, and 100–1,000 MFWs.

For each iteration, the distances between test set and training set were saved into a matrix. At the end of the process, mean values were calculated. In all 16 configurations, Ritter and Musil are the only candidates for authorship of the TSZ articles. This evidence has been corroborated by the discovery of a document in the Kriegsarchiv in Wien, which confirms that Ritter was part of the TSZ editorial team (see Fig. 3).



Protokoll

Über"Zu viel verrechnete Inseratengebühren aus dem Konto
Grabner."

Aufgenommen am 22.April 1917 im Dienstzimmer der Soldaten-Zeitung

Anwesend waren:Oblt.Dr.Robert Musil und Oblt.Dr.Albert Ritter der Soldaten-Zeitung.E.F.Tit.Ptf.Rudolf Grabner.Oblt.Audit. Dr.Karl Triebenbacher,Justizreferent d.H.G.Kdos.FM.Freiherr v.Conrad.

Figure 3. Ritter in the TSZ team. Source: Kriegsarchiv, Wien

The stylometric results are synthetized by Fig. 4, which represents only the distances between Musil's and Ritter's signals. For highlighting the distinctions, measures were normalized to a range between –1 and +1. A general trend is evident: while, for the articles published in 1916 (articles 1–14), figures point quite clearly to Musil's authorship, the picture is less clear for the articles published in 1917 (articles 15–28). In no case, however, Ritter's signal is clearly dominant. Musil thus appears as the most likely author, with the following caveats: First, the combinatory design, while having shown the dominance of Musil's signal, may have suppressed different, minor signals. Second, Musil, in the role of chief editor, may have altered many articles in the journal, thus intermixing his authorial signal with those of others. By consequence, results that question Musil's authorship are as a tendency more substantial than those that corroborate it.

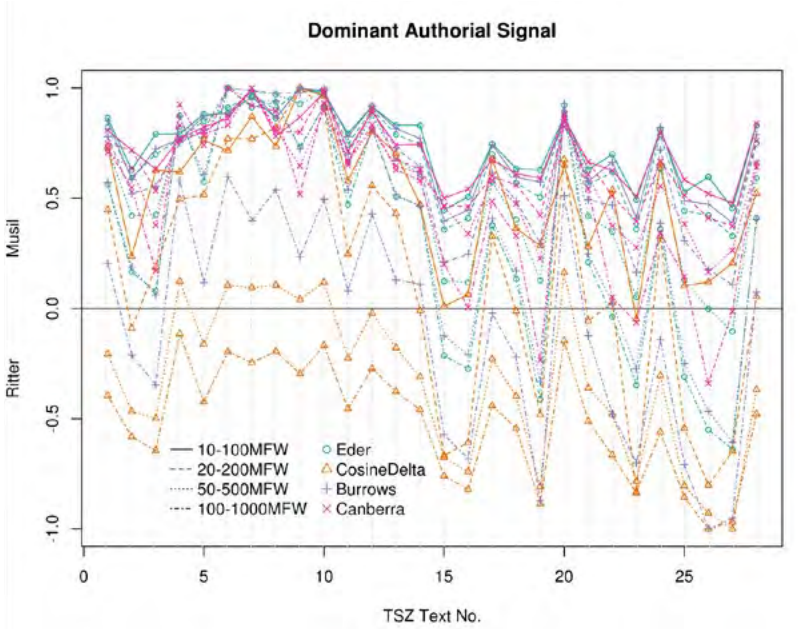


Figure 4. Experiment results (full test set)

In a second experiment, we split the corpus in two, applying the same experimental set-up. The first sample just contained those texts that were **not attributed to Musil by at least two distance measures** (N=14). Here, Ritter appeared as dominant throughout the whole selection. The second sample contained the texts that had **been relatively clearly attributed to Musil** in the first round (N=14): results show that here, Musil's signal was even more dominant, with all values close to +1 (see Fig. 5).

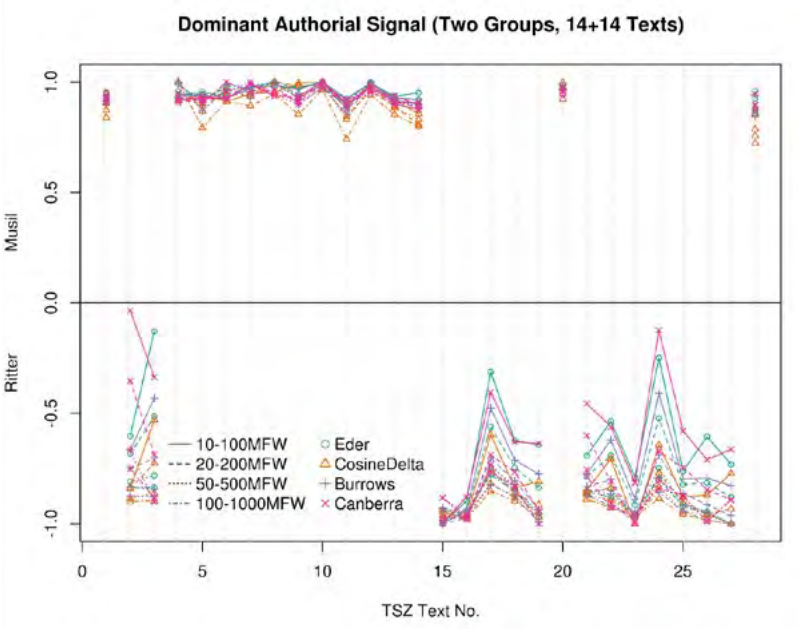


Figure 5. Experiment results (split test set)

When further reducing the selection to 9 texts (those for which all classifiers scored less than –0.5 in the previous round, see Fig. 5), all texts were attributed to Ritter with a stronger probability, while the graph generated by the remaining 19 texts was still confined to Musil's region (see Fig. 6). In answer to our research question, results suggest that Musil attribution may be disproved with a high level of confidence for texts No. 15, 16, 18, 19, 22, 23, 25, 26, and 27 (see Table 1 for details). At the same time, our analysis proposes that Ritter may be the author of these 9 articles.

Figure 6. Experiment results (split test set)

Future research

Future expansion of our research should define new training sets to validate the results and increase the test set. Both, however, will require an extensive digitization effort: most of the useful texts (i.e., propagandistic WWI writings) are not available in a clean plain-text format. In addition, other software should be tested on the already defined corpus, e.g., JGAAP (Juola et al., 2008) and CLEF/PAN (Stamatatos et al., 2015), as these consider features and methodologies excluded from the present experiment, such as character n-grams and machine learning.

With our study, we hope to have laid the groundwork for a research that can have long-lasting consequences on the historiography of German literature, evidencing at the same time how quantitative methods are not in opposition, but complementary to the qualitative strands (Herrmann, 2017) of literary history.

References

Amann, K., Corino, K. and Fanta, W. (2009). *Robert Musil, Klagenfurter Ausgabe.* Klagenfurt: Robert Musil-Institut der Universität Klagenfurt.

Arntzen, H. (1980). *Musil-Kommentar sämtlicher zu Lebzeiten erschienener Schriften außer dem Roman "Der Mann ohne Eigenschaften"*. München: Winkler.

Corino, K. (1973). Robert Musil, Aus der Geschichte eines Regiments. *Studi Germanici*, 11: 109–15.

Corino, K. (2003). *Robert Musil: eine Biographie*. Reinbek bei Hamburg: Rowohlt.

Corino, K. (2010). Klaviersonnen über Schluchten des Gemüts. Robert Musil und die Musik. *Das Plateau*, 120: 4–21.

Dinklage, K. (1960). *Robert Musil. Leben, Werk, Wirkung*. Zürich: Amalthea Verlag.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1): 107–21.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2): 167–82.

Fontanari, A. and Libardi, M. (1987). *La guerra parallela*. Trento: Reverdito.

Halvani, O., Winter, C. and Pflug, A. (2016). Authorship verification for different languages, genres and topics. *Digital Investigation*, 16: 33–43.

Herrmann, J. B. (2017). In test bed with Kafka. Introducing a mixed-method approach to digital stylistics. *Digital Humanities Quarterly*, [in press].

Juola, P., Noecker, J., Ryan, M. and Zhao, M. (2008). JGAAP3.0 – authorship attribution for the rest of us. *Digital Humanities 2008: Book of Abstracts*. Oulu: University of Oulu, pp. 250–51.

Juola, P. (2015). The Rowling case: a proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30: 100–13.

Koppel, M. and Winter, Y. (2014). Determining if two documents are by the same author. *JASIST*, 65(1): 178–87.

Roth, M.-L. (1972). *Robert Musil. Ethik und* Ästhetik. München: List.

Schaunig, R. (2014). *Der Dichter im Dienst des Generals. Robert Musils Propagandaschriften im Ersten Weltkrieg*. Klagenfurt: Kitab.

Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M. and Stein, B. (2015). Overview of the Author Identification Task at PAN 2015. *CLEF 2015: Working Notes*. http://ceur-ws.org/Vol-1391/inv-pap3-CR.pdf [accessed 26.11.2017].

Urbaner, R. (2001). "... daran zugrunde gegangen, daß sie Tagespolitik treiben wollte"? Die "(Tiroler) Soldaten-Zeitung" 1915-1917. *eForum zeitGeschichte*, 3/4. www.eforum-zeitgeschichte.at [accessed 26.11.2017].

Creating and Implementing an Ontology of Documents and Texts

Peter Robinson
peter.robinson@usask.ca
University of Saskatchewan, Canada

Outline

The application of computing methods to scholarly editing was one of the first major areas to be explored in the nascent discipline of "Humanities Computing": the direct ancestor of what we now know as "Digital Humanities". The highly-structured nature of scholarly editions, with their formal links between text and carefully crafted apparatus, and the promise that complex patterns in historical texts might be usefully explored by computer methods, made them obvious targets for the early application of computer methods to humanities materials (Hockey; see also early work by Dearing and Petty and Gibson). The first formal attempt at systematic computer representation of texts, the Text Encoding Initiative, analyzed these structures into a formal encoding scheme, itself building on the principles set out in De Rose et al's landmark 1990 publication, "What is Text, Really".

While the TEI encodings created from 1990 onward proved a solid foundation for many scholarly editions in digital form, from the very beginning scholars recognized a fundamental problem in the TEI encodings when applied to scholarly editions. At its most basic: one wants to see the text of a manuscript on screen as it appears in the manuscript: page by page, line by line. But also one wants to see that text not as it appears in the manuscript, but according to its logical structure as an act of communication: that is, as composed of segments (Acts and scenes; or stanzas and lines; or chapters, paragraphs and

sentences). Because these two views almost never correspond, we have what is usually termed the problem of "overlapping hierarchies": paragraphs cross page boundaries; manuscripts contain multiple works, distributed in complex ways across the pages.

Many papers have addressed this issue of the "overlapping hierarchy" (De Rose; Sperberg-McQueen and Huitfeldt), and this author has wrestled with this issue across multiple editions and operating systems. In 2010, the author commenced work on a new system for collaborative online scholarly editing, "Textual Communities". A key aim was that this system would seek a robust and fundamental solution to the problem described as "overlapping hierarchies". Accordingly, the first task was to rethink exactly what we mean by the terms "document", "work" and "text". For this, the author went to textual scholarship, which has been considering the meaning of these terms for centuries. In a series of articles (2013a, 2013b, 2017) the author has explored their meaning, with the 2013a article most clearly anchoring his perceptions in the traditions of textual scholarship. In summary, these terms are defined as follows:

1. A text is an act of communication instanced in a document
2. The act of communication is composed of an ordered hierarchy of objects (Acts and scenes; or stanzas and lines; or chapters, paragraphs and sentences): hence, a tree
3. The document is composed of an ordered hierarchy of objects: the volume, divided into quires, divided into leaves, divided into recto and verso pages, divided into columns, divided into lines (or, surfaces, divided into zones, etc): hence, a tree

In this analysis, every text is composed of two distinct and independent hierarchies: one tree for the document, and one for the act of communication. Both trees are essential. An act of communication cannot exist unless it is physically instantiated in a document. If the document does not present an act of communication, then it is simply marks on paper, without lexical meaning.

Textual communities formalized these definitions in an ontology. The naming system used by this ontology is based on the well-known Kahn-Wilensky system (1995), commencing with a naming authority (in this example, TC:CTP) and then using a sequence of property/value pairs to specify each object. In this case, we are describing that part of paragraph 291 of the Parson's Tale ("PA") which appears in line 40 of folio 232v of the Hengwrt manuscript of the Canterbury Tales:

1. The document hierarchy: TC:CTP/Document=Hg:Page=232v:Line=40
2. The act of communication hierarchy: TC:CTP/Entity=CTP:Part=PA:ab=291
3. The text, combining both hierarchies: TC:CTP/Entity=CTP:Part=PA:ab=291: Document=Hg:Page=232v:Line=40

In this formulation, every text is composed of a sequence of leaves, with every leaf shared by two distinct trees. Thus the "leaf" of text contained in line 40 of folio 232v occupies TC:CTP/Document=Hg:Page=232v:Line=40; that same text is part of TC:CTP/Entity=CTP:Part=PA:ab=291. The power of the system can be readily appreciated. First, one may travel through the document hierarchy to show the text page by page, line by line. Second, one may travel through the act of communication ("entity" in our system) hierarchy to find the different versions of paragraph 291 in multiple manuscripts and compare them. In this analysis, what we term "overlapping hierarchies" is a symptom, a result of the underlying system of distinct trees sharing leaves.

Theory is one thing; implementation is another. We wanted a system that could be updated in real time. (Here, I speak of "we" as the progressing work became more and more a collaborative project). That is: a manuscript page could be transcribed, the order and structure of the text on the page rearranged, deleted, replaced, and the results written near-instantly to a storage system and available immediately to others. Over a long text (20,000 lines of the Canterbury Tales) in many manuscripts (88 for the Tales, some 30,000 pages in total) this is rather challenging. One may compare this with removing leaves from the trees, rebuilding the branches to which they were attached, and then reattaching the leaves, all in a howling gale. A brief attempt to use an XML database (in this case, XML DB, now maintained by Oracle) revealed substantial performance problems. For several years, we used a MySQL relational database. But the tables linking the distinct trees rapidly became so complex, and the queries required to manipulate them so unwieldy, that we abandoned it. Finally we moved to representing all documents in JSON form, and then storing and retrieving them through a JSON document store (MungoDB). This has proved complex, but very fast and effective. We are able to represent the two hierarchies precisely, in a manner which permits realtime updating and retrieval, within the JSON store. Indeed, one could extend the model we apply beyond two hierarchies: a text could be composed of as many hierarchies as one likes.

The first public version of Textual Communities (after seven years of work) will be released in the first half of 2018, and the author will propose a workshop on the system at the conference. This paper will show the full system briefly. It is designed to be easy to use, to the point that a textual scholar with no special computer expertise will be able to use it to create an edition. Further, the implementation of the underlying database in JSON, and of javascript throughout the system, should make it possible for computer programmers expert in Javascript (and with no expertise in XML) to make complex critical editons. The system also contains tools to allow management of a large collaborative edition, with management of transcription page by page. The sophisticated Collation Editor, developed by the New Testament Greek edition projects in Birmingham, England and Munster, Germany, itself built on CollateX, is also integrated.

This work raises many questions. XML is currently used for basic document input, and for transcription page-by-page. However, the inability of XML to fully represent more than a single hierarchy in a single document is a serious impediment to Textual Communities. In essence, the textual model we implement in Textual Communities is more powerful than XML can provide. Our hope is that others will take up this challenge, to find ways to move past this weakness in XML. Indeed, we offer Textual Communities not as, in any sense, a definitive system. It is a first attempt to implement the ontology of text and document upon which it is built. We hope and expect others will do better than this system.

References

Dearing, Vinton A. 1962. *Methods of Textual Editing*. Clarke Library Lecture. University of California.

DeRose, S., David Durand, Elli Mylonas, and Allen Renear, 1990. "What is Text, Really?" *Journal of Computing in Higher Education*, pp. 3-26.

DeRose, S. 2004.Markup overlap: A review and ahorse. *Proceedings of the extrememarkuplanguages 2004*. Rockville: Mulberry Technologies.

Hockey, Susan. 1980. *Guide to Computer Applications in the Humanities*, Duckworth and Johns Hopkins.

Kahn, Robert E., and Robert Wilensky, 1995. *A Framework for Distributed Digital Object Services*. Available at http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.

Petty, George R. and William M. Gibson, 1970. *Project OCCULT: The OrderedComputerCollation of Unprepared LiteraryText.* New York: New York University Press.

Robinson, P., 2013a. "The Concept of the Work in the Digital Age." In Barbara Bordalejo (ed.), *Work, Text and Document in the Digital Age*. *Ecdotica*, 10, 13- 41.

Robinson, P., 2013b. "Towards A Theory of Digital Editions." *Variants* 10, 105- 132.

Robinson, P. 2017. 'Some principles for making collaborative scholarly editions in digital form.' *Digital Humanities Quarterly* 11.2.

Sperberg-McQueen, C. M. and ClausHuitfeldt, 2004 "GODDAG: A Data Structure for Overlapping Hierarchies". In *DDEP/PODDP*, pp. 139-160.

Detección y Medición de Desequilibrios Digitales a Escala Local Relacionados con los Mecanismos de Producción y Distribución de Información Cultural

Nuria Rodríguez-Ortega
nro@uma.es
University of Málaga, Spain

Research Context

Through different projects, the iArtHis_Lab (www.iarthis-lab.es) research group has been analyzing the mechanisms for the production and distribution of cultural information on the Internet since 2015. Furthermore, this research group has considered the possibilities of reusing this information in the form of structured data to generate new knowledge and value through computational processing techniques. The interest of this research is derived from two main motivations. First, there is the need to examine and unveil the informational gaps that the digital society is producing in the cultural field. Secondly, there is the need to evaluate the potential connections between such inequalities and the development of the so-called creative economy.

The creative economy has been defined as the role played by artistic creation, creative sectors and cultural institutions in the articulation of new production models that catalyze the generation of wealth, quality of life and social welfare (European Commission, 2010 UNESCO, 2013, Unctad, 2010). According to this understanding, it seems evident that the mechanisms for the production and distribution of cultural information are closely related to the possibilities of its promotion and development, since they are able to mobilize the flow of people and communities to certain locations[1], to highlight the distributed cultural heritage and so on (Casacuberta et al., 2008). Therefore, our starting point is the assumption that the digital practices associated with the production and distribution of cultural information are part of the systemic and structural factors that must be taken into consideration when evaluating the possibilities of certain territories to project their creative potential and add value to their own cultural richness.

Examining this scenario from a local perspective has been previously proposed by UNESCO in its *Report for the Creative Economy*, the UNDP and other authors (for example, O'Connor, 2008 and Florida et al., 2008). This is one of the objectives of the research project *Data Methodologies Applied to the Analysis of Art Exhibitions for the Development of the Creative Economy*[2], which focuses its attention on the region of Andalusia located in the south of Spain and on the subsector of the exhibition activity. The exhibitions represent a key component of the creative economy as they are a first-order element in the promotion of socio-cultural and economic dynamics in the terri-

1 According to the information provided by the Ministry of Education, Culture and Sport of the Government of Spain, 14.7% (Spanish) and 12.6% (foreigners) of the travelers in 2016 were motivated by cultural interests. Likewise, 39.4% of Spanish citizen participated in activities organized by museums, galleries, art centers, etc., which is an increase of 1.5% compared to the previous year.

2 This paper is part of the results of the project funded by the Centro de Estudios Andaluces of the Junta de Andalucía (Spain) [*Metodologías de datos aplicadas al análisis de las exposiciones artísticas para el desarrollo de la economía creativa*, PRY128-17].

tories. Furthermore, they favor the valuation of creativity as an expression of territorial and cultural diversity. Within this research framework, this paper will focus on the issues related to the inequalities and imbalances detected with respect to the digital presence and informational/communicational practices (recording, documentation, curation, dissemination of contents) associated with the institutions and organizations holding exhibitions in the Andalusian cultural system.

Methodology

To undertake this research, specific work methodologies and technological devices were developed, which allow us to analyze these aspects from different perspectives. In particular, a repository of the structured and semantically enriched data related to all of the cultural institutions and organizations located in Andalusia holding exhibitions since 2000 was built. This repository is accessible through the Expofinder system (www.expofinder.es). Currently, the corpus size is 1,917 institutions and 12,414 metadata, with an average of 12.98 metadata per institution.

The designed methodology for the analysis combines a double evaluation system with quantitative and qualitative analysis, which employ different diagnostic indicators to measure the two types of variables mentioned: digital presence and informational/communication practices associated with the recording, documentation and dissemination of cultural-exhibition activities.

To measure the digital presence of the cultural institutions in the whole Andalusian region, we have used the number and distribution of the URIs (information included in the Expofinder data model) associated with these cultural institutions as a main indicator, which function as information sources of their exhibition activities. We chose to use URI as an indicator instead of the webpage, which has been mostly used in the traditional approach of this type of analysis, as this parameter not only allows us to quantify the digital presence of cultural entities (according to the number of URIs), but also to project a prospective vision about their potential to distribute cultural information in the digital space[3].

On the other hand, the results of previous projects have demonstrated that cultural information, especially the information related to art exhibitions, is heterogeneous, discontinuous, unstructured and dispersed (Rodríguez Ortega, 2016, Rodríguez Ortega, in press). Thus, we will complement this analysis with the qualitative evaluation of the information systems used by cultural institutions to record and document their exhibition activities in accordance with an analysis model developed ad hoc for this project.

This valuation model integrates the traditional approaches used in the analysis of the digital communication of cultural institutions and organizations (factors inherent to the webpage, such as the usability, design, navigation, presence in RRSS, etc.), although this model also focuses its attention on the two dimensions that are more scarcely taken into account until now: (a) the type and degree of structuring of the information about art exhibition and its flow dynamics; and (b) the digital practices related to the production and communication of the cultural information associated with each type of cultural organization analyzed, which are shaped by very diverse interests, infrastructures, objectives and actors. In the first case, the evaluation is performed using a measurement system based on a qualitative indicator, which is being carried out by peer reviewers. The results of this measurement system were expressed in a data matrix. The second case is conducted according to the Latourian approach (Latour, 2007). This analysis is based on a direct interaction with the actors involved through surveys and interviews. These latter results are provided as the variants of singularity that contextualize the former data.

Analysis and discussion

The analysis of all these data allow us to obtain an accurate picture of the degree of digital presence of cultural institutions and organizations in Andalusia. Furthermore, we have been able to visualize the diversity of the digital practices associated with the production and dissemination of art exhibition activities. The results mainly reveal that both variables constitute one of the problems and weaknesses of the cultural system of the region.

As an example of the results that will be shown during the presentation, Figure 1 shows the magnitude of the black hole that characterizes the region of Andalusia as there are cultural institutions that do not have any type of web source or have a high precariousness.

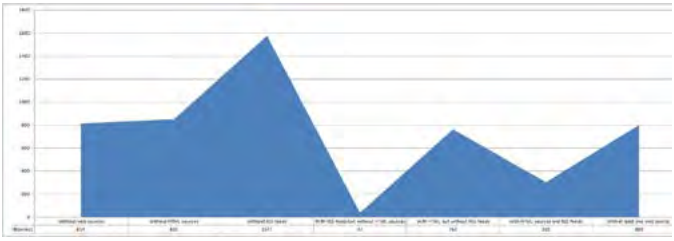


Figure 1. Dimensions of the "black hole", which exceeds the space occupied by web sources. Likewise, it is observed that the weakest point is the scarce use of RSS feeds, which is significant considering that they are more efficient in the process of distributing information by digital means due to being highly structured.

By combining the geolocation layers of institutions and web sources (as rendered in the map of Figure 2), we

3 Keep in mind that a webpage may have associated or non-specific web sources (URI) for the distribution of information about cultural activities; and that the same web site can bring together different URIs devoted to the distribution of cultural information, which multiplies its capacity to produce and distribute information.

also observed that the greatest number of entities without web sources are found in the areas peripheral to the provincial capitals.

Figure 2. Territorial distribution of cultural entities (colored by provinces) and URIs (black). Source: Expofinder. Software: Carto

By the time of the presentation at DH2018, these results will be completed and other statistical and geospatial results will be collected, based on the correlation of demographic, economic, cultural and social data. Furthermore, they will be connected with the results of the qualitative evaluation of the information systems and digital practices.

The discussion will delve into the interpretive keys that underlie the observed trends. In general terms, the inquiry will be guided by the following questions: given that there is an association of the degree of digital presence with the size of the cities and as their location in the regional and provincial hierarchy, what is the incidence of public policies and cultural strategy programs endorsed by the government of each city in the configuration of this scenario? In what way do the policies aimed at promoting certain territories within the framework of the development of the creative economy generate new territorial peripheries by concentrating these efforts in certain contexts? Are there significant differences between public and private institutions/organizations?

References

Casacuberta, D., García Alba J. et al. (2008). *Industrias culturales en la web 2.0*. Fondo Multilateral de Inversiones, BID.

Comisión Europea (2010). *Libro Verde. Liberar el potencial de las industrias culturales y creativas.* Bruselas: Comisión Europea.

Cruces Rodríguez, A. (2017). *Sistema de gestión de datos Expofinder. V. 1.2*. (PDSDT-0102). Documento técnico disponible en: http://exhibitium.com/documentacion/ [Consulta: 24-11-2017].

Florida, R.; Mellander, C. & Stolarick, K. (2008). Inside the black box of regional development. Human capital, the creative class and tolerance. *Journal of Economic Geography*, 8(5): 615-649.

Latour, B. (2007). *Reensamblar lo social. Una introducción a la teoría del actor-red,* Buenos Aires: Manantial.

Ministerio de Educación, Cultura y Deporte (2017). *Plan de fomento de las industrias culturales y creativas,* Madrid: Subdirección General de Documentación y Publicaciones.

O'Connor, J. (2008). *The Cultural and Creative Industries: A Review of the Literature. A report for Creative Partnerships*. London: Arts Council of England.

Rodríguez Ortega, N. (en prensa). Exposiciones y proyectos curatoriales: entre la preservación de la memoria y la explotación de sus datos». In Roseras Carcedo, E. (coord.), *Datos abiertos vinculados y gestión integral de la informació*n en los centros patrimoniales. Vitoria: Artium.

Rodríguez Ortega, N. (2016). Construcción y uso de terminologías, categorías de descripción y estructuras semánticas vinculadas al patrimonio en la sociedad global de datos. In *El lenguaje sobre el patrimonio. Estándares documentales para la descripción y gestión de colecciones*. Madrid: Ministerio de Educación, Cultura y Deporte, pp. 115-130.

Unctad (Conferencia de las Naciones Unidas sobre Comercio y Desarrollo) (2010). *Creative Economy Report.* Ginebra-Nueva York: Unctad.

UNESCO (2013). *Creative Economy Report 2013 Special Edition. Widening Local Development Pathways*. New York: Programa de las Naciones Unidas para el Desarrollo (PNUD) and (UNESCO)

#SiMeMatan Será por Atea: Procesamiento Ciberactivista de la Religión como Parte del Canon Heteropatriarcal en México

Michelle Vyoleta Romero Gallardo
michelle.romero@flacso.edu.mx
Multiversidad Mundo Real Edgar Morin, Mexico

El 3 de mayo de 2017, el cuerpo sin vida de una mujer de 22 años fue encontrado dentro de un campus universitario de la capital mexicana. En su difusión por Twitter de los avances del caso, la Procuraduría General Judicial de la Ciudad de México difundió detalles como el nombre de la joven (Lesvy Berlín) y realizó afirmaciones con respecto a que ella vivía con su pareja, que habría interrumpido sus estudios dos años atrás, que antes de hacerlo no habría aprobado varias asignaturas y que de forma inmediatamente previa a su fallecimiento la mujer y su pareja habrían consumido bebidas alcohólicas y drogas en las instalaciones de la universidad. En Twitter, el recibimiento de la difusión del caso inmediatamente se transformó en una recriminación ciudadana contra la Procuraduría General Judicial, al compartirse generaliza-

damente la impresión de que la forma en que la instancia presentó a Lesvy Berlín generaba un discurso justificatorio de su muerte y que, en última instancia, hacía recaer en ella la responsabilidad de haber perdido la vida. La crítica social ante este discurso se canalizó hacia el hashtag #SiMeMatan, al cual las y los usuarios de redes sociales le acompañaron de descripciones de sus propias vidas que, en el hipotético caso de que se les encontrase sin vida, las instituciones de procuración de justicia potencialmente podrían difundir para crear una impresión de lo "justificado" de su fallecimiento.

Por una parte, el surgimiento y difusión en torno a #SiMeMatan puede estudiarse como una muestra de ciberactivismo o activismo[1] que tiene lugar exclusivamente en Internet y/o que se sirve de la red para reforzar acciones que tendrán lugar en entornos no virtuales (McCaughey y Ayers, 2003). En segundo término, es posible analizar la manera en que esta acción (virtual) de participación en la vida social plasmó en la forma de textos a comportamientos, valores y actitudes percibidos como disruptivos de los cánones de normalidad vigentes en la sociedad mexicana contemporánea. Así, cada vez que las y los usuarios emplearon el hashtag #SiMeMatan acompañado de una descripción de las "disrupciones" por las que las autoridades podrían culparles de su propia muerte, los ciberactivistas contribuyeron al registro de dos conjuntos de discursos sociales interdependientes: el de las "faltas" a la normalidad y, en una imagen invertida por el espejo, el discurso de los comportamientos, valores y actitudes colocados en una posición de hegemonía por su conformación con las normas sociales más ampliamente aceptadas. El análisis discursivo del uso de #SiMeMatan perite el rastreo de "grandes temas" o ejes en torno a los cuales se evidenciaron las tensiones de la polaridad *aceptable/no aceptable* percibida por los usuarios de redes sociales. Fundamentalmente, se trató de los terrenos en disputa del cuerpo, el uso del tiempo libre, los vínculos entre varones y mujeres, la orientación sexual, así como el dualismo autonomía/conformismo. Al condensar las características movilizadas en los discursos de lo desafiante[2] y lo desafiado en ocasión de la muerte de Lesvy Berlín, es posible apreciar que la imagen a trazos amplios de "la normalidad" identificada por los ciberactivistas del caso coincide con una descripción de heteropatriarcalidad, en tanto que elementos situados por fuera del despliegue modesto/tradicionalista del cuerpo femenino y las preferencias heterosexuales fueron señalados críticamente como motivos latentes de "merecimiento de consecuencias negativas" (desde la discriminación a la pérdida de la vida). Sin embargo, en este caso de ciberactivismo también puede analizarse la manera en que los usuarios que emplearon el hashtag #SiMeMatan integraron a la religión como parte del canon de heteropatriarcalidad que hicieron objeto de sus críticas: mensajes como "#SiMeMatan será porque no creía en Dios, usaba falda y fumaba", "#SiMeMatan será porque no era católica, era lesbiana y salía a bailar", "#SiMeMatan será porque no iba a misa y decía groserías", "#SiMeMatan será porque no me confesaba y no tenía relaciones estables" o "#SiMeMatan será porque era atea, me emborrachaba y salía sola con hombres", son algunos de los ejemplos que permiten conformar como objeto de estudio una percepción social según la cual el desafío a la "normalidad religiosa" se concibe como parte integral del conjunto de desafíos a la heteropatriarcalidad vigente en la sociedad mexicana.

La presente discusión tiene como objetivo analizar el procesamiento ciberactivista de la religión como parte del canon heteropatriarcal en México en ocasión de la muerte de Lesvy Berlín y el surgimiento del hashtag #SiMeMatan en la Ciudad de México a inicios de 2017. Este objeto de estudio haya su justificación en la forma en que el caso contribuye al mapeo de las percepciones sociales vigentes en torno a la "normalidad y el desafío de la normalidad" en los comportamientos, valores y actitudes de varones y mujeres jóvenes del México contemporáneo. De igual manera, este análisis brinda información sobre la manera en que la no-conformidad con un canon religioso está entrelazado dentro del discurso social de no-conformidad con otros cánones en materia de sexualidad y de expectativas en torno al comportamiento femenino. Finalmente, el caso también brinda oportunidad para debatir la categoría de ciberactivismo cuando se le aplica a reacciones coyunturales de los usuarios de redes sociales. El análisis se construye a partir del análisis discursivo de cien comentarios marcados con el hashtag #SiMeMatan, seleccionados por muestreo teórico, difundidos por la red social Twitter durante el día 3 de mayo de 2017. En un primer momento se contextualiza la dinámica con la que se desenvolvió el tratamiento oficial en redes de la muerte de Lesvy Berlín y al que se dio respuesta con el hashtag #SiMeMatan. A continuación, se discute por qué los comentarios que se sumaron a esta expresión en Twitter pueden calificarse como ciberactivismo en particular y una acción colectiva en términos generales. Para ello, se recurre a la propuesta teórica de Snow (1986) en lo que respecta a los procesos de vinculación de los intereses, valores y creencias de los individuos como pilar de la acción colectiva. A partir de esas bases, se usa la metodología de análisis del discurso para procesar las evidencias de percepciones de norma-

1 "El activismo es una categoría de acción para participar en la política que es 'públicamente declarada y una abierta contribución a la vida política'" (Yeatman, 1998, p. 33, citado por Demetrious, 2013, p. 34). Esta definición de activismo presenta la limitación de haber sido generada desde una perspectiva según la cual la acción colectiva tiene como su objetivo fundamentalmente a los planteamientos políticos, sin tomar en cuenta motivos como el reconocimiento de identidades marginalizadas, la lucha por el medio ambiente o cualquiera de las causas asociadas a los nuevos movimientos sociales de los contextos postmaterialistas. Pese a lo anterior, de Yeatman y Demetrious se retoma la vocación del concepto "activismo" para enunciar la acción participatoria de un campo de la vida social (à la Melucci: fuere para modificarlo o conservarlo).

2 Aquello por lo cual las autoridades construirían discursos de ruptura con la normalidad en caso de perder la vida, de acuerdo con los usuarios que utilizaron el hashtag #SiMeMatan.

lidad y no-normalidad movilizados en las respuestas de #SiMeMatan en Twitter. De ese modo, se muestra la conformación del canon de normalidad en torno al cuerpo, la modestia tradicionalista y la heterosexualidad criticado por los ciberactivistas del caso y se brindan evidencias de la integración de una idea de "normalidad religiosa" dentro de ese conjunto de valores hegemónicos criticados.

References

Checa Godoy, A. (2008). *Historia de la comunicación: de la crónica a la disciplina científica*. La Coruña, España: Netbiblio.

Demetrious, K. (2013). *Public Relations, Activism, and Social Change: Speaking Up*. Nueva York, Estados Unidos: Routledge.

McAdam, D. (1999). Oportunidades políticas: Orígenes terminológicos, problemas actuales y futuras líneas de investigación. En D. McAdam *et al*. (Eds.). *Movimientos sociales: perspectivas comparadas* (pp. 49-70). Madrid, España: Istmo.

McCarthy, J. D., y Zald, M. N. (1977). Resource Mobilization and Social Movements: A Partial Theory. *American Journal of Sociology, 82*, 1212-1242.

McCaughey, M. y Ayers, M. (2003). *Cyberactivism: Online Activism in Theory and Practice*. Nueva York, Estados Unidos: Routledge.

Melucci, A. (2003 [1996]). *Challenging codes. Collective action in the information age*. Nueva York, Estados Unidos: Cambridge University Press.

Snow, D. A., Rochford, E., B. Jr., Worden, S., K., y Benford, R. D. (1986). Frame Alignment Processes, Micromobilization, and Movement Participation. *American Sociological Review, 51*, 464-481.

Turner, E. (2013). New Movements, Digital Revolution, and Social Movement Theory. *Peace Review*, *25*(3), 376-383.

Edición literaria electrónica y lectura SMART

Dolores Romero-López
dromero@filol.ucm.es
Universidad Complutense de Madrid, Spain

Alicia Reina-Navarro
areina.ali@gmail.com
Universidad Complutense de Madrid, Spain

Lucía Cotarelo-Esteban
luzia_cotarelo@hotmail.com
Universidad Complutense de Madrid, Spain

José Luis Bueren-Gómez-Acebo
joseluis.bueren@bne.es
Biblioteca Nacional de España

En respuesta a los nuevos hábitos de lectura y escritura que están surgiendo a raíz de la edición digital, el proyecto eLITE-CM, Edición Literaria Electrónica, (H2015/HUM-3426), nace –bajo los auspicios de la Comunidad de Madrid, en colaboración con la Biblioteca Nacional de España y co-liderado por el grupo de investigación LOEP–, con el objetivo de crear tres colecciones de libros enriquecidos pertenecientes a la denominada Otra Edad de Plata de la Literatura Española (Ena, 2013): 1) Colección *Literatura Infantil*; 2) Colección *Madrid en la Literatura*; 3) Colección *La Mujer Moderna*. Se trata de textos publicados entre 1868 y 1936 que, bien por motivos ideológicos, bien por 'rareza' estética, han quedado relegados a los márgenes de la Modernidad. Las colecciones son accesibles a través de la *Biblioteca Digital Mnemosine* (Romero, Bueren y Gayoso: 2017).

El proyecto, que cuenta con un equipo interdisciplinar de investigadores expertos en la época y jóvenes especialistas en edición electrónica, se propone así rescatar textos olvidados, pero de gran calidad literaria, para fomentar su relectura en línea, digitalizados y enriquecidos mediante narraciones transmedia (hipertextualidad, multimedialidad e interactividad). En esta comunicación se presentarán los resultados de las tres colecciones, llevadas a cabo mediante el software Madgazine: una plataforma de edición que permite combinar imagen, vídeo, texto e hipervínculos con el fin de crear contenido multimedia. La visualización pública de estas colecciones será accesible a través del portal de la Biblioteca Nacional de España (trabajo en curso), dado que los textos digitalizados pertenecen al fondo de dicha institución. De esta manera la Biblioteca Nacional de España y los investigadores colaboran en editar, enriquecer y compartir en formato digital libros pertenecientes a nuestro legado histórico con el fin de enriquecer didácticamente los usos del pasado.

Fruto de esta experiencia de edición interactiva ofreceremos unas reflexiones teóricas. Para sacar el máximo provecho a la lectura digital tenemos que *deconstruir* culturalmente el artefacto libro tal y como hoy lo conocemos y *reconstruirlo* con un nuevo formato interactivo que integre hipertexto e hipermedia (Bleeker, 2010). Los beneficios que la lectura digital aporta al individuo a nivel cognitivo siguen los cinco principios de la lectura inteligente: (1) *Simplicidad* en el uso de los dispositivos, herramientas y recursos, (2) *Motivación* hacia la innovación y la creatividad, vectores del conocimiento humano (3) *Accesibilidad* a un mundo interconectado, (4) *Reciclaje* de contenidos digitales a través del enriquecimiento cultural y (5) *Transferencia* a la comunidad global para permitir nuevas relecturas de la historia literaria. Este modelo de lectura *Smart* que proponemos puede alcanzar un gran desarrollo si se vinculan contenidos entre distintos libros interactivos que fomenten tanto el aprendizaje significativo (Garita Sánchez, 2001 y Moreira, 2005) como la conciencia literaria (Zyngier, Chesnokova, Viana, 2007)

References

Bleeker, E. (2010*). On Reading in Digital Age*. Lezen: Amsterdam.

Ena Bordonada, Á. ed., (2013). *La otra Edad de Plata. Temas, géneros y creadores.* Madrid: Ediciones Complutense.

Garita Sánchez, G. (2001). "Aprendizaje significativo: de la transformación en las concepciones acerca de las formas de interacción". *Revista de Ciencias Sociales*, vol. I, núm. 94, 19-34.

Moreira, M. A. (2005). "Aprendizaje significativo crítico". *Indivisa. Boletín de Estudios* e investigación, núm. 6, 83-102.

Romero-López, D. Bueren-Gómez-Acebo, J. L., Galloso-Cabada , J. (2017). "Modelling Colecciones de Literatura de Quioscos for Mnemosine Digital Library" , *Kiosk Literature in Silver Age Spain: Modernity and Mass Culture*, eds. J .Zamostny y S .Larson ,Reino Unido , Intellect Books, 397-418.

Zyngier, S., Chesnokova, A. Y Viana, V. (eds.) (2007). *Acting and connecting: Cultural approaches to Language and Literature*. Kommunikation und Kulturen/ Cultures and Communication. Münster: LIT Verlag.

(A) Enlaces generales:

1.-Proyecto de Edición Literaria Electrónica (eLITE): https://www.ucm.es/edicionliterariaelectronica

2.-Grupo de Investigación La Otra Edad de Plata: proyección cultural y legado digital (LOEP): https://www.ucm.es/loep

3.-*Mnemosine,* Biblioteca Digital de La Otra Edad de Plata (1868-1936): http://repositorios.fdi.ucm.es/mnemosine/

(B) Enlaces a un ejemplo de edición interactiva. Las URL definitivas estarán alojadas en la Biblioteca Nacional de España (aún no disponibles).

Colección: Literatura Infantil

1.- Cuento "Plaga de dragones"
http://www.madgazine.com/revista/?h=46f-185c3185976675&r=187722709

2.- Cuento de "El veraneo estropeado"
http://www.madgazine.com/revista/?h=46f-185c3185976675&r=191510705

3.- Revista: "El libro de los dragones"
http://www.madgazine.com/revista/?h=46f-185c3185976675&r=191677579

4.- Revista: "La editorial Calleja"
http://www.madgazine.com/revista/?h=46f-185c3185976675&r=191849852

Colección: Madrid en la Literatura

5.-"Cinematógrafo" de Carranque de Ríos http://www.madgazine.com/revista/?h=46f-185c3185976675&r=204998279

6.- "Geolocalización de Madrid en la Literatura http://www.madgazine.com/revista/?h=46f-185c3185976675&r=207222621

7.- "El señor director" http://www.madgazine.com/revista/?h=46f185c3185976675&r=192719988

8.- "El método" http://www.madgazine.com/revista/?h=46f185c3185976675&r=192722885

9.- Revista vida de Carranque de Ríos http://www.madgazine.com/revista/?h=46f-185c3185976675&r=207136164

Colección: La Mujer Moderna

10.- "El Kodak" de Carmen de Burgos http://www.madgazine.com/revista/?h=46f-185c3185976675&r=230036753

Para la(s) historia(s) de las mujeres en digital: pertinencias, usabilidades, interoperabilidades

Amelia Sanz
amsanz@filol.ucm.es
Complutense University, Spain

Proponemos evaluar los retos que plantean los recursos digitales disponibles para el estudio de las escritoras europeas a lo largo de los siglos, atendiendo a la interoperabilidad y usabilidad de tales recursos, así como a la cantidad y pertinencia de sus datos en cuestiones de género. Para ello partimos de nuestra experiencia en la coordinación del grupo de trabajo de DARIAH-EU, *Women Writers in History* (*Mujeres escritoras en la Historia*, en adelante *WWH-WG*), así como de nuestro trabajo en un entorno virtual para la investigación como es *NEWW* (*New Approaches to Women Writers/Nuevas aproximaciones a las mujeres escritoras*) con el fin de presentar una evolución que hoy parece necesaria al conjunto de la comunidad científica y a los ciudadanos en general.

Es cierto que, en las últimas décadas hemos asistido, de un lado, a un proceso a gran escala de digitalización de fuentes en Europa, tales como periódicos, inventarios de bibliotecas, correspondencias privadas, etc. , que contienen mucha información sobre el impacto de las actividades de las mujeres escritoras y lectoras, y ahora están disponibles en línea; por otro lado, contabilizamos más de 50 bases de datos, repertorios o bibliotecas consagrados a estas autoras europeas a nivel local, nacional o transnacional. Todo ello, si bien permite empezar a desarrollar nuevas perspectivas sobre el lugar de las mujeres en la historia literaria y cultural, también plantea preguntas sobre la pertinencia de tantos y tan dispersos recursos digitales, más aún sobre la posibilidad de una historia atenta a las categorías de género gracias a las herramientas electrónicas y a los recursos digitales.

Por eso, abordaremos de forma exhaustiva y completa, en primer lugar, las fortalezas y las debilidades que esos más de cincuenta espacios virtuales presentan tanto en lo que se refiere a la interoperabilidad y usabilidad según los estándares que se nos están imponiendo como globales, como a la calidad y pertenencia de sus datos de acuerdo con las preguntas locales que formula cada grupo de investigación en cada caso.

En segundo lugar, trazaremos la evolución de un caso particular para su estudio como es ***NEWW*** : de ser una base de datos sobre la recepción de escritoras en los Países Bajos a convertirse en un entorno de trabajo virtual. Explicaremos la necesidad de insertar NEWW dentro de una plataforma que permita integrar macro y micro perspectivas desde un punto de vista teórico y metodológico y, en consecuencia, la personalización de recursos y recorridos, no sólo para la investigación sino también para la enseñanza y al aprendizaje, esto es para audiencias más amplias.

Así podremos identificar algunas de las categorías que todas estas bibliotecas y repertorios digitalizados o digitales han logrado constituir en forma de trozos de información (esto es, limitados, nombrados, diferenciados) con el fin de satisfacer (o no) las condiciones de una historia literaria con marcas de género representables (o no) como pueden ser la superación de los marcos nacionales y de las oposiciones binarias masculino-femenino, la representación de las relaciones entre género y género literario o de la dinámica de inclusión-exclusión en contextos sociales y políticos precisos. Con ello mostraremos si los recursos digitales analizados pueden incorporar, e incorporan de facto, las aportaciones de la historiografía, la crítica literaria y los estudios de género en las últimas décadas.

Presentaremos así nuestras propias conclusiones durante un año de andadura dentro del *WWH-WG* desde su constitución en Berlín (abril 2017) hasta la presentación de sus líneas de trabajo y buenas prácticas en París (mayo 2018).

La pertinencia de las categorías y de los datos, la interoperabilidad y la usabilidad de todas las bibliotecas y repositorios, bases de datos y páginas analizados son retos que exigen respuestas no solo a escala global, sino también local.

Burrows' Zeta: Exploring and Evaluating Variants and Parameters

Christof Schöch
schoech@uni-trier.de
University of Trier, Germany

Daniel Schlör
daniel.schloer@informatik.uni-wuerzburg.de
University of Würzburg, Germany

Albin Zehe
zehe@informatik.uni-wuerzburg.de
University of Würzburg, Germany

Henning Gebhard
s2hegebh@uni-trier.de
University of Trier, Germany

Martin Becker
becker@informatik.uni-wuerzburg.de
University of Würzburg, Germany

Andreas Hotho
hotho@informatik.uni-wuerzburg.de
University of Würzburg, Germany

Introduction

The research presented here concerns methodological issues surrounding Zeta, a measure of distinctiveness or keyness initially proposed by John Burrows (2007). Such measures are used to identify features (e.g. words) that are characteristic of one group of texts in comparison to another (Scott 1997), a fundamental task that many standard tools support, e.g. WordCruncher (Scott 1997), AntConc (Anthony 2005), TXM (Heiden et al. 2012) or stylo (Eder et al. 2016). Widely used methods include the log-likelihood ratio (where observed frequencies are compared to expected frequencies; see Rayson and Garside 2000) and Welch's t-test (where two frequency distributions are compared; see Ruxton 2006). Zeta, by contrast, is based on a comparison of the degrees of dispersion of features (see Lyne 1985, Gries 2008). Zeta appeals to the Digital Literary Studies community because it is mathematically simple and has a built-in preference for highly interpretable content words. Indeed, Zeta has been successfully applied to various issues in literary history (e.g. Craig & Kinney 2009, Hoover 2010, Schöch 2018). However, its statistical properties are not well understood, as important work on evaluating measures of distinctiveness (Kilgariff 2004, Lijfijt et al. 2014) has not included Zeta. Therefore, we submit two key aspects of Zeta to exploration and evaluation: (a) variations in the way Zeta is calculated and (b) variations of key parameters. We gain a more precise understanding of how Zeta works and propose a new variant, "log2-Zeta", that shows more stable behavior for different parameters than Burrows' Zeta.

Deriving Zeta variants and key parameters

Zeta is calculated by comparing two groups of texts (G1 and G2). From each text in each group, a sample of n segments of fixed size with m word tokens is taken. For each term (t) in the vocabulary (e.g., consisting of lemmatized words), the segment proportions (sp) in each group are calculated, i.e. the proportion of segments in which this term appears at least once (binarization). Zeta of t results from subtracting the two segment proportions:

$$zeta_t = sp_t(G_1) - sp_t(G_2)$$

From this formalization, we can derive several variants of Zeta: applying division instead of subtraction; using relative frequencies (rf) instead of segment proportions (sp); and applying a log-transformation to the values rf and sp instead of using them directly. This results in eight variants of Zeta (Table 1).

	segment proportions		relative frequencies	
	normal	log2	normal	log2
subtraction	sd0	ds2	sr0	sr2
division	dd0	dd2	dr0	dr2

Table 1: The eight variants of Zeta with their labels; "sd0" corresponds to Burrows' Zeta.

The formalization also points to two major parameters of Zeta: segment sampling strategy (using all possible consecutive segments, or sampling n segments per text to overcome text length imbalances) and segment size (segments with m tokens, influencing the granularity of the dispersion measure). We expect the segment size to be of particular importance, as choosing extreme values affects the calculations very strongly: using a segment size of 1 token is equivalent to relative term frequencies; using unsegmented texts is equivalent to document frequencies. Because Burrows (2007) gives no theoretical justification for his particular formulation of Zeta, a systematic exploration and evaluation is called for.

Text collection, code and raw data

Experiments have been performed using two very different text collections:

- A collection of French Classical and Enlightenment Drama (1630-1788): 150 comedies and 189 tragedies (from the Théâtre classique collection; Fièvre 2007-2017).
- A collection of Spanish novels (1880-1940): 24 novels from Spain and 24 from Latin America (from the CLiGS textbox: Henny 2017 and Calvo Tello 2017).

For reasons of space, we only report results for the Spanish novels. Texts, metadata, code, results and figures are available on Github: https://github.com/cligs/projects2018/tree/master/zeta-dh.

Methods and hypotheses

To obtain a better understanding of Zeta and its variants, we first visually explore the relation between segment size and the resulting zeta scores. We expect both Zeta variants and segment size to have visible consequences in this setting. Secondly, we evaluate the distinctiveness of words selected by different Zeta variants by using the highest ranked words as features in a classification task for distinguishing texts into two previously defined classes. This captures the degree to which the different Zeta variants and parameters identify words distinctive of these two classes. Note that we calculate Zeta scores from the complete set of documents. While this is not valid for a real-world classification task, it allows us to better judge the level of distinctiveness of the selected words. We expect better performance in the classification task with some of the new variants, compared to the classic "Burrows Zeta" (sd0). We also expect extreme segment lengths to significantly impact classification performance. We primarily aim at a methodological contribution here, so we do not attempt include a discussion of our results from a literary perspective (but see Schöch 2018 for such a contribution).

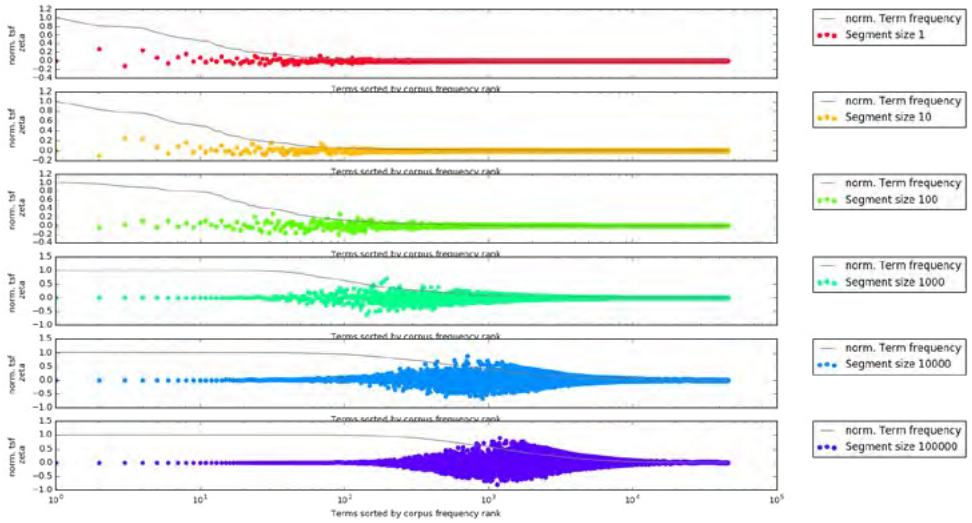
Exploratory approaches to Zeta variants and parameter variation

First, we take a closer look at the relationship between overall frequency and Zeta scores as it evolves with increasing segment size.

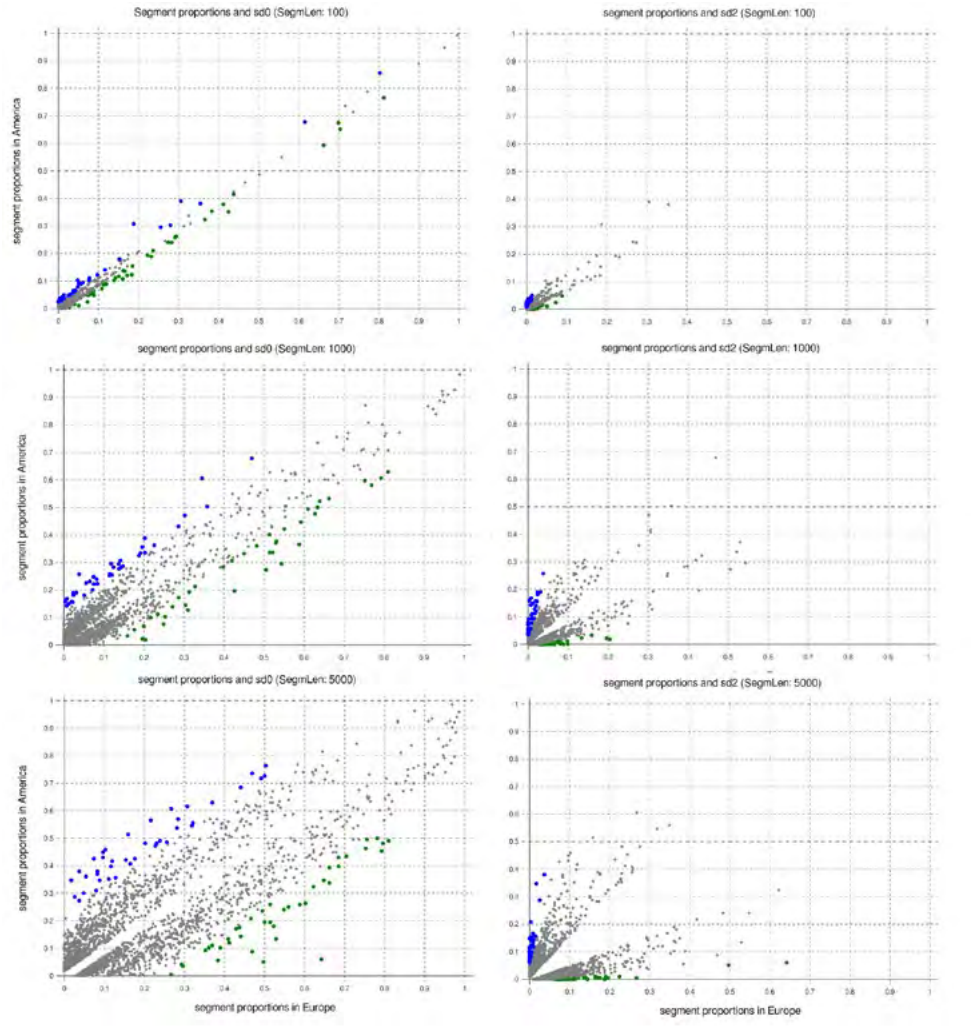
Figure 1 shows that when using very short segments, only highly frequent words (such as function words) can get high Zeta scores. With longer segments, words that are somewhat less frequent overall (well-interpretable content words) can also reach high Zeta scores; a desirable effect.

Additionally, we explore the influence of segment length and Zeta variant on the relation between segment proportions and zeta scores.

Figure 2 shows that with increasing segment size, Zeta scores generally increase because segment proportions increase. It also shows that in Burrows' Zeta (left), terms with low segment proportions can never gain high Zeta scores. This limitation motivates the log2 and division variants that alleviate this effect: here, words to the bottom and left of the plots can also obtain extreme Zeta scores.



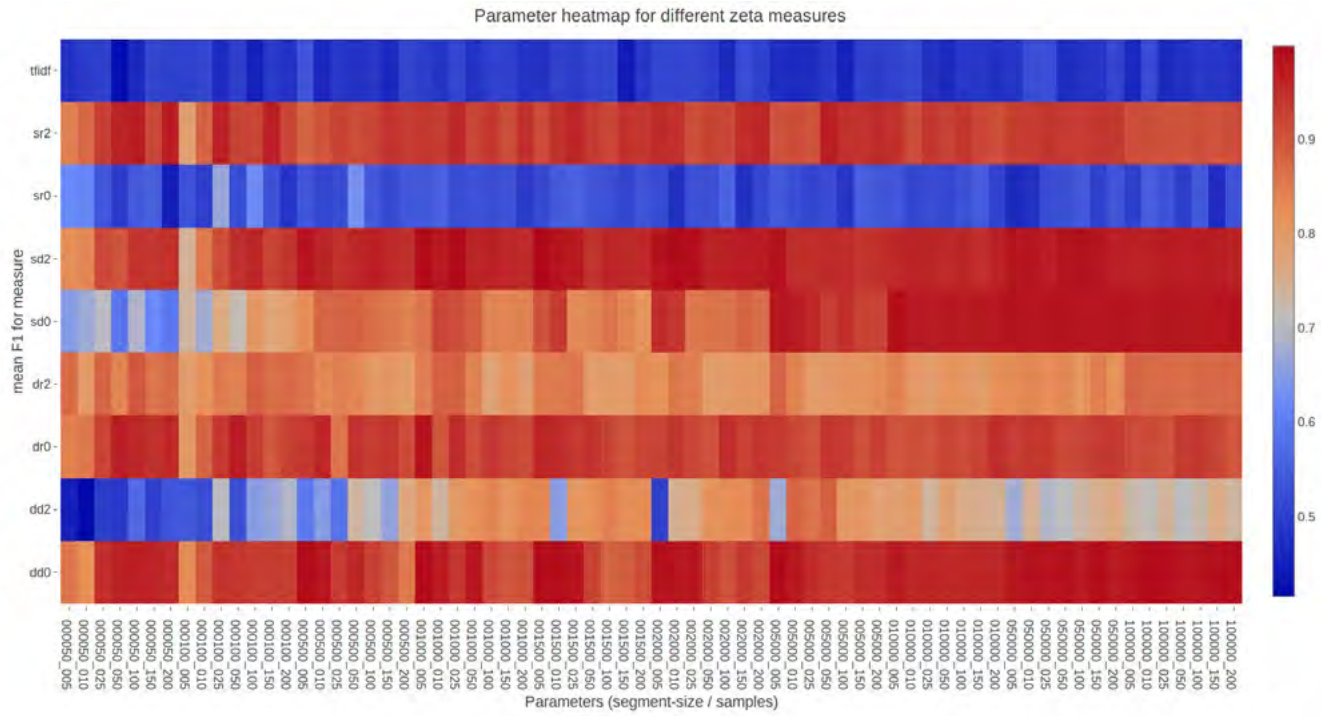
Distribution of Burrows Zeta (sd0) scores depending on segment size. Each dot is one word, ordered by descending frequency. The x-axis is log-scaled.



Segment proportions and Zeta scores for two Zeta variants (left: Burrows Zeta; right: log2-Zeta) and three segment sizes (100, 1000, 5000). Each dot is one word, 500 top Zeta words are shown. Colors indicate the words with the 40 highest (green) and lowest (blue) Zeta scores.

Evaluation of Zeta variants and parameters

Evaluating the performance of Zeta variants and different parameters is non-trivial, because it is impossible to define a human-annotated gold standard for distinctive words. Therefore, we use a classification task (with a Linear-SVM classifier and 3-fold cross-validation) for evaluation (the Spanish novels have to be classified by their continent of origin: America and Europe). The baseline of classification performance is F1=0.49 on average across all conditions and has been obtained using the top-80 most frequent words weighted with TF-IDF (see Robertson 2004).



Classification performance depending on Zeta variant, segment size and number of segment-samples. We report mean F1-score over 15x3 folds.

Figure 3 shows that, as expected, most Zeta variants outperform the baseline. Segment size also influences performance: Burrows Zeta ("sd0") performs particularly poorly with small segments (50, 100) and particularly well with large segments (>10000). Contrary to our expectation, large segment sizes do not generally have a negative impact on performance. The log2-Zeta variant ("sd2") performs better than Burrows' Zeta and is more robust with respect to segment size. In addition, we evaluate the parameter sampling size (number of segments randomly sampled for each document). For Burrows' Zeta ("sd0"), we observe a better classification performance for small samples.

Discussion: Interpretability

While improved performance and robustness are welcome, another important characteristic of Burrows' Zeta should not be forgotten, namely the high interpretability of the most distinctive words it identifies. The question is whether the gain in performance obtained with log2-Zeta comes at the expense of interpretability of the most distinctive words. Currently, we can merely offer some preliminary observations on this issue: First, the interpretability of distinctive words could be operationalized in a first approximation as the proportion of content words (nouns, verbs and adjectives) in the list of the most distinctive words, as opposed to function words and named entities. Second, segment size and Zeta variant both appear to influence the types of words Zeta that determines to be distinctive: for example, very small segment sizes favor highly frequent function words, while very large segment sizes lead to place and person names taking up a considerable space in the word list. Also, some Zeta variants, including log2-Zeta, produce lists of words containing high proportions of place and person names even at intermediate segment lengths, that is wordlists that are less interpretable (see annex). These preliminary observations point to a possible trade-off between performance and interpretability that requires further, systematic investigation.

Conclusions and Future Work

Our experiments have allowed us to gain a much more detailed understanding of how Zeta works, mathematically and empirically. Additionally, we have identified at least one Zeta variant ("log2-Zeta") that selects more distinctive words with regard to our classification task and is more robust against variation in segment length than Burrows Zeta.

As future work, we plan to conduct an investigation into the notion of "interpretability" and its relation to classification performance. Also, we plan to build an interactive visualization for our results to support a dynamic exploration of Zeta variants, key parameters and their influence on classification accuracy and distinctive words obtained. A larger agenda item is the evaluation of a substantial number of measures of distinctiveness, including Zeta, in a common framework.

Annex

For reasons of space, the wordlist annex can be found at: https://github.com/cligs/projects2018/blob/master/zeta-dh/annex.pdf.

References

Anthony, L. (2005). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*. 7–13.

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22(1): 27–47 doi:10.1093/llc/fqi067.

Calvo Tello, J. (ed.) (2017). Corpus of Spanish Novel from 1880-1940. (CLiGS Textbox). Würzburg: CLiGS. https://github.com/cligs/textbox/tree/master/spanish/novela-espanola.

Craig, H. and Kinney, A. F. (eds). (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis. *The R Journal*, 16(1): 1–15. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Fièvre, P. (ed). (2007). *Théâtre classique*. Paris: Université Paris-IV Sorbonne. http://www.theatre-classique.fr.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. I*nternational Journal of Corpus Linguistics*, 13(4): 403–37. doi:10.1075/ijcl.13.4.02gri.

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro, R., et al. (eds), *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24*. Sendai: Waseda University, 389–98. https://halshs.archives-ouvertes.fr/halshs-00549764/en.

Henny, U. (ed.) (2017). Collection of 19th Century Spanish-American Novels (1880-1916). (CLiGS Textbox). Würzburg: CLiGS. https://github.com/cligs/textbox/tree/master/spanish/novela-hispanoamericana.

Hoover, D. L. (2010). Teasing out Authorship and Style with t-tests and Zeta. *Digital Humanities Conference*. London http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1): 97–133. doi:10.1075/ijcl.6.1.05kil.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. and Mannila, H. (2014). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31(2): 374–97. doi:10.1093/llc/fqu064.

Lyne, A. A. (1985). *Dispersion. The Vocabulary of French Business Correspondence*. Paris: Slatkine / Champion, pp. 101–24.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora*. Hong Kong: ACM, 1–6.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5): 503–20.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology*, 17(4): 688–90. doi:10.1093/beheco/ark016.

Schöch, C. (2018). Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie. In Bernhart, T., et al. (eds.), *Quantitative Ansätze in der Literatur- und Geisteswissenschaften*. Berlin: de Gruyter. 77-94. https://www.degruyter.com/viewbooktoc/product/479792.

Schöch, C., Calvo Tello, J., Henny-Krahmer, U. and Popp, S. (under review). The CLiGS textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI. *Journal of the Text Encoding Initiative*. Preprint: http://cligs.hypotheses.org/files/2017/09/Schoech-et-al_2017_Textbox.pdf.

Scott, M. (1997). PC Analysis of Key Words and Key Key Words. System, 25(2): 233–45.

From print to digital: A web-edition of Giacomo Leopardi's Idilli

Desmond Schmidt
desmond.allan.schmidt@gmail.com
Queensland University of Technology, Australia

Paola Italia
paolaitalia3@gmail.com
Università di Bologna, Italy

Milena Giuffrida
milenagiuffrida@gmail.com
Università degli studi di Catania, Italy

Simone Nieddu
nieddu.sim@gmail.com
La Sapienza, Università di Roma, Italy

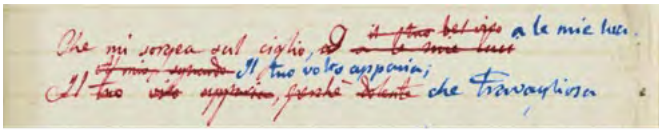
Giacomo Leopardi (1798-1837) was a significant Italian romantic poet best known for a volume of poetry, *Canti*, published in 1835 in Naples, one copy of which he manually corrected shortly before his death. Among his poems, the series entitled *Idilli* 'Idylls' (1819-1821) comprises *Alla Luna, L'infinito, Lo spavento notturno, La sera del giorno festivo, Il sogno* and *La vita solitaria*. In the earliest extant form in the Naples Notebook the poems were written in three separate phases (Idylls 1-3, 4-5 and 6), then corrected by the author in identifiably different pens, before being copied into the Visso manuscript, where they were again revised. Their first publication was in the review 'Nuovo Ricoglitore' in 1825/1826, then in the Bologna edition of *Versi* (1826). Except for the third Idyll: *Lo spavento notturno*, the other poems were collected in the Florence edition of the *Canti* (1831), then reprinted in the Naples edition (1835), in which *Lo spavento notturno* appears as one of the *Fragments* (n. XXXVII).

This textual history of Leopardi's *Idilli* is interesting because much of Italian Philology of the Author (*Variantistica*) has been based upon it. After Francesco Moroncini's first critical edition of the *Idilli* (1927), three other editions have appeared: Peruzzi (1981), De Robertis (1984) and the latest by Franco Gavazzeni, printed by Accademia della Crusca (Leopardi, 2009). This is both a critical edition of the manuscripts and the print-versions. It contains detailed textual notes recording all handwritten changes including marginal alternative readings, which are characteristic of Leopardi's method of correction. Four different pens have been identified in 'Naples notebook' of the *Idilli,* before their copying into the Visso manuscript: A, B, C and D, which were used both for writing the new texts and for correcting the earlier idylls. For example, Leopardi wrote idylls 4 and 5, then corrected idylls 1, 2 and 3. With pen C he then wrote idyll 6, and then finally corrected all the preceding ones.

A wiki edition of Leopardi was also produced as part of an advanced course in Italian Literature in 2016-2017 at the University of Rome La Sapienza (Giuffrida and Nieddu, 2017; Caterino and Nieddu, 2017) and was based on Gavazzeni's critical print edition of the *Canti*.

The technical limitations of the wiki software used (MediaWiki), led the group of editors to seek better ways to encode and present the text. They were also interested in producing editions of other authors, and to have the capacity to expand the Leopardi edition in future. As a result of a lecture given by Paul Eggert in October 2017 at the University of Bologna, Eggert suggested the use of the Ecdosis editing system which had already been developed for the Australian romantic poet Charles Harpur. Other tools, such as Tapas (Bauman et al., 2017), EVT (**Rosselli Del Turco et al., 2014**), CollateX (Dekker and Middell, 2017) and Juxta (n.d.) had also been considered but in spite of their individual strengths none provided the comprehensive web-based editing system for modern manuscripts that the editors were seeking.

Leopardi's Idylls were chosen as the basis for the pilot project. Although it was designed as a general set of web-tools, this was the first time that Ecdosis had been applied to another editorial project. Conversion of the print and wiki editions into Ecdosis took only 12 days of part-time work. Since Ecdosis's own WYSIWYG editor was still incomplete it was decided to encode the text in XML and import it. The XML files were created by copying from the wiki edition and from a PDF of the print edition (Italia, 2018). The detailed textual notes of the print edition were used to create separate files for each of the identifiable versions. Within each version changes were encoded with the usual <add> and <del> codes and our own <undeleted> code for earlier alternatives that were not cancelled. The importation process split the corrections and their contexts into separate layers, amalgamating local levels of correction into coherent sub-versions. Although these layers were never written by the author they are still useful as a storage mechanism to record local changes within a version. In this way individual layers remain simple, needing only a few codes to denote changes in format, such as lines, headings and stanzas, since all deletions and insertions have already been converted into layers. Figure 1 shows an example of how this process works for a segment of Idyll 3, La Luna, from the Naples notebook. Finally, the separated versions and layers of each poem were stripped of their remaining markup, which in Ecdosis is stored separately and only recombined with the text for display, so reducing each version/layer to a readable plain text file. This greatly simplifies all subsequent text processing such as searching, comparing and hyphenating when compared to the complexity of the original XML.



A-layer-1: a le mie luci | Il tuo viso apparia, perchè dolente
A-layer-final: il tuo bel viso | Al mio sguardo apparia, perchè dolente
B-layer-final: a le mie luci | Il tuo volto apparia; che travagliosa

Figure 1: Falsely colourised portion of Naples notebook showing pen A (red) and pen B (blue)

Hyphenation of XML encoded texts is quite difficult due to the variety of tags that may occur between two halves of a word. (Bauman, 2016). Our tests of search engines on major digital scholarly editions revealed that literal searches (often more useful than keyword searches) do not work across internal variant boundaries (<add>,<del>,<rdg> etc). Comparison between TEI-XML files con-

taining inline variants is still an open problem that requires human intervention (Bleeker, 2017).

During importation the page-images were also linked to the text. The manuscript images are copyright to the National Library, Naples, but it is anticipated that permission to publish these with the edition will be obtained soon. For the moment the site is protected by the same passsword as for the wiki. The linking produces a list of images which scrolls in sync with the text to keep the top and bottom of each page-image aligned with its corresponding position in the text. To provide a smooth transition and accurate alignment between text and image when scrolling all other possible alignment positions are calculated in proportion to these fixed points. Changing the version loads the relevant images on the left hand side of the screen and the corresponding text on the right. Layers within a version can be changed by clicking on a tab above the text, and changes between layers are highlighted. This removes altogether the need for a 'diplomatic' display where changes are displayed awkwardly above or below the line using inline formats.

In Compare View differences between versions and layers are shown at the character level: deletions on the left-hand version/layer in red and additions on the right-hand version/layer in blue. When displaying a layer the invariant text is shown in grey to indicate that this is not a true version, but true versions and final layers are displayed in black. Scrolling is also synchronous and aligned left to right, regardless of differences in length between the two currently displayed versions.

Table View resembles the traditional critical apparatus. Differences and similarities between versions/layers are arranged in columns. Versions/layers can be excluded from comparison and the table rebuilt. Also versions can be moved up or down the display to explore specific clusters of variation for editorial purposes. The establishment of a reading text from this information could be encoded as another version and added to each poem as a default text.

The pilot edition has mostly been a success (Leopardi, 2018). There is still some difficulties with the encoding of marginal alternatives which cannot be placed with certainty in the text. These will probably be encoded as annotations instead. Another problem is that the modules in Ecdosis and the website itself currently resemble too closely those of the Charles Harpur Critical Archive (Eggert, 2018) and will therefore need significant customisation. The possibility to export the entire contents of the edition to a simple collection of files in nested folders, encoded in standard HTML and plain text, has mitigated initial concerns about 'lock-in'. It is hoped to use other Ecdosis tools to expand the edition in future by increasing the number of poems and placing them in the context of research into the issues and people of Leopardi's day.

References

Bauman, S. (2016). The Hard Edges of Soft Hyphens. https://www.balisage.net/Proceedings/vol17/html/Bauman01/BalisageVol17-Bauman01.html Accessed 2 May 2018).

Bauman, S., Clark, A., Quinn, B, Flanders, J., Hamlin, S. and Zoller, E. (2017). Tapas Project. http://www.tapasproject.org/ (accessed 30 April 2018).

Bleeker, E. (2017). Scholarly Intervention in Automated Collation Software, ESTS Book of Abstracts 2017 https://textualscholarship.files.wordpress.com/2017/11/book_of_abstracts_ests_2017.pdf (accessed 2 May 2018).

Caterino, M. and Nieddu, S. (2017). Wiki Leopardi. http://wikileopardi.altervista.org/wiki_leopardi/index.php?title=NR26_Edizione_critica, user: wiki_leop, password: leopardi (accessed 30 April 2018).

Dekker, R. and Miiddell, G. (2017). CollateX – Software for Collating Textual Sources. https://collatex.net/ (accessed 30 April 2018).

Eggert, P. (2018). Charles Harpur Critical Archive, http://charles-harpur.org (accessed 30 April 2018).

Italia, P. (2008). I tre tempi degli «Idilli» Leopardiani (con un'edizione del quaderno napoletano), Filologia Italiana, 4, 173-213.

Juxta. http://www.juxtasoftware.org/.

Leopardi, G. (2009). *Canti e Poesie disperse*, edizione diretta da Franco Gavazzeni, edited by Paola Italia, Florence: l'Accademia della Crusca.

Giuffrida, M. and Nieddu, S. (2017). Wiki Critical Editions: a sustainable philology, in AIUCD 2017 Conference, 23-28 January, pp. 257-258. http://aiucd2017.aiucd.it/wp-content/uploads/2017/01/book-of-abstract-AIUCD-2017.pdf (accessed 30 April 2018).

Rosselli Del Turco, R., Buomprisco, G., Di Pietro, C., Kenny, J., Masotti, R. and Pugliese, J. (2014). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions, JTEI 8. https://jtei.revues.org/1077 (accessed 30 April 2018).

Leopardi, G. (2017). Idilli di Giacomo Leopardi. http://giacomo-leopardi.me (accessed 30 April 2018).

Designing Digital Collections for Social Relevance

Susan Schreibman
susan.schreibman@gmail.com
Maynooth University, Ireland

Digital Humanities, and by extension digital humanists, tend towards a culture of open access, interdisciplinary collaboration, and a maker ethos. These disciplinary values position the digital humanities for high impact reaching beyond disciplinary boundaries into more public fora. One might argue that this public-facing ethos is a natural extension of web-based scholarship.

Yet, simply putting resources on the web does not necessarily engage the public or publics they wish to reach. With institutions, research bodies, and funding agencies expecting greater impact from research (see, for example, Watermeyer, Ozanne, Reale), digital humanities scholarship is increasingly being viewed as an answer to that perpetually thorny crisis in the humanities. In his 2010 article, 'The Engaged Humanities: Principles and Practices of Public Scholarship and Teaching', Gregory Jay believes that public scholarship and community engagement will become central to revitalizing the humanities in the 21st century. He argues that the future of the humanities depends upon two interrelated innovations: the organized implementation of project based engaged learning and scholarship, and the continued advancement of digital and new media learning and scholarship (51)

Further, he writes, 'efforts to connect humanities research and teaching with projects to advance democracy, social justice, and the public good might take advantage of the latest episode of the crisis in the humanities and even represent a new direction for revival (51). A means for advancing new values within our teaching and research is through the development of projects of social relevance which engage the public in their design and implementation. These go by various names: crowdsourcing, participatory engagement, and social engagement.

Crowdsourcing is a popular term that has been used for over a decade. Defined by Mia Ridge in *Crowdsourcing our Cultural Heriage* as '...the act of taking work once performed within an organisation and outsourcing it to the general public through an open call for participants' (1). Crowdsourcing structures public participation more as what the public can do for my project as opposed to understanding why the public might choose to spend their creative capital on my research. This may lead to a framing of public participation as project work that gets outsourced to those with lesser skills than individuals within the academy and/or on the project team. This can lead to a bifurcation of a them vs us mentality, with the 'them' (the public) not as educated, talented, or resourceful as the 'us' within the academy.

This very point was articulated in a thread on the Text Encoding Initiative list in February 2016. After some 15 positive responses about successful crowdsourced projects, one respondent asked why were we not hearing about failures. This quickly morphed into another thread with the subject line 'Crowdsourcing Transcription Failures'. In this thread respondents posted a number of issues and challenges in carrying out these projects (as well as some advice by others on the list in how to address these). One respondent, however, indicated that it was not feasible to think about public participation at the same level as that from those on the project team. And while in principle, this is a reasonable assumption, the articulation of this particular post starkly drew the us vs them line with the us not having the 'patience' to clean up the mess left by the them:

> what we call 'failure' may simply be a matter of impatience. If we expect to do the equivalent of a 'barn-raising' in digital humanities, where a large number of people come together and do a lot of tedious work quickly, we have to expect a lot of 'mopping up' to do afterwards. And we're limited by our level of patience, in how far we want to go to train people to mark texts in thoughtful, observant, well-informed ways. (Beshero-Bondar).

This attitude reinforces public perceptions of academics existing in an ivory tower in which research can only be undertaken by a scarce, specialised work force. Participatory engagement projects, on the other hand, begin to debunk notions of us (the experts) against them (the amateurs [or the public]). Participatory engagement, on the other hand, frames involvement on the project differently. It is a political and a design issue, in addition to being a research decision. The participatory process is one that generates new thinking about the research process, audience, and value while a vehicle to challenge limiting beliefs of who our scholarship is for and the role of the humanities as a public good in society. The National Centre for Public Engagement identifies the public as not simply an extended work force, but active players in the design process:

> Public engagement describes the myriad of ways in which the activity and benefits of higher education and research can be shared with the public. Engagement is by definition a two-way process, involving interaction and listening, with the goal of generating mutual benefit.[3]

In his 2010 book *Cognitive Surplus*: *Creativity and Generosity in a Connected Age* Clay Shirky observes how the Internet changes the way we spend our spare time. The so-called 'cognitive surplus' that used to be spent on passive activities (notably watching television) can now be used in a profoundly different way, for new kinds of creativity and problem-solving. He writes, 'the wiring of humanity lets us treat free time as a shared global resource, and lets us design new kinds of participation and sharing that can take advantage of that resource'. (39)

Participatory Engagement projects provide us with opportunities to rethink our roles as researchers and as teachers, about our obligations to those in society who have not had the same opportunities as we have, and last, but not least, how to build meaningful, socially relevant, digital collections for our own and future generations.

Yet, these types of projects have different lifecycles, require different staffing and skills, and come with different obligations than more traditional DH projects, and many project teams are not prepared for this. This talk will explore the motivations of the public in participating

3 https://www.publicengagement.ac.uk/explore-it/what-public-engagement

in our scholarship,our responsibilities in inviting them as collaborators, and new ways to think about the goals, motivations, and audiences for our research.

References

Boshero-Bondar, Elisa. (2016) 'Crowdsourcing Transcription Failures. TEI-L@Listserv.brown.edu . https://listserv.brown.edu/archives/cgi-bin/wa?A2=ind1602&L=tei-l&F=&S=&P=86710

Jay, Gregory. (2010) 'The Engaged Humanities: Principles and Practices of Public Scholarship and Teaching' *Imagining America*. 3:1 p.51-63. http://surface.syr.edu/ia/15

Ozanne, J.L. et al. (2017) 'Assessing the Societal Impact of Research: The Relational Engagement Approach'. ^Journal of Public Policy and Marketing. 36:1. P1-14. https://doi.org/10.1509/jppm.14.121

Reale, E, et al. (2017) 'A review of literature on evaluating the scientific, social and political impact of social sciences and humanities research'. *Research Evaluation.* P1-11. https://doi.org/10.1093/reseval/rvx025

Ridge, Mia. (2014) *Crowdsouring our Cultural Heritage*. Farmham. Surrey: Ashgate.

Shirky, Clay. (2010) *Cognitive Surplus*: *Creativity and Generosity in a Connected Age*. New York: Penguin Books.

Watermeyer, Richard. (2014) 'Impact in the REF: issues and obstacles'. *Studies in Higher Education*. 41:2. p199-214.

The Digitization of "Oriental" Manuscripts: Resisting the Reinscribing of Canon and Colonialism

Caroline T. Schroeder
carrie@carrieschroeder.com
University of the Pacific, United States of America

The past decade has witnessed a wave of manuscript digitization projects initiated by museums, libraries, and individual scholars. This paper will address digitization of some primary sources essential for the study of late antiquity and Byzantine history and religion. Many of these initiatives will advance the study of Greek and Latin texts, as well as Hebrew—the primary languages of the Christian canon and the early to medieval Christian tradition in the West. Research in Syriac, Coptic, and Christian Arabic, however, are essential for understanding the development of religion in the late antique and early Medieval or Byzantine periods. The digitization of their sources has lagged behind. Focusing specifically on Coptic manuscripts—the texts of early and medieval Christian Egypt—this paper will explore the role of colonialism in in the history of Coptic archives and how to resist reinscribing both colonial epistemologies and traditional notions of "canon" after the "digital turn" in archival studies.

Digitization has been heralded as a means of increasing access and availability of texts that may be inaccessible for various reasons, including the dispersal or dismemberment of the original archives or repository. Technology is seen as a possible means to reassemble these dismembered texts and archives, to reunite fragments of papyri and codices virtually online. It is also heralded as a way to save texts that still reside in the Middle East, in zones of political, military, or cultural conflict. Finally some scholars hope it will bring more exposure to traditions that up until now have been seen as marginal to the dominant Greek and Latin traditions. This paper will interrogate two premises: first, that digitization can "recover" or "reconstruct" an original, now dismembered ancient or medieval archive; second, that current digitization efforts are disrupting the dominant canonical paradigms in the study of late antique, Byzantine, and Medieval religious history. The paper will argue that digitization cannot fully repatriate, reconstruct, or save damaged or dispersed physical archives. But the digital can transform our relationships with the sources of early Christanities if we pay critical attention to the limits of the digital, so as not to reify colonial archaeological, archival, and canonical practices in the digital realm.

er will first discuss the original collection of Coptic manuscripts in the context of colonial occupation of Egypt, excavations in Egypt, and the antiquities trade. It will then examine the progress, possibilities, and potential problems of digitization initiatives at specific libraries and museums with significant Coptic collections: British Library, Vatican, Bibliothèque Nationale, Österreichische Nationalbibliothek, etc. The paper will also analyze the work of specific digital humanities projects in Coptic (particularly Coptic Scriptorium at the University of Pacific and Georgetown University, PATHs at Sapienza University in Rome, and the virtual Hill Museum and Manuscript Library) as well as the efforts of the Coptic cultural heritage organization the St. Shenouda Foundation to collect microfilms and digital images for diasporic Coptic cultural heritage preservation.

The paper draws on insights from post-colonial digital humanities, Native American digital humanities (especially regarding issues of repatriation and digitization of cultural heritage), archival theory, Coptic Studies, and manuscript studies.

References

el Salam, Shadi Abd. A*l Mummia (The Night of Counting the Years)*, 1969.

Adler, Eric. *Classics, the Culture Wars, and Beyond*. Ann Arbor: University of Michigan Press, 2016.

Bell, Joshua A., Kimberly Christen, and Mark Turin. "Introduction: After the Return." *Museum Anthropology Review* 7.1-2 (2013): 1–21. Print.
Brier, Bob. *Egyptomania: Our Three Thousand Year Obsession with the Land of the Pharaohs*. New York: Macmillan, 2013.
Davis, Stephen J., Gillian Pyke, Elizabeth Davidson, Mary Farag, and Daniel Schriever. "Left Behind: A Recent Discovery of Manuscript Fragments in the White Monastery Church." *Journal of Coptic Studies*, 16 (2014): 69–87.
Derrida, Jacques. Archive Fever: *A Freudian Impression. Translated by Eric Prenowitz*. Chicago: University Of Chicago Press, 1995.
Earhart, Amy E. *Traces of the Old, Uses of the New: The Emergence of Digital Literary Studies*. Ann Arbor: University of Michigan Press, 2015.
Emmel, Stephen, and Cornelia Eva Römer. "The Library of the White Monastery in Upper Egypt/Die Bibliothek Des Weißen Klosers in Oberägypten." In *Spätantike Bibliotheken: Leben Und Lesen in Den Frühen Klöstern Ägyptens, by Harald Froschauer and Cornelia Eva Römer*, 5–25. Nilus: Studien zur Kultur Ägyptens und des Vorderen Orients 14. Vienna: Phoibos Verlag, 2008.
Hering, Katharina. "Digital Historiography and the Archives." *Journal of Digital Humanities*. N.p., 1 Nov. 2014. Web. 13 Nov. 2014.
---. "Provenance Meets Source Criticism." *Journal of Digital Humanities*. N.p., 4 Aug. 2014. Web. 13 Nov. 2014.
Gad, Usama. "The Digital Challenges and Chances: The Case of Papyri and Papyrology in Egypt." Ed. Monica Berti and Franziska Naether. *Egyptology, Papyrology and beyond proceedings of a conference and workshop in Leipzig*, November 4-6, 2015 (2016): n. pag. www.qucosa.de. Web. 19 May 2016.
Goehring, James E. "The Dark Side of Landscape: Ideology and Power in the Christian Myth of the Desert." In T*he Cultural Turn in Late Ancient Studies: Gender, Asceticism, and Historiography*, edited by Dale B. Martin and Patricia Cox Miller. Durham, NC: Duke University Press, 2005.
Koh, Adeline. "Inspecting the Nineteenth-Century Literary Digital Archive: Omissions of Empire." *Journal of Victorian Culture* 19, no. 3 (July 3, 2014): 385–95.
Maspero, Gaston. Fragments de la version thébaine de l'Ancien testament. Vol. 1. 5 vols. *Mémoires publiés par les membres de la Mission Archéologique Française au Caire 6*. Paris: Ernest Leroux, 1892.
———. "Rapport Sur La Trouvaille de Deir El Bahari." *Bulletin de l'Institut Egyptien 2 Series 2 (1881)*: 129–69.
Putnam, Lara. "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast." *The American Historical Review* 121/2 (2016): 377–402.
Schroeder, Caroline T., and Amir Zeldes "Raiders of the Lost Corpus" *Digital Humanities Quarterly* 10.2 (2016)
Schroeder, Caroline T. "The Digital Humanities as Cultural Capital: Implications for Biblical and Religious Studies," *Journal of Religion, Media, and Digital Culture* 5:1 (2016)
Shetty, Sandhya, and Elizabeth Jane Bellamy. "Postcolonialism's Archive Fever." *Diacritics 30/1 (2000)*: 25–48.
Stoler, Ann Laura. *Along the Archival Grain: Epistemic Anxieties and Colonial Common Sense*. Princeton: Princeton University Press, 2010.
Wharton, Annabel Jane. *Architectural Agents: The Delusional, Abusive, Addictive Lives of Buildings*. Minneapolis: University of Minnesota Press, 2015.

A Deep Gazetteer of Time Periods

Ryan Shaw
ryanshaw@unc.edu
University of North Carolina at Chapel Hill, United States of America

Adam Rabinowitz
arabinow@utexas.edu
University of Texas at Austin, United States of America

Patrick Golden
ptgolden@email.unc.edu
University of North Carolina at Chapel Hill, United States of America

> "All of us, even when we think we have noted every tiny detail, resort to set pieces which have already been staged often enough by others."
> — W. G. Sebald, *Austerlitz*

Periods are the set pieces of history, and their staging is a strategy for making change over time meaningful and understandable. Periodization structures not only histories themselves, but also the ways those histories are organized in libraries, the ways teachers of history organize syllabi and textbooks, and the ways historians organize themselves in academic institutions. Like the histories they structure, periodizations are also imposed on the conquered by their conquerors. Periodization itself is a legacy of colonialism, grounded in a linear ontology of time that has forced aside indigenous understandings of temporality. Periodization is also a perennial topic for reflection in the humanities, as scholars cast a critical eye on the categories that organize their work. But like a linear conception of time, periodization is both easily critiqued and difficult to relinquish.

Critique is not the only response to periodization: several scholars have suggested alternative approaches to conceptualizing historical temporality. Wishart (2004: 313), responding to histories of the Plains Indians that

"fold their ethnographies into periods that are derived from American, not indigenous, realities," suggests as alternatives periodizations grounded in economic cycles or patterns of population change. Dimock (2001: 758) proposes abandoning the "decades and centuries" scale of conventional literary periods in favor of a "deep time" of "extended and nonstandardized duration." Others explicitly consider the role of the digital humanities in realizing alternatives to periodization. Brooks (2012) claims that "the digital world is moving in concert with Indigenous literary traditions" (312) and foresees that, as scholars embrace digital media, "the measuring tape of time will become decreasingly useful and, perhaps, increasingly (self)destructive" (309). Underwood (2013) argues that the penchant for periodization among literary scholars stems not from a desire to neatly sort history into standardized bins, but from a disciplinary identity rooted in theories of discontinuity and rupture. He sees the digital humanities as challenging that identity by providing tools and vocabulary for describing gradual, continuous change.

Besides critique and the imagining of alternatives, a third response to periodization is to document it. This is the motivation behind PeriodO (http://perio.do), a gazetteer of scholarly definitions of historical periods. Gazetteers are typically directories of place names, but understood broadly, any reference tool documenting named concepts that can be spatiotemporally located is a gazetteer (Shaw, 2016: 58). The PeriodO gazetteer documents specific published assertions about periods, including their names, their extent in space and time, and when, where, and by whom these assertions were made. Unlike gazetteers focused primarily on standardization, PeriodO is a *deep* gazetteer which attempts to document a range of perspectives taken and judgments made (Shaw, 2016: 58–60). Hence there is not a single "Bronze Age" in PeriodO, but hundreds (Figure 1).

PeriodO is published as "linked data," providing for the documented concepts stable identifiers in the form of URLs, which can then be resolved into sets of "triples"—subject-predicate-object structures representing assertions about those concepts. As of November 2017, there are over five thousand periods documented in PeriodO, from more than one hundred sources in over twenty languages. For each of these periods, the assertions documented include structured bibliographic data describing the source, temporal extent as delimited by up to four points in time, and spatial coverage via links to places in other linked data gazetteers. PeriodO has been designed to be collaboratively edited by a community of scholars, regardless of whether they have any knowledge of or experience with linked data technologies (Shaw et al., 2015). Anyone with a free ORCID personal identifier (Haak et al., 2005) can immediately submit proposed additions to the gazetteer, without any additional barriers to contribution.

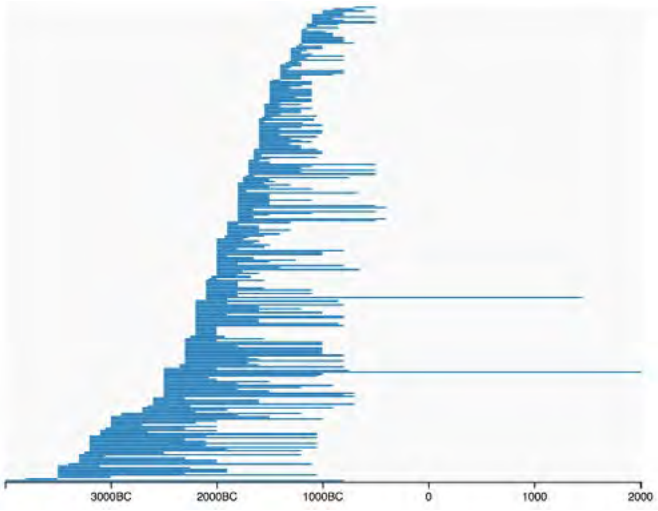


Figure 1. Visualizing the temporal extents of hundreds of different "Bronze Age" concepts

Through its public domain identification and documentation of period concepts, PeriodO provides a means by which curators of periodized data can resolve ambiguous period terms and bridge datasets employing different interpretations of the same period term. This is much like the service provided by the Pleiades gazetteer of ancient places (Elliott and Gillies, 2008). Pleiades uses PeriodO URLs to identify its period concepts, just as PeriodO uses URLs from place gazetteers like Pleiades to identify its place concepts. Pleiades plays an integral role in an increasingly fecund ecology of linked scholarly data projects, many of them incubated by the community-driven Pelagios initiative (Isaksen et al., 2014). The Peripleo spatiotemporal search and visualization tool, built to search over and visualize data produced by the projects participating in this initiative, indexes PeriodO URLs and can use PeriodO data to translate between period terms and spatiotemporal locations (Simon et al. 2016). PeriodO is also used by the ARIADNE archaeological research data infrastructure to document the more than 600 period concepts it employs (Niccolucci and Richards, 2013).

The PeriodO team would like to engage those gathered at the 2018 Digital Humanities conference in Mexico City for several reasons. First, we would like to present the results of four years' iterative development of the PeriodO dataset and tools, funded by consecutive grants from the National Endowment for the Humanities and Institute of Museum and Library Services. This will include the outcomes of two workshops focused on periodization and spatiotemporal knowledge organization, held in August 2016 and December 2017. We hope that an overview of PeriodO's design and implementation will be of interest not only to those working with periodized data, but anyone interested in the architecture of scholarly infrastructure.

Second, we hope to inspire others to use the PeriodO data for purposed other than data curation. Visualizations of PeriodO data could be used to help students understand the nature and politics of periodization, or to make arguments about the history of historiography. Advocates of alternatives to periodization may find PeriodO's documentation useful in the spirit of "know your enemy." A large collection of multilingual descriptions of temporal extent and their corresponding interpretations as numerical ranges may be useful for natural language processing. There are undoubtedly other possibilities, and if the data in its current state is not adequate for exploring them, we'd like to figure out how we can make it so.

Most importantly, we hope to catalyze collaborations with a broader range of scholars interested in documenting periodizations. The majority of period concepts documented in PeriodO originated in archaeology, art history, and the authority files of libraries and museums. We would like to have far more documentation of periodizations from areas such as literary studies, social history, and intellectual history—areas with far less consensus on periodization than archaeology and art history. And while PeriodO documents period concepts associated with places around the world, the majority of its scholarly sources are still American and European—another defect we'd like to correct. A primary goal of PeriodO is to enable contrast of and comparison between different interpretations of the past, and this requires broad collaboration.

As broad as that collaboration may become, PeriodO will always be limited by the framework of linear time that it employs as a means of making temporal extents comparable. Still, there is no reason that PeriodO could not connect with other projects exploring alternative temporalities, in the vein of Drucker's (2009) experiments with relational temporal modeling, Brooks' "spiral" time, or even Underwood's probabilistic, gradient time. Though it may not be possible to directly compare the temporal entities or processes registered by these various alternative conceptualizations, they might still be interlinked and hence more readily brought into dialogue with one another. We hope that our colleagues at DH 2018 will have some ideas about how that could happen, or insights into why it might not.

References

Brooks, Lisa. (2012). The primacy of the present, the primacy of place: navigating the spiral of history in the digital world. *PMLA*, 127 (2): 308–16. doi:10.1632/pmla.2012.127.2.308.

Dimock, Wai Chee. (2001). Deep time: American literature and world history. *American Literary History*, 13 (4): 755–75. doi:10.1093/alh/13.4.755.

Drucker, Johanna. (2009). Temporal modeling. In *SpecLab*, 37–64. University of Chicago Press. doi:10.7208/chicago/9780226165097.003.0003.

Elliott, Tom, and Sean Gillies. (2011). Pleiades: an un-GIS for ancient geography. In *Digital Humanities Conference 2011*. Stanford. http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-192.xml.

Haak, Laurel L., Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, 25 (4): 259–64. doi:10.1087/20120404.

Isaksen, Leif, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares. (2014). Pelagios and the emerging graph of ancient world data. In *Proceedings of the 2014 ACM Conference on Web Science - WebSci '14*, 197–201. New York: ACM Press. doi:10.1145/2615569.2615693.

Niccolucci, Franco, and Julian D. Richards. (2013). ARIADNE: advanced research infrastructures for archaeological dataset networking in Europe." *International Journal of Humanities and Arts Computing*, 7 (1–2): 70–88. doi:10.3366/ijhac.2013.0082.

Shaw, Ryan. (2016). Gazetteers enriched: a conceptual basis for linking gazetteers with other kinds of information. In *Placing Names: Enriching and Integrating Gazetteers*, edited by Merrick Lex Berman, Ruth Mostern, and Humphrey Southall. Bloomington, Indiana: Indiana University Press.

Shaw, Ryan, Adam Rabinowitz, Patrick Golden, and Eric Kansa. (2015). A sharing-oriented design strategy for networked knowledge organization systems. *International Journal on Digital Libraries*, 17 (1), 49–61. doi:10.1007/s00799-015-0164-0.

Simon, Rainer, Leif Isaksen, Elton Barker, and Pau de Soto Cañamares. (2016). Peripleo: a tool for exploring heterogeneous data through the dimensions of space and time. *code{4}lib*, no. 31. http://journal.code4lib.org/articles/11144.

Underwood, Ted. (2015). *Why Literary Periods Mattered.* Stanford University Press.

Wishart, David. (2004). Period and region. *Progress in Human Geography*, 28 (3): 305–19. doi:10.1191/0309132504ph488oa

Feminismo y Tecnología: Software Libre y Cultura Hacker Como Medio Para la Apropiación Tecnológica

Martha Irene Soria Guzmán
idem24@gmail.com
UAM-Xochimilco, Mexico

Pensar la mancuerna mujeres y tecnología en tiempos de capitalismo cognitivo, vigilancia masiva, violencia en línea, hiper mediación, control por parte de las empresas proveedoras de servicios y fabricantes de *software* y más aún, atravesado por el planteamiento político que supone el feminismo y las humanidades digitales, requiere de una problematización en torno al dominio de la técnica y cómo ésta ha sido negada ancestralmente a las mujeres,

los obstáculos para un conocimiento profundo de la tecnología y los alcances de una posible autonomía tecnológica en el movimiento feminista del siglo XXI, que libra una nueva lucha en el entorno digital.

Para plantear las diversas problemáticas, comenzaré hablando sobre algunos postulados que sostienen que existe una relación estrecha entre los instrumentos usados tradicionalmente por ambos sexos, la división del trabajo y el dominio de los hombres sobre las mujeres.

En este sentido, Paola Tabet sostiene que a las mujeres les fue negada ancestralmente la posibilidad de extenderse más allá de sus propias fuerzas físicas, de la capacidad de sus manos, de prolongar su cuerpo y sus brazos en instrumentos complejos que acrecienten su poder sobre la naturaleza, lo cual ha sido condición necesaria para que sean "usadas" materialmente en el trabajo y la reproducción. (Tabet, 2005: 67)

Posteriormente, en el siglo XVI, cuando ya el modelo de explotación capitalista fue una realidad que respondió a la crisis del sistema feudal, aparecen nuevos adelantos técnicos vinculados a la minería y a la devastación natural, así como con la aparición de una excesiva preocupación por el tiempo y el espacio. El capitalismo trae como consecuencia que el tiempo sea dinero, el dinero sea poder y el poder exija fomentar el comercio y poseer los medios de producción (Ruiz Ordóñez, 1998: 43). Capitalismo y técnica son interdependientes. Las máquinas de este período de invención y producción, producen beneficios a particulares (44), es decir, a los dueños de los medios de producción .

En este mismo sentido, Silvia Federici explica cómo para el desarrollo del capitalismo, fue fundamental la construcción de un nuevo orden patriarcal, que hacía que las mujeres fueran sirvientas de la fuerza de trabajo masculina. La acumulación primitiva y la subsecuente división sexual del trabajo fue, "sobre todo, una relación de poder, una división dentro de la fuerza de trabajo, al mismo tiempo que un inmenso impulso a la acumulación capitalista" (Federici, 2010: 176).

Lo anteriormente desarrollado, deviene en una primer problemática, que supone que dentro del sistema capitalista, el alejamiento de las mujeres del dominio de la técnica, es evidente y marcado, lo cual eventualmente implicó que las mujeres fuéramos despojadas de determinado tipo de "saber".

Luego de la invención de las máquinas alimentadas por carbón, seguidas por las alimentadas por vapor y finalmente por electricidad, surge un nuevo tipo de "máquina" a principios del siglo XX que parte del surgimiento de la cibernética hacia los años cuarenta. La cibernética busca entender cómo piensan los seres humanos para hacer luego "pensar" a la máquina. Esto da origen al cómputo moderno en los cuarenta, el cual requiere de una serie de pasos para ejecutar una acción: los algoritmos. Es hasta poco antes de la década de los sesenta que dentro del trabajo de la comunidad científica, particularmente de matemáticos y físicos, se crean los primeros programas de cómputo, los cuales representan un importante cambio de paradigma en el desarrollo y ejecución de la técnica, y que marcan el rumbo de la tecnología digital que usamos hoy en día.

En este punto, es necesario comprender qué es un *software* o programa de computadora. Richard Stallman (2004) usa la metáfora de una receta de cocina, ya que el *software* es un conjunto de recetas minuciosamente detalladas para la solución de un determinado problema, que puede ir desde hacer una suma hasta escribir una carta, crear un vector o editar un video. Dichas recetas están escritas de una manera muy parecida a cómo la música se escribe usando notaciones propias, a lo cual podemos llamarle lenguajes formales que son con los que están escritos los programas.

En la década de los setenta, era muy común compartir programas entre las personas programadoras y con ello, pedir y ofrecer parte del **código fuente** para mejorarlo entre todos. Sin embargo, en la década de los ochenta, algunas empresas pioneras de computación crearon programas que no pudieron compartirse. Algunas computadoras modernas de la época comenzaban a tener su propio sistema operativo para el cual, se necesitaba firmar un acuerdo de confidencialidad para obtener una copia ejecutable (Stallman, 2004: 21)

Paulatinamente, la cultura del uso y compartición del código fuente se transformó en su privatización, volviéndolo cerrado y desembocando luego en la creación y uso de patentes. Comenzó con ello la era de la comercialización de un *software* que no tenía el código fuente abierto para que todas las personas usuarias lo conocieran y estudiaran, por el contrario, se trataba de un código cerrado que representó un velo que impidió saber cómo fue hecho. Se trata del inicio de la era del uso de una caja negra como herramienta tecnológica.

Éstas y otras implicaciones (como el hecho de que este tipo de *software* es propiedad privada) han hecho que un grupo de personas opten por llamarle *software* **privativo** para subrayar la definición de un programa que "priva" libertades y ofrece ciertas limitaciones. Lo que es cierto es que se ha convertido en un *software* comercial y hegemónico usado de manera muy habitual, normalizado y poco cuestionado.

Por el contrario, algunas personas han decidido crear y optar por el camino del *software* **libre**, el cual permite no sólo que los usuarios conozcan las líneas de código con el que se ha hecho el programa, sino que sea estudiado, copiado, distribuido y mejorado, convirtiéndose así en una alternativa para conocer el interior de la caja negra, para concebir la tecnología desde la perspectiva del **código abierto**, donde los 'saber-hacer' y por lo tanto la técnica está al descubierto, donde cualquier persona que pueda leer o estudiar el código también pueda modificarlos; volviendo a la metáfora de la receta de cocina, la posibilidad de saber los ingredientes y los pasos para la elaboración nos permitiría contrarrestar la dependencia al *software* hegemónico y quizá fomentar la autonomía tecnológica.

En este sentido, existe un grupo de personas en el que observo una autonomía tecnológica altamente desarrollada, así como un conocimiento profundo de su equipo de cómputo. Esta autonomía les ofrece una posibilidad de liberarse de la mediación tecnológica del *software* privativo o hegemónico y la empresa que lo fabrica. Me referiré a este grupo de personas como *hackers*. Cabe señalar que se usará el término "*hacker*" no desde el punto de vista del "pirata informático" con el que popularmente se asocia, sino tomando como referencia a Steven Levy (1984) quien desglosa de manera extensa la "cosmovisión" de la cultura *hacker*, o la ética *hacker*, en su libro: *Hackers: heróes de la revolución informática* (Wolf, 2016). Las personas que denominamos *hacker* desarrollan una autonomía tecnológica y habilidades computacionales especializadas gracias a que conocen el funcionamiento de su equipo de cómputo. Esto ha llevado a considerar a la figura del, *hacker* como disidente en un mundo tecnológico normado, con la capacidad de decidir diversos aspectos de la tecnología que usa. No es gratuito entonces que en la comunidad *hacker* sea muy habitual el uso y programación de *software* libre, ya que ambos son manifestaciones tecnopolíticas que históricamente han ido de la mano.

Mientras que las y los ingenieros surgen como figuras que dominan la técnica y hacen funcionar las máquinas en la evolución del capitalismo, la figura del *hacker,* tal como dice Guiomar Rovira "propone hacer ingeniería inversa para conocer cómo funcionan las máquinas que el mercado ofrece como cerradas, para darles otras terminaciones y usos […] Es por ello que la figura del *hacker* se contrapone a la del ingeniero" (Rovira, 2017: 110).

El alejamiento de las mujeres del dominio de la técnica y por lo tanto, el despojo de cierto saber-hacer del que ya he hablado anteriormente, continúa hasta nuestros días, tanto fuera como dentro de la hegemonía tecnológica, ya que la figura masculina del *hacker* es mucho más común y visible que la mujer-*hacker.* Sin embargo, la propuesta *hacker* implica perder el miedo a la máquina que es entendida sólo por unos pocos "implica reapropiarse de las tecno-logías para volverlas técnicas a nuestra disposición y no lógicas de sometimiento (Rovira, 2017: 111) lo cual podría ser condición necesaria para que las mujeres, en tanto sujetos del feminismo, regresemos al dominio de la técnica en pro de una lucha que hoy se libra también en el terreno tecnológico digital. Dicho regreso a la técnica implica que las mujeres nos apropiemos de la tecnología que usamos y ejerzamos autonomía, más allá del papel de "usuarias" o consumidoras pasivas que hemos jugado durante mucho tiempo y eso implicaría el uso de herramientas tecnológicas que puedan ser modificadas, estudiadas, comunitarias y *hackeables*, es decir, *software* libre, pero ahora inmerso en una cultura **hackfeminista**.

References

Federici, S. (2010), *Calibán y la bruja. Mujeres, cuerpo y acumulación primitiva*, Madrid, España, Traficantes de sueños

Hache, A., et all. (2013). Yo programo, tú programas, ella hackea: mujeres hackers y perspectivas tecnopolíticas. *Internet en Código Femenino, teorías y prácticas.* Buenos Aires, La crujía ediciones, pp. 77-90. Hache, A., et all. (2011). *Mujeres programadoras y mujeres hackers. Una aproximación desde Lela Coders*. http://www.rebelion.org/docs/141550.pdf consultado 14 de febrero de 2018.

Levy, S. *(1984) Hackers: Heroes of Computer Revolution, Nueva York, Anchor Press Doubleday.*

Pujol, J. and Montenegro, M. *(2015). Technology and Feminism: A Strange Couple, Revista de Estudios Sociales, no. 51, Universidad de los Andes, enero, 2015, pp. 173–185. https://revistas.uniandes.edu.co/doi/full/10.7440/res51.2015.13, consultada 18 de octubre de 2017.*

Rovira, G. (2017). *Activismo en red y multitudes conectadas, comunicación y acción en la era de Internet,* Ciudad de México, Icaria Editorial / UAM, 2017. Ruiz Ordoñez, Y. (1998) *Lewis Mumford, una interpretación antropológica de la técnica,* Castellón, España, Universitat Jaume I. 1998.

Stallman, R. (2004) *Software Libre para una Comunidad Libre,* Madrid, Traficantes de Sueños.

Tabet, P. (2005). *Las manos, los instrumentos y las armas, Ochy Curiel, Jules Falquet (comp.), El patriarcado al desnudo: tres feministas materialistas, Buenos Aires, Brecha Lésbica, pp. 130-175.* Wolf, G. *(2016). Cifrado e identidad, no todo es anonimato, Irene Soria (coord.) Ética hacker, seguridad y vigilancia, Cuidad de México, UCSJ, 2016, pp.19-65.*

Interpreting Difference among Transcripts

Michael Sperberg-McQueen
cmsmcq@acm.org
Black Mesa Technologies LLC, United States of America

Claus Huitfeldt
Claus.Huitfeldt@uib.no
University of Bergen, Norway

Introduction

The semantics and logic of transcription have received attention from a number of digital humanists, some starting from the practice of digital editions [Pierazzo 2011, 2015], some from a consideration of markup languages [Robinson 1994, Huitfeldt 1995], some from a critical exa-

mination of the foundations of digital humanities [Caton 2009, 2013a, 2013b, 2014].

Attempts to describe how transcripts provide information about their exemplars [Huitfeldt/Sperberg-McQueen 2014, Sperberg-McQueen et al. 2017] have focused on individual transcripts, not multiple divergent transcripts of the same exemplar. Here we describe ways in which transcripts of the same exemplar can differ and we sketch a model of transcription which accounts for such differences.

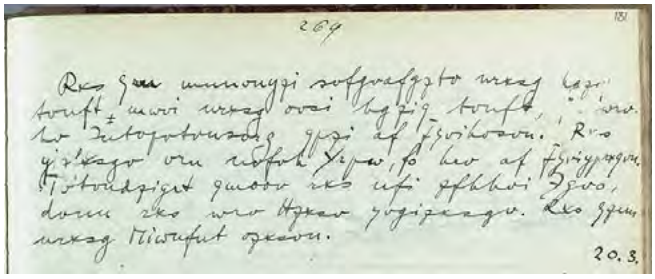
Examples

Our catalog of ways in which transcripts differ and disagree takes the form of examples, many illustrating exceptions to the general rule that a transcript reflects "the exemplar, the whole exemplar, and nothing but the exemplar" and that competent transcribers will agree on the reading of the exemplar [Sperberg-McQueen et al. 2014].

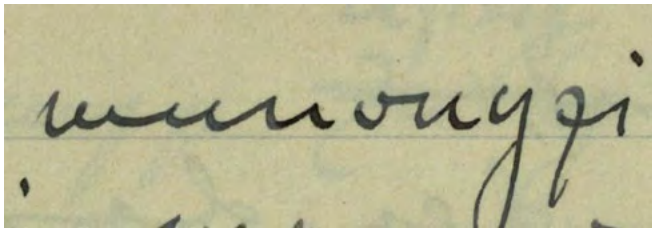
For brevity, examples often consider only single words; discussions refer to the first of several transcripts as A, the second as B, an arbitrary transcript as T, and the exemplar as E. Some transcripts were constructed for this paper.

Transcripts which differ and disagree

Example LW: what type does this token instantiate?



E is a word from Ludwig Wittgenstein's notebook 117 (p. 269), written in a simple substitution code [Wittgenstein, n.d.]. A and B, ignorant of the cipher, transcribe it as "munonyqi" and "wunouyqi", respectively. C, better informed, has "muuvnyzi" ("offenbar").



The transcripts reflect contradictory readings of the token in E; at most one can be correct.

Here all transcripts agree on which marks in E are tokens, but disagree on the types they instantiate. We infer that a transcript's mapping from tokens in E to types is a salient feature for modeling.

Example MCN: which marks are tokens?

E is a tombstone from northwestern Britain [Collingwood/Wright 1965-1990, no. 932].

A [Lafleur 2010 p. 28f.] reads the mark between some word pairs as a punctum:

DIS
MANIB · M · COCCEI
NONNI · ANNOR · VI
HIC SITVS EST

B is similar except for the last line: 'HIC · SITVS · EST'.

Here A and B do not disagree over the reading of any tokens; they disagree on what marks in E are tokens. A formal account must distinguish the identification of tokens from the mapping of tokens to types.

Example TE: what is the structure of this text?

At the eastern end of Magdeburg cathedral lies the Tumba Edithae (tomb of Edith), with an inscription part of which is shown here.

A [Neugebauer/Brandl 2012] begins reading on the south:

(Südseite:) DIVE · REGINE · RO[MA]NOR[UM] · EDIT · ANGLIE · REGIS · ECMVNDI · FILIE · HIC · OSSA · CO[N]DVNTVR · CVIVS · RELIGIOSI (Ostseite:) AMORIS (Nordseite:) INPVLSV · HOC · TE[M]PLV[M] · AB · OTHONE · MAGNO · DIVO · CAESARE · CONIVGE · FV[N]DATV[M] · EST · OBIIT · AN[N]O · CHRISTI (Westseite:) DCCCCX-LVII ·

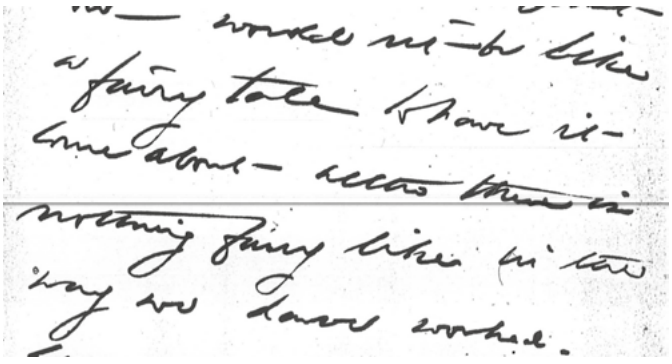
B begins reading on the north with "INPVLSV" but otherwise resembles A.

A and B agree in their readings of each individual character and word and also in identifying which marks in E are writing; they differ only in the higher-level structure(s) compounded from words and characters. A model of transcription must include such higher-level structural organization as a substantive part of transcription; so similarly [Huitfeldt et al. 2010].

Transcripts which differ without disagreeing

That some differences between transcripts do not signal disagreements about E goes (almost) without saying. A and B can differ in pagination and running heads without disagreeing on how to read E: page furniture is normally an exception to the general rule that everything in T transcribes something in E.

Example JA: literal transcription and marked corrections



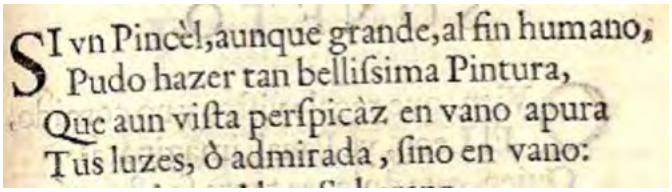
E is a word from a letter of Jane Addams [Hajo et al. 2015-].

A writes "altho", B "altho[*ugh*]" (bracketed italics mark editorial additions), and C "[although]" (brackets mark editorial interventions). A and B thus differ in content but agree that E has "altho". B and C provide the same normalized spelling but provide different (albeit compatible) information about E.

B and C assign special meaning to brackets and bracketed material: unlike other characters, they transcribe nothing in E.

An account of transcription must specify which tokens in the transcript are to be interpreted as transcribing tokens in E and which not.

Example SJ: long and short s



E is one word from a sonnet by Sor Juana Inés de la Cruz [Sor Juana 1700 p. 163].

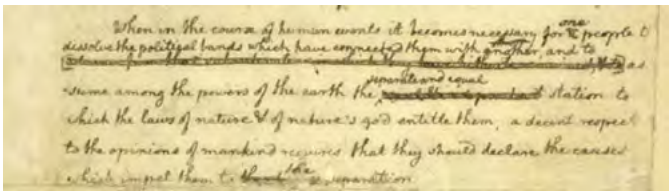
A (a web site presenting Sor Juana's work) writes "vista", B "viſta".

A and B differ but do not disagree. Both identify the third character of this word as a lower-case S; B further specifies a long S. If we take A to be ambiguous (the S could be long or short), then A subsumes B: B provides additional but not contradictory information.

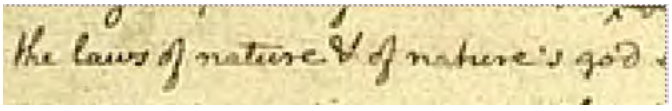
In E, however, long and short S are allographs in complementary distribution; typographic context determines which appears. In vi_ta, S will always be long not short. So in reality A provides the same information as B, not less.

Many differences without disagreement arise where one transcript preserves allographic differences and the other preserves graphemes. Arguments on the topic involve no disagreements about E, only about the choice of type system. A model should clarify the role of type systems in transcription.

Example TJ: word-level and character-level fidelity



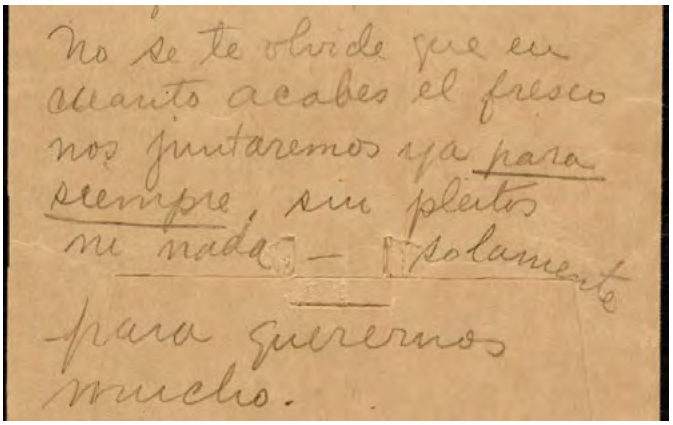
E is from Thomas Jefferson's draft of the U.S. Declaration of Independence.



A: the laws of nature & of nature's god
B: the laws of nature and of nature's God
A [Boyd 1950- p. 1:423] preserves and reinstantiates the type of each character, while B [Harrison 1993 p. 39] preserves and reinstantiates the type of each word but not each character. (That is, it normalizes the spelling of words.)

Here the type systems of A and B diverge even more sharply than above. A formal account must be able to describe type-system differences of this kind.

Example FK: typographic rendering of inscription details



E is the words "para siempre" in a letter from Frida Kahlo to Diego Rivera [Kahlo]. A and B differ only in rendering the underscoring in E as italics or underscoring.

Typographic features of T often convey information about E, but different transcribers use different conventions. A formalization must account for such conventions.

Example FE: completeness and incompleteness

E is a grave marker (Naples, fourth/fifth century CE) now in the Jewish Museum, New York (JM3-50).

A:

HIC POSITVS
EST FLAES EBR
EVS

B:

שלם

HIC POSITVS
EST FLAES EBR
EVS

B transcribes all of E, A [Lafleur 2010 p. 144]] only the Latin text. Any model must describe how we know which material in E is transcribed and which (if any) is not.

JLM: Transcripts which disagree without differing

It is hard to find plausible examples of this class of phenomena. But an imaginary example may illustrate it. If A uses italics to mark editorial insertions, and B to represent underlining in E, then

John *loves* Mary.

will mean different and contradictory things in A and B.

Differences in typographic conventions and type system can lead to conflicting interpretations of T. A model must describe how such conflicts arise.

Formal model

Space constraints limit us to a sketch.

We assume the concepts *type*, *token*, and *document*. Types and tokens are not limited to graphemes or words but include larger structures. The document itself is typically a compound token, and its text a compound type.

A set of mutually exclusive types we call a type inventory. Tokens instantiate exactly one type in an inventory: a letter is an I or a J, but not both. Transcriptions commonly involve not one type inventory but several. ("I" is both a letter and a word.) A set of type inventories is a type system.

A reading of a token k with respect to a type inventory I maps k to a type p in I; we write (k, I, p) for such a reading.

A reading of a document D identifies a set K of tokens in D and maps them to types. We write R = (D, K, P, M), where P is a type system and M a set of triples (k, I, p) where k ∈ K, I ∈ P, p ∈ I. Every k in K maps to at least one type; none maps to two types in the same inventory.

Examples MCN and TE illustrate differences in K, examples SJ and TJ differences in P, example LW differences in M.

Transcription policies determine which tokens in E are transcribed (normal) and which not (special); similarly which tokens in T transcribe E (normal) and which do not (special). They also constrain the type system by distinguishing some types and equating some with each other. A transcription policy is thus a triple (SE, ST, Q), where SE and SE are predicates true of all and only the special tokens in E and T respectively, and Q is a set of type equivalences.

Examples FE and JLM illustrate differences in SE, example JA a difference in ST, and example FK a difference in Q.

From a reading of T we can reconstruct a reading of E based on an assumed transcription policy; this allows readers of T to have information about E without examining E directly.

References

Boyd, J. P., et al. (eds). (1950-). *The Papers of Thomas Jefferson.* Princeton: Princeton University Press. Quoted from (Stevens/Burg 1997), p. 81.

Caton, P. (2009). Lost in Transcription: Types, Tokens, and Modality in Document Representation. Paper given at DH 2009, held June 2009 at College Park, University of Maryland.

Caton, P. (2013a). Pure transcriptional markup. Paper given at DH 2013, held July 2013 at the University of Nebraska in Lincoln.

Caton, P. (2013b). On 'text' in Digital Humanities. *Literary & Linguistic Computing* 28.1 (2013): 209-220.

Caton, P. (2014). Six terms fundamental to modelling transcription. Paper given at DH 2014, held July 2014 at the University of Lausanne. Short version on the Web at http://dharchive.org/paper/DH2014/Paper-780.xml.

Collingwood, R. G., Wright, R. P. (eds). (1965-1990). *The Roman inscriptions of Britain (RIB).* Vol. 1 Oxford: Oxford Univ. Press; Vol 2 Gloucester: Alan Sutton. Image and transcript of RIB 932 reproduced from (Lafleur 2010), pp. 28f.

Driscoll, M. J. (2006). Levels of transcription. In (Unsworth 2006). On the web at http://www.tei-c.org/About/Archive_new/ETE/Preview/driscoll.xml.

Hajo, C. M., et al. (eds). (2015-). *Jane Addams Digital Edition.* Mahwah, NJ: Ramapo College of New Jersey. https://digital.janeaddams.ramapo.edu. The letter cited is from Jane Addams to Florence Kelley, February 1, 1901. https://digital.janeaddams.ramapo.edu/items/show/64.

Harrison, M., and Gilbert, S. (eds). (1993). *Thomas Jefferson Word for Word.* La Jolla: Excellent Books. Quoted from (Stevens/Burg 1997), p. 82.

Hayford, H., Parker, H., and Tanselle, G. T. (eds). (1988). *Moby Dick, or, The Whale.* Vol. 7 of *The Writings of Herman Melville.* The Northwestern–Newberry Edition. Evanston [Ill.]: Northwestern University Press; Chicago : Newberry Library. Rpt. 1994, 1997.

[Huitfeldt, C]. (1993). *MECS-WIT, A registration standard for the Wittgenstein Archives at the University of Bergen.* [Bergen]: Wittgenstein Archives, 1993. Currently on the Web at http://folk.uib.no/fafch/oldstuff/mecswit.html.

Huitfeldt, C. (1995). Multi-dimensional texts in a one dimensional medium. *Computers and the Humanities* 28: 235-241.

Huitfeldt, C. (2006). Philosophy case study. In (Unsworth 2006). On the web at http://www.tei-c.org/About/Archive_new/ETE/Preview/huitfeldt.xml.

Huitfeldt, C., Marcoux, Y., and Sperberg-McQueen, C. M. (2010). Extension of the type/token distinction to document structure. Paper presented at Balisage: The Markup Conference 2010, Montréal, Canada, August 3 - 6, 2010. In Proceedings of Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies, vol. 5 (2010). doi:10.4242/BalisageVol5.Huitfeldt01. On the Web at http://www.balisage.net/Proceedings/vol5/html/Huitfeldt01/BalisageVol5-Huitfeldt01.html.

Huitfeldt, C., and Sperberg-McQueen, C. M. (2017). Transcriptional implicature: Using a transcript to reason about an exemplar. Paper given at DH 2017, held August 2017 at the McGill University, Montréal. Short version on the web at https://dh2017.adho.org/abstracts/235/235.pdf.

Kahlo, F. (1940). Letter to Diego Rivera, 1940. Emmy Lou Packard Papers 1900-1990, Archives of American Art, Smithsonian Institution. Facsimile of letter on the Web at https://www.aaa.si.edu/collections/items/detail/frida-kahlo-letter-to-diego-rivera-739

Kline, M.-J. (1987). *A guide to documentary editing.* Baltimore and London: Johns Hopkins University Press; second edition 1998.

Lafleur, R. A. (2010). *Scribblers, sculptors, and scribes: A companion to Wheelock's Latin and other introductory textbooks.* [New York]: Collins Reference.

Mommsen, T., et al. (eds). (1863-). *Corpus Inscriptionum Latinarum.* Berlin: Georg Reimber. Image and transcript of CIL 12 498 reproduced from (Lafleur 2010)xs, p. 14.

Neugebauer, A., and Brandl, H. (2012). ubi sancta requiescit Aedith. Das Grabmal der Königin Editha im Magdeburger Dom. In Meller, H., et al. (eds), *Königin Editha und ihre Grablegen in Magdeburg.* (Archäologie in Sachsen-Anhalt, Sonderband 18.) Halle, pp. 33-54.

Pierazzo, E. (2011). A rationale of digital documentary editions. *Literary & Linguistic Computing* 26.4: 463-477.

Pierazzo, E. (2015). *Digital scholarly editing: Theories, models and methods.* Aldershot: Ashgate , 2015.

Robinson, P. (1994). *The transcription of primary textual sources using SGML*. Office for Humanities Communication Publications, Number 6. [Oxford: OHC].

Robinson, P., and Solopova, E. (1993). Guidelines for the transcription of the manuscripts of the Wife of Bath's Prologue. In Blake, N., and Robinson, P. (eds), *The Canterbury Tales Project Occasional Papers Volume I*. Office for Humanities Communication Publications, Number 5. [Oxford: OHC], 1993.

Sanger, M. (1914). *The Woman Rebel.* Vol. 1 No. 1. From Katz, E., Hajo, C. M., and Engelman, P. C. (eds). *The Margaret Sanger Papers.* Sample from the MSP in the Model Editions Partnership at http://modeleditions.blackmesatech.com/mep/.

Sor Juana Ines de la Cruz. (1700). *Fama y obras posthumas del fenix de Mexico, Decima musa, poetisa americana Sor Juana Ines de la Cruz, Religiosa profesa,* [ed.] Don Juan Ignacio de Castorena y Visua. Madrid: Manuel Ruiz de Murga. Digitized version by the University of Bielefeld on Web at http://ds.ub.uni-bielefeld.de/viewer/image/1592397/1/; page 163 is at http://ds.ub.uni-bielefeld.de/viewer/image/1592397/153/as.

Sperberg-McQueen. C. M., Marcoux, Y., and Huitfeldt, C. (2014). Transcriptional implicature: A contribution to markup semantics. Paper given at DH 2014, held July 2014 at the University of Lausanne. Short version on the Web at http://dharchive.org/paper/DH2014/Paper-61.xml.

Stevens, M. E., and Burg, S. B. (1997). *Editing historical documents: A handbook of practice.* Walnut Creek, Ca.: Altamira Press, published in cooperation with the American Association for State and Local History, the Association for Documentary Editing, and the State Historical Society of Wisconsin.

Unsworth, J., O'Brien O'Keeffe, K., and Burnard, L. (eds) (2006). *Electronic textual editing.* New York: MLA.

Vander Meulen, D. L, and Tanselle, G. T. (1999). A system of manuscript transcription. *Studies in Bibliography* 52: 201-212.

Wittgenstein, L. (n.d.). Wittgenstein Source. Curator: Alois Pichler, Wittgenstein Archives at the University of Bergen (WAB). http://wittgensteinsource.org/. Includes material from the Bergen Electronic Edition (BEE) of Wittgenstein's Nachlaß.

Modelling Multigraphism: The Digital Representation of Multiple Scripts and Alphabets

Peter Anthony Stokes
peter.stokes@kcl.ac.uk
King's College, London, United Kingdom

Digital approaches to palaeography – the study of historical or ancient handwriting – have received significant attention in recent years. Projects like ORIFLAMMS (Stutzmann, 2016) and DigiPal (Brookes, 2015) have focussed on this, as well as projects aimed more at the book or written object in general such as HisDoc and Diva DIA, work on the Cairo Genizah (e.g. Wolf et al., 2011), work at the Centre for the Study of Manuscript Cultures (CSMC) in Hamburg; and many more. Although taking different approaches and addressing different aspects of palaeography, these projects have all made significant and important advances. However, relatively few have addressed explicit semantic modelling of handwriting itself. One such model was developed for the DigiPal project (Stokes, 2012) and has since been implemented in open-access and freely-available software called Archetype (2017). The model was developed initially for the Latin alphabet, but it has proven to be much more versatile than anticipated, with application to Hebrew and decoration (Brookes et al., 2015), bilingual Greek-Latin inscriptions, and experiments with Chinese, Cuneiform, Mayan, and others (see Figure 1: Archetype applied to Chinese script, showing search results for characters (graphs) containing the component and Figure 2: Demonstration of Archetype applied to Mayan hieroglyphics).

Figure 1: Archetype applied to Chinese script, showing search results for characters (graphs) containing the component

Figure 2: Demonstration of Archetype applied to Mayan hieroglyphics

Although applied successfully to different scripts, the DigiPal model (like most palaeographical method) assumes a homogeneous corpus comprising samples of the same alphabet and writing-style. Although convenient, this homogeneity is in fact very limiting, as throughout history many, perhaps even most, societies employed a range of different writing styles and systems. In Western Europe, for example, the best-known example is probably the Gothic script-system, comprising *Textura*, *cursiva* and so on; but similar patterns can be observed elsewhere in Hebrew, Arabic, Chinese, Tibetan and beyond. Furthermore, different alphabets or writing systems were often used together. Egyptian scribes use hieroglyphic, Hieratic and Demotic contemporaneously; Greek and Latin inscriptions are found across much of the Roman Empire, sometimes with third languages and alphabets; the Dunhuang materials contain a wide range of languages and scripts including Chinese, Tibetan, Sanskrit, Arabic and Sogdian; materials in four writing-systems survive from medieval Sicily; and so on. All of this suggests very strongly that people often – perhaps usually – could and did write in different alphabets or writing-systems. Identifying such cases would give us much important information about areas like education (how many Arab scribes also learned Hebrew?), cultural influence (were Sogdian annotations written at the same time and place as the main text in Chinese?) and so on. However, this requires an approach that allows for comparison across different scripts, discussion of which has really only just begun in both "digital" and "non-digital" palaeography (Stokes, 2017).

In principle, the DigiPal model provides such a framework. It specifies that characters are made up of components, defined as structural elements which recur across different letters (such as the ascenders in **b**, **h**, **l** and so on: Stokes, 2012). If one can map between components of different writing-systems then it is relatively easy to search for graphs (i.e. instances of letters written on the page) which share those components, and this allows for comparison. A proof-of-concept is illustrated in Figure 3: Screenshot from proof-of-concept cross search for example letters (graphs) with ascenders, where six instances of the Archetype system are searched simultaneously via the software's web API; a further example is given by Stokes (2017).

Figure 3: Screenshot from proof-of-concept cross search for example letters (graphs) with ascenders

However, multigraphic contexts require revising some basic assumptions of the model. For instance, DigiPal (and much palaeographical discussion) assumes that "etically" same is also "emically" same: that things that look the same mean the same. **H** and **H** look identical and therefore are normally assumed to represent the same letter: this is normally valid and indeed essential for communication in a monographic context. However, it may not hold in general: if we write **H**ELLO and I**H**ΣOYΣ then it becomes clear that the first is the Latin capital H (Unicode U+0048) and the second a Greek capital Eta (Unicode U+0397: cf. Bugarski, 1993). In context one can categorise these as separate characters, but a user searching the database for palaeographically comparable forms would presumably want to find both. Similar are apparently unambiguous characters used for different functions, perhaps deliberately to echo a different writing system. For example, in "pseudo-fonts" like GRΣΣK, the linguistic context shows that the two central letters function as the English grapheme E, but they are represented by the Greek capital Sigma. Comparable examples are widespread, such as the use of Greek letters in the Latin text of the ninth-century Lindisfarne Gospels illustrated in Figure 4: Detail from the Lindisfarne Gospels (London, British Library, Cotton Nero D.iv), showing use of Greek letters for writing Latin. In a monographic context this could be addressed as grapheme E with allograph "sigmoid" or something similar, but in a multigraphic context a palaeographer would presumably want to be able to find examples of both capital Sigma and "sigmoid" E with a single search and without necessarily anticipating the coincidence of forms in advance.

Figure 4: Detail from the Lindisfarne Gospels (London, British Library, Cotton Nero D.iv), showing use of Greek letters for writing Latin

In order to address this, some changes to the DigiPal model are proposed. The first is to change the central hierarchy of Character-Allograph-Idiograph-Graph presented by Stokes (2012), separating the linguistic/emic from the graphic/etic in a many-to-many relationship as shown in Figure 5 and Figure 6. This allows users to search by form, component or linguistic function.

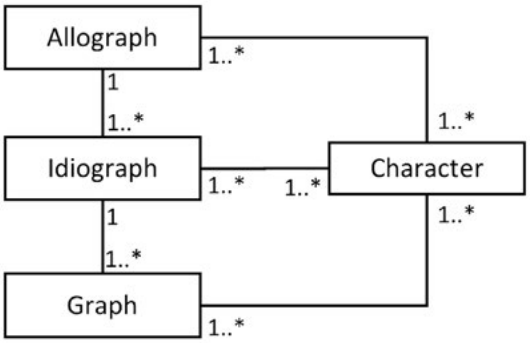


Figure 5: Revised (extract of) the DigiPal model of script

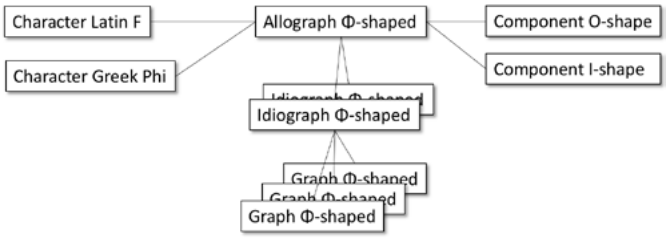


Figure 6: Example of the model applied to the form Φ used for Latin **F** as shown in Figure 4: Detail from the Lindisfarne Gospels (London, British Library, Cotton Nero D.iv), showing use of Greek letters for writing Latin

A further more practical change is to add sub-components. The use of components works well for alphabets and abugidas but much less so for more ideographic or hieroglyphic writing systems. In these cases, characters are made up of components which can themselves be further characters which contain further components, and so on. With sub-components, a component may be a discrete entity in its own right, or it may be made up of a set of further components. One might then describe the Korean glyph 곶 as a character having ス as a component which has フ as a sub-component; this would allow retrieval of all instances of any graphs (character blocks) containing ス, or idiographs containing フ, and so on (cf. Stokes, 2014). The same approach can be used for writing systems with subscript letters or character stacks, such as Myanmar or Tibetan (e.g. the second element in ကြွေးကျွေးကြောင်း which comprises four distinct characters: Flynn, 2015), or even ligatures and conjoined letters such as ﷺ. This then allows searching for more complex components that reoccur across writing systems, such as the Korean sub-component ス (U+110C) which also appears in Japanese *katakana* as a distinct character (ス, U+30B9) and as a sub-component in numerous Chinese ideographs.

One limitation of this model is the lack of linguistic context. DigiPal treated characters as distinct entities with no direct relationship between each other, but in practice letters normally appear in a broader linguistic context, namely the text, and this becomes essential in a multigraphic environment. Archetype goes some way towards addressing this, as it allows for including the text on an image, and the text itself can be marked up in XML and then linked in turn to sections of the image (Figure 7: Screenshot of Archetype used to provide Hebrew transcription and English translation linked to manuscript image, courtesy of Stewart Brookes and Figure 8: Screenshot of Archetype implemented for the Models of Authority project, showing text and image isolated by XML markup (here for salutatio clauses in medieval charters)). From this it is relatively trivial to detect the section of text in which a given graph is found, and if the XML markup specifies the language then this would allow one to find (for example) occurrences of a given grapheme or allograph within a specific linguistic context without needing to specify the language of each individual graph.

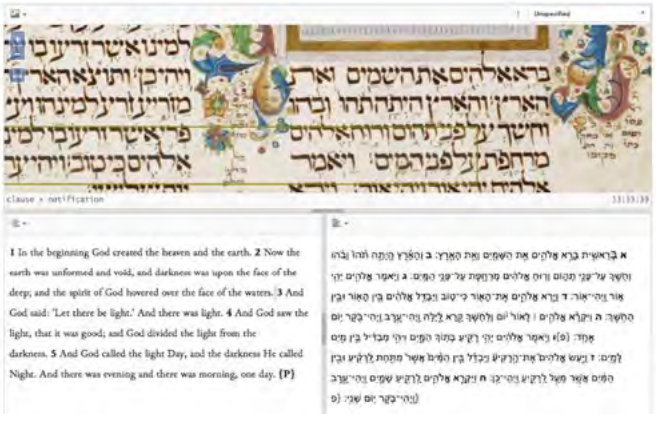


Figure 7: Screenshot of Archetype used to provide Hebrew transcription and English translation linked to manuscript image, courtesy of Stewart Brookes

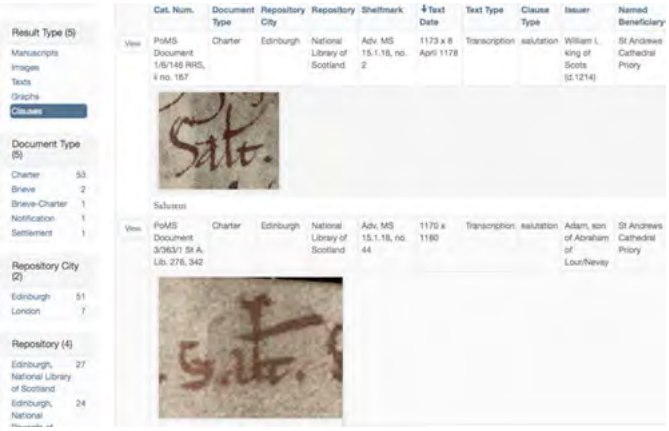


Figure 8: Screenshot of Archetype implemented for the Models of Authority project, showing text and image isolated by XML markup (here for *salutatio* clauses in medieval charters)

These extensions and changes to the model are not sufficient in themselves to fully address the challenges raised here. For instance, they still assume that different scripts can be reduced to common "atomic" units but it is not clear that this always holds, particularly for scripts that have entirely different genealogies such as Chinese and Sogdian, or when different writing implements and different directionality of script is involved (but see Stirnemann and Oszlowy-Schlanger, 2012). It also assumes that researchers can agree on what these "atoms" might be, and in principle link between them using the Semantic Web or similar. In some cases these units are evident, such as ascenders or descenders in many scripts, or the radical in Chinese and so on, but as Petrucci in particular has pointed out different scholarly viewpoints will necessarily produce different descriptions, and each of these different views is potentially valid and important (2001: 70–1), but it is not evident that they can necessarily be compared. This problem that extends well beyond the present discussion to encompass data interchange in general, though in practice it may be that if different scholars have such different approaches then perhaps uniting them is not meaningful, and an "ecosystem" of alternative viewpoints may be more appropriate.

Nevertheless, work to date suggests that the approach described here can provide a useful entry into the problems of multigraphism, particularly when combined with further refinements such as the distinction between components of allographs as envisaged by the scribe "in mind's eye" and traces of graphs actually executed on the page, as well as the ability to compare "essential elements" like components as well as "elements of style" in DigiPal's Features (Stokes, 2017 and Parkes, 2008). By building further in this direction, and most likely also adding machine vision and other approaches to searching, it seems likely that good further progress can be made.

References

Archetype (2017). London: King's College, https://archetype.ink (accessed 26 November 2017).

Brezina, D. (2013) Balkan Sans. *Typographica*, http://typographica.org/typeface-reviews/balkan-sans/ (accessed 26 November 2017).

Brookes, S.J. et al. (2015). The DigiPal Project for European scripts and decorations. In Conti, A., O. da Rold and P. Shaw (eds), *Writing Europe 500–1450: Texts and Contexts.Essays and Studies*n.s. 68: 25–58

Bugarski, R. (1993). Graphic relativity and linguistic constructs. In Scholes, R.J. (ed.), *Literacy and Language Analysis*. Hillsdale, NJ: Erlbaum, pp. 5–18.

Centre for the Study of Manuscript Cultures. University of Hamburg, https://www.manuscript-cultures.uni-hamburg.de/index_e.html (accessed 26 November 2017).

DigiPal. (2010–14). *Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic*. London: King's College, http://www.digipal.eu (accessed 26 November 2017).

Diva DIA. University of Fribourg, https://diuf.unifr.ch/main/hisdoc/divadia (accessed 26 November 2017).

Flynn, C. (2015). Encoding model of the Tibetan script in the UCS. *The Tibetan and Himalayan Library*, http://www.thlib.org/tools/#wiki=/access/wiki/site/26a34146-33a6-48ce-001e-f16ce7908a6a/encoding%20model%20of%20the%20tibetan%20script%20in%20the%20ucs.html (accessed 26 November 2017).

HisDoc. University of Fribourg, https://diuf.unifr.ch/main/hisdoc/hisdoc2 (accessed 26 November 2017).

Models of Authority (2017): Scottish Charters and the Emergence of Government 1100–1250. London: King's College, https://www.modelsofauthority.ac.uk (accessed 24 April 2018).

Parkes, M.B. (2008). *Their Hands Before Our Eyes: A Closer Look at Scribes*. Ashgate: Aldershot.

Petrucci, A. (2001). *La descrizione del manoscritto: storia, problem, modelli*. 2nd ed. Roma: Carroci.

Stirnemann, P. and Olszowy-Schlanger, J. (2008). The Twelfth-century trilingual psalter in Leiden. *Scripta* 1, pp. 103–112.

Stokes, P.A. (2012). Modeling medieval handwriting: A new approach to digital palaeography. *Digital Humanities 2012: Book of Abstracts*. Hamburg: University of Hamburg, pp. 382–5, http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-medieval-handwriting-a-new-approach-to-digital-palaeography (accessed 24 April 2018).

Stokes, P.A. (2014). Describing handwriting, part VII: Chinese (Han) script. *DigiPal: Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic*. London: King's College, http://www.digipal.eu/blog/describing-handwriting-part-vii-chinese-han-script/ (accessed 24 April 2018).

Stokes, P.A. et al. (2016).The Models of Authority Project: Extending the DigiPal Framework for Script and

Decoration. *Digital Humanities 2016: Conference Abstracts*. Krakow: pp. 896-99, http://dh2016.adho.org/abstracts/387 (accessed 24 April 2018).
Stokes, P.A. (2017). Scribal attribution across multiple scripts: A digitally-aided approach.*Speculum*92: S65–85.doi:10.1086/693968 (accessed 24 April 2018).
Stutzmann, D. (2016) *Compte-rendu de fin de projet: ORIFLAMMS: Programme Corpus, données et outils*, https://f.hypotheses.org/wp-content/blogs.dir/1267/files/2017/04/Oriflamms-Compte-rendu-final.pdf (accessed 24 April 2018).
Wolf, L., et al. (2011). Automatic paleographic exploration of genizah manuscripts. InFischer, F. (ed), *Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*. Norderstedt: Books on Demand, 157–179, http://kups.ub.uni-koeln.de/4348/ (accessed 24 April 2018).

Chinese Text Project
A Dynamic Digital Library of Pre-modern Chinese

Donald Sturgeon
djs@dsturgeon.net
Harvard University, United States of America

Introduction

Traditional full-text digital libraries, including those in the field of pre-modern Chinese, have typically followed top-down, centralized, and static models of content creation and curation. In this type of model, written materials are scanned, transcribed by manual effort and/or Optical Character Recognition (OCR), then corrected manually, reviewed, annotated, and finally imported into a system in their final, usable form. This is a natural and well-grounded strategy for design and implementation of such systems, with strong roots in traditional academic publishing models, and offering greatly reduced technical complexity over alternative approaches. This strategy, however, is unable to adequately meet the challenges of increasingly large-scale digitization and the resulting rapid growth in potential corpus size as ever larger volumes of historical materials are digitized by libraries around the world.

The Chinese Text Project (https://ctext.org) is a full-text digital library of pre-modern Chinese written materials which implements an alternative model for creation and curation of full-text materials, adapting methodologies from crowdsourcing projects such as Wikipedia and Distributed Proofreaders (Newby and Franks 2003) while also integrating them with full-text database functionality. In contrast to the traditional linear approach, in which all stages of processing including correction and review must be completed before transcribed material is ingested into a database system, this approach works by immediately ingesting unreviewed materials into a publicly available, managed system, within which these materials can be navigated and used, as well as improved through an ongoing collaborative correction and annotation process. From a user perspective, this has the consequence that the utility of the system does not rest upon prior expert review of materials, but instead derives from provision to individual users of the ability to interact directly and effectively with primary source materials and verify accuracy of transcription and annotation for themselves. Combined with specialized Optical Character Recognition techniques leveraging features common to pre-modern Chinese written works (Sturgeon 2017a), this has enabled the creation of a scalable system providing access to a long tail of historical works which would otherwise not be available in transcribed form. The system is highly scalable and currently contains over 25 million pages of primary source material while being used by over 25,000 users around the world every day.

Creating transcriptions

The most fundamental type of material contained in the Chinese Text Project consists of digital facsimiles of pre-modern published works. These are typically ingested in bulk through collaboration with university libraries which have created high quality digital images of works in their collections. After ingestion, the next step in making these materials more useful to users is creation of approximate transcriptions from page images using OCR. Producing accurate OCR results for historical materials is challenging due to a number of issues, including variation in handwriting and printing styles, varying degrees of contrast between text and paper, bleed-through from reverse sheets, complex and unusual layouts, and physical, water or insect damage to the materials themselves prior to digitization. In addition to these challenges which are common to OCR of historical documents generally, OCR for premodern Chinese works faces additional difficulties in extracting training data due to the large number of distinct character types in the Chinese language. Most OCR techniques apply machine learning to infer from an image of a character which character type it is that the image represents, and these techniques require comprehensive training data in the form of clear and correctly labeled images in the same writing style for every possible character. This is challenging for Chinese due to the large number of character types needed for useful OCR (on the order of 5000); unlike historical OCR of writing systems with much smaller character sets, it is not feasible to simply create this data manually. Instead, training data is extracted through an automated procedure (Sturgeon 2017a) which leverages knowledge about existing transcriptions of other texts to assemble clean labeled character images extracted from historical works for every character to be recognized (Figure 1). Together with image processing and language modeling tailored to pre-modern Chinese, this significantly reduces the error rate in comparison with off-the-shelf OCR software.

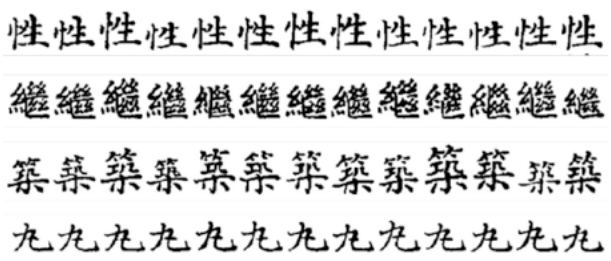


Figure 1. OCR training data is extracted automatically from handwritten and block-printed primary source texts.

Navigating texts and page images

Once transcriptions of page images have been created, they are directly imported into the public database system. The system represents textual transcriptions as sequences of XML fragments, in which markup is used to express both the relationship between transcribed text and the page image to which it corresponds, as well as the logical structure of the document as a whole. This facilitates two distinct methods of interacting with the transcribed material: firstly, as a single document consisting of the transcribed contents of each page concatenated in sequence to give readable plain-text with logical structure (divisions into chapters, sections, and paragraphs); secondly, as a sequence of page-wise transcriptions, in which a direct visual comparison can be made between the transcription and the image from which it is derived (Figure 2). In both cases, an important contribution of the transcription is that it enables full-text search; the primary utility of the page-wise view is that it enables efficient comparison of transcribed material with the facsimile of the primary source itself. As these two views are linked to one another and created from the same underlying data, this makes it feasible to read and navigate a text according to its logical structure, and at any stage of the process jump to the corresponding location in the sequence of page images to confirm accuracy of the transcription.

Figure 2. Full-text search results can be displayed in context in a logical transcription view (left), as well as aligned directly together with the source image in an image and transcription view (right).

Crowdsourced editing and curation

As initial transcriptions are created using OCR, they inevitably contain mistakes. Users of the system have the option to correct mistakes they identify, as well as to annotate texts in a number of ways. Two distinct editing interfaces are provided: a direct editor, which enables direct editing of the underlying XML representation, and a visual editor allowing simplified editing of page-level transcriptions, which edits the same underlying content but does not require direct understanding or modification of XML. Regardless of which mechanism is used to submit an edit, all edits are committed immediately to the public system. Edits are versioned, allowing visualization of changes between versions and simple reversion of a text to its state at an earlier point in time. At present, the system receives on the order of 100 edits each day, representing much larger numbers of corrections, as editors frequently choose to correct multiple errors and sometimes entire pages in a single operation.

Further visual editing tools supplement these mechanisms to enable crowdsourcing of more complex information. Illustrations are entered by the user drawing a rectangular box on the page image to indicate the location of the illustration, then filling in a simple form describing various aspects of it (Figure 3). This results in an XML fragment describing the illustration, which can simply be inserted into the text at the appropriate location to represent it. This allows the illustration to be extracted from its context on the page and represented in the full-text transcription view as well as in the page-wise view. It also facilitates illustration search functionality, where illustrations can be searched by caption across all materials contained in the system (Figure 4). A similar visual editing interface is used to enable the inputting of rare

and variant characters which do not yet exist in Unicode. These characters are no longer in common use, but occur in many historical documents. The visual editing interface for rare character input also uses metadata provided by the user to identify whether a given character is the same as any existing character known to the system, and if so, assigns a common identifier so that data about these characters can be aggregated, and text containing such characters searched.

Figure 3. Identification and markup of illustrations within source materials are crowdsourced using purpose-designed visual editing tools which convert user input into XML.

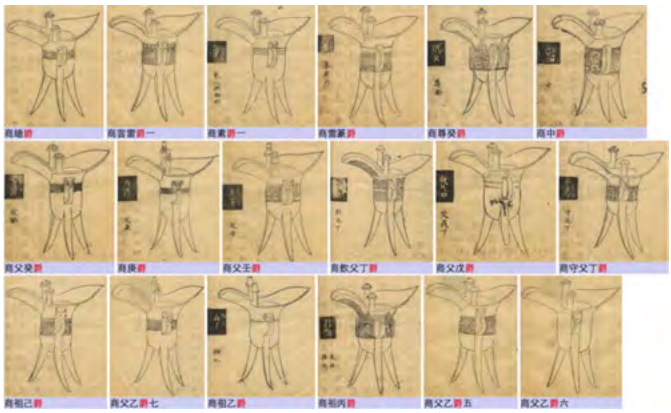


Figure 4. Image search: individual images are extracted from (and linked to) the precise locations at which they occur in source materials, and can be searched by caption.

Exporting data and integrating with external systems

In addition to the main user interface, a web-based Application Programming Interface (API) provides machine-readable access to data and metadata stored in the system. This facilitates text mining applications, as well as integration with other online systems. An example of the latter is the MARKUS textual markup system (De Weerdt et al. 2016), which can use the API to search for texts and load their transcriptions directly into this externally developed and maintained tool. An XML-based plugin system for the Chinese Text Project user interface also enables users to define and share extensions to the web interface which can be used to create connections to external projects and resources. This allows third-party tools such as MARKUS to integrate directly into the web interface, facilitating seamless connections between separately developed online projects. Text mining access is further facilitated by the provision of a Python module capable of accessing the API (Sturgeon 2017c), which is already in use in teaching and research (Sturgeon 2017b)

References

Newby, G. B. and Franks, C. (2003). Distributed proofreading. *2003 Joint Conference on Digital Libraries, 2003. Proceedings.* pp. 361–63 doi:10.1109/JCDL.2003.1204888.

Sturgeon, D. (2017a). Unsupervised Extraction of Training Data for Pre-Modern Chinese OCR. *Florida Artificial Intelligence Research Society. Proceedings.*

Sturgeon, D. (2017b). Classical Chinese DH: Getting Started. *Digital Sinology* https://digitalsinology.org/classical-chinese-dh-getting-started/ (accessed 27 November 2017).

Sturgeon, D. (2017c). Chinese Text Project API wrapper https://pypi.python.org/pypi/ctext/ (accessed 27 November 2017).

Sturgeon, D. (2017d). Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities* doi:10.1093/llc/fqx024. https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqx024/4583485 (accessed 27 April 2018).

Weerdt, H. D., Ming-Kin, C. and Hou-Ieong, H. (2016). Chinese Empires in Comparative Perspective: A Digital Approach. *Verge: Studies in Global Asias*, **2**(2): 58–69 doi:10.5749/vergstudglobasia.2.2.0058.

Handwritten Text Recognition, Keyword Indexing, and Plain Text Search in Medieval Manuscripts

Dominique Stutzmann
dominique.stutzmann@irht.cnrs.fr
Institut de Recherche et d'Histoire des Textes CNRS, France

Christopher Kermorvant
kermorvant@teklia.com
Teklia, France

Enrique Vidal
evidal@prhlt.upv.es
Universitat Politècnica de València, Spain

Sukalpa Chanda
s.chanda@rug.nl
Rijksuniversiteit Groningen, The Netherlands

Sébastien Hamel
sebastien.hamel@irht.cnrs.fr
Institut de Recherche et d'Histoire des Textes CNRS, France

Joan Puigcerver Pérez
joapuipe@upv.es
Universitat Politècnica de València, Spain

Lambert Schomaker
l.r.b.schomaker@rug.nl
Rijksuniversiteit Groningen, The Netherlands

Alejandro H. Toselli
ahector@prhlt.upv.es
Universitat Politècnica de València, Spain

Introduction

Artificial Intelligence has unlocked the access to the text of medieval manuscripts! The partners of the European research project HIMANIS implemented, for the first time, the indexing and plain text querying of thousands of pages of medieval manuscripts. The large scale of the corpus and the possibility to search in plain text for handwritten sources are unheard of in medieval studies, so that the results present a major shift for historians. The challenge of multilingualism, script variation and abbreviations, which are crucial for HTR on medieval sources, has been successfully met.

Context

Millions of medieval manuscripts, charters and archival documents are preserved worldwide, and centuries of scholarship and text editions could naturally not exhaust the wealth of these resources. Digital libraries (BVMM, Gallica, e-Codices, Manuscripta Mediaevalia, etc.) and archives (Monasterium) are amassing **reproductions of medieval manuscripts and archives, often with scarce metadata**. However, while Optical Character Recognition technologies allow to easily "distant read" several millions of books (Moretti, 2013; Crane, 2006; Clement et al., 2008; GDELT Project, 2015), **medieval manuscripts and archives remain difficult to access, read and understand**. Handwritten Text Recognition (HTR) systems cannot offer sufficiently accurate transcripts on historical documents. Therefore continuous efforts are made in Europe. After tranScriptorium (McNicholl and Miles-Board, 2015), the EU has funded HIMANIS and also funds Recognition and Enrichment of Archival Documents (READ, 2016-2019) under the H2020 program, with few medieval sources. MONK, developed by the University of Groningen, is also a well-known infrastructure, including some medieval resources as those from Stadsarchiev Leuven.

HIMANIS: consortium, corpus, method, and evaluation

Handwritten Text Recognition (HTR) is the focus of the European cross-disciplinary research project HIMANIS (Historical MANuscript Indexing for user-controlled Search), funded by the JPI Cultural Heritage. The partners applied HTR technologies for multilingual medieval manuscripts and demonstrated the feasibility of an accurate and meaningful automated text indexing of large collections of hitherto untranscribed text images.

The partners build a **cross-disciplinary consortium**. The principal investigator is a researcher in the Humanities (Institut de Recherche et d'Histoire des Textes, CNRS) and the project gathered several research teams in engineering sciences, both in the private and public sectors : A2iA (France), University of Groningen (The Netherlands) and Polytechnic University of Valencia (UPVLC, Spain). Cultural Heritage institutions provided support and datasets: Archives Nationales (France), Bibliothèque nationale de France. UPVLC and University of Groningen are partners in or host institutions for the above mentioned READ and MONK developments.

As a challenging and particularly interesting corpus, the partners chose the large collection of registers and formularies produced by the French royal chancery in the 14th and 15thc., encompassing **199 volumes, representing 83'000 pages, with 64'830 royal charters in 175 registers, and 24 formularies and related resources**. This large and iconic collection bears witness to the rationalization of late medieval administration and is a key source to our understanding of medieval Europe and the rise of centralized nation states on the continent as consequence of the long lasting wars between France and England. While HTR on medieval sources is notoriously highly difficult given the greatly variable handwriting styles, this corpus is even more challenging because of its multilingual content and the large number of abbreviated words.

A first work package consisted in creating the corpus, formatting available metadata, and authority data. The Archives Nationales digitized the corpus in several batches during the project. The metadata on French chancery registers are diverse. In increasing order of information quality, there are: (1.) medieval tables of content copied in autonomous inventories in the 18th and 19th c.; (2.) index cards with reference to shelfmarks (and rarely folio number) containing person and place names; (3a.)printed systematic inventories, including some already converted to EAD without their indexes, as well as (3b.) handwritten systematic inventories which were only accessible *in situ*, and (3c.) printed geographic or thematic inventories; (4.) partial, rarely scholarly, editions. The partners devised an integrated TEI format to accommodate all four types of metadata, and converted all metadata to this format, including the handwritten inventories on which the partners applied HTR-technologies to recognize the text abstracts and the index entries (Stutzmann et al., 2017). In total, after more

than 150 years of systematic research, inventories only covered 28'000 charters, that is ca. 43% of the register corpus and only one formulary was edited (Odart Morchesne, 2005; Guyotjeannin and Lusignan, 2011). Authority data encompass linguistic dictionaries and gazetteers, which were used to produce a lemmatized search engine.

A second work package consisted in training and applying a robust "optical model", capable of dealing with the variability and abbreviations in medieval, multilingual handwritings. The existing editions were first "aligned" on the available "text images" at a line level, applying techniques developed by the partners in the project Oriflamms (Leydier et al., 2014; Stutzmann et al., 2015; Bluche et al., 2016; Oriflamms, 2017). Basing on this alignment and using deep neural networks (CNN/RNN), machines could learn to "read" (Bluche et al., 2017). Learning on the monumental, modernizing, and very regularizing edition by P. Guérin (normalized punctuation, expanded abbreviations) (Guérin and Celier, 1881), the system created so-called "character lattices" which included abbreviations, so that the system was also able to read and expand abbreviations (fig. 1).

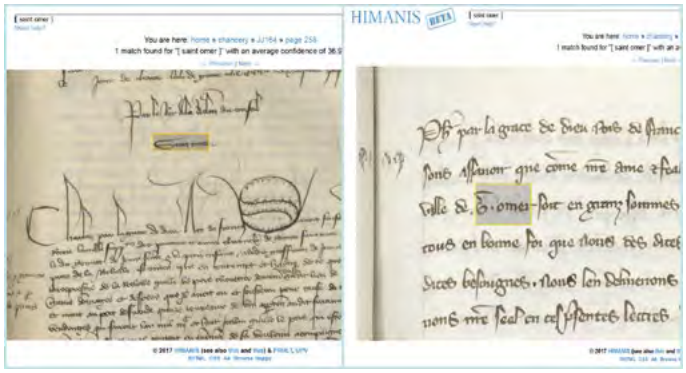


Figure 1: Text indexing. The query "[Saint Omer]" retrieves both abbreviated and unabbreviated strings in different volumes and different handwritings. Images: Paris, Archives Nationales, JJ 35 and JJ 164.

The decoding process produces different "hypotheses" for each spot on the image (typically from one to ten variant readings) and rated them according to their confidence levels according to inner statistical models and linguistic ones. The index has been "pruned" (i.e. reduced by removing) from the most unlikely readings, but still contains more than 28 bn index entries, 3 bn lines, 44 bn "words" and "pseudo-words".

The search engine was published online as a beta version (http://prhlt-kws.prhlt.upv.es/himanis/) and is being transferred to https://himanis.org/ where images and texts are accessible through the IIIF protocol and as IIIF annotations (fig. 2).

Figure 2: Indexed words as IIIF annotation in the IIIF compliant viewer Mirador (image: Paris, Archives Nationales, JJ 137, page 14)

Like in the tranScriptorium model, the users can set the confidence level for the search to reduce the noise or maximize the hit list value, a functionality that is the information science equivalent to the performance measure in computer science through "precision" (number of correct occurrences in the hitlist) and "recall" (number of correct occurrences compared to an error-free edition).

Qualitative and quantitative measures demonstrates that the HIMANIS system has obtained a very high level of precision, more than 85%, even increased to 99% through lemmatization (Stutzmann, 2017a-c) (fig. 3).

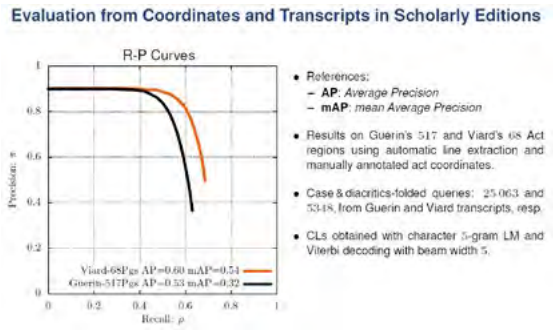


Figure 3: Measuring the performance of text indexing (Recall/Precision Evaluation)

Additional challenges: writer identification, granularity, crowdsourcing…

In parallel to HTR, another focus was automated writer identification. It allows a preliminary, nonetheless novel, analysis on the organization of the French chancery. These are among the first measured and convincing results produced for medieval handwritings, not represented in international competitions (Fiel et al., 2017; Andreu Sánchez et al., 2017). Based on the Quill feature (Brink et al., 2012) and validated on a partial ground-truth established by a paleographer, the system clustered hitherto unstudied page images, attributing them to 204 hypothetical writers.

Fig. 4 illustrates the calculated presence of writers in each volume, giving a first insight into possible collaborations between scribes within the chancery across time.

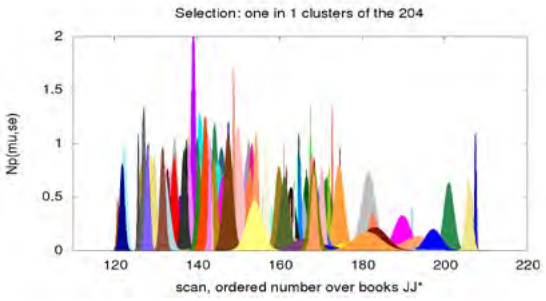


Figure 4: Writer Identification: Visualization of the different "hands" and their likeliness in the different volumes (representation of Gaussian standard deviation for style clustering, with 204 clusters)

The partners tackle additional challenges. The text and structure of registers containing multiple charters impose to combine different granularities and intertwined both physical (page) and intellectual levels (one/several charter(s) on one/several page(s)). The integration of authority data and gazetteers allows new access, but with possible errors in text indexing and identification of named entities, measuring the applicability and usefulness of text indexing is an important methodological new task. Crowdsourcing results are currently negatively biased, because of the implemented ergonomics and users' strategies. They tend to "suggest corrections" in order to improve their future search experience, rather than to validate correct spots (fig. 4). Nevertheless, it helps measuring impact, adequacy, precision, and usefulness.

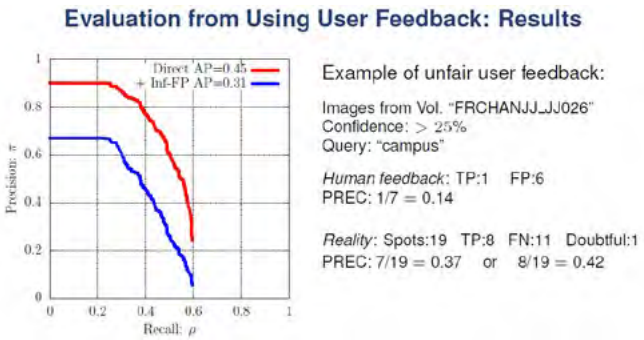


Figure 5: Crowdsourcing as biased feedback

Perspectives

The results of HTR are setting a new standard and make the digitized images a new source. Yet, they must obviously not be mistaken for a scholarly edition. HIMANIS participated in the current trend of Digital Humanities and uses images as data, both for text and for writer identification (Kestemont et al., 2017). Deep indexing represents new challenge for "distant reading" addressing topics and charter contents, because there is not an even number of hypotheses for all image spots. In new funded projects HOME (History Of Medieval Europe) and HORAE (Hours: Recognition Analysis, Editions), the partners are working on a methodology to create truthful and trustworthy results from uncertain and uneven, automatically generated data.

References

Andreu Sánchez, J., Romero, V., Toselli, A. H., Villegas, M. and Vidal, E. (2017). ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset. *14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017*. Kyoto: CPS, pp. 1383–88 doi:DOI 10.1109/ICDAR.2017.226.

Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Stutzmann, D., Toselli, A. H. and Vidal, E. (2017). Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project. *14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017*. Kyoto: CPS, pp. 312–17 doi:DOI 10.1109/ICDAR.2017.59.

Bluche, T., Stutzmann, D. and Kermorvant, C. (2016). Automatic Handwritten Character Segmentation for Paleographical Character Shape Analysis. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. pp. 42–47 doi:10.1109/DAS.2016.74.

Brink, A. A., Smit, J., Bulacu, M. L. and Schomaker, L. R. B. (2012). Writer identification using directional ink-trace width measurements. *Pattern Recognition*, 45(1): 162–71 doi:10.1016/j.patcog.2011.07.005.

Clement, T., Steger, S., Unsworth, J. and Uszkalo, K. (2008). How Not To Read A Million Books Harvard University, Cambridge, MA http://www.people.virginia.edu/~jmu2m/hownot2read.html (accessed 25 March 2017).

Crane, G. (2006). What Do You Do with a Million Books?. *D-Lib Magazine*, 12(3) doi:10.1045/march2006-crane. http://www.dlib.org/dlib/march06/crane/03crane.html (accessed 25 March 2017).

European Commission (2016). CORDIS : Projects & Results Service : Recognition and Enrichment of Archival Documents *CORDIS Community Research and Development Information Service* http://cordis.europa.eu/project/rcn/198756_en.html (accessed 14 October 2016).

Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Stamatopoulos, N. and Gatos, B. (2017). ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI). *14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017*. Kyoto: CPS, pp. 1377–82 doi:DOI 10.1109/ICDAR.2017.225.

GDELT Project (2015). Google BigQuery + 3.5M Books: Sample Queries *GDELT Blog* https://blog.gdeltproject.org/google-bigquery-3-5m-books-sample-queries/ (accessed 23 November 2017).

Guérin, P. and Celier, L. (1881). *Recueil des documents concernant le Poitou contenus dans les registres de la chancellerie de France*. 14 vols. (Archives historiques du Poitou). Poitiers: Société des archives historiques du Poitou http://gallica.bnf.fr/ark:/12148/bpt6k209478j (accessed 25 April 2014).
Guyotjeannin, O. and Lusignan, S. (2011). Introduction au formulaire d'Odart Morchesne *Le Formulaire d'Odart de Morchesne* http://elec.delisle.enc.sorbonne.fr/morchesne/ (accessed 23 November 2017).
Kestemont, M., Christlein, V. and Stutzmann, D. (2017). Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts. *Speculum*, 92(S1): S86–109 doi:10.1086/694112.
Leydier, Y., Eglin, V., Bres, S. and Stutzmann, D. (2014). Learning-Free Text-Image Alignment for Medieval Manuscripts. *2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. pp. 363–68 doi:10.1109/ICFHR.2014.67.
McNicholl, R. and Miles-Board, T. (2015). tranScriptorium : Computer-Aided, Crowd-Sourced Transcription of Handwritten Text (for Repositories). *10th International Conference on Open Repositories (OR2015)*.
Moretti, F. (2013). *Distant Reading*. London ; New York: Verso.
Odart Morchesne (2005). *Le Formulaire d'Odart Morchesne : Dans La Version Du Ms BnF Fr. 5024*. (Ed.) Guyotjeannin, O. & Lusignan, S. (Mémoires et Documents de l'École Des Chartes 80). Paris: École des chartes.
Oriflamms, C. (2017). Compte-rendu final du projet ORIFLAMMS / ORIFLAMMS Final report Billet *Écriture Médiévale & Numérique | Écritures Médiévales et Lecture Numérique. Carnet Du Projet ORIFLAMMS (Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts)* http://oriflamms.hypotheses.org/1592 (accessed 25 May 2017).
Stutzmann, D. (2017a). 99 pilgrimages / 99 'pelerinage(s)': accuracy and historical research Billet *Himanis* https://himanis.hypotheses.org/171 (accessed 23 November 2017).
Stutzmann, D. (2017b). Saint-Omer in the registers of the French royal chancery (99.6% precision!) Billet *Himanis* https://himanis.hypotheses.org/195 (accessed 23 November 2017).
Stutzmann, D. (2017c). The Royal Highness / L'altesse royale / regalis celsitudo : une thématique de préambule pour mesurer l'utilité d'HIMANIS Billet *Himanis* https://himanis.hypotheses.org/246 (accessed 23 November 2017).
Stutzmann, D., Bluche, T., Lavrentiev, A., Leydier, Y. and Kermorvant, C. (2015). From Text and Image to Historical Resource: Text-Image Alignment for Digital Humanists. Sydney http://dh2015.org/abstracts/xml/STUTZMANN_Dominique_From_Text_and_Image_to_Histor/STUTZMANN_Dominique_From_Text_and_Image_to_Historical_R.html (accessed 29 June 2015).
Stutzmann, D., Moufflet, J.-F. and Hamel, S. (2017). La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique = Full Text Search in Medieval Manuscripts : Issues and Perspectives of the HIMANIS Project for Electronic Publishing. *Médiévales : Langue, Textes, Histoire*, 73.

Estudio exploratorio sobre los territorios de la biopirateria de las medicinas tradicionales en Internet : el caso de America Latina

Luis Torres-Yepez
luisyepez13@gmail.com
Université Paris 8 Vincennes - Saint-Denis, France

Khaldoun Zreik
zreik@univ-paris8.fr
Université Paris 8 Vincennes - Saint-Denis, France

Resumen

El fenómeno de la biopirateria es una problemática que afecta principalmente a las comunidades indigenas y sus conocimientos tradicionales. De esta manera, las medicinas tradicionales son afectadas por la practica de la biopirateria. Este estudio exploratorio propone desarrollar una cartografía sobre la biopirateria de las medicinas tradicionales en la red social Twitter utilizando los métodos digitales. De esta manera, se presentan una serie de gráficos donde se describe el fenómeno y se identifican los diversos productos que son mencionados en Twitter. Concluimos que los métodos digitales nos permiten interpretar el fenómeno de la biopirateria de las Medicinas tradicionales en Twitter e identificar los productos que son mencionados.

Introducción

Se considera que el término biopirateria es un concepto contemporáneo (RAFI, 1994), pero también se encuentra que es una práctica antigua, con orígenes colonialistas (Boumediene, 2016; Shiva, 1997). Aubertin y Moretti mencionan sobre el término biopirateria.

> La Coalición contra la biopirateria (etcGroup) define la biopirateria como la apropiación general mediante los derechos intelectuales de recursos genéticos, de conocimientos locales y de culturas tradicionales pertenecientes a campesinos o comunidades indigenas quienes han desarrollado y mejorado el uso de recursos naturales. La biopirateria incluye la bioprospeccion, las patentes sobre genes y moléculas y la comercialización de los conocimientos culturales (2013, p. 91).

Al mismo tiempo que surge el término biopirateria, emerge un movimiento que busca la legitimación de los conocimientos locales e indigenas. De hecho, la principal problemática de la biopirateria se encuentra en la falta de legitimación de los conocimientos tradicionales por parte de gobiernos e instituciones. Como consecuencia, la problemática se extiende a la apropiación de conocimientos y recursos biológicos por parte de empresas e instituciones con la ayuda de otras instituciones que gestionan los derechos de autor y la propiedad intelectual (OMPI). Se encuentran diversos ejemplos, el caso del Neem y la Curcuma en la India, la Ayahuasca, la Quinoa, la Maca y la Mayacoba en America Latina y la Hoodia y la Rooibos en Africa (Aubertin et al., 2007; Delgado, 2002; Dumesnil, 2012; IEPI, 2016).

El caso de la biopirateria de los conocimientos alrededor de las medicinas tradicionales ha sido ignorado por gobiernos, instituciones y empresas por mucho tiempo, los casos de biopirateria como muestra el etc-Group (RAFI, 1994) son variados a lo largo del tiempo. En la actualidad por fin se encuentran discusiones sobre el tema en instituciones como la la Organización Mundial de la Propiedad Intelectual (OMPI)[1] y por otro lado en el Convenio sobre la Diversidad Biológica[2] y el protocolo de Nagoya (CBD, 2012). De la misma forma, es importante un rol mas activo de la UNESCO en la protección de los conocimientos locales y las medicinas tradicionales como Patrimonio Cultural Inmaterial (Unesco 2003).

Cartografía de internet

La cartografía de Internet permite analizar los perfiles, los comportamientos y asimismo analizar las relaciones entre los actores, las comunidades y las tendencias (Bastard, et al., 2017; Diminescu, 2012; Severo & Venturini, 2016). Nos interesa el análisis y la cartografía de la red social Twitter, ya que es una red abierta donde se encuentran conversaciones sobre fenómenos sociales y políticos.

El objetivo de esta investigación exploratoria es analizar los textos de las publicaciones en Twitter y desarrollar una cartografía de los productos relacionados con las medicinas tradicionales que son mencionados en las publicaciones. Por tanto, mediante el uso de los métodos digitales se desarrolla una cartografía del fenómeno de la biopirateria en Twitter.

Métodos digitales

Los llamados métodos digitales son una serie de métodos, técnicas y herramientas que permiten realizar estudios sobre la redes sociales e Internet en general (Diminescu, 2012; Rieder, 2013; Rogers, 2013; Severo and Venturini, 2015).

El método se desarrolló en cuatro etapas: la extracción, el análisis y la clasificación, la cartografía y finalmente la interpretación.

1. En la primer etapa, la extracción de datos en Twitter se realizó mediante la herramienta TAGS[3] y la opción de búsqueda: biopiracy OR biopiraterie OR biopirateria OR bio-pirateria OR bio-piracy OR bio-piraterie.

2. En la segunda etapa, con la herramienta OPEN REFINE[4] se desarrolló el análisis, la clasificación y la limpieza general de la base de datos. Con el objetivo de analizar todos los tweets registrados se decidió desarrollar la cartografía a partir de la relación entre los usuarios (@) y los tweets. Para el tratamiento y el análisis de los tweets se realizó un agrupamiento de los textos, por lo cual se agrupo en un mismo tipo de mensaje una publicación normal y un RT, así como otras con ciertas variaciones en el texto (Fig. 2). Asimismo se analizó cada texto para encontrar pistas sobre los productos de las MT y se creo una variable en la BD con el nombre del producto.

from_user	text	text_Clustering
gfc123	Mexicans Protest Law That Will Amount to Biopiracy for #Indigenous Communities https://t.co /BwQ1jpgO8W #mexico https://t.co /WzQ1XRsNTo	RT @telesurenglish: Mexicans protest law that will create biopiracy for Indigenous communities https://t.co /PitbhJ3QZa https://t.co /FxEMJER...
mariannenicols4	RT @RussDiabo: Mexicans Protest Law That Will Amount to Biopiracy for Indigenous Communities | News | teleSUR English https://t.co /7FXv19Qj...	RT @telesurenglish: Mexicans protest law that will create biopiracy for Indigenous communities https://t.co /PitbhJ3QZa https://t.co /FxEMJER...
skylightpix	Mexicans protest law that will create biopiracy for Indigenous communities https://t.co/hMKnf5d4xo v @telesurenglish https://t.co /nOlvxW75bl	RT @telesurenglish: Mexicans protest law that will create biopiracy for Indigenous communities https://t.co /PitbhJ3QZa https://t.co /FxEMJER...
pvlpfction	RT @telesurenglish: Mexicans protest law that will create biopiracy for Indigenous communities https://t.co /PitbhJ3QZa https://t.co/FxEMJER...	RT @telesurenglish: Mexicans protest law that will create biopiracy for Indigenous communities https://t.co /PitbhJ3QZa

1 http://www.wipo.int/meetings/en/details.jsp?meeting_id=42302
2 https://absch.cbd.int/es/

3 https://tags.hawksey.info/
4 http://openrefine.org/

from_user	text	text_Clustering
ugabhsi edit	RT @SusanSanchez19: Treaty to stop biopiracy threatens to delay flu vaccines https://t.co /mP8H5fOlcY @OneHealth	RT @GlobalBioD: Treaty to stop biopiracy threatens to delay flu vaccines https://t.co /LtfsPSn5pk via @NatureNews https://t.co /tXiCzKkRDA
AFHSBPAGE	Treaty to Stop #Biopiracy Threatens to Delay #Flu Vaccines https://t.co /95TkJ53ylc #Influenza #Military #PublicHealth	RT @GlobalBioD: Treaty to stop biopiracy threatens to delay flu vaccines https://t.co /LtfsPSn5pk via @NatureNews https://t.co /tXiCzKkRDA
SusanSanchez19	Treaty to stop biopiracy threatens to delay flu vaccines https://t.co /mP8H5fOlcY @OneHealth	RT @GlobalBioD: Treaty to stop biopiracy threatens to delay flu vaccines https://t.co /LtfsPSn5pk via @NatureNews https://t.co /tXiCzKkRDA
ProstateCell	Treaty to Stop Biopiracy Threatens to Delay #Flu Vaccines https://t.co /DtJf0PDuwh https://t.co /xcK6l8YUjt	RT @GlobalBioD: Treaty to stop biopiracy threatens to delay flu vaccines https://t.co /LtfsPSn5pk via @NatureNews https://t.co /tXiCzKkRDA

Fig. 1. Ejemplos de agrupamiento de los textos de las publicaciones

3. En Gephi1 se desarrolló la cartografía utilizando el algoritmo "Force Atlas" para visualizar la red espacialmente y se aplicaron los valores "in-degre" y "out-degre" para identificar las nodos con mayor valor.

4. En la etapa de la interpretación, se realizó el diseño visual de la cartografía utilizando también otras herramientas de visualización de datos.

Análisis de la biopirateria en Twitter

El resultado de la muestra de datos extraídos de Twitter entre las fechas 06/02/2017 – 06/10/2017 consistió en 3,995 publicaciones (Tweets) de los cuales se encontraron 494 repetidos y 1 con error, por lo cual se realizó el análisis con una muestra total de 3,500 Tweets. Asimismo se encontraron 581 tweets sin hashtags ni marcas de usuarios mencionados (@usuario).

En la Fig. 2 se presenta una red donde se aplicó la variable "Out degree", y se observa el contexto general de las conversaciones sobre la biopirateria según el idioma seleccionado por los usuarios. Esto no quiere decir que los usuarios publiquen únicamente en la lengua en la que se registraron en Twitter. El idioma mayoritario es el español, pero los usuarios mas activos se observan en la región en francés. La relación con las publicaciones se observa en la Fig. 3, donde se presentan los 20 tweets mas difundidos.

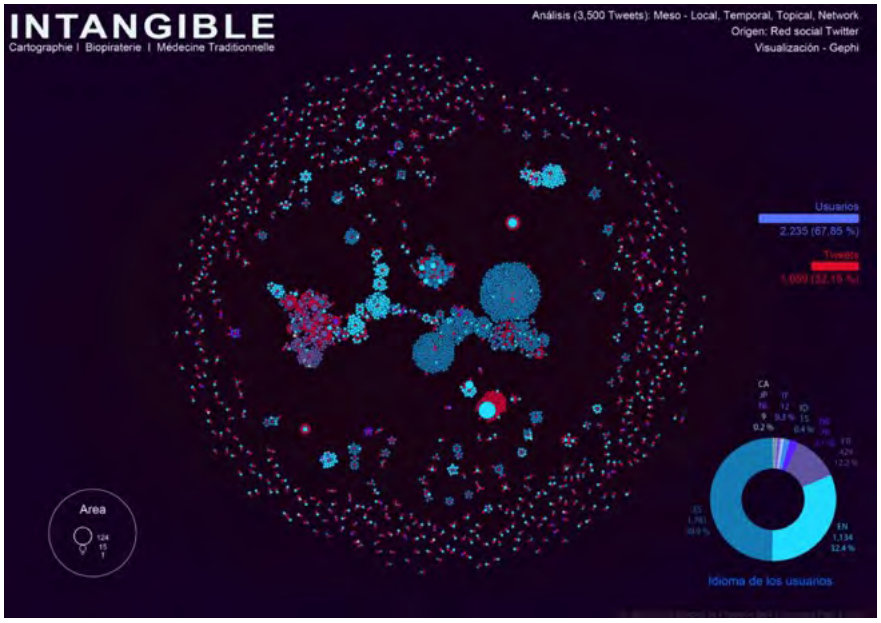


Fig. 2. Cartografía de usuarios por idioma

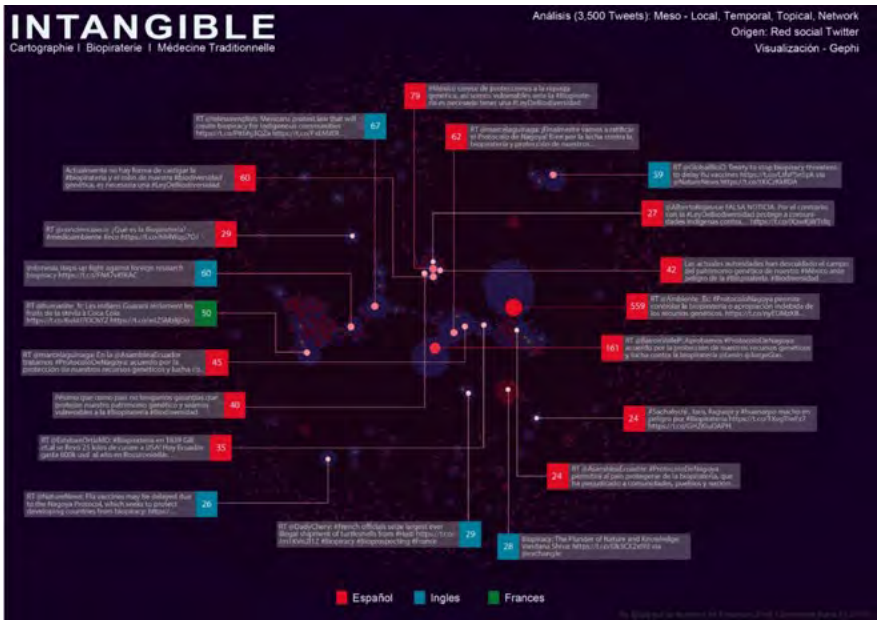


Fig. 3. Cartografía de los tweets mas difundidos

En la figura 3 se presenta una red donde se aplicó la variable "In degree" para darle mas relevancia a los tweets. Los tweets 559 y 161 son difundidos principalmente por bots y hablan sobre la implementación del protocolo de Nagoya en Ecuador (24, 45, 62). El Tweet mas difundido en ingles (67) menciona el movimiento contra la #LeyDeBiodiversidad en México que se relaciona con la ratificación del protocolo de Nagoya (27, 40, 42, 60, 62, 79). Se observa en este caso una disputa entre los ciudadanos que están en contra y los bots del gobierno. Los tweets 26 y 59 mencionan un articulo publicado en la revista @NatureNews sobre los riesgos del protocolo de Nagoya para el desarrollo de la vacuna contra la influenza (flu). El tweet 28 menciona a Vandana Shiva quién es una luchadora relevante contra la biopirateria. Se observa asimismo dos tweets que mencionan situaciones particulares sobre la biopirateria en Indonesia y en Haiti (29, 60). Finalmente, se observan tres tweets importantes los cuales mencionan algunos productos de las medicinas tradicionales como el Curare (35), la Estevia (60) y la Sacha inchi, Tara, Aguaje y Huanarpo macho (24).

A continuación se presenta la cartografía con los 16 productos encontrados que se relacionan con las medicinas tradicionales.

Fig. 4. Cartografía de las plantas y productos de las medicinas tradicionales

Observamos las diversas plantas y productos mencionados en Twitter, de los cuales la Stevia , la Sacha Inchi, el Aguaje, el Huanarpo macho, la Tara, la Maca, el Curare, la Quassia, el Maqui, la Hoja de coca y el Kambô son de origen latinoamericano. Después se observa el Kakadu plum de origen australiano, luego el Neem y el Turmeric de origen indio y entre los menos mencionados la Ixempra y el Rooibos de origen africano. Entre todos ellos la Ixempra no se encuentra su uso en las medicinas tradicionales, pero se menciona como un caso de biopirateria.

Conclusiones

Esta investigación exploratoria nos permitió analizar el contexto de la biopirateria en Twitter e identificar los diferentes productos que son mencionados. La cartografía presentada es en sí misma una interpretación y por lo tanto es una visión particular de la biopirateria.

Twitter es una red social abierta en la cual los usuarios participan libremente, pero se decidió no mostrar sus nombres en la cartografía para mantener en respeto su privacidad.

Este estudio exploratorio nos permite tener una base de conocimiento para seguir una investigación más específica sobre cada uno de los productos encontrados.

References

Aubertin, C. and Moretti, C. (2013) La biopiraterie, entre illégalité et illégitimité. In: *Les Marchés de la Biodiversité. Institut de recherche pour le développement*, p. 269.

Aubertin, C., Pinton, F. and Boisvert, V. (2007) *Les marchés de la biodiversité*. Recherche. IRD Editio. Paris: Institut de Recherche pour le Développement.

Bastard I., Cardon, D., Charbey, R., et al. (2017) Facebook, pour quoi faire ? *Sociologie* 8(1): 57–82. DOI: 10.3917/socio.081.0057.

Boumediene, S. (2016) *La colonisation du savoir : une histoire des plantes médicinales du 'Nouveau Monde' (1492-1750)*. Vaulx-en-Velin: Les éditions des mondes à faire.

CBD. (2012) Protocole de Nagoya sur l'accès aux ressources génétiques et le partage juste et équitable des avantages découlant de leur utilisation relatif à la Convention sur la diversité biologique. *Convention sur la diversité biologique Nations Unies*. Montreal. Available at: http://www.cbd.int/abs/doc/protocol/nagoya-protocol-fr.pdf.

Delgado, G. C. (2002) Biopi®acy and Intellectual Property as the Basis for Biotechnological Development : The Case of Mexico. *International Journal of Politics, Culture and Society* 16(2): 297–318.

Diminescu, D. (2012) Introduction: Digital Methods for the Exploration, Analysis and Mapping of e-Diasporas. *Social Science Information* 51(4): 451–458. DOI: 10.1177/0539018412456918.

Dumesnil, C. (2012) Les savoirs traditionnels médicinaux pillés par le droit des brevets ? *Revue internationale de droit économique* XXVI(3): 321–343. DOI: 10.3917/ride.257.0321.

IEPI. (2016) *Primer informe sobre biopiratería en el ecuador*. Quito.

RAFI. (1994) COPs... and Robbers... Transfer-Sourcing Indigenous Knowledge. Pirating Medicinal Plants. *Occasional Paper Series* 1(4): 20. Available at: http://www.etcgroup.org/content/volume-1-4-pirating-medicinal-plants.

Rieder, B. (2013) Studying Facebook via Data Extraction: The Netvizz Application. In: *Proceedings of WebSci '13, the 5th Annual ACM Web Science Conference, 2013*, pp. 346–355. DOI: 10.1145/2464464.2464475.

Rogers, R. (2013) *Digital Methods*. Massachusetts: MIT Press.

Severo, M. and Venturini, T. (2015) Intangible cultural heritage webs : Comparing national networks with digital methods. *New Media and Society* 18(8): 1616–1635. DOI: 10.1177/1461444814567981.

Severo, M. and Venturini, T. (2016) Enjeux topologiques et topographiques de la cartographie du web. *Reseaux* 1(195): 85–105. DOI: 10.3917/res.195.0085.

Shiva, V. (1997) *Biopiracy : the plunder of nature and knowledge*. South end Press.

Unesco. (2003) *Convention pour la Sauvegarde du Patrimoine Culturel Immateriel*. Paris.

In Search of the Drowned in the Words of the Saved: Mining and Anthologizing Oral History Interviews of Holocaust Survivors

Gabor Toth
gabor.toth@yale.edu
Yale University, United States of America

The experiences of six million victims of the Holocaust perished with them. This paper will discuss the ways text and data mining technology has helped to recover fragments of lost experiences out of oral history interviews with survivors. The paper will also present how a data-driven anthology of these fragments has been built. The first part situates the challenge of uncovering experiences of the voiceless in historiography. The second part shows how text and data mining techniques have been applied to recover fragments of lost experiences from a big corpus of English language interview transcripts in the collection of the United States Holocaust Memorial Museum (USHMM). The third part demonstrates how web technology and visualization are used to render these fragments in a digital anthology.

The ethical and theoretical problem of narrating the experience of those who did not survive the Holocaust

has been often addressed. Primo Levi has argued that survivors cannot tell the experience of those who did not survive because the Saved and the Drowned are "two particularly well differentiated categories among men." The Saved lived in a morally questionable "grey zone" that compromises their testimony (Levi, 2018). Others have pointed out how trauma inhibits survivors from recalling their own experiences (Felman, Laub 2013; Lacapra, 2014; Hartman, 2015). Others have argued that testimonies are shaped by narrative and discoursive processes (Bernard-Donals, Glejzer 2001; Rosen, 2009). Survivors' testimonies are therefore often used to study memory, and the underlying mediative processes (Langer, 2007). In short, there are gaps between the experience of the Saved and the Drowned, and between experiences recalled in a testimony and the original experience in the past.

This paper argues that despite these gaps, in testimonies there is a set of rudimentary experiences that are shared by both the Saved and the Drowned. They are basic physical and emotional states, as well as actions, that are cross-cultural; they are not the expression of post-traumatic states or any discursive, narrative, and linguistic mediation, but the very original experience. "Children crying for their parents" or "feeling ashamed at the moment of being forced to undress" are examples of these rudimentary experiences. Beyond their rudimentary nature, experiences shared by the Drowned and the Saved have another feature: given a reasonably large collection of testimonies, they recur in narration of victims who had very different fates. Epistemologically, the recurrent rudimentary experiences in testimonies by the Saved are the likely experiences of the Drowned. This however overlooks - on purpose - the realm of suppressed memories.

The first computational goal of this work was to retrieve textual fragments expressing similar rudimentary experiences in a corpus of 1571 randomly selected interview transcripts (approximately 27 million tokens) in the USHMM. The retrieval of textual fragments expressing similar experiences is a text mining task that has two differences from text reuse and plagiarism detection (Alzahrani et al, 2012; Büchler et al., 2014). First, non–native speakers are likely to use different vocabulary, as well as different grammatical constructions, to describe the same experience. Second, while plagiarism and text reuse detection aim to discover any repeating sequence in a text, this project has sought to discover only rudimentary experiences. In addition to the fact that plagiarism and text reuse detection tools could not offer solutions to the problems above, the project had to face another core difficulty: inference of meaning from longer sequences of words requires substantial further research in Text Mining.

In order to retrieve fragments describing experiences that are recurrent and rudimentary, a specific pipeline involving both algorithmic and human supervised stages has been designed by the author. Prior to the implementation of the pipeline, the data underwent a standard linguistic pre-processing, including detection of multiword expressions. The document frequency of all verbs in the corpus was computed, and verbs with document frequencies above the median (0.14) were labelled as "recurrent" and were investigated by a human agent. From this list of recurrent verbs, those expressing rudimentary physical and emotional experiences (for instance, "cry", "yell", "fear") were selected. Focus on verbs is explained by the fact that they are the most natural form to express experience. As a second step, a word embedding model was trained on the data, and synonyms of the pre-selected verbs were identified. The word embedding model broadened the initial focus on verbs since less frequently used adverbial and adjectival expressions were also identified (for instance, "undress", "barefooted" and "naked"). This resulted in an array of recurrent synonym sets. As a third step, from all textual contexts in which members of a given synonym set occur, document collections were constructed, and trained with a TF-IDF based LDA (Blei et al, 2003). The LDA model resulted in groups of words, also known as topic words, that tend to co-occur in a collection of textual contexts, as well as those textual contexts that are the most likely to be close to the group. As a short evaluation, the context based application of LDA was efficient to analyze the tendency of larger unit of words to co-occur. Traditional metrics to measure strength of association give less efficient results with units longer than bigrams. Furthermore, they cannot capture synonymy while LDA can do in certain cases. The last stage of the pipeline was the analysis of groups of words, and the textual contexts close to them, by a human agent. This was meant to investigate whether a given word combination, uncovered by LDA, actually referred to an experience, and to capture complete phrases, or "fragments," that express the experience. The result of the modelling process was a collection of approximately 200 fragments expressing 30 rudimentary experiences, though the model continues to identify additional experiences. In short, the pipeline has helped to detect sets of "sub-experiences" associated with a given rudimentary experience (shooting as sub-experience in the domain of nakedness), as well as textual fragments expressing them. At the same time, the model features limitations: it cannot for instance detect metaphorical expressions.

The second computational task was to find prototypical episodes in the domain of a rudimentary experience without supervision. For this purpose, the document collection of textual contexts underlying a given synonym set was trained with paragraph vector model (Le and Mikolov, 2014), and clustered with affinity propagation (Frey et al, 2007). This produced not only clusters but specific contexts that are the centers or the prototypical members of clusters. These prototypes are seen as typical episodes in the domain of a rudimentary experience.

Using these findings, a digital anthology that renders fragments expressing rudimentary experiences, prototypical instances of rudimentary experiences along with

transcripts and audio / video recordings is currently being developed. This anthology will support a hierarchical tree visualization in which branches represent core rudimentary experiences and leaves represent either prototypical instances or sub-experiences in the domain of rudimentary experiences. It will answer three important requirements of scholarship. First, it will uncover the multiplicity of contexts and ways in which the very same rudimentary experiences could take shape. Second, it will enable the investigation of a testimony both as a text and as an audio / video record. Third, the anthology enables the reading, listening or watching of experiences, which were retrieved and selected not by drawing on a historical preconception. Instead, a "let the data speak" approach was implemented in the pipeline described above. The retrieval and selection process was guided by features (recurrence, prototypically, characteristic word combinations) inherent in the data set, which gives rise to a data-driven anthology. As a whole, the anthology does not aim to present hitherto unknown or surprising experiences. Instead, the goal is to challenge the implicit banality of experiences such as "children crying for their parents" by letting survivors talk about them (where and how they happened; most importantly what and how they felt). The contribution of the anthology is the offering of a wide-scale overview of a large variety of experiences - narrated by victims themselves and retrieved with a bottom-up approach - which would not be accessible by reading individual testimonies.

The goal of this work can be summarized with an analogy. Original works of Pre-Socratic philosophers vanished forever; nonetheless, their intellectual world have remained accessible and investigable through hundreds of fragments recovered from later works (Kirk et al., 1957). Individual experiences of millions perished, but their likely experiences continue to live through fragments in testimonies. Our contemporary understanding of the Holocaust is by large based on archival sources produced by perpetrators. These sources can help to investigate the process through which victims went through, but not the way victims experienced the process. The anthology of recovered fragments wants to impact scholarship by presenting the perspective of the victim from less studied angles. The overall goal is to let those who did not survive speak through recovered fragments.

References

Alzahrani S.M, Salim N and Abraham A (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Trans Syst Man Cybern Pt C Appl Rev IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 42(2): 133–49.

Bernard-Donals, M. F. and Glejzer, R. R. (2001). *Between Witness and Testimony: The Holocaust and the Limits of Representation (UPCC Book Collections on Project MUSE)*. State University of New York Press.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4/5): 993–1022.

Büchler, M., Burns, P. R., Müller, M., Franzini, E. and Franzini, G. (2014). Towards a Historical Text Re-use Detection.

Felman, S. and Laub, D. (2013). *Testimony: Crises of Witnessing in Literature, Psychoanalysis and History.* Florence: Taylor and Francis http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=1486841 (accessed 26 April 2018).

Frey BJ and Dueck D (2007). Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315(5814): 972–76.

Hartman, G. (2015). *Scars of the Spirit: The Struggle against Inauthenticity*. New York: St. Martin's Press http://rbdigital.oneclickdigital.com (accessed 26 April 2018).

Kirk, G. S. and Raven, J. E. (1957). *The Presocratic Philosophers: A Critical History with a Selection of Texts*. Cambridge, England: University Press.

LaCapra, D. (2014). *Writing History, Writing Trauma*. Baltimore: Johns Hopkins University Press.

Langer, L. L. (2007). *Holocaust Testimonies: The Ruins of Memory*. New Haven: Yale University Press.

Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *ArXiv:1405.4053 [Cs]* http://arxiv.org/abs/1405.4053 (accessed 26 April 2018).

Levi, P. (2018). *Drowned and the Saved.* London: ABACUS.

Rosen, A. (2009). *Sounds of Defiance: The Holocaust, Multilingualism, and the Problem of English*. Lincoln, Neb.; Chesham: University of Nebraska Press.

LitViz: Visualizing Literary Data by Means of text2voronoi

Tolga Uslu
uslu@em.uni-frankfurt.de
Goethe University of Frankfurt, Germany

Alexander Mehler
mehler@em.uni-frankfurt.de
Goethe University of Frankfurt, Germany

Dirk Meyer
dirk-meyer1@gmx.net
Goethe University of Frankfurt, Germany

Abstract

We present LitViz, a webbased tool for visualizing literary data which utilizes the text2voronoi algorithm to map natural language texts onto voronoi diagrams. These diagrams can be used, for example, to visually differentiate between (groups of) authors. Text2voronoi utilizes the paradigm of text visualization to reconstruct text classification (e.g., authorship attribution) as a task of image

classification. This means that, in contrast to conventional approaches to text classifiction, we do not directly use linguistic features, but explore visual features derived from the texts' visualizations to perform operations on texts. We illustrate LitViz by means of 18 authors, each of whom is represented by 5 literary works.

Introduction

In this paper we present a new tool, called LitViz, for the visual depiction of literary works. To this end, we utilize the text2voronoi algorithm (see Mehler et al. (2016b)) which maps natural language texts to image representations. The idea is to generate images of texts which can be used instead of these texts' symbolic information to characterize them, for example, in terms of authorship, topic or genre. Text2voronoi is in line with the paradigm of text visualization to reconstruct text classification (e.g., authorship attribution) as a task of image classification. In contrast to conventional approaches to text classification, we therefore do not directly use linguistic features, but explore visual features derived from the texts' visualizations in order to identify, for example, their authors. We exemplify LitViz by means of 18 authors each of whom is represent by 5 literary works. LitViz allows for interacting with the visualizations of these works in two modes: two- and three-dimensionally (see Figure 1 and 2).

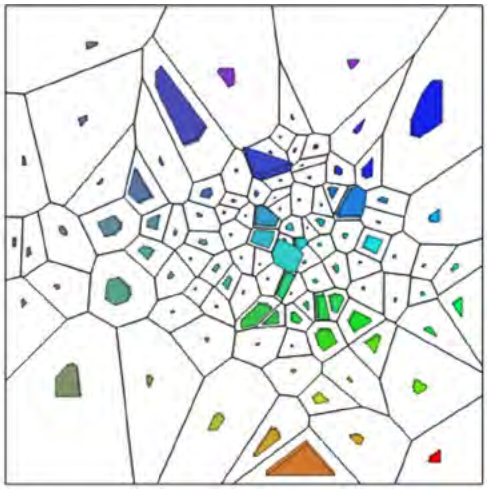


Figure 1: Visual depiction of E.T.A. Hoffmann's Das steinerne Herz

Related Work

The idea of visualizing literature was inspired by Martin Wattenberg's The Shape of Song1 (Wattenberg, 2001; Wattenberg, 2002). Wattenberg explores identical or otherwise repetitive passages of a composition to visually depict them. This is done by means of semicircles, which combine repeated and repetitive positions in such a way that the micro- and macro-structure of a composition becomes visible. Our idea is to transpose this idea to the visualization of literary data.

Kucher and Kerren (2015) give an overview of state-of-theart techniques of text visualization and present a website that allows for differentiating between these techniques. Cao and Cui (2016) provide a systematic review of many advanced visualization techniques and discuss the fundamental notion of information visualization.

Mehler et al. (2016a) present a web tool called Wikidition which allows for automatically generating large-scale editions of text corpora. This is done by using multiple text mining tools for automatically linking lexical, sentential and textual data. The output is stored and visualized using a MediaWiki. Thus, any Wikidition is extensible by its readers based on the wiki principle.

Rockwell and Sinclair (2016) present a detailed web tool, called Voyant tools, for visualizing texts. Unlike Voyant, our focus is on non-standard techniques of visualizing textual data that go beyond histograms, scatterplots, line charts and related tools.

Generally speaking, text visualization supports distant reading as introduced and exemplified by Moretti (2013), Rule et al. (2015) and Michel et al. (2011). These approaches show how visualizations that support distant reading may look like to get overviews of documents by just looking at the final visualizations. LitViz is a tool following this tradition: it utilizes text2voronoi to extend the set of techniques mapping textual data. In this way, it combines Wattenberg's approach with distant reading techniques from the point of view of text visualization.

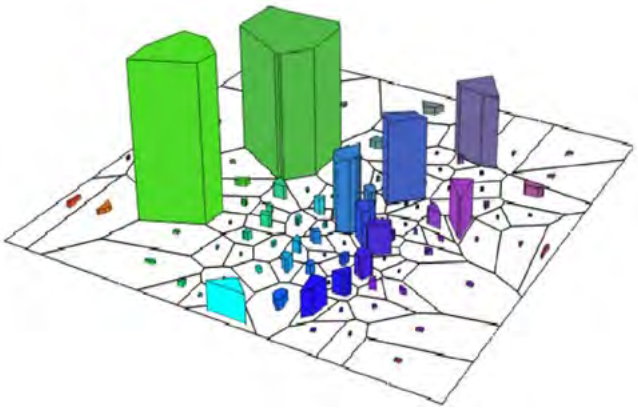


Figure 2: 3D visualization of Franz Kafka's Der Kübelreiter.

Model

Our goal is to generate images from literary works in a way that text classifiers can be fed by the features of these iconic representations in order to perform classification experiments, for which usually linguistic features are explored. This is the task of the text2voronoi algorithm, which calculates image representations of texts in four

steps Mehler et al. (2016b): In the first step, the input text is analyzed by means of TextImager Hemati et al. (2016) to extract linguistic features in the usual way, that is, features, spanning a vector space of linguistic data. In the second step, the resulting vector space is used to compute embeddings for each of the extracted linguistic features. Embeddings are produced by means of word2vec (Mikolov et al., 2013). In the third step, a voronoi tessellation of the embedded features is computed. As a result, each lexical feature is mapped onto a separate voronoi cell whose neighborhood reflects the feature's syntagmatic and paradigmatic associations with other features of the same space. The topology of the voronoi cells spans a voronoi diagram that visually represents the input text. Each of these cells is characterized by its filling level, transparency and height (third dimension) thereby reflecting its co-occurrence statistics within the input text, while the position and size of a cell is determined by the embedding of the corresponding feature – for the mathematical details of this algorithm see Mehler et al. (2016b). Finally, the text2voronoi algorithm extracts visual features from the voronoi diagrams to feed classifiers performing classifications of the input texts.

LitViz utilizes the first three steps of this algorithm. Unlike the classical text2voronoi procedure, it does not address the final step of classification. Rather, it gives access to voronoi diagrams of input texts via a two-dimensional graphical interface, which can be transformed into a three-dimensional one by means of user interaction. These two- and threedimensional text representations can be used by the user of LitViz to interact with the underlying input texts in order to highlight single voronoi cells, to change her or his reading perspective or to visually compare voronoi diagrams of different texts. In this way, LitViz paves the way to a kind of a comparative distant reading by making accessible the visual depictions of different texts in an interactive manner.

The LitViz Tool

We have selected 18 authors of German literature each of whom is represent by 5 literary works. The works are taken from the Project Gutenberg (https://www.gutenberg.org/) and visualized by means of the text2voronoi algorithm. Any of these examples is made accessible by the front page of LitViz (see Figure 3). When hovering over a voronoi cell of the voronoi diagram of a sample work, information about the underlying linguistic feature represented by this cell is displayed. According to Mehler et al. (2016b), we call these images VoTes: Voronoi diagram of a Text. LitViz presents VoTes via a graphical user interface for two- and three-dimensional interactive graphics. In this way, we go beyond Wattenberg's 2D depictions of musical pieces.

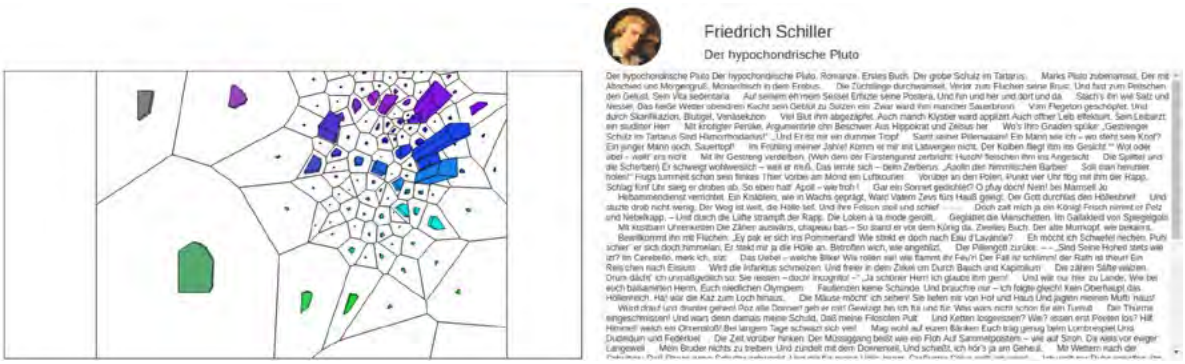


Figure 3: Front page of LitViz.

The second page (tab) of LitViz gives access to the comparison tool. Here the user first selects the number of VoTes to be compared. Then the user selects a subset of works of the authors to be compared. In the example in Figure 4, we compare four VoTes of two authors: two VoTes of two works of Heinrich Heine (top) and two VoTes of Heinrich Mann (bottom). It is easy to see that these VoTes fall into two classes, depending on the underlying authorship. Heinrich Mann's two VoTes are organized around a center that is composed of many small cells, while there is a small subgroup of peripheral cells that are large. In contrast to this, the two VoTes of Heinrich Heine do not display such a center and are more evenly distributed in terms of their size. It is a main task of LitViz to allow for such comparisons. In this way, that is, by interacting with the texts' image representations and by using the mouse-over technique, the user can study single features and how they are related to other features of the same representational space.

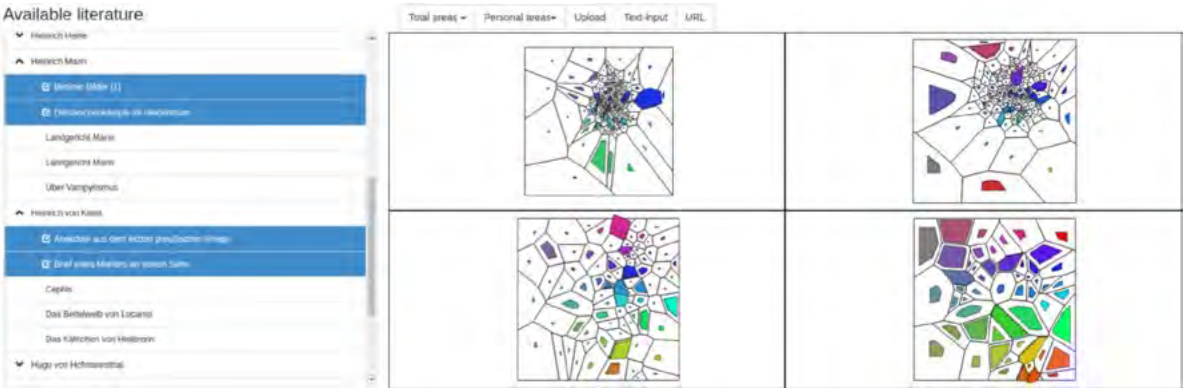


Figure 4: Comparison tool: Heinrich Heine (top) in comparison to Heinrich Mann (bottom).

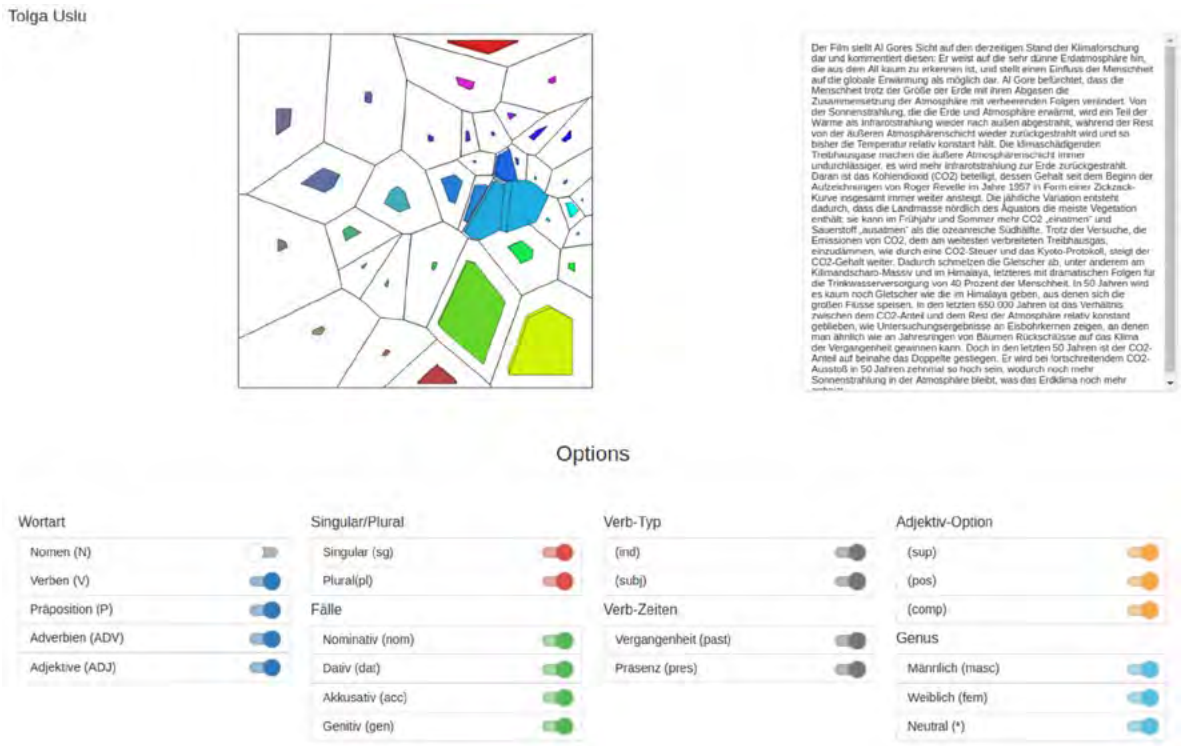


Figure 5: Custom VoTe with filter options.

Conclusion

We introduced a novel web tool, called LitViz, for visually depicting natural language texts based on the text2voronoi algorithm. LitViz enables the comparison of the visualizations of different texts. This allows, for example, for comparing the styles of the underlying authors visually. In this way, we extend the existing tool palette of distant reading. LitViz can be accessed via: http://alba.hucompute.org/text2voronoi

References

Cao, N. and Cui, W. (2016). *Introduction to Text Visualization. Atlantis Briefs in Artificial Intelligence*. Atlantis Press.

Hemati, W., Uslu, T., and Mehler, A. (2016). TextImager: a distributed uima-based system for NLP. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 59–63.

Kucher, K. and Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community in sights. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific*, pages 117–121. IEEE.

Mehler, A., Gleim, R., vor der Bruck, T., Hemati, W., Uslu, ¨ T., and Eger, S. (2016a). Wikidition: Automatic lexiconization and linkification of text corpora. *Information Technology*, pages 70–79.

Mehler, A., Uslu, T., and Hemati, W. (2016b). Text2Voronoi: An image-driven approach to differential diagnosis. In *Proceedings of the 5th Workshop on Vision and Language (VL'16) hosted by the 54th Annual*

Meeting of the Association for Computational Linguistics (ACL), Berlin.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint *arXiv*:1301.3781.
Moretti, F. (2013). *Distant reading*. Verso Books.
Rockwell, G. and Sinclair, S. (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press.
Rule, A., Cointet, J.-P., and Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112(35):10837–10844.
Wattenberg, M. (2001). *The shape of song*. Website http://www.turbulence.org/Works/song/mono.html.
Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 110–116. IEEE.

Lo que se vale y no se vale preguntar: el potencial pedagógico de las humanidades digitales para la enseñanza sobre la experiencia mexicano-americana en el midwest de Estados Unidos

Isabel Velázquez
mvelazquez2@unl.edu
University of Nebraska, United States of America

Jennifer Isasi
jennifer.isasi@huskers.unl.edu
University of Nebraska, United States of America

Marcus Vinícius Barbosa
mbarbosa@huskers.unl.edu
University of Nebraska, United States of America

Objetivo. La meta de esta presentación es describir el potencial y las limitaciones de las humanidades digitales para la enseñanza sobre la experiencia social de colectivos multilingües poco estudiados. Basamos nuestro argumento en el caso de la construcción de un archivo digital que reúne la correspondencia y otros documentos personales de una familia que emigró de Zacatecas, México, a Colorado y a Nebraska, Estados Unidos, en la primera mitad del siglo XX. Durante los últimos cuatro años, el repositorio que aquí se describe ha servido como espacio de aprendizaje activo para estudiantes de licenciatura y posgrado de una universidad pública del Midwest. Mediante su participación en el proyecto, los estudiantes aprenden sobre la experiencia sociolingüística de las comunidades hispanohablantes en EU transcribiendo, traduciendo, digitalizando, marcando y analizando estos objetos.

Contexto. Varios autores han destacado la necesidad de sortear el abismo entre la investigación y la enseñanza en las humanidades digitales (Hawkins, Mannan, Iantorno y Bistline, 2014; Hirsch, 2012). En su introducción a un número especial dedicado a la pedagogía y las humanidades digitales, Iantorno (2014) se pregunta: ¿Cómo enseñar habilidades en humanidades digitales a estudiantes de licenciatura en un curso tradicional? ¿y a estudiantes de posgrado en un seminario de literatura? (140). Nuestra respuesta es un modelo que intenta servir como puente entre la investigación y la enseñanza, y que proporciona a los estudiantes la oportunidad de aprender haciendo.

Este archivo digital difiere de otros proyectos basados en la experiencia de las comunidades hispanohablantes en Estados Unidos porque va más allá de la preservación y descripción (Iowa, 2004); de la historia oral (Foulis); del uso de audio y video para la enseñanza del español (Spanish in Texas); y de la inclusión de monografías dentro de colecciones mayores (Kansapedia). Es un esfuerzo deliberado por integrar tres componentes: la investigación lingüística, social e histórica; una cara abierta a los miembros del público general interesados en saber más sobre los latinos en el Midwest, y una sección con materiales, actividades y guías de clase para maestros de español interesados en usar los objetos de la colección en su salón de clase.

Descripción. **Family Letters** es un proyecto interdisciplinar y bilingüe que surgió a partir de una búsqueda aparentemente sencilla. Lo que originalmente se pensó como un trabajo de traducción, terminó expandiéndose a un proyecto digital colaborativo entre el Centro para la Investigación Digital en las Humanidades y el Departamento de Lenguas y Literaturas Modernas de la Universidad de Nebraska-Lincoln, y la comunidad local. Esta colección, que cuenta con 713 objetos digitalizados y marcados en TEI en su mayoría (225 cartas, 199 documentos, 270 fotografías y 19 objetos personales de diversa naturaleza), supone una excelente fuente de información sobre la vida diaria de una familia de inmigrantes en Estados Unidos. Puesto que los artefactos fueron escritos tanto en español como en inglés, la colección nos permite examinar fenómenos de contacto lingüístico, rasgos de lengua no estándar, y procesos de pérdida y mantenimiento lingüístico cultural descritos en voz de sus protagonistas.

Aspectos técnicos. A pesar de que el enfoque de nuestra presentación es el potencial pedagógico de un archivo digital de este tipo, mencionaremos de manera breve algunos aspectos técnicos que son de particular importancia para un proyecto de este tipo. Por ejemplo,

los detalles de codificación y la construcción de la personografía, entre los que se incluyen los retos que presenta la codificación de nombres personales que a veces aparecen en la colección hasta con 27 iteraciones que cambian por idioma, nivel de escolaridad, inestabilidad ortográfica o convención social.

Lo que se vale y no se vale preguntar. Este proyecto posibilita la construcción de un conocimiento amplio sobre la experiencia de migración mexicana en el Medio Oeste. Los materiales permiten el desarrollo de investigaciones interdisciplinarias, desde la microhistoria hasta la sociolingüística. Es posible observar la potencialidad del proyecto para la elaboración de una pedagogía en tres niveles. El primer nivel, centrado en la manipulación más inmediata de los objetos digitales, permite al estudiante de licenciatura que es hablante de español como lengua de herencia reconocer elementos de su propia experiencia y desarrollar habilidades lingüísticas de forma comparada. El segundo nivel, enfocado en la sistematización de los materiales de la colección, permite que los estudiantes de licenciatura y posgrado hagan investigación en una primera aproximación a las Humanidades Digitales. El tercer nivel, enfocado en la creación de materiales pedagógicos permite a los maestros usar los materiales de la colección en el aula.

El uso de un archivo digital para la enseñanza del español. Los materiales pedagógicos diseñados a partir de los materiales de la colección tienen como enfoque a los estudiantes de español de nivel intermedio/avanzado de high school y universidad. El objetivo de este esfuerzo es permitir que los estudiantes utilicen los recursos de la colección para desarrollar sus habilidades lingüísticas y su competencia cultural mientras aprenden más sobre la experiencia de las familias mexicano-americanas a principios del siglo XX. Estas actividades fueron pensadas para utilizarse en todo o en parte en el salón de clase con el fin de ayudar a los estudiantes a fortalecer sus habilidades de escritura, desarrollo de vocabulario, ortografía y gramática. Adicionalmente, estas actividades permiten a los maestros abordar temas centrales en la experiencia sociolingüística de las comunidades latinas en EEUU, tales como la pérdida y el mantenimiento intergeneracional del español, la variación dialectal y los fenómenos de contacto lingüístico (Beaudrie y Potowski, 2014).

Al trabajar directamente con los documentos de una familia mexicano-americana, los estudiantes no solo están expuestos a la lengua: se abren aquí las puertas a una experiencia que posibilita la construcción de una identidad social concebida de manera compartida. El principal objetivo pedagógico es que al manipular estos objetos digitales, los estudiantes establezcan una conexión que les permita reflexionar en su propio papel como actores sociales. A partir de la historia particular de esta familia, los estudiantes pueden dimensionar los eventos cotidianos en la escala más amplia de la experiencia de los inmigrantes en Estados Unidos.

Invitación a la colaboración, direcciones futuras. Uno de nuestros principales objetivos es garantizar el acceso a estos materiales a aquellos investigadores interesados en la inmigración, el contacto de lenguas y temas afines. Por otro lado y, sobre todo, ofrecemos nuestro proyecto a aquellos profesores en México, Estados Unidos y otros países que estén interesados en compartir con sus alumnos una faceta de la experiencia mexicano-americana de principios del XX. Por descontado, pretendemos la ampliación de este proyecto digital en el futuro, con colecciones de familias provenientes de otros países de Latinoamérica y residentes hoy en día en Nebraska.

References

Beaudrie, S.M., Ducar, C. and Potowski, K., 2014. *Heritage language teaching: Research and practice*. New York, NY: McGraw-Hill Education Create.

Foulis, E., 2017. HISTORIAS: LATIN@ VOICES IN OHIO. *alter/nativas, latin american cultural studies journal*, (7).

Hawkins, A.R., Mannan, J., Iantorno, L., Bistline, E. and Haileselassie, S., 2014. Perspectives on Work and Workflow in Digital Humanities.

Hirsch, B.D., 2012. : *Digital Humanities and the Place of Pedagogy*.

Iantorno, L.A., 2014. Introducing Digital Humanities Pedagogy. *CEA Critic*, *76*(2), pp.140-146.

Kansapedia nd, Kansas Historical Society, accessed 23 April 2018, <http://www.kshs.org/kansapedia/kansapedia/19539>

Spanish in Texas Project, 2010, University of Texas at Austin, accessed 23 April 2018, <http://spanishintexas.org/>

The Mujeres Latinas Project nd, Iowa Women's Archives, The University of Iowa, accessed 23 April 2018, <https://www.lib.uiowa.edu/iwa/mujeres/>

Solving the Problem of the "Gender Offenders": Using Criminal Network Analysis to Optimize Openness in Male Dominated Collaborative Networks

Deb Verhoeven
deb.verhoeven@uts.edu.au
University of Technology Sydney, Australia

Katarzyna Musial
katarzyna.musial-gabrys@uts.edu.au
University of Technology Sydney, Australia

Stuart Palmer
ststuart.palmer@deakin.edu.au
Deakin University, Australia

Sarah Taylor
sarahtaylor247@gmail.com
RMIT Univerrsity, Australia

Lachlan Simpson
ladatakid@gmail.com
Independent Researcher

Vejune Zemaityte
vzemaity@deakin.edu.au
Deakin University, Australia

Shaukat Abidi
shaukat.abedi@gmail.com
University of Technology Sydney, Australia

Statistics describing the inequitable conditions for women in global film industries have been gathered and circulated for more than 30 years. These statistics have barely deviated despite the development and application of a range of equity policies. In some instances the participation of women has become marginally worse. Furthermore, the repeated release of poor equity data has given the industry's structural misogyny an air of inevitability. This situation is not unique to the film industry.

Our project uses newly available forms of data and data analysis to proose innovative strategies for redressing the systemic and frequently personal bias against women in two different "merit based" industries – the film industry and academic grants schemes. Using data derived from the Australian, Swedish and German film industries as well as data from two different Australian research grant schemes, we propose, compare and evaluate several approaches to controlling collaborative network evolution in order to increase network openness. Our approach is informed by the findings of a major longitudinal study which found that "female actors have a higher risk of career failure than do their male colleagues when affiliated in cohesive networks, but women have better survival chances when embedded in open, diverse structures." (Lutter 2015)

This project rests on two inter-related manoeuvres then. Firstly, it flips the object of analysis. If we are going to make these industries a better place for women and other minorities then we need to understand the specific operations of gatekeeping that maintain the dominance of white, cis men. The second aspect of the project is to use the data we have collected about specific collaboration networks to propose an innovative course of action to change male dominated, exclusionary environments.

This data, on creative roles in films and on researchers receiving grants, contains not only information about the characteristics of projects and all the people involved but also, equally importantly, relational data that enables us to look into the connections within and across teams working on films or research projects respectively. Social network analysis (SNA) provides methods for visualising these group relationships, and through quantitative measures that characterise network structure, provides methods for identifying strategically important components and participants in the network. It also therefore points to ways in which these networks can be most effectively "dismantled" or opened up.

Network visualizations are useful for observing the implicit structure in the collaboration data, for understanding the scale of the problem and for identifying the key connected players. By adding the dimension of gender to these network visualizations we can clearly see the influence of gender on patterns of domination. In addition to making the existing network patterns visible our further concern was to see beyond these patterns and look for ways in which the data could suggest the most effective interventions for challenging and changing the status quo.

In this regard, network visualizations enable us to quickly identify outliers, and easily demonstrate the discrepancies between a given network and the more open reference network we would like to achieve. By depicting changes in both the network structure and its components, visualizations can facilitate the process of testing different policy proposals for achieving social change in organisational or industrial settings and can be used to monitor the emergence of new patterns (especially unwanted ones).

There is some precedent in approaching network visualizations in this way. Crime experts and counter terrorist specialists have used "criminal network analysis" for example to identify opportunities to undermine the coherence of dominant groups.

Drawing on the literature on the use of social network analysis to characterize criminal networks and identify key nodes whose removal would disrupt the network (i.e., Borgatti, 2006; Rostami & Mondani, 2015; Schwartz & Rouselle, 2009; Réka A., Hawoong J., & Barabási A-L., 2000), we investigated the network of male-only producers and other creatives in the film industry and male-only networks of researchers in the university sector. We investigated the impact of key players in these networks, and the hypothetical impact of removing different key players.

Specifically, we used Borgatti's network fragmentation factor (F) (equation 4 in Borgatti, 2006) as a quantitative measure of network disruption. In this equation, the F value is 0 when there is no fragmentation in a network (all nodes connected in a single component), and is 1 when all nodes in a network are isolated. Using an iterative script, F was calculated for the initial network, a node was removed, and then F was recalculated to assess the impact of the node removal on network fragmentation.

At each iteration, the increase in F obtained from removing a range of male producer or male researcher nodes from the initial networks was calculated and compared. The large(est) male producer/researcher node in the centre of a given network suggests itself as a node whose removal would significantly increase the network fragmentation, and it was indeed the case that this chan-

ge yielded the largest increase in network fragmentation. Those male producer or researcher nodes whose removal from the initial network yielded relatively large increases in network fragmentation were also observed to have relatively high values of 'betweenness centrality', as computed for the initial network. Node betweenness centrality measures how often a node appears on shortest paths between nodes in the network. A high betweenness centrality in the initial network provides a heuristic for identifying candidate nodes for removal that would significantly increase the network fragmentation.

This paper will present the project's findings on the best strategies for dismantling domination patterns and behaviours in collaborative networks, one of them being removing the nodes with the highest betweenness centrality, or in the case of male dominated collaboration networks, removal of the men we call the "gender offenders".

References

Borgatti, S. P. (2006). Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory, 12*(1), 21-34. doi:10.1007/s10588-006-7084-x

Lutter M., (2015). Do Women Suffer from Network Closure? The Moderating Effect of Social Capital on *Gender* Inequality in a Project-Based Labor Market, 1929 to 2010. *American Sociological Review* Vol. 80(2) 329–358 doi: 10.1177/0003122414568788

Réka A., Hawoong J., & Barabási A-L. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378–382, doi:10.1038/35019019

Rostami, A., & Mondani, H. (2015). The Complexity of Crime Network Data: A Case Study of Its Consequences for Crime Control and the Study of Networks. *PLOS ONE, 10*(3), e0119309. doi:10.1371/journal.pone.0119309

Schwartz, D. M., & Rouselle, T. (2009). Using social network analysis to target criminal networks. *Trends in Organized Crime, 12*(2), 188-207. doi:10.1007/s12117-008-9046-9

"Fortitude Flanked with Melody:" Experiments in Music Composition and Performance with Digital Scores

Raffaele Viglianti
rviglian@umd.edu
University of Marylands, United States of America

Joseph Arkfeld
jarkfeld@umd.edu
University of Marylands, United States of America

Concert goers are growing accustomed to performers, particularly small ensembles, who bring a tablet computer on stage and place it on the music stand before they begin to play. Perhaps less noticed, a silent tap on the side of a screen, or the nimble pressing of a foot pedal has replaced the rustle of page turning. Notwithstanding an observable increase in the use of digital devices to study and perform music, the transactions between score and performers are only mildly affected: the digital score primarily acts like its paper counterpart: a static, largely reliable instrument for learning and performing a given work of music. The flexibility of the digital medium, as opposed to something fixed on paper, calls for a more modern concept of the score, one that undermines its prescriptiveness by making room for material and features targeted at supporting performers in their interpretation and advocacy of a musical work. What would such a truly digital score look like? And more importantly, what are the critical instruments necessary to understand its impact on performance and music making at large?

Scholarly digital editions, through carefully encoded music notation, are leading the re-definition of the digital music score and are tackling some of these questions, but these publications often take the form of rich websites apt to careful study, but less for performance practice. The role of dynamic digital scores for performance still demands investigation. This paper will present three experiments based on the same technical application, two of which involve newly composed music, that require a performer to use a digital score as it changes its shape based on factors out of the performer's control.

The experiments

In these experiments, we use location-based weather data obtained from the web to introduce variation; as a factor that neither the composer or the performer can control, weather data provides a mechanism for driving change in the music notation. *Meteomozart*, the earliest of our experiments, is a dynamic score of Mozart's Piano Sonata No.13 in B major, K.333/315c. At the time of writing, the score only includes the first theme of the first movement (about 60 measures). *Meteomozart* adjusts the score based on the weather at the performer's location (or at a location set by the user). Different slurs and dynamics are shown, taken from four sources: Mozart's 1783 manuscript, the first printed edition (1784) and two performing editions by Bartók (1911) and Saint-Saëns (1915). This sonata is often discussed as an example of the lack of clarity in Mozart's slurring, George Barth, for instance discusses how slurs in the sonata were changed substantially from the autograph manuscript to the first edition, and more dynamics were introduced. In *Meteomozart*, certain weather condition (e.g. sunny, cloudy, rainy, etc.), obtained via the Dark Sky API, will change the score to show slurs and dynamics from one of the four sources. For example, Mozart's autograph comes up in clear weather, while Bartók will show up in stormy weather. While there is no scholarly reason to associate weather conditions with specific

editions, the driving idea behind the experiment is to take away some control from the performer, or rather, to make it more obvious that some control is always taken away in a printed edition. Editors typically build an understanding of the historical contingencies that makes one version better than the other; also, it makes sense to provide a clean text that performers can pick up and play. But often there are very good reasons for not making decisions for the performer (Chopin, for example, published versions of his works in three countries all with minor differences, all published around the same time; which one is "correct"?). *Meteomozart* presents a slightly unpredictable text instead of a clear one. It is obvious from audio recordings that the same work of music can be performed in more than one way; likewise, this experiment tries to make it obvious that the score may have more than just one "text". What would happen if performers took control of the variants as opposed to having the weather determining them?

Meteomozart's software was re-purposed for a collaboration between composer Joseph L. Arkfeld and the digital humanities scholar Raffaele Viglianti. The first experimental composition resulting from this collaboration was a piano piece, *Chance of Weather*, a modern take on a piano program piece that takes inspiration from weather conditions. Rather than focusing on a specific condition or setting, such as Debussy's *Jardins sous la pluie*, Chance of Weather engages with the weather that is currently affecting the performance environment. The piece invites the outside world into the performance space, which is usually sterile to the elements. If the audience had to walk through a windy and rainy evening to get to a windowless, temperature controlled performance space, they will find that Chance of Weather evokes the gusts and the dampness of their day. Likewise, if they came on a pleasant warm afternoon, the piece will reflect their recent experience.

A second composition, *Blue Bird*, deploys this technique with new parameters by setting to music a fragmentary and variant-rich poem by Emily Dickinson. The unfinished and fragmentary state of Dickinson's late poetry itself provides a point of departure for dynamic notation. Text-setting for a digital dynamic score presents its own unique challenges apart from instrumental music that must be addressed in order for the music to be singable. One of the most directly accessible permutations would be for the text to change between iterations, which would demand a non-traditional text. Dickinson's text at the heart of *Blue Bird* was located on Marta L. Werner's database *Radical Scatters: Emily Dickinson's Late Fragments and Related Texts*. Spanning Dickinson's final years, this collection includes facsimiles and transcriptions of letters, full compositions, drafts, and manuscript copies found after her death. Werner identifies one of these documents as a "trace fragment," and, akin to leitmotifs, these fragments exist both autonomously and as parts of larger compositions. Like many of her published works, the constellation of fragments selected for this piece was not titled by Dickinson; *Blue Bird* is titled after the longest fragment's subject out of 6 total. These fragments lend themselves naturally to a musical style that kaleidoscopically hovers around a sound space, given their contradictory instability and tight interrelation. *Blue Bird* not only sets to music Dickinson's text, but also its textual condition.

To determine what texts will be shown for the performers, the score will use the past 24 hours' apparent temperature and cloud cover data from the location of the performance obtained via the Dark Sky API. The musical setting communicates the relationship between weather and the chosen fragment of text. While it would be possible to use more data points from the API, Using more data points with relatively limited text would produce different settings of the same text, rathern than the more tightly through-composed aleatoric composition that is *Blue Bird*.

The dynamic scores: encoding and presentation

The scores of the experiments described above are encoded with the Music Encoding Initiative XML format, which provides a number of strategies for encoding textual variance and ambiguity. Specifically, the encoding uses elements defined for encoding variants across textual sources and critical apparatus. We define a list of "sources" that correspond to predetermined weather patterns; these correspond to <rdg> elements throughout the text that are grouped within the <app> element when certain musical text changes depending on the weather condition.

```
<app>
  <rdg source="#dry-dayC0 #dry-nightC0"/>
  <rdg source="#dry-dayC0to10 #dry-nightC0to10">
    <layer n="1">
      <note dur="1" oct="4" pname="g" stem.dir="up"/>
    </layer>
  </rdg>
  <rdg source="#dry-nightC11to25">
    <layer n="1">
      <chord dur="1" stem.dir="down">
        <note dur="1" oct="4" pname="g" />
        <note dur="1" oct="5" pname="g" />
      </chord>
    </layer>
  </rdg>
  <rdg source="#dry-dayC11to25">
    <layer xml:id="m-734" n="1">
      <chord xml:id="m-735" dur="1" stem.dir="down">
        <note xml:id="m-736" dur="1" oct="4" pname="f">
          <accid xml:id="m-737" accid="s"/>
        </note>
        <note xml:id="m-738" dur="1" oct="5" pname="f">
          <accid xml:id="m-739" accid="s"/>
        </note>
      </chord>
    </layer>
  </rdg>
  <!-- etc. -->
</app>
```

Ex. 1: Example use of <app> and <rdg> in *Chance of Weather*.

Besides including text in direct alternation, variation is also included by translating musical ideas to different locations in the piece. In order to do this effectively and avoid encoding the same music notation multiple times, these units will be encoded in separate files and included via XInclude operations.

The dynamic scores are published as a dedicated website by using Verovio, an engraving engine for MEI. By producing SVG output that maps directly to the underlying MEI encoding, Verovio makes it possible to locate the pre-composed variants in the text and manipulate the score according to weather data

On Alignment of Medieval Poetry

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de, Germany
Leipzig University

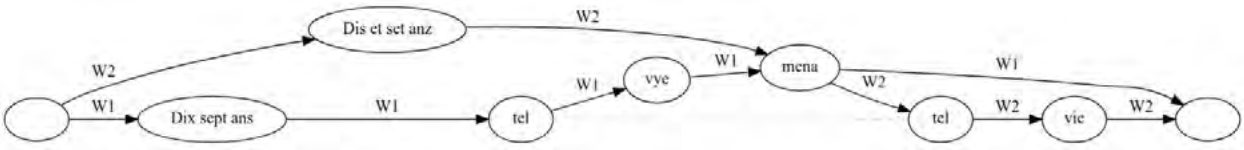
David Joseph Wrisley
djw12@nyu.edu
New York University Abu Dhabi, United Arab Emirates

What constitutes an alignment, different varieties of alignment, or even different degrees of alignment is a topic in need of further interdisciplinary discussion. The co-authors of this paper have been working on the computational alignment of medieval poetry, an exchange that has resulted in the design of a visual analytics system for the exploration of complex textual traditions. The purpose of this paper is twofold: first, to describe how we arrived at the user centered design of the VA system (Heuwing et al., 2016) and second, to introduce an alternative means of alignment, that of Sequence-to-Sequence Models based on recurrent neural networks, that does not oblige the user to adopt a parameter driven approach, but still allows for discovery of baseline potential alignment for subsequent human scoring.

Pre-modern writing exhibiting both textual and performative forms of instability is challenging for alignment. Twentieth-century print editions employed synoptic style layouts for textual traditions where line-level interpolation and excision were most common, as well as rough stanza-to-stanza numbering based on narrative cues in the poem, as in the case of the mid-century edition of the *Chanson de Roland* (Mortier, 1940-44). Alignment in print could not be more granular on account of the highly complex patterns of textual recombination found across different redactions.

Sequence alignment algorithms were originally developed in bioinformatics to identify and analyze functional or evolutionary relationships between genome sequences. Unfortunately, these algorithms are not straightforwardly adaptable to the computational alignment of textual traditions rife with orthographic and transpositional variance (Dekker and Middell, 2011). A number of algorithms have been developed and implemented in user centered design models to examine intertextual similarities, but none of them delivers fully satisfactory results for medieval vernacular poetry (Jänicke and Wrisley, 2017a).

Our computational alignment compares each line of one edition to each line of another edition, marking all significantly similar line pairs as alignment candidates. Whereas for the human reader such candidates are obviously valid alignments, they are not easy to detect by purely computational means. For example, using CollateX (https://collatex.net/) for aligning a pair of lines from the tradition of the *Vie de saint Marie l'Egyptienne* (**Anon_RenartContre1325**: Dix sept ans tel vye mena | **Rutebeuf_SteMarie**: Dis et set anz mena tel vie) yields the following result:



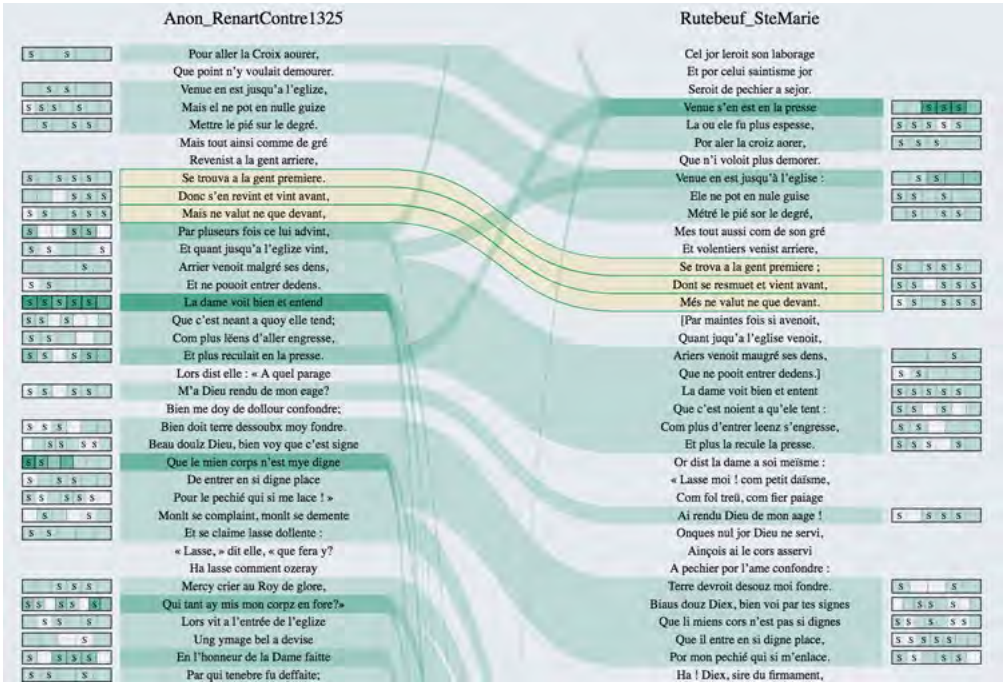
Having only one word match and one transposed word, the pair of lines would not be classified as an alignment candidate. Whereas morpho-syntactic tagging could be helpful in surmounting the problem of orthographic variance, we are still faced with the problem of word order.

In previous work, we have implemented a user defined parameter system in order to achieve initial alignment results, with subsequent scoring by a specialized user. We developed the "white box" alignment system *iteal* (http://iteal.vizcovery.org/) that uses a set of user-configurable parameters to steer the alignment procedure (Jänicke and Wrisley, 2017b):

- **Edit distance:** With orthographically unstable language, variant spellings needed to be taken into account. We define two words as spelling variants if they have the same first letter, and if the string similarity of the remaining substrings is higher than a user-configurable threshold.
- **Coverage:** In order to ensure that a specific proportion of words of both lines are aligned, the user can configure a minimum coverage value of the line.
- **N-grams:** The user can configure the minimum required n-gram size *n* that is the largest number of subsequent word matches of both lines.
- **Broken n-grams:** Quite often, the only difference between two lines is a single word in the middle of

a line that is either inserted, synonymous, or a transposed stopword. Large n-grams, from this perspective, are not achieved. Thus, we allow the user to consider broken n-grams.

Indeed, a parameter-driven approach has suggested many possible sequential alignments. Traditional scenarios of intertextual expansion or contraction of poetry are visualized quite clearly. Take, for example, the condensation of episodes of Rutebeuf's *Vie de saint Marie l'Egyptienne* in the *Renart le Contrefait* that exhibits a conservatism in replication of whole lines or excision of whole lines:



Different redactions of the epic poem the *Chanson de Roland* illustrate a more complex, recombinatory intertextuality. The Venice 7 version is double the length of the oldest extant version known as the Oxford version and the Lyons manuscript is 75% the length of the Oxford. Alignment in this case depends heavily on the use of broken n-grams and edit distance since the versions vary significantly in orthography, word choice and order:

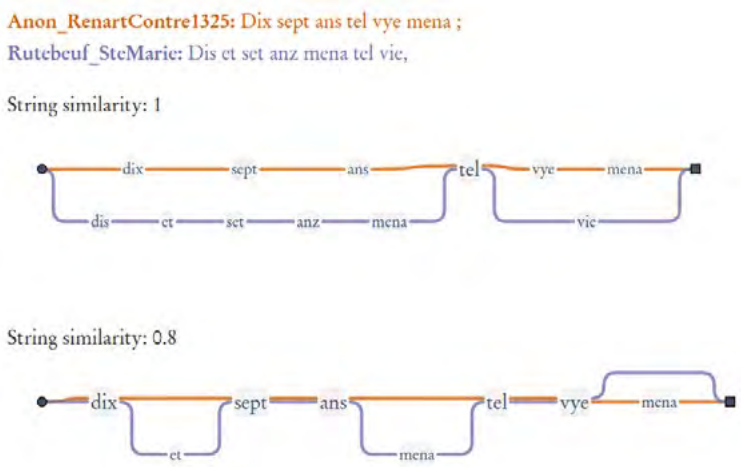
Oxford: Ki est de France, si est mult riches hom
Venice 7: Bien est de Franse, mult par est riches hon

and

Oxford: Ja cil d'Espaigne n'avrunt de mort guarant
Venice 7: Ja cil d'Espeigne de mort n'aront garant.

Sequentially the lines above are divergent, and yet semantically they are nearly identical.

Using the aforementioned example from the *Vie de saint Marie l'Egyptienne*, an alignment example is considered an alignment candidate by *iteal* using a combination of several parameter settings, e.g., a string similarity of 80%, a coverage of 40% and allowing for broken 4-grams:



Different parameter settings yield very different initial alignments for consideration and scoring by the specialized user. Too liberal or too strict of a choice in settings yields either too many possible alignments or almost none at all.

In oral literatures textual reuse is not limited to full-line intertextuality, however, but rather exists along a continuum: from small formulaic expressions to partial and full line reuse. It is on this point that *iteal* does not allow for more granular scoring of partial line alignments or multi-line segment alignment, as in the examples that follow:

Oxford: Je vos plevis, ja returnerunt Franc.
Venice 7: Je vos plevis, ja sera il tornez,
Lyons: je vos plevis sempres ert retornant

and

Anon_RenartContre1325: Ainsi paist comme beste mue.
Rutebeuf_SteMarie: Si comme une autre beste mue.

To make matters more complex, rewriting of medieval texts engages with different genres and prosodies as well as jumping back and forth between poetry and prose. *Iteal* does not perform optimally yet with different forms. Our research, thus far, has focused on poetry, where the common denominator across textual redactions is the poetic line. Below we see some examples of alignments across versions of the *Vie de saint Alexis* (one written in octosyllablic verse and the other in decasyllablic), 3-grams matches produce simply too many false alignments to be valid. Alignments based on 4-gram point to common narrative leitmotifs within the text, such as the force against which the saint resists, his father's home as a setting:

AlexisOctP: Treire par force et par engin
AlexisPRI: Il me prendront par force et par poeste

and

AlexisOctP: Que il laissa en la maison son pére
AlexisPRI: Enz la meson son pére issi.

Whereas we implemented the calculation (or exclusion) of alignments using a medieval French stopword list, this is not necessarily valid across our samples, as the proposed alignments below illustrate:

AlexisPQ: Adonc le fist son pére de l'escole partir
AlexisP11: Il le nonçat son pedre Eufemien

and

AlexisPQ: Ad un des porz qui plus est pres de Rome
AlexisP11: Li uns des pers de Romme c'on nommoit Contantin.

Whereas the latter set of aligned lines satisfies a computational condition of a broken 4-gram and minimum coverage of 40%, ultimately the alignment seems silly to a human reader for the collapsing of two substantives, *porz* [seaport] and *pers* [great men].

A parameter-driven "white box" system might seem appealing for its algorithmic transparency in the alignment of medieval text versions, however, we are now turning to an alternative "black box" solution that employs Sequence-to-Sequence Models based on recurrent neural networks (Sutskever et al., 2014). While this idea was not implemented initially, as it makes it difficult to backtrack, our work has begun to migrate to such models (Cho et al., 2014; Bengio et al., 2015). As opposed to a parameter set with its concomitant results, the recurrent neural network system functions with requisite semi-automated training indicating which alignments are appropriate, and which ones are not. While taking into account the contexts in which words appear, the neutral network suggests alignment candidates. We can deliberately map *Line-i-of-Edition-A* to a certain hash value, and likewise its variant *Line-j-of-Edition-B*, thereby training the neural network to find the candidate *Line-k-of-Edition-C* automatically, in turn mapping similar lines to the same hash value.

The potential of this computational shift is that we can further nuance the palette of possible alignments, without remaining bound to the traditional starting parameters. We plan to move beyond our original model of line to line comparisons and to accommodate other units of comparison. By presenting the results of work in progress in the final section of our paper, we intend to explore whether recurrent neural networks produce results for similar text genres and prosodies.

References

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15*. Cambridge, MA: MIT Press, 2015, pp. 1171-79.

Cho, K., van Merrienboer, B., Gülcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1724–34.

Dekker, R. H. and Middell, G. (2011). Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. Supporting Digital Humanities 2011. University of Copenhagen, Denmark. 17-18 November 2011.

Heuwing, B., Mandl, T. and Womser-Hacker, C. (2016). Methods for User-Centered Design and Evaluation of Text Analysis Tools in a Digital History Project. In *Proceedings of the Association for Information Science and Technology*, 53(1):1–10.

Jänicke, S. and Wrisley, D. J. (2017a). Visualizing Mouvance: Towards a Visual Analysis of Variant Medieval Text Traditions. *Digital Scholarship in the Humanities* 32.suppl_2: ii106-ii123.

Jänicke, S. and Wrisley, D. J. (2017b). Interactive Visual Alignment of Medieval Text Versions, In *IEEE Visual Analytics Science and Technology (VAST) Proceedings, Phoenix, Arizona, 1-6 October 2017*.

Mortier, R., ed. (1940-44). *Les textes de la "Chanson de Roland."* Paris: Éditions de la Geste francor.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS*. MIT Press: Cambridge, MA, USA, pp. 3104–12.

Short Papers

Archivos digitales, cultura participativa y nuevos alfabetismos: La catalogación colaborativa del Archivo Histórico Regional de Boyacá (Colombia)

Maria Jose Afanador-Llach
mj.afanador28@uniandes.edu.co
Universidad de los Andes; Fundación Histórica Neogranadina, Colombia

Andres Lombana
alombana@cyber.law.harvard.edu
Berkman Klein Center for Internet and Society, Harvard University, United States of America

Este artículo explora las prácticas colaborativas de catalogación y creación de metadatos en archivos localizados en contextos de escasa conectividad, acceso tecnológico limitado, e incipiente desarrollo de nuevos alfabetismos. Tomando como ejemplo el proyecto de Catalogación Colaborativa del Fondo Notaría Segunda del Archivo Histórico Regional de Boyacá en la ciudad de Tunja, Colombia, analizamos cómo una plataforma digital y una comunidad de práctica pueden suplir las necesidades de acceso a tecnología, información y conocimiento a través de la "producción entre pares" o "*peer production*" (Benkler 2006; Benkler, Shaw, & Hill 2015) y la cultura participativa (Jenkins et al. 2006; Jenkins 2010). Dada la desigualdad de acceso a recursos tecnológicos, culturales y humanos para proyectos de digitalización y catalogación documental, en este artículo identificamos estrategias para acceder a tecnologías abiertas, y desarrollar nuevos alfabetismos (Lankshear and Nobel 2006, 2007; Dussel 2009; Jenkins et al. 2006; Jenkins 2010) que faciliten la producción colectiva de conocimiento y la construcción de culturas participativas desde el sur global.

En Colombia, la situación de numerosos archivos históricos regionales, se ha caracterizado por la carencia de una organización sistemática de sus colecciones, contribuyendo a que permanezcan subutilizados por parte de los investigadores y del público general (Marín 2004). Los procesos de digitalización presentan entonces una oportunidad no solamente para la preservación de archivos sino también para la catalogación y creación de metadatos de calidad que garanticen el acceso y usabilidad a futuro, y para el fomento de una cultura participativa. Sin embargo, existen numerosos archivos privados con colecciones patrimoniales que carecen de acceso a los recursos para llevar a cabo procesos de digitalización, catalogación y creación de metadatos. Tal es el caso del Archivo Histórico Regional de Boyacá (AHRB), en Tunja, Colombia, un archivo privado con colecciones que van desde 1539 hasta 1850, sin acceso a los recursos y apoyos de la red pública de archivos, y carente de catálogos para algunas de sus colecciones documentales.

A partir de un experimento de construcción colaborativa de catálogos para el AHRB en este artículo abordamos la siguiente pregunta: ¿De qué manera puede el uso de tecnologías digitales y en red por expertos y aficionados ampliar el alcance de la investigación en las humanidades digitales en contextos de escasa conectividad, acceso tecnológico limitado e incipiente desarrollo de nuevos alfabetismos? A través del análisis de las motivaciones y prácticas socioculturales desarrolladas por los participantes del proyecto AHRB, elaboramos una reflexión sobre los retos y oportunidades que la producción colaborativa de información y conocimiento, o „producción entre pares" (*peer production*), ofrece a los procesos de migración de materiales culturales a formatos digitales, particularmente en contextos donde el acceso a recursos tecnológicos es limitado. En dicho experimento, el proceso de catalogación del Fondo Notaría Segunda permitió a un grupo de expertos y aficionados conformar una comunidad de práctica (Wenger 1998), desarrollar nuevos alfabetismos relacionados a la paleografía y participar en un proceso de producción entre pares.

Existen diversos proyectos de *crowdsourcing* en las humanidades digitales que han sido objeto de análisis en el mundo angloparlante (Terras 2016). Sin embargo, los retos de la colaboración abierta distribuida en el sur global están conectados a factores culturales, económicos y de acceso a tecnología, que han sido poco estudiados. Nuestro análisis del proyecto del AHRB permite apreciar cómo la comunidad de práctica conformada para la catalogación de documentos históricos le ofrece a los participantes no solo la oportunidad de contribuir a la construcción de la memoria pública (Owens 2012) sino también desarrollar nuevos alfabetismos como el trabajo en red entre pares y la inteligencia colectiva (Jenkins et. al. 2006). A pesar de las brechas digitales existentes en algunos contextos locales, el proceso de catalogación colaborativa permite crear puentes de acceso a tecnología, tejer redes entre expertos y aficionados, y cultivar una cultura participativa, a la vez que contribuye a la conservación y promoción del patrimonio cultural.

References

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Heaven, Connecticut: Yale University Press.

Benkler, Y., Shaw, A and Hill, B.M. (2015) "Peer Production: A Form of Collective Intelligence." In *Hand- book of Collective Intelligence*, edited by Thomas Malone and Michael Bernstein. MIT Press, Cambridge, Massachusetts.

Dussel, I. (2009) "Los nuevos alfabetismos en el siglo XXI: desafíos para la escuela", *conferencia en Virtualeduca, 2009*. http://www.virtualeduca.info/Documentos/veBA09%20_confDussel.pdf

Jenkins, H. (2010) Afterword: Communities of readers, clusters of practices. In M. Knobel and C. Lankshear (Eds) *DIY Media: Creating, Sharing and Learning with New Technologies*. New York: Peter Lang, pp. 231–53.

Jenkins, H. et al. (2006) *Confronting the Challenges of a Participatory Culture: Media Education for the 21st Century*. Chicago: The MacArthur Foundation.
Lankshear, C., & Knobel, M (2007) "Sampling the New' in New Literacies." In Lankshear, C., & Knobel, M. *A new literacies sampler*. New York : P. Lang.
Lankshear, C., & Knobel, M. (2006). *New literacies: Everyday practices and classroom learning*. 2nd ed. Maidenhead, UK: Open University Press.
Marín, M. "Elementos de la archivística colombiana para la historia de los orígenes de la provincia." En *Theologica Xaveriana*, 152 (2004), 707-718.
Owens, T. (2012a). Crowdsourcing Cultural Heritage: The Objectives Are Upside
Down. http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/.
Terras, M. (2016) "Crowdsourcing in the Digital Humanities," in Schreibman, S., Siemens, R., and Unsworth, J. (eds). *A New Companion in the Digital Humanities*, Blackwell Companions to Literature and Culture Series, Wiley
Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

The Programming Historian en español: Estrategias y retos para la construcción de una comunidad global de HD

Maria Jose Afanador-Llach
mj.afanador28@uniandes.edu.co
Universidad de los Andes, Colombia

The Programming Historian comenzó en el año 2008 como una publicación de acceso abierto que publica tutoriales revisados por pares dirigidos a humanistas para aprender una amplia gama de herramientas, técnicas computacionales y flujos de trabajo útiles para investigación y pedagogía. El proyecto está liderado por un equipo de doce editores voluntarios de seis países con el objetivo de crear una comunidad colaborativa y una audiencia de carácter global. Desde agosto de 2016, el equipo editorial de contenidos en español de PH comenzó el proceso de traducción de los más de 50 tutoriales publicados en el portal del proyecto en inglés. Al la fecha se han traducido alrededor de 30 tutoriales a partir de la participación de alrededor de 15 colaboradores de países como Argentina, España, Colombia y México.

La expansión de una comunidad de práctica de humanidades digitales en el mundo hispanoparlante plantea preguntas sobre acceso y diversidad. La brecha digital, en su dimensión de uso y aprovechamiento de tecnologías de la información y el desarrollo de competencias digitales, implica serios retos para la producción de conocimiento sobre HD en el sur global. PH, una publicación en línea de acceso abierto bilingüe, ha desarrollado un modelo para afrontar el problema del acceso global y lingüístico a recursos, metodologías y herramientas digitales para las humanidades. Este compromiso con la diversidad lingüística y geográfica en las humanidades digitales significa comprender los límites y posibilidades de los contextos institucionales, históricos, culturales y económicos en el mundo hispanoparlante.

Estamos en un momento de expansión del campo de las humanidades digitales en España y América Latina. Ya existen programas de posgrados en HD en universidades Latinoamericanas (Universidad de los Andes, Universidad del Claustro de Sor Juana), que se suman a las ofertas ya existentes en España (por ejemplo, LINHD). En este contexto de expansión, PH representa un proyecto colaborativo de servicio académico voluntario, que se sostiene en la conformación de redes globales de conocimiento abierto. El proyecto ha enfrentado los retos que suponen encontrar voluntarios que quieran revisar, traducir y crear tutoriales del inglés al español. Lo anterior, teniendo en cuenta la falta de reconocimiento y validez académica dada la carencia de mecanismos de evaluación de productos de investigación digital (Galina Russell 2016). De igual manera, ha resultado un reto garantizar la calidad de los contenidos desde un punto de vista lingüístico. Por último, el proyecto afronta el reto de combinar una aproximación global, que al mismo tiempo respete la diversidad local y que no reproduzca prácticas colonizadoras. Estos retos además se alinean con la misión de PH crear recursos sustentables con una prioridad por el Acceso Abierto y los recursos libres y de código abierto.

Esta presentación es una reflexión sobre la experiencia del equipo de contenidos en español de *The Programming Historian* en relación al panorama general de las humanidades digitales en el mundo hispanoparlante. En primer lugar, se pretende analizar las estrategias de divulgación del proyecto y evalúa las experiencias de uso de los tutoriales de PH en el salón de clase y en talleres. En segundo lugar, analizamos el comportamiento del tráfico de usuarios del portal de PH en español desde su lanzamiento en comparación con la evolución del tráfico en el portal en inglés. (Ver muestra de datos en las Figuras 1 y 2) Se analizará también cuáles han sido los tutoriales más visitados y los menos visitados, los lugares de mayor acceso y el tiempo promedio de los usuarios en los tutoriales. En tercer lugar, nos gustaría reflexionar, asimismo, sobre los retos de construir una comunidad de colaboradores que además de hacer traducciones, produzca contenidos sobre herramientas y metodologías de trabajo digital para las humanidades en español.

Las estrategias de divulgación del proyecto y de construcción de una comunidad de colaboradores se ha llevado a cabo mayoritariamente a través de redes sociales, encuestas en línea, listas de correos y ocasionalmente charlas presenciales. Mientras se consolidan espacios institucionales que apoyen la investigación desde

las humanidades digitales, consideramos que será difícil que los países de habla hispana produzcan contenidos y tutoriales en español. Sin embargo, los esfuerzos de traducción son esenciales para impulsar una comunidad de práctica en el sur global. A futuro, esperamos que los investigadores en el mundo hispanoparlante y otras partes del mundo contribuyan a la producción de nuevas metodologías, herramientas y flujos de trabajo digital que reflejen las particularidades sociales, culturales e históricas de las humanidades de los contextos de Latinoamérica y España, y el sur global.

País	Enero 2017	Mayo 2017	Octubre 2017
México	163	472	2100
Colombia	85	252	1200
España	454	912	2300
Argentina	94	188	891
Brasil	392	585	878

Figura 1. Número de sesiones en portal de PH desde países hispanoparlantes, 2017

País	Enero 2017	Mayo 2017	Octubre 2017
Estados Unidos	10,000	13,000	23,000
India	3,900	4,500	7,300
Alemania	1,500	1,900	2,100
Reino Unido	2,400	2,200	5,800

Figura 2. Número de sesiones en portal de PH desde países angloparlantes, 2017

La Sala de la Reina Isabel en el Museo del Prado, 1875-1877: La realidad aumentada en 3D como método de investigación, producto y vehículo pedagógico

Eugenia V Afinoguenova
eugenia.afinoguenova@marquette.edu
Marquette University, United States of America

Chris Larkee
christopher.larkee@marquette.edu
MarVL: Marquette University Visualization Laboratory, United States of America

Giuseppe Mazzone
gmazzone@nd.edu
School of Architecture, Notre Dame University, United States of America

Pierre Géal
pierre.geal@univ-grenoble-alpes.fr
Université Grenoble Alpes, France

http://prado.nfshost.com

Hacia 1875-1877, el fotógrafo francés Jean Laurent retrató la Sala de la Reina Isabel del Museo del Prado en Madrid (Fig. 1). Ocupando un espacio absidial en el centro del edificio que el arquitecto Juan de Villanueva había planeado un siglo antes como un salón de juntas, la Sala de la Reina Isabel reunía los cuadros que entonces se consideraban las "perlas" de la colección. De modo similar a la Tribuna de la Galería de los Uffizi o el Salon Carré del Louvre, la colocación de los cuadros propiciaba comparaciones estéticas. En 1893, el espacio fue reformado y en 1899 se convirtió en la Sala Velázquez.

En 2015-2017, a partir de la fotografía de Laurent, nuestro equipo interdisciplinar emprendió una reconstrucción digital en 3D de este espacio que todavía existe, pero ha sido profundamente transformado. Para reconstruir la estructura original, hemos utilizado las medidas que se encuentran en el proyecto de la reforma fechado en 1887 y las pruebas de color recientemente hechas en las paredes del museo. Un bosquejo original de Federico de Madrazo fue utilizado para reconstruir los banquillos.

Una vez que el modelo estaba hecho, había que "colgar" los cuadros. Pero la fotografía original solamente recogía una parte de la sala. La tarea de reconstruir la exposición transformó el trabajo de visualización en un proyecto de investigación. Al analizar el posicionamiento de la cámara fotográfica, se llegó a la conclusión de que la cámara no fue centrada y, además, tenía una inclinación. Este análisis permitió averiguar la superficie de las paredes en que se debía poner los cuadros restantes. Sabíamos qué obras eran debido al trabajo previo de Géal (2001 y 2005: 495-515), quien había utilizado las guías decimonónicas para establecer una lista de obras que se encontraban en la Sala de la Reina Isabel.

Nuestro plan era terminar la identificación de los cuadros en la foto, llegar a una hipótesis sobre los criterios subyacentes en la colocación de los cuadros y aplicar estos criterios para encontrar un sitio para las obras restantes. Para lograrlo, tuvimos que superar dos desafíos: 1) no sabíamos exactamente en qué orden estaban los cuadros y 2) no había sitio para todos los cuadros que, según las guías, estaban en la sala. La solución para el primer problema vino en forma de la guía de España publicada en 1878 (Ford 1878: 57-59), que menciona gran parte de los cuadros expuestos en la Sala. Infiriendo el movimiento de la descripción comparando el texto con la foto, extrapolamos el orden de la mención a otras paredes. La misma guía nos permitió establecer una lista mínima de las obras expuestas.

La atribución y los marcos han cambiado considerablemente desde 1875-1877. Nuestro proyecto reconstruye los

marcos de aquel entonces a base de las placas de cristal hechas por Laurent en la misma época. Mientras las tablillas reproducen la atribución decimonónica, los usuarios pueden activar las anotaciones que reflejan la atribución actual.

La resultante reconstrucción existe en tres versiones, cada una diseñada para un público y usos diferentes. La versión inicial fue ideada como espacio inmersivo interactivo para una "cueva" de proyección en 3D (Fig. 2). Este espacio, de 20 pies de ancho, se utiliza para clases y conferencias que crean una experiencia extremadamente detallada, en algunos aspectos superior a una visita al museo, generada a partir de los programas Blender e Unity. Para abrir la experiencia a un mayor número de usuarios, hemos creado una versión optimizada para teléfonos móviles Samsung Galaxy S6 y gafas de RV. Esto nos hizo buscar soluciones ingeniosas en cuanto a las texturas y la iluminación para reducir los requisitos técnicos sin sacrificar el detalle y el efecto. Dado el éxito de esta versión, decidimos buscar aún mayor accesibilidad, llevando la experiencia interactiva de "realidad aumentada" a cualquier ordenador o dispositivo móvil a través del buscador de la red: en una proyección en 2D para todos y en RV para los que tienen las gafas. Debido al gran volumen de datos necesario para exponer y anotar 104 cuadros y la gran variedad de dispositivos, se decidió rechazar la opción más obvia, Unity WebGL y usar, en su lugar, un nuevo instrumento A-Frame que se utiliza en juegos interactivos. A través de un código QR, los visitantes que acuden ahora al Museo del Prado podrán utilizar sus dispositivos para proyectar la reconstrucción sobre las paredes actuales de la sala y ver los cambios en la arquitectura y el uso del espacio sin tener que descargar ninguna aplicación adicional (Fig. 3).

Así, la reconstrucción permite reflexionar, a cualquier distancia de Madrid y 140 años después, sobre los criterios de "comparación estética" y las ideas museísticas, estudiar los cuadros y entender los fundamentos intelectuales de la exposición. Por ejemplo, nos hace preguntar si los criterios nacionalistas no formaban parte de la confrontación entre las obras incluso en esta sala, a pesar de haber sido diseñada para ofrecer un paréntesis en el recorrido por un museo ordenado por escuelas nacionales. O nos hace comprender la influencia que ejercían los patrones del ornato de los templos (en el ábside) y los retratos en las casas particulares (en las paredes a los dos lados de la entrada). Esto indica que, en un museo como el Prado, la exposición de obras maestras contribuía al pensamiento nacionalista mientras sugería paralelismos con la esfera pública confesional y, a la vez, la esfera privada.

Esta presentación demuestra que una reconstrucción en 3D a base de datos incompletos puede constituir un proyecto de investigación que no sólo permite cotejar diversas fuentes para producir, refinar y compartir hipótesis, sino también se convierte en una exposición visitable *in situ* y remotamente que, a su vez, genera otras hipótesis y abre nuevas líneas de investigación.

Fig. 1. Fotografía original de Jean Laurent, 1875-77

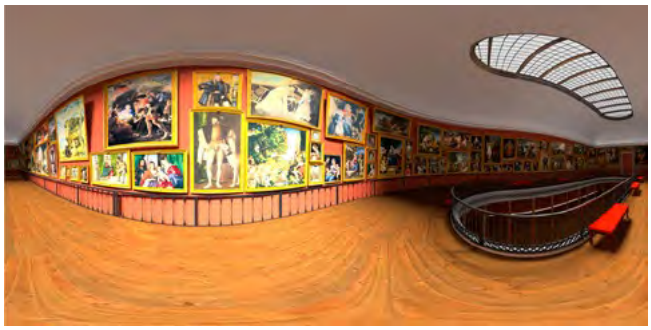


Fig. 2. Rendición de la cueva 3D

Fig. 3. Reconstrucción para cualquier dispositivo

Referencias

Ford, Richard (1878). *A Handbook for Travellers in Spain*. 1845. 5th edition. London, Murray.

Géal, Pierre (2001). "El Salón de la Reina Isabel en el Museo del Prado (1853-1899)." *Boletín del Museo del Prado*, XIX: 37, 143-72.

— (2005). *La Naissance des musées d'art en Espagne (XVIIIe–XIXe siècle)*. Madrid, Casa de Velázquez.

A Digital Edition of Leonhard Euler's Correspondence with Christian Goldbach

Sepideh Alassi
sepideh.alassi@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Tobias Schweizer
t.schweizer@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Martin Mattmüller
martin.mattmueller@unibas.ch
Bernoulli Euler Center, University of Basel, Switzerland

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Helmut Harbrecht
helmut.harbrecht@unibas.ch
Bernoulli Euler Center, University of Basel, Switzerland

Introduction

The edition of the works of Leonhard Euler (1707-1783), entitled *Leonhardi Euleri Opera omnia* (LEOO), is a monument of scholarship known to most historians of science. Leonhard Euler's *Opera omnia* consists of 81 volumes, 76 of which have already been published in paper format as four series of books. Volume IV, LEOO IV, of the fourth series contains the correspondence between Leonhard Euler and the German mathematician Christian Goldbach, encompassing 200 letters sent over 35 years(Martin Mattmüller, 2015). The aim of our project is to present this volume to researchers in science and history as a digital edition via the Bernoulli-Euler Online Platform, BEOL(Tobias Schweizer, 2017). BEOL is implemented using Knora(Benjamin Geer, 2017), a generic virtual research environment for the humanities. In this environment, scientists have access to all edited materials of LEOO IV, and can also annotate and edit material in their private workspace and share the results of their research with others. In Knora, the contents of the LEOO IV volume can be represented as a directed graph providing an overview of the network of different entities (letters, persons, bibliographic items, etc.). The tools provided in this environment are intended to facilitate research on the origin of ideas and findings.

Technical steps

LEOO IV consists of two parts: one with transcriptions of the letters in the original languages (Latin and German), and another with English translations of the letters. LaTeXis used to edit both text and mathematical formulas. The volume also contains an index of persons, a bibliography of cited works by Euler, and a general bibliography. The project aims to import all this content into Knora, which represents data as RDF graphs using OWL ontologies(Pascal Hitzler, 2012). Therefore, ontologies are created to describe the structure of the texts and entities of this edition. The data itself must then be converted to XML and imported into Knora.

Specifying the structure of the data

The data model specifying the structure of the data to be imported must be given in the form of OWL ontologies.[1] All bibliographical items, as well as persons in the name index of the edition, are represented internally as RDF triples. For example, every person is represented as an RDF resource belonging to the OWL class beol:Person, which has properties such as beol:hasFamilyName. The property beol:hasIAFIdentifier refers to the IAF/GND dataset maintained by German national library[2], and ensures the uniqueness of each person mentioned in the BEOL platform.

Figure.1 illustrates a part of the generic bibliography ontology, which we have defined to describe all the bibliographical information needed in the BEOL platform (publication types, manuscripts, publishers, etc.). The prefix biblio refers to this ontology, beol refers to the ontology of BEOL-specific entities, and knora-base is the standard Knora ontology, which defines the basic data structures that Knora works with. Ellipses represent types or classes of resources, arrows semantically defined properties attached to them, and rectangles their literal values.

In Knora, a text document (stored in a knora-base:TextValue) can contain markup as well as text. Internally, markup is stored separately from the text, using an RDF-based standoff format[3]. A project such as BEOL defines a mapping between XML and Knora's standoff/RDF markup; texts can then be imported from XML into standoff and exported from standoff back into identical XML[4]. Standoff/RDF markup can contain links to other resources, such as a person or a bibliographical entity mentioned in a text. The Knora API server ensures that the target of the link exists. Standoff links are directed statements, but can easily be queried as incoming links to a given resource.

1 A user interface for designing these ontologies is under development.
2 Integrated Authority File, Deutsche National Bibliothek, http://www.dnb.de/EN/Standardisierung/GND/gnd_node.html
3 Text with Standoff Markup, http://www.knora.org/documentation/manual/rst/knora-ontologies/knora-base.html#text-with-standoff-markup
4 Creating a Custom Mapping, http://www.knora.org/documentation/manual/rst/knora-api-server/api_v1/create-a-mapping.html#creating-a-custom-mapping

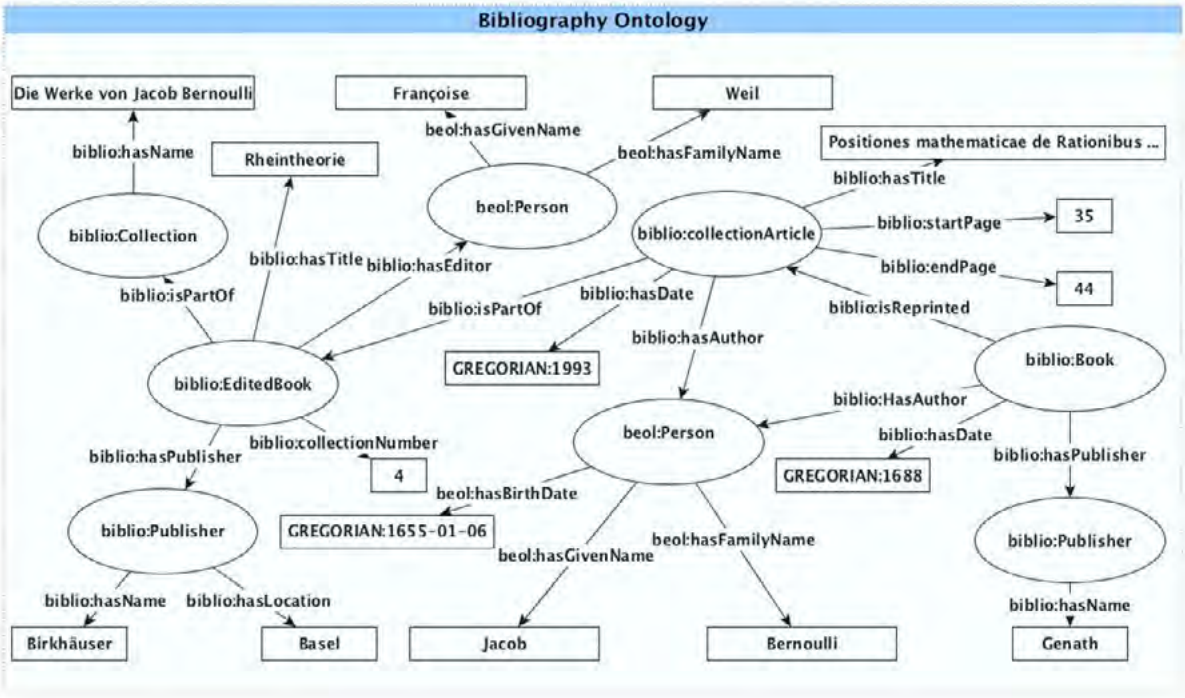


Figure 1. OWL ontology for bibliographical data

We have also defined a data model for letters and their metadata such as author, recipient, date, etc., which provides a network of the correspondence included in the edition. Figure.2 illustrates an excerpt from ontology of the whole LEOO IV project.

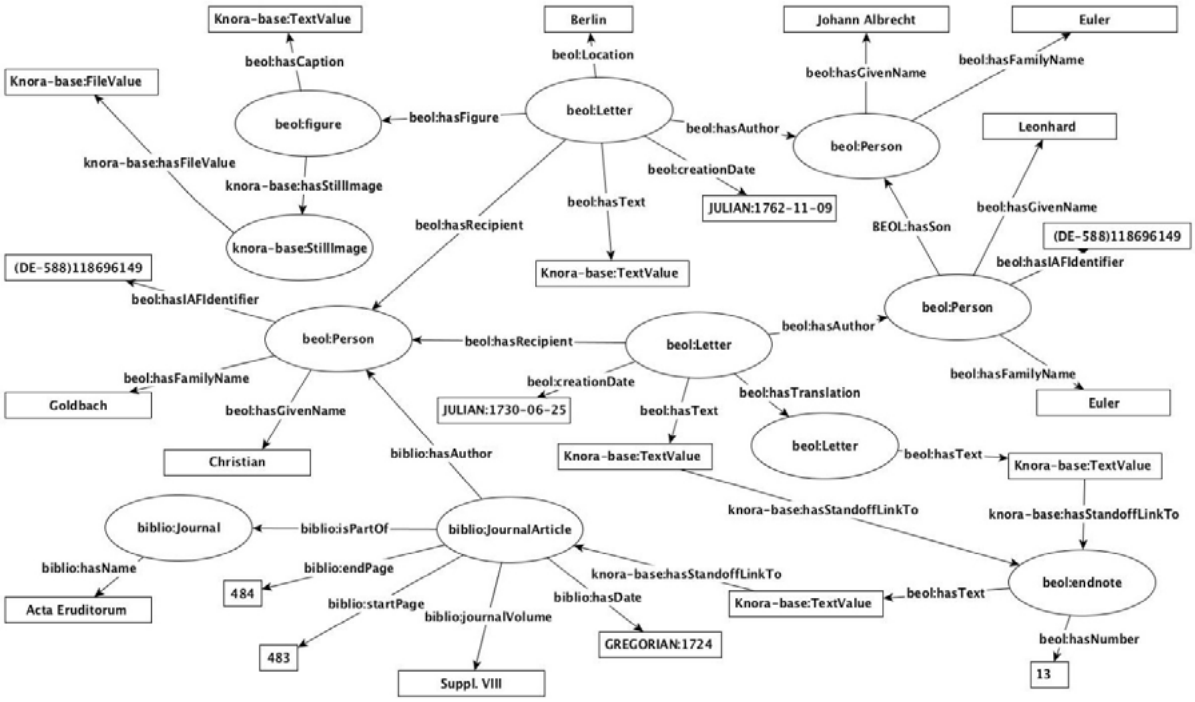


Figure 2. Excerpt from the LEOO IV project data model

Importing data into the BEOL platform

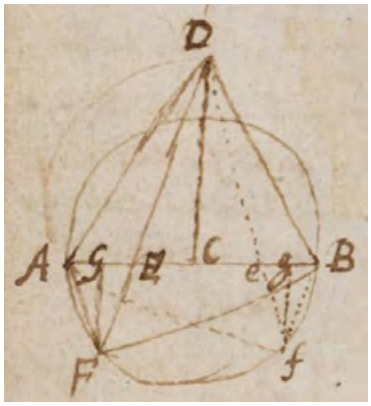
First, the index of persons and the bibliographical items of LEOO IV are written in XML format, using XML schemas that are automatically generated by the Knora API server, based on the ontologies defined for the project. This XML data is then validated against these schemas. After validation, the data can be imported in a single API request (an HTTP POST request to the Knora API server).

Second, the text of the letters is imported using a similar process. Although the text has been transcribed in LaTeX, these transcriptions are first converted to XML to ensure the homogeneity of texts from different editions, and to make it possible to present texts as TEI/XML by applying XSL transformations. The LaTeXML tool(Miller, 2017), with the addition of some BEOL-specific Perl scripts, is used to convert LaTeXto XML. All references to persons and bibliographical items within the text of the letters are replaced with references to the corresponding resources in BEOL, making them queryable via the Knora API. The XML representing the letters is then imported using the same process as for the bibliographical data.

Future work

Since we have developed the methodology for this type of digital edition in a generic way, we expect to be able to integrate all the other recent volumes of Leonhard Euler's *Opera omnia*, which have also been edited using LaTeX. The older volumes in printed form should be scanned, their text should be recognized via OCR, and their structure should be defined with markup.

Most of the older volumes contain figures that are reproduced from scanned letters. We are working on a machine learning algorithm to interpret these figures as well as their labels, so they can be automatically redrawn as vector graphics, see Figure. 3.



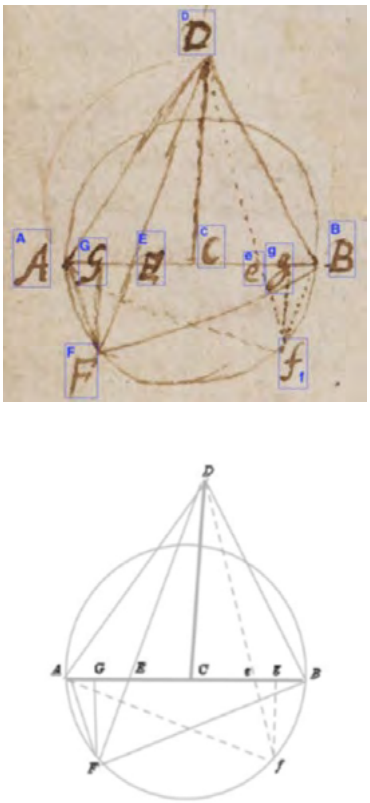


Figure 3. Original figure, detected labels, and reconstructed figure

References

Benjamin Geer, et al (2016). *Knowledge, Organization, Representation, and Annotation*. Digital Humanities Lab http://www.knora.org/.

Martin Mattmüller, F. L. (ed). (2015). *Leonhardi Euleri Opera Omnia: Correspondence of Leonhard Euler with Christian Goldbach*. Vol. IVA/4. Basel.

Miller, B. R. (2017). *LaTeXML: A Latex to Xml/Html/Mathml Converter*. http://dlmf.nist.gov/LaTeXML/.

Pascal Hitzler, et al (2012). *OWL 2 Web Ontology Language Primer (Second Edition)*. W3C http://www.w3.org/TR/2012/REC-owl2-primer-20121211/.

Tobias Schweizer, et al (2017). Integrating historical scientific texts into the Bernoulli-Euler online platform. *Digital Humanities 2017*. https://dh2017.adho.org/abstracts/147/147.pdf.

Bridging the Divide: Supporting Minority and Historic Scripts in Fonts: Problems and Recommendations

Deborah Anderson
dwanders@sonic.net
UC Berkeley, United States of America

Introduction

Today, users of many modern minority and historic scripts in Unicode are not able to reliably send text electronically, because Unicode-enabled fonts and software are not available.[1] In addition, some communities have access to Unicode fonts, but the fonts aren't used, because they do not provide features deemed necessary, such as positioning of characters (e.g., Egyptian Hieroglyphs [Richmond and Glass, 2016]) or variant glyphs (e.g., Old Italic [Anderson, 2017]). Instead, images are used, which are not searchable or, alternatively, "hacked" fonts are employed, which require each person to have the same, non-standard font to send text. Keyboards or other input mechanisms are also not available for many of these same scripts. As a result, the promise that Unicode will "enable people around the world to use computers in any language" (Unicode Consortium, 2018a), does not yet ring true for some communities.

This short paper will highlight font-related problems with specific examples and will provide suggestions on how to address them.

Problems

- Creating a Unicode-enabled font for a language is often not a simple task, especially when the script for the language includes combining marks (which require correct positioning), or if the script has special rendering behavior, such as the consonant clusters found in South Asian scripts (Evans, 2017).
- Font creation is made more challenging when typographic details on the script (and language) are not available. Since many recently approved scripts in Unicode are not well known, information on the typography is not readily available. Unfortunately, fine details are often not included in Unicode proposals for the scripts.
- Interaction with the user community is critical in developing a suitable font, but some communities are difficult to contact. In addition, there can be differing views on the preferred shapes of glyphs. For a set of 51 Tamil numbers and fractions, for example, the community took 8 years to come to agreement on the preferred representative shapes. Specific cases will be cited, based on the author's experience, including discussion of how to connect user communities with font providers.

Technical Issue: Glyph Variants

- For some script users, access to glyph variants is important. This is true, for example, for the Old Italic Unicode block which unified several related alphabets of Italy, dating from approximately the 8 until 1c BCE. In Old Italic, the glyph in a particular alphabet may vary from that shown in the Unicode Standard. The Old Italic block was encoded with the understanding that different fonts would be used for the different languages and alphabets (Unicode Consortium, 2017). How should the two forms of Faliscan (above) be handled in the same font then? How should a pan-Old Italic font handle the different alphabets (which use the same code points)?

This paper will describe the pros and cons of different options available, including use of:

- Code points in Unicode's Private Use Area (with the caveat that these code points would not be reliable for general interchange) (Unicode Consortium, 2018c).
- A Unicode variation sequence, when a distinction needs to be captured in plain-text (Unicode Consortium, 2018d).
- An OpenType font feature, such as character variants, stylistic alternates, stylistic sets, or localized forms (Microsoft Typography, 2018).
- Language-specific fonts (i.e., Faliscan1 and Faliscan2 fonts for the two forms above).

Suggested Solutions

- Incorporate font creation as a part of the overall script encoding effort, such as: including a font item in the budget to pay for a font designer to develop a font; provide information on how to create a font for users; fund a font-creation workshop within the community.
- Encourage user communities to submit a list of the basic repertoire of characters and auxiliary characters to the Common Locale Data Repository (Unicode Consortium, 2018b), since this information is used for by font and software developers worldwide. In addition, provide information on the shapes of the needed letters and variants, citing reference works (i.e., a book or website) on a publicly accessible webpage.
- For handling glyph variants, short-term and long-terms approaches should be considered:
  - If a given variant is deemed by users to be necessary in plain-text, submit a Unicode proposal
  - If OpenType features are used in a font, lobby software vendors to provide better support for the features in applications (as support for some features is still spotty [4])
  - For the short-term, PUA or separate fonts may be necessary.

For font designers:

- Use language tags from ISO 639 (SIL International, 2017), BCP 47 (Phillips and Davis, 2009), and

1 Especially true for scripts in Unicode versions 6.0 to 9.0 (2010 – 2016), where over 40% of the scripts have no fonts. (Unicode version 10.0 was released in June 2017, so support in fonts would not yet be expected). The Google Noto project aims to provide fonts for all approved scripts, but release of fonts is only up to fonts for Unicode version 6.2, released in 2012.

OpenType language/script tags (Microsoft Typography, 2017a; Microsoft Typography, 2017b) in the font internals. If a language (or script) is missing a tag, a new tag should be registered. According to Roozbeh Pournader, an expert at implementation of fonts, these tags are the way the fonts communicate with other software today.

- Encourage users to review the glyphs in alpha versions of any forthcoming or any released Noto fonts, and submit comments to the Noto project (Google.com, n.d.).

Conclusion

Access to a Unicode font is critical for users of lesser-used scripts, in order to participate more fully in the digital world. Unicode fonts make the user's text interchangeable, discoverable, and able to be preserved for the long-term in a stable format. Recognition of font-related issues is a small step towards addressing the problem. Input from the audience will be encouraged in order to identify other potential approaches.

Funding

This work was supported by the National Endowment for the Humanities [grant number PR-253360-17].

References

Anderson, D. (2017). Dealing with Variants in Historic Scripts. Presentation at *41st Internationalization and Unicode Conference*, Santa Clara, California, October, 2017.

Evans, L. (2017). Beyond Unicode Proposals: Encoding Characters and Scripts is Not Enough! Presentation at *41st Internationalization and Unicode Conference*, Santa Clara, California, October 2017.

Google.com. (n.d.). *Google Noto Fonts*. https://www.google.com/get/noto/ (accessed April 17, 2018).

Microsoft Typography. (2017a). *Language system tags*. https://www.microsoft.com/typography/otspec/languagetags.htm (accessed April 17, 2018).

Microsoft Typography. (2017b). *Script tags*. https://www.microsoft.com/typography/otspec/scripttags.htm (accessed April 17, 2018).

Microsoft Typography. (2018). *OpenType® specification*. https://www.microsoft.com/en-us/Typography/OpenTypeSpecification.aspx (accessed April 17, 2018).

Phillips, A., and Davis, M. (2009). *Tags for Identifying Languages*. https://tools.ietf.org/html/bcp47 (accessed April 17, 2018).

Richmond, B. and Glass, A. (2016). *Proposal to encode three control characters for Egyptian Hieroglyphs. Proposal submitted to the Unicode Technical Committee*. http://www.unicode.org/L2/L2016/16018r-three-for-egyptian.pdf (accessed April 17, 2018).

SIL International. (2017). *ISO 639-3: ISO 639 Code Tables*. http://www-01.sil.org/iso639-3/codes.asp (accessed April 17, 2018).

Unicode Consortium. (2017). Old Italic. In: Unicode Consortium, *The Unicode Standard, Version 10.0.0*. Mountain View, CA: The Unicode Consortium, 349-351. http://www.unicode.org/versions/Unicode10.0.0/ (accessed 24 April 2018).

Unicode Consortium. (2018a). *The Unicode Consortium website*. http://unicode.org/ [accessed April 17, 2018).

Unicode Consortium. (2018b). *CLDR - Unicode Common Locale Data Repository. Unicode Consortium website*. http://cldr.unicode.org (accessed April 17, 2018).

Unicode Consortium. (2018c). *Private-Use Characters, Noncharacters & Sentinels FAQ. Unicode Consortium website*. http://www.unicode.org/faq/private_use.html (accessed April 17, 2018).

Unicode Consortium. (2018d). *Variation Sequences. Unicode Consortium website*. http://www.unicode.org/faq/vs.html (accessed April 17, 2018).

Unwrapping Codework: Towards an Ethnography of Coding in the Humanities

Smiljana Antonijevic Ubois
smiljana@smiljana.org
The Pennsylvania State University, United States of America

Joris van Zundert
joris.van.zundert@huygens.knaw.nl
Royal Netherlands Academy of Arts and Sciences,
The Netherlands

Tara Andrews
tara.andrews@univie.ac.at
University of Vienna, Austria

Code and codework share many properties with text and writing, and code can be seen as an argument, corresponding to Galey and Ruecker's (2010) understanding of the epistemological status of graphical user interfaces as argument. From an epistemic point of view, the practice of a programmer is no different from the practice of a scholar when it comes to writing (Van Zundert, 2016). Both are creating theories about existing epistemic objects (e.g. text and material artifacts, or data) by developing new epistemic objects (e.g. journal articles and critical editions, or code) to formulate and support these theories. However, as expressions of a technē whose inner workings are opaque to most humanities scholars, code and codework are all too often treated as an invisible hand, influencing humanities research in ways that are not transparent. The software used in research is treated as a black box in the

sense of information science—expected to produce a certain output given a certain input—but at the same time often mistrusted precisely for this lack of transparency.

The digital humanities (DH) does not generally engage with the code and coding parts of programming in an explicit and critical manner, which is necessary for opening up black boxes of code. The invisibility and un-critiqued use of code means that the scholarly quality and contribution of codework goes both uncredited and unaccounted for. Black-boxing the code results in neglect of its epistemological contributions and imperils one of the key components of knowledge production in the DH. Much more insight into code and codework in the humanities is needed, including how coders approach their tasks, what decisions go into its production, and how code interacts with its environment.

The purpose of this paper is to provide some of those insights in the form of an ethnography of codework, wherein we observe the decisions that programmers make and how they understand their own activities. Our study follows in the footsteps of ethnographies of technoscientific practice (see: Forsythe, 2001; Coleman, 2013), Critical Code Studies (see: Marino, 2014), and reflections on coding and tool development in the DH (see: Schreibman and Hanlon, 2010; Ramsey and Rockwell, 2012). The study does not aspire to be representative of the DH coding practice, but to initiate a debate about some still overlooked elements of that practice.

This exploration applies Latour's (1998) first rule of method to the context of narrative creation through codework, looking at the practices, dilemmas, and decisions of programmers. To do that, we use analytical autoethnography (cf. Anderson, 2006), combined with collaborative ethnography (cf. Lassiter, 2005). In our methodological design, the team ethnographer first formulated a set of ten questions aimed at generating reflexive accounts and examples of DH coding in the making. Each of the team DH programmers then individually answered the questions in a written form, providing elaborate, semi-formal accounts of his or her DH programming practice. Thus generated written accounts became the basis for a series of team discussion, both written and oral, which eventually formed the results of this contribution. This methodological design enabled us to return from the final outputs of DH coding to scholarly uncertainties and resolutions that preceded them. Through such reconstruction, we were able to document some of the key phases in epistemological construction of coding artifacts, and to identify methodologically significant moments in stabilization of those artifacts. In other words, we relied on the experiences of scholars proficient in both humanities research and coding seeking to make explicit what DH coders themselves know, maybe tacitly, about why and how they code.

We have grouped our observations into the categories known as the five canons of rhetoric, proposed in Cicero's *De Inventione*. Although originally developed for public speaking, these canons have proven to be equally potent heuristic in analyzing written and, more recently, digital discourse (Gurak and Antonijevic, 2009). Our contribution sought to extend this heuristic to the analysis of coding as argumentation, not in an attempt to fit codework and its elements into a pre-defined ontology, nor to suggest that it fully conforms or matches classical rhetoric. Rather, it was a way of presenting our experiences and claims in a form that we expected to facilitate interpretation by scholars well versed in text production but likely less so in codework.

Our study showed that codework reflects humanistic discovery (*inventio*) in that humanities- specific research questions drive coding, and tasks specific to the humanities research motivate software development. Similarly, crafting and organizing code resonates with development and arrangement of a scholarly argument (*dispositio*)—a programmer writes lines of code and makes many decisions on how to arrange these pieces into larger, meaningful constructs that influence the epistemological and methodological structure of research. Our study also illustrated that, like any humanities scholar, an author of software has her own style (*elocutio*) in the aesthetics of code and in her way of working to create code, and this style develops through both individual and norms of coding communities. We also showed that, parallel to books or libraries, code and codework serve as memory systems (*memoria*) that embed theoretical concepts in order to augment research methodology and create new theory, where code can be regarded as a performative application or explanation of theory. Finally, our ethnography illustrated how codework *actio* compares to the publication and reception of the software, where DH programming is still not recognized as a locus of humanities expertise, and it is hard for humanities programmers to have their code academically evaluated as digital output.

The insights of our study demonstrate that both code as an epistemic object and coding as an epistemic practice increasingly shape research in the humanities and must be given a proper theoretical and methodological recognition in the DH, with both the consequences and the rewards that such a recognition bears. Therefore, a strategy for making code and codework visible, understandable, trustworthy and reputable within humanities scholarship is needed. Such a strategy should be comprehensive, both in the sense of accounting for the source code and the executed result of software. While we agree with Ramsay and Rockwell (2012) that providing source code is not sufficient for understanding the underlying theoretical assumptions, we disagree in viewing the "dependence on discourse" as a feature that relativises epistemic and communicative capacities of code and codework. We argue in contrast that interdependence of code and text should be embraced as a means of acknowledging their distinctive yet corresponding methods of knowledge production and communication. Just as code enhances text making it amenable to methodological and epistemological approaches of DH, text enhances

code making it more visible and intelligible for the humanities community. Evaluating code and DH programming in a disengaged way would thus be similar to the literary criticism enacted on a novel without reading it. Yet currently it is practice to "criticize" software and code based only on a journal article that derived from it. As much as possible, coders should support the involved evaluation of code as opposed to its disengaged criticism. We believe that theoretical discussions of codework should become an established trajectory in the humanities, along with the development of methods for documenting, analyzing, and evaluating code and codework.

One important element of that strategy is understanding codework as necessarily shaped by its social context, which influences the attitude and perception that both coders and other scholars hold towards their work. Too often, DH programmers are treated as service instead of research focused scholars, which results in a number of negative consequences. A necessary step in the direction of a real change in how codework is received into the humanities is recognition and reward for peer-reviewed digital outputs, including code, as research outputs (cf. Nowviskie, 2011; Presner, 2012; American Historical Association, 2015). A precondition for this is to start grassroots procedures for peer review of code (Fitzpatrick, 2011), and to regard the code as alternative expressions of research or epistemologies with equal research value and validity instead of subordinating code and codework to 'humanities proper' (cf. Burgess and Hamming, 2011 and Ramsay and Rockwell, 2012). There is a need for peer review and critical examination of actual code, which is hardly even present in DH (Zundert and Haentjens Dekker, 2017). Also, open publishing of code in verifiable ways can be easily facilitated through existing public code repositories or institutionally-run versions of the same repositories, but it is not common practice throughout the humanities to publish code. Finally, reflexive accounts on (digital) humanities codework and ethnographic studies of actual work can help us understand how code and codework are changing the humanities (Borgman, 2009). We believe that an important step in illuminating the process and results of DH programmers' codework is to develop and explicate reflexive insights into its key epistemological, methodological, and technical aspects. Explaining, for instance, what kind of research questions give impetus to one's codework and how new research insights co-evolve during code development can help both DH programmers and their traditionally trained colleagues recognize the important epistemological connections between humanistic theory and scholarly programming.

References

American Historical Association, A. H. C. on P. E. of D. S. by H. (2015). *Guidelines for the Professional Evaluation of Digital Scholarship in History*. Draft http://bit.ly/1PC1tDL (accessed 8 November 2017).

Anderson, L. (2006). Analytic Autoethnography. *Journal of Contemporary Ethnography*, 35(4): 373–95 doi:10.1177/0891241605280449.

Borgman, C. (2009). The Digital Future is Now: A Call to Action for the Humanities. *DHQ: Digital Humanities Quarterly*, 3(4) www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

Burgess, H. J. and Hamming, J. (2011). New Media in Academy: Labor and the Production of Knowledge in Scholarly Multimedia. *DHQ: Digital Humanities Quarterly*, 5(3) http://digitalhumanities.org/dhq/vol/5/3/000102/000102.html (accessed 2 September 2016).

Coleman, E. G. (2013). *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton (US), Woodstock (UK): Princeton University Press http://gabriella-coleman.org/Coleman-Coding-Freedom.pdf (accessed 8 November 2017).

Fitzpatrick, K. (2011). Peer Review, Judgment, and Reading. *Profession*(6): 196–201 doi:-prof.2011.2011.1.196.

Forsythe, D. and Hess, D. J. (2001). *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford, CA: Stanford University Press.

Galey, A. and Ruecker, S. (2010). How a prototype argues. *Literary and Linguistic Computing*, 25(4): 405–424 doi:10.1093/llc/fqq021.

Gurak, L. and Antonijevic, S. (2009). Digital Rhetoric and Public Discourse. In Lunsford, A. A., Eberly, R. A. and Wilson, K. H. (eds), *The Sage Handbook of Rhetorical Studies*. London, Thousand Oaks: SAGE Publications, Inc., pp. 497–508.

Lassiter, L. E. (2005). *The Chicago Guide to Collaborative Ethnography*. (Chicago Guides to Writing, Edi). Chicago, London: University of Chicago Press http://bit.ly/2iLCmGY.

Latour, B. (1988). *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA, USA: Harvard University Press.

Marino, M. C. (2014). Field Report for Critical Code Studies, 2014. *Computational Culture—A Journal of Software Studies*(4) http://computationalculture.net/article/field-report-for-critical-code-studies-2014%E2%80%A8 (accessed 10 June 2015).

Nowviskie, B. (2011). Where Credit Is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship. *Profession*(6): 169–181 doi:-prof.2011.2011.1.169.

Presner, T. (2012). How to Evaluate Digital Scholarship. *Journal of Digital Humanities*, 1(4) http://journalofdigitalhumanities.org/1-4/how-to-evaluate-digital-scholarship-by-todd-presner/.

Ramsay, S. and Rockwell, G. (2012). Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 75–84 http://dhdebates.gc.cuny.edu/debates/text/11.

Schreibman, S. and Hanlon, A. M. (2010). Determining Value for Digital Humanities Tools: Report on a Sur-

vey of Tool Developers. *DHQ: Digital Humanities Quarterly*, 4(2) http://digitalhumanities.org/dhq/vol/4/2/000083/000083.html (accessed 9 November 2017).
Zundert, J. J. van (2016). Author, Editor, Engineer — Code & the Rewriting of Authorship in Scholarly Editing. *Interdisciplinary Science Reviews*, 40(4): 349–375 doi:http://dx.doi.org/10.1080/03080188.2016.1165453.
Zundert, J. J. van and Haentjens Dekker, R. (2017). Code, Scholarship, and Criticism: When is Coding Scholarship and When is it Not? *Digital Scholarship in the Humanities*, 32(Suppl_1): i121–i123 doi:https://doi.org/10.1093/llc/fqx006.

Conexiones Digitales Afrolatinoamericanas. El Análisis Digital de la Colección Manuel Zapata Olivella

Eduard Arriaga
earriaga@alumni.uwo.ca
University of Indianapolis, United States of America

Las manifestaciones afrolatinoamericanas y sus conexiones con el mundo digital han comenzado a generar un creciente interés en diversos campos de estudio: las humanidades digitales, los estudios culturales, literarios y antropológicos entre otros. A pesar del interés, el estudio de tal intersección se encuentra en una etapa inicial debido a factores como a) las limitaciones de acceso a herramientas digitales por parte de algunos agentes y comunidades identificadas y auto-identificadas como afrolatinoamericanas/afrolatinas; b) limitaciones en la consecución de derechos de autor de algunas piezas y manifestaciones cuya distribución e intercambio digital se hace más difícil; y c) falta de innovación en la forma de clasificar piezas y manifestaciones que, en muchos casos, no coinciden con la tradición letrada que subyace al proceso de archivo ya sea digital o no. Tales limitaciones han hecho más difícil la consolidación de propuestas analíticas que, desde las humanidades digitales, den cuenta del estado y evolución de las culturas afrolatinoamericanas, así como de sus aportes a nivel de conocimiento en espacios locales, regionales y globales.

Algunas formas de revertir dichas limitaciones ha sido el desarrollo de iniciativas y colecciones digitales por parte las mismas comunidades afrolatinoamericanas en cooperación con entidades académicas, agencias multilaterales, gubernamentales, intergubernamentales y no gubernamentales. Tales iniciativas muestran la diversidad de manifestaciones generadas desde dichas comunidades; manifestaciones que son fundamentales para su identificación, visibilización y, sobre todo, consideración dentro de un modelo de justicia social que, como el contemporáneo, se centra en el reconocimiento de los derechos humanos. Asimismo, dichas adaptaciones tecnológicas se convierten en una forma de lo que Steve E. Jones determina como 'eversion" (Jones, 2016) o la consolidación de unas realidades híbridas entre lo digital, lo análogo y lo performático. Algunos de los proyectos más importantes en este ámbito son, entre otros, Digital Portobelo, Mueseu Afro Digital Río de Janeiro o Proyecto Afrolatin@, a partir de los cuales se hacen evidentes diversas formas de ser afrolatinoamericano, así como diversas formas de representación y expresión de sujetos cuya identificación intersecta varios espacios discursivos, políticos y de acción. Algunos de los puntos positivos de dichas plataformas y colecciones es que a) son espacios en constante construcción –actuales y constantemente actualizados- y b) permiten ver procesos de acceso, creatividad, justicia simbólico-social que las comunidades están persiguiendo y han perseguido por largo tiempo. Sin embargo, el carácter de construcción constante de dichas plataformas es, al mismo tiempo, un aspecto negativo dado que el flujo de información se convierte en un desafío para unas humanidades digitales cuyo modelo se ha centrado en la digitalización y análisis de información canónica, única, extraordinaria (Manovich, 2016). Las plataformas generadas por parte de esas comunidades afrolatinoamericanas, por el contrario, registran el flujo de la cultura en el presente que no ha sido propiamente abordado por las humanidades ya sean análogas o digitales. En el caso de la intersección entre estudios afrolatinoamericanos y estudios digitales, el proceso de análisis ha estado mucho más rezagado no solo por la falta de bases de datos o de construcción de archivos digitales, sino por la falta de interés y apoyo para construirlos y, a partir de allí, desarrollar metodologías innovadoras de análisis (Gomez, 2011).

De acuerdo con el panorama descrito, esta presentación corta dará cuenta del proceso de investigación e implementación metodológica llevado a cabo a partir de *Manuel Zapata Olivella Collections*, una colección digital desarrollada por la biblioteca de la Universidad de Vanderbilt. Manuel Zapata Olivella fue uno de los escritores y activistas afrolatinoamericanos más importantes del siglo XX, cuya obra y pensamiento han influido al movimiento afrolatinoamericano contemporáneo. Sus cartas, manuscritos y documentación personal como escritor, artista y activista habían quedado en un archivo personal manejado por su familia. Sólo hasta el 2008 la Universidad de Vanderbilt adquirió el fondo y desarrolló una colección digital en el cual se hacen visibles varios de sus documentos y proyectos tanto etnológicos como antropológicos. Entre los archivos digitalizados se encuentran los documentos –cartas, panfletos, memorias, comunicaciones personales, fotografías y audios- del *Primer Congreso de Cultura Negra de las Américas*, realizado en Colombia en 1978. El proyecto, llevado a cabo con apoyo de la Universidad de Indianápolis, consistió en el análisis digital de dicha documentación y del Congreso como uno de los nodos centrales de la acción política, literaria y cultural afrolatinoamericanas del siglo XX y XXI. El proyecto buscaba a) responder preguntas tales

como: ¿Cuáles fueron las redes artísticas y textuales que permitieron la emergencia del Congreso?, ¿Cuáles fueron los discursos socio-culturales latinoamericanos con los cuales el congreso desarrolló un diálogo y logró establecer su propio conjunto de valores y códigos para explicar lo afrolatinoamericano?, ¿Cuáles de los valores políticos y estrategias estéticas creadas y adoptadas por el Congreso devinieron patrones de acción y fueron transmitidas al movimiento afrloatinoamericano de la era digital?. Asimismo, el proyecto buscaba b) desarrollar propuestas metodológicas digitales para comenzar a entender la complejidad e interconexión –en tiempo y espacio- del movimiento afrolatinoamericano. Esta última actividad se desarrolló a través de la implementación de mapas de tópicos y el uso de plataformas digitales para visualizar la información de forma inter-relacional –Vg. Scalar, Wandora, Gephi, etc.-, considerando la diversidad de materiales en el ecosistema informativo de la tradición afrolatinoamericana.

La presentación entonces mostrará los resultados de esa investigación a través del mapeo de textos, de agentes, instituciones y sistemas de valores relacionados para, finalmente, conectarlo con las propuestas ideológicas fundamentales del movimiento afrolatinoamericano surgido de la Conferencia Mundial Contra el Racismo realizada en Durbán en 2001. A través de esta presentación se discutirán no solamente los hallazgos de la investigación en particular sino, sobre todo, las perspectiva de unas humanidades digitales afrolatinoamericanas que, aunque se incluyan en las discusiones regionales (RedHD, Humanidades digitales en Latinoamérica) intentan ir más allá, en busca de la conexión entre activismo e investigación académica con un objetivo claro: la justicia social y la descolonización del conocimiento.

References

Gómez F. P. (2011). La colección Manuel Zapata Olivella. *Revista de estudios colombianos*, 37-38: 117-118.

Jones E. S. (2016). The Emergence of the Digital Humanities. *Debates in the Digital Humanities,* University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/52

Manovich, L. (2016). The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. *Journal of Cultural Analytics*. Doi: *10.22148/16.004*

Dal Digital Cultural Heritage alla Digital Culture. Evoluzioni nelle Digital Humanities

Nicola Barbuti
nicola.barbuti@uniba.it
Dipartimento di Studi Umanistici DISUM - Università degli Studi di Bari Aldo Moro, Italy

Ludovica Marinucci
lud.marinucci@gmail.com
Scuola a Rete per la Formazione nel Digital Cultural Heritage, Arts and Humanities - DiCultHer, Italy

Introduction

Digital has transformed the way to produce, transmit and share knowledge. The increasingly widespread diffusion of digital methods and techniques in all the social and cultural levels of the communities, in fact, brings an unheard democratization of knowledge and culture, making the citizen a privileged and intelligent actor in the sustainable development of the new smart societies which are based on the process of digitization, digital co-creation and digital design.

The **art. 2** of the UE "**Council conclusions of 21 May 2014 on cultural heritage as a strategic resource for a sustainable Europe" (2014/C 183/08)** states: "Cultural heritage consists of the resources inherited from the past in all forms and aspects - tangible, intangible and digital (born digital and digitized), including monuments, sites, landscapes, skills, practices, knowledge and expressions of human creativity, as well as collections conserved and managed by public and private bodies such as museums, libraries and archives. It originates from the interaction between people and places through time and it is constantly evolving. These resources are of great value to society from a cultural, environmental, social and economic point of view and thus their sustainable management constitutes a strategic choice for the 21st century".

It is therefore inevitable to rethink digital and digitization as social and cultural expressions of the contemporary age. This implies the need to rethink data as cultural entities and no longer as mere tools for simplifying administration management, or as extemporary surrogates for enhancing the fruition of tangible and intangible cultural heritage.

The current process for archiving and storing data, although they generate from the awareness of the need to preserve them, don't solve the problem of their both current and historical reuse, because they are still strongly conditioned by the instrumental function that presides over their production and use.

Towards a first classification of Digital Culture

This paper aims to provide a new definition of methodological and technological approach to digital and digitization, with the goal to guarantee data stability, sustainability, usability and reusability so as to foster their long term preservation.

The research originates from observing that, in the human evolution, the survival, preservation and permanence over time of any entity has always been strictly linked to its identification as cultural heritage, because of its value of historical witness which conveys knowledge.

For several years, authoritative scientific voices have highlighted how long term digital preservation is the real emergency to be faced worldwide. In 2015, Vinton Cerf raised the alarm about the risk that the Twenty-First Century will become for posterity the first black hole in human evolution since the establishment of intelligent communication. The alarm resumed what was debated in the 2012 UNESCO Conference held in Vancouver with the significant title "The Memory of the World in the Digital Age: Digitization and Preservation".

In order to start a serious and shared process for cultural identification of digital and digitization, it is therefore essential to recognize data as cultural entities, defining a clear and regulated position in the contemporary cultural scene. In fact, several existing digital entities could be considered contemporary **Digital Cultural Heritage (DCH)**, expression of the **Digital Culture** of the Twenty-First Century smart societies.

A first useful identification could come out from a classification of **digital cultural entities**, which can be traced back to the following three basic categories in which the Digital Culture could be declined:

- **Digital FOR Cultural Heritage**: process, methods and techniques aimed at co-creation of digital artifacts reproducing in their contents tangible and intangible cultural heritage: e.g., digital objects, digital libraries, virtual museums, demo-etno-anthropological databases.
- **Digital AS Cultural Heritage**: approach, process, methods and techniques aimed at recognizing and preserving both digital artifacts reproducing intangible and tangible cultural heritage, and dematerialization as expression of contemporary cultural *facies* to be known, safeguarded, preserved and transferred in time as witness and memory of the current **Digital Age**.
- **Born Digital Heritage**: process, methods and techniques aimed at co-creating and managing digital entities that record the current activities of contemporary communities, to be safeguarded, preserved and transferred to future generations as witness and memory of Twenty-First Century culture and societies.

Digital Culture as identity of contemporary age

According to the above classification, Digital Culture could therefore be defined as implementation of integrated cultural and training approaches, processes, methods, and techniques aimed at co-creating an ecosystem of aware digital knowledge. This, in fact, will be enabled to trigger processes for the construction of networks to safeguard, preserve, sustain, transfer, reuse **DCH** through awareness of its identity as historical memory of the contemporary age and, therefore, as source of knowledge for future generations.

So, starting from the analysis and co-design of a digital entity, whatever it is – one digital artifact, a digital library, a management system for Public Administrations or an app for Augmented Reality –, the focus on preservation is primary to define it a digital cultural entity. It will determine and regulate both the co-creation process, and the methodological and technological approaches, systems, information, metadata schema, digital image content structures, data description, complex data set, and their any further development and sustainability. This approach can only exist in an ecosystem of aware digital culture, in which digital and digitization with their processes are recognized as DCH.

In this regard, our opinion is that what differentiates DCH from the non-cultural digital artifacts are the descriptive metadata for indexing digital object. Above all, it is primary the correct proportion between:

- **quantity**: it is the correct ratio among exhaustiveness of information, knowledge to be provided, number of metadata elements and attributes necessary to retrieve, reuse and store it;
- **quality**: it is the correct ratio among choice of the informative and cognitive level to be given both to each descriptor and to set of descriptors, and the variables of information and cognitive need of the users, according to whether they are current or future.

Descriptive metadata as sources of Digital Cultural Heritage

The issue is addressed with regard to the preservation of **Digital AS Cultural Heritage**. The case study object of the research is the metadata schema co-created for the digitization project "Historical Archive of the G. Laterza & Figli Publishing House", undertaken at the end of 2015 and today publishing in the Puglia Digital Library of the Puglia Region.

The metadata schema used for managing and indexing the digital artifacts scanned from the original documents has been co-created with reference to the Italian national METS-SAN standard structured by the National Archival System.

The preservation of both the process of digital co-creation and of the digital resources themselves has been the focus of the project. So, attention has been focused on descriptive metadata both of the project as a whole, and of each section of the original Archive (series, sub-series, etc.), and of each one digital artifact. The starting point was the awareness that, at the state of the art, the images present great difficulty for long term digital preservation. The planning and structuring of the metadata schema has therefore been focused not only on the needs of contemporary users, but above all on the cognitive and informative needs of future users about our

contemporary culture. So, metadata will be the only sources of knowledge on both the digital artifacts we produce today, and the processes by which we co-create them.

We preferred to use "granular" indexing, describing each digital document with its metadata.

In structuring the metadata schema, we considered the tag sequence as an organic structure composed of forms entities (elements and attributes) and descriptive information. The narrative contents have been articulated hybridizing methods and techniques of archival description with cataloguing solutions, and they have been written with stylistic criteria deduced from the storytelling methodology, providing information on both the whole project and the detail of each section and, inside the sections, of each partition.

In each metadata, the <header> section, after the namespaces (<xlmns: --->) enbeds the descriptors related to:

- project: body responsible for the project, owner of original Archive, editor of digital resources;
- history of the original Archive;
- structure of the original Archive;
- historical/biographical profile of the owner of the original Archive;
- rights that regulate the use of original documents.

The <desc> section has been divided into two sub-sections:

1. context: it enbeds the data relating to entities involved in the ownership and management of the original documents;
2. description: it describes the consistency of the subfund to which the resource described in the sub-section <File> belongs.

The <File> section dedicated to single document describes:

- the original document represented in the image: subject, text abstract, creator, contributors, chronic date, topical date, support, language;
- the physical position of the original in the Archive;
- the editor who creates the descriptions.

The section on rights follows, which describes:

- ownership of the digital artifact;
- accessibility and reuse of the digital artifact;
- ownership and accessibility of the original document.

The schema closes with the technical metadata describing the different image formats in which each digital objects relating to the respective pages of a document have been reproduced, with their structural components.

Conclusion

Starting from the art. 2 of the UE "Council conclusions of 21 May 2014 on cultural heritage as a strategic resource for a sustainable Europe" (2014/C 183/08), the paper focuses on the need to rethink digital and digitization process for long term digital preservation, aiming to redefine them as the new Cultural Heritage of the contemporary era.

This new way to observe digital artifacts and their co-creation process is the indispensable prerequisite for co-creating aware Digital Culture and for giving due importance to digitization and dematerialisation, whose process, from the planning stages, need an approach focused on data preservation and, to this goal, on the decisive role that the descriptive metadata play.

The case study was the digitization project of the "Historical Archive of the Giuseppe Laterza & Figli Publishing House". In particular, the attention to preservation focused on structuring the schema of metadata and, above all, on descriptive writing, with regard to the choice of tags, elements and attributes, and to draft descriptive information of each digital artefact. In fact, our opinion is that they constitute the main source for the knowledge of both the single digital artifact, and the full project and its evolution, thus configuring itself as fundamental elements to validate and certify the data, guaranteeing quality, authenticity and sustainability as witness and memory of the contemporary Digital Age, with the aim to increase the knowledge of future generations about Twenty-First Century.

References

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XG0614%2808%29

http://www.interpares.org/

http://www.pugliadigitallibrary.it/

Agenzia per l'Italia Digitale (AgID), Presidenza del Consiglio dei Ministri, *Linee guide sulla conservazione dei documenti informatici*,, Versione 1.0 – dicembre 2015, pp. 45 ss. (http://www.agid.gov.it/sites/default/files/linee_guida/la_conservazione_dei_documenti_informatici_rev_def.pdf).

L. Bailey, *Digital Orphans: The Massive Cultural Black Hole On Our Horizon*, Techdirt, Oct 13th 2015 (https://www.techdirt.com/articles/20151009/17031332490/digitalorphans-massive-cultural-blackhole-our-horizon.shtml).

S. Cosimi, *Vint Cerf: ci aspetta un deserto digitale*, Wired.it, 16 febbraio 2015 (http://www.wired.it/attualita/2015/02/16/vint-cerf-futuro-medievale-bit-putrefatti/).

T. Di Noia, A. Ragone, A. Maurino, M. Mongiello, M. P. Marzocca, G. Cultrera, M. P. Bruno, *Linking data in digital libraries: the case of Puglia Digital Library*, in A. Adamou, E. Daga, L. Isaksen, "Proceedings of the 1st Workshop on Humanities in the Semantic Web,

co-located with 13th ESWC Conference 2016 (ESWC 2016)", Anissaras, Greece,May 29th, 2016 (http://ceur-ws.org/Vol-1608/paper-05.pdf).
L. Duranti, E. Shaffer (ed. by), *The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage*, UNESCO Conference Proceedings, 26-28 September 2012, Vancouver (http://ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf)
V. Gambetta, *La conservazione della memoria digitale*, [Rubano], Siav, 2009.
P. Ghosh, *Google's Vint Cerf warns of 'digital Dark Age'*, BBC News, Science & Environment, 13 February 2016 (http://www.bbc.com/news/science-environment-31450389).
M. Guercio, *Gli archivi come depositi di memorie digitali*, "Digitalia", Anno III, n. 2, ICCU Roma, 2008, pp. 37-53.
M. Guercio, *Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali*, Roma-Bari, Laterza, ed. 2013.
M. Guercio, *Conservazione delle e-mail: le raccomandazioni del progetto InterPares* (http://www.conservazionedigitale.org/wp/wp-content/uploads/2014/12/Guercio-8-Conservare-documenti-email.pdf)
Joint Steering Committee for Development of RDA, *Resource Description and Access (RDA)* (http://www.iccu.sbn.it/opencms/export/sites/iccu/documenti/2015/RDA_Traduzione_ICCU_5_Novembre_REV.pdf)
W. Kool, B. Lavoie, T. van der Werf, *Preservation Health Check: Monitoring Threats to Digital Repository Content*, OCLC Research, Dublin (Ohio), 2014 (http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-preservation-health-check-2014.pdf).
B. Lavoie, R. Gartner, *Preservation Metadata (2nd edition), DPC Technology Watch Report*, 03 May 2013, DPC Technology Watch Series (http://www.dpconline.org/docman/technology-watch-reports/894-dpctw13-03/file).
Library of Congress, *PREMIS – Preservation Metadata: Implementation Strategies*, v. 3.0 (http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf)
G. Marzano, *Conservare il digitale. Metodi, norme, tecnologie*, Milano, Editrice Bibliografica, 2011.
Mellon Foundation and Digital Preservation Coalition Sponsor Formation of Task Force for Email Archives, 1 November 2016 (https://mellon.org/resources/news/articles/mellon-foundation-and-digital-preservation-coalition-sponsor-formation-task-force-email-archives/).
OCLC. *PREMIS (PREservation Metadata: Implementation Strategies) Working Group*, 2005 (http://www.oclc.org/research/projects/pmwg/).
S. Pigliapoco, *Conservare il digitale*, Macerata, EUM, 2010.
David S. H. Rosenthal, *Emulation & Virtualization as Preservation Strategies*, Report commissioned by The Andrew W. Mellon Foundation, October 2015 (https://mellon.org/media/filer_public/0c/3e/0c3eee7d-4166-4ba6-a767-6b42e6a1c2a7/rosenthal-emulation-2015.pdf).
Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (F. Berman and B. Lavoie, co-chairs), La Jolla, February 2010 (http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf).
F. Tomasi, M. Daquino, *L'uso delle ontologie per la preservazione dei dati*, in presentazione al Convegno AIUCD 2017, Roma, 26-28 gennaio 2017.
M. Zane, *Per una nuova pedagogia del patrimonio*, "Giornale delle Fondazioni" (http://www.ilgiornaledellefondazioni.com/content/una-nuova-pedagogia-del-patrimonio).

Mesurer Merce Cunningham: une expérimentation en «theatre analytics»

Clarisse Bardiot
clarisse_bardiot@mac.com
University of Valenciennes, Belgium

Theatre studies is a largely under-discussed topic in digital humanities research projects. It's lagging behind the first wave of digital humanities scholarship, « focus[ing] on large-scale digitization projects and the establishment of technological infrastructure» (Presner, 2010). Theatre studies remains on the fringe of a growing phenomenon: culture analytics. In the context of big and complex datasets, culture analytics «is the data-driven analysis of culture» (IPAM, 2016). I suggest the expression «theatre analytics» (Bardiot, 2017). To paraphrase the culture analytics definition, theatre analytics is the data-driven analysis of theatre, whether it concerns theatre history (Caplan, 2016), drama or mise-en-scène. To understand what quantitative methodologies can bring to the knowledge of theatre, I propose a case study of Merce Cunningham. What can we learn about Merce Cunningham, one of the most influential *choreographers* of the 20th century, thanks to theatre analytics? A leader of the American avant-garde throughout his seventy year career from 1938 to 2009, he establishes in 2000, in the twilight of his career, the Merce Cunningham Trust, in order to preserve the integrity of his work. At the same time, he decides to dissolve the Merce Cunningham Dance Company (MCDC) two years after his death and a legacy tour. This is an unprecedented initiative. On one hand, it demonstrates exceptional effort and dedication to document the works. On the other hand, it challenges the ephemeral nature of

performing arts : 86 out of 183 choreographies are documented with "digital Dance Capsules" "so that it may be performed in perpetuity"(Dance Capsules, n.d.). By the way, two groups of works are defined: the canon (key works with extensive documentation in order to perform them again and again); the auxiliary (minor works with no documentation available to the public and *de facto* impossible to replay).

The data was collected from the Merce Cunningham Trust website. It concerns theatre production and cast, Dances Capsules documentation and the history of the MCDC. The dataset contains 183 works from 1938 to 2009 (including 86 Dance Capsules) and 347 people. We can identify three main data categories: people, works and documentation. What can we infer from beyond the data about the MCDC history, Cunningham's aesthetics and documentation strategies?

Measuring means measuring instruments. I used various and complementary tools in order to vary the approaches and analysis of the same dataset : Gephi for network analysis; Palladio for geographic and temporal representation; spreadsheet (Excel, Open Office, Datamatic) for statistics analysis. This paper will present the first results of this research, part of it conducted with students during a graduate «introduction to digital humanities» course. Statistical diagrams show three different periods of Cunningham's work; a stylistic signature with a preference for pieces that are 30 minutes long, and for soli, sextets and works with 13 to 15 dancers; a general trend towards more dancers and more length; the special place of soli in order to articulate the canon and the auxiliary; the organization of documents in the Dance Capsules. Network analysis let me define two different ways of collaboration, the «star» and the «spiral», and raises awareness on pivotal dancers. Geographic representation highlights relations between Europe and the United States.

In a wider historical perspective, it would be interesting to compare these preliminary results with other datasets. One example : two patterns have been identified in the Cunningham collaborations network : the star (figure 1), with discontinuous, centralized collaborations and groups separated from each other; the spiral (figure 2), with continuous, collective collaborations and one group growing organically. The change from the star to the spiral takes place when the company is created. Do these patterns characterize other choreographers and directors careers ? Is the creation of the company the main factor causing the evolution from the first pattern towards the second one ? While a well-worn issue – we do know that the creation of a company plays a crucial role in a career – the fact remains that "theatre analytics" let us visualize the patterns this break constitutes (or maybe not) and define different ways of collaborations.

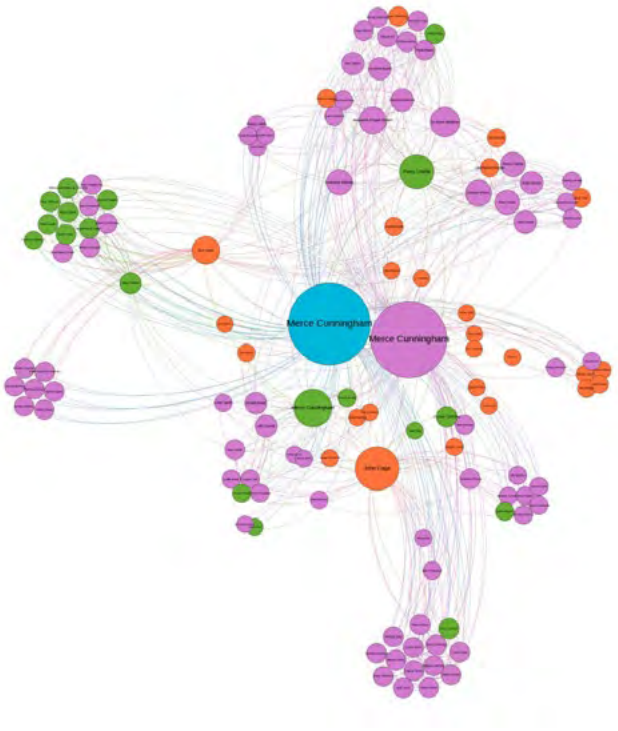


Figure 1 : Merce Cunningham's collaborations network before 1954. The star pattern.
Pink, dancers; orange, composers ; green, stage designers ; blue, choreographer.

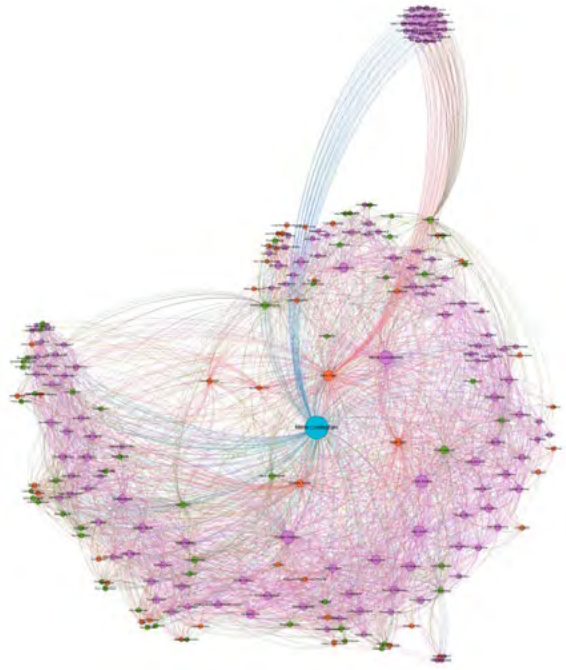


Figure 2 : Merce Cunningham's collaborations network after 1954. The spiral pattern.

References

Dance Capsules - Merce Cunningham Trust https://mercecunningham.org/film-media/dance-capsules/ (accessed 29 May 2018).

Merce Cunningham Dance Capsules http://dancecapsules.merce.broadleafclients.com/about.cfm (accessed 29 May 2018).

Bardiot, C. (2017). Arts de la scène et culture analytics. (Ed.) Galleron, I. *Revue d'historiographie du Théâdre. Etudes théâtrales et humanités numériques*(4): 11–20.

Caplan, D. (2016). Reassessing Obscurity: The Case for Big Data in Theatre History. *Theatre Journal*, 68(4): 555–73.

Tangherlini, T. R. (ed). (2016). *Culture Analytics : White Papers*. http://www.ipam.ucla.edu/wp-content/uploads/2016/09/Culture_Analytics_WhitePapers.pdf.

Presner, T. (2010). Digital Humanities 2.0: a report on knowledge. *Connexions Project*.

Is Digital Humanities Adjuncting Infrastructurally Significant?

Kathi Inman Berens
kathiberens@gmail.com
Portland State University, United States of America

The question of when "digital humanities" will drop the "digital" modifier and become "humanities" has special resonance for adjunct instructors. Digital humanities senior scholars might bridge the gap between tenured working conditions and adjunct working condition when crafting field infrastructures: not just because adjuncts merit both employment protections and what I call "microbenefactions" (more on that below), but because adjuncts are the invisible mass of humanities faculty buttressing every kind of institution, from community college to elite research-1 university. Adjuncts shoulder the humanities enterprise, teaching the general education classes that free researchers to pursue critical questions that advance the field.

This talk examines the infrastructural causes of DH adjunct invisibility and proposes two remedies: to motivate DH adjunct self-identification by convening DH adjunct-specific prizes and bursaries; and to invite senior DH faculty to perform "microbenefactions" that cost little effort and give adjuncts access to prize-worthy work opportunities or other benefits, such as renewable funding.

When "digital" humanities becomes just humanities, what's to stop "adjunctification" from converting DH tenure lines into part-time or other tenure-ineligible work, as has happened pervasively in other sub-specialties? In 2012, Stephen Ramsay problematized DH as "the hot thing." It's a skepticism shared by many in the field, including panelists of the DH 2017 Conference panel "Challenges for New Infrastructures and Paradigms in DH Curricular Program Development," which openly wondered whether graduate students were well served by DH certificate programs.[1] Miriam Posner notes that DH's "sexiness" today obscures the "widespread understaffing" of many DH initiatives"[2] This is an analog to adjunctification, the "shortsighted" boom/bust cycles of "soft" money quickly depleted which then require maintenance with a precarious budget. Amy Earhart has documented the unsustainability of early DH passion projects, websites whose hand-built archives rusticate when the faculty author retires or moves institutions.[3] Startups are sexy, but maintenance is not. When today's senior DH faculty retire in ten or twenty years, what infrastructures of care will be in place to stop those vacated tenure lines from being converted to part-time positions? The gender politics of "sexy," "hot" DH cast a pall over the field when one factors in that the majority of adjuncts are women. "As a woman of color," Liana M. Silva wonders, "I am especially interested to know what the women in contingent ranks look like. According to the Education Department's 2009 report, 51.6 percent of contingent faculty are women. The same report says 81.9 percent of contingent faculty are white. To what extent is contingent labor a problem for white women? Or, from another angle, to what extent is this a white labor issue, where class is meant to trump race?"[4] These questions about race, gender and contingent labor are digital humanities questions.

Awarding DH Adjuncts

In its mentoring, promotion, and awards structures, the humanities professoriate is legacy-bound, oriented to a tenure system that pertains to only one quarter of people working in the field.[5] If, as James F. English contends in

1 See the DH 2017 panel abstract here: https://dh2017.adho.org/abstracts/176/176.pdf. Ryan Cordell pointedly observes in published version of his DH 2017 talk that "completing the hours required for our robust [DH graduate] certificate program requires students to decide their path almost immediately upon admission, and the decision to pursue the certificate dictates very particular routes through the larger Ph.D. program." See Cordell's "Abundance and Usurpation While Building a DH Curriculum" posted to his blog: http://ryancordell.org/research/abundance/

2 Miriam Posner, "Money and Time," http://miriamposner.com/blog/money-and-time/

3 Earhart, *Traces of the Old, Uses of the New*.

4 Liana M. Silva: https://chroniclevitae.com/news/1017-how-many-women-are-adjuncts-out-there; National Center for Education Statistics 2009 report to which Silva refers: https://nces.ed.gov/pubs2011/2011150.pdf See also: "Women as Contingent Faculty: The Glass Wall," published by the American Association of University Professors http://archive.aacu.org/ocww/volume37_3/feature.cfm?section=1 and New Faculty Majority's "Women and Contingency" project: http://www.newfacultymajority.info/women-and-contingency-project/

5 "Adjunctification" is well documented by adjunct advocacy organization like New Faculty Majority and Adjunct Nation; professio-

The Economy of Prestige, the key indicator of any contemporary cultural phenomenon entering the mainstream is the creation of a prize (2), then perhaps it is time for digital humanists to create criteria of DH excellence specific to DH adjunct working conditions because adjuncting is the instructional mainstream. Doing so would motivate adjunct DHers to identify their work as DH and contribute recognizably toward DH research and pedagogy field development. Lack of access to an adjunct-specific DH prize reinforces adjunct invisibility, making it highly unlikely that even very good research will attain the recognition necessary to vault the scholar out of adjuncting. Most of the seven DH adjuncts I interviewed don't necessarily identify themselves as "digital humanists" because they are not hired specifically to teach DH, though their methods are consistent with DH pedagogy practices.[6] "Imposter syndrome" is intensified by employment insecurity and DH definitional heterogeneity.[7]

How to give adjuncts access to prize-worthy work opportunities? Senior scholars are key. In my talk, I will discuss microbenefactions senior scholars gave me when I adjuncted (2011-2014). Those invitations gave me access to nationally-visible projects and let me train myself in techniques that are now a core part of my tenure track job.

"Microbenefactions" is a term I invented to signify the opposite of microaggressions. They are small actions that shift the balance of power, the order of operations, that give adjuncts access to prestige or information otherwise inaccessible to them. Note that I use the singular here: "an" adjunct. These acts of inclusion are do-able as a one-off or in the course of a given term, not the Herculean efforts of adjunct advocates such as New Faculty Majority President Maria Maisto, Adjunct Nation, and the PrecariCorps collective who publish PrecariTales, 300-500 word anonymously authored adjunct stories.[8]

Unlike state-mandated employment protections, microbenefactions are individual and hyperlocal. They layer adjuncting's transactional dyad with the more branching, collegial conceptualization of value typical of tenure-track employment. This is human-centered DH infrastructure. We acknowledge that humans are not widgets, and that DH teaching is not a dissemination of knowledge. The medium is the message. If the medium is adjuncting, then the message our students imbibe is that learning is transactional. Microbenefactions disrupt neoliberal infrastructure that shrinks learning and collegiality to transactions.

What is a microbenefaction? It's action by a tenured or tenure-track scholar who

- writes funding for adjunct salary into grant proposals
- advises and mentors adjuncts
- seeks input from adjuncts about student-centered pedagogy
- aids adjuncts in finding university resources or paid extra work
- invites adjuncts to meetings
- co-authors with adjuncts
- doesn't eliminate adjunct applications when deciding awards and honors
- authorizes support for adjunct professional development, such as conference travel
- pays to license adjunct-authored course materials after the adjunct leaves the institution
- writes letters of recommendation for adjuncts

Microbenefactions enact DH's ethical ambit, which the Global Outlook::Digital Humanities special interest group articulates as a recognition "that excellent work is being done around the world,"[9] even in elite first-world institutions that rely on adjunct labor but largely eliminate that labor from tenure and promotion consideration.

References

Berens, Kathi Inman. "Judy Malloy's Seat at the (Database) Table: A Feminist Reception History of Early Electronic Literature Hypertext." *Literary and Linguistic Computing*, Volume 29, Issue 3, 1 September 2014, pages 340–348, https://doi.org/10.1093/llc/fqu037.

nal groups such the AAUP and the Modern Language Association (2014); intra-university studies such as George Mason's, which surveyed 240 GMU adjuncts and "has been hailed as the most comprehensive study of a university's contingent faculty working conditions to date" (2014); trade journals like *Inside Higher Education* and *The Chronicle of Higher Education*; and the popular press. I am struck by *The Atlantic Monthly*'s occasional series (2013-present) that features titles like "There's No Excuse for How Universities Treat Adjuncts" and "The Cost of an Adjunct." See also Kathi Inman Berens and Laura E. Sanders, "DH and Adjuncts: Putting the Human Back in the Humanities."

6 A note about method. My university's Human Subjects Research Review Committee determined an IRB was not required for me to conduct informational interviews with adjuncts. I used a common set of questions with each adjunct. The conversations veered to the specifics of their own particular cases.

7 The authors of the "Alternate Histories of DH" panel note in their abstract: "Matthew Kirschenbaum's identification of the digital humanities in 2014 as a 'discursive construction' that ignores the 'actually existing projects' of the field set the stage for scholars to rethink how the digital humanities conceptualizes its work and its history ('What Is' 48). More recently, in the introduction to *Debates in the Digital Humanities 2016*, Matthew Gold and Lauren Klein use the scholarship of Rosalind Krauss who, in 1979, described art history as emerging as 'only one term on the periphery of a field in which there are other, differently structured possibilities.'"

8 https://precaricorps.org/about/true-stories/ The pinned story at time of writing details an adjunct who's taught at the same university for ten years and has been hired to revise materials for a large enrollment course. One chair made sure she got paid the first lump sum; the replacement chair didn't with the second, and she's still waiting with "no recourse except to wait." The Twitter hashtags #AdjunctLife and #RealAcademicBios also gather adjunct (but don't curate) stories.

9 Global Outlook::Digital Humanities is a special interest group of the Alliance of Digital Humanities Organization. See http://www.globaloutlookdh.org/

_____. "Want to Save the Humanities? Pay Adjuncts to Learn Digital Tools" in *Disrupting the Humanities: Digital Edition*, 05 January 2015, http://www.disruptingdh.com/want-to-save-the-humanities-pay-adjuncts-to-learn-digital-tools/. Accessed 27 November 2017.

_____. "Sharing Precarity: Adjuncts, Global Digital Humanities, and Care," in *Debates in Digital Humanities 2017*, eds. Lauren F. Klein and Matthew K. Gold. Minneapolis: University of Minnesota Press. In press.

Berens, Kathi Inman and Laura E. Sanders. "Putting the Human Back in the Humanities: Adjuncts and Digital Humanities" in *Disrupting Digital Humanities: Print Edition*, eds. Dorothy Kim and Jesse Stommel. New York: Punctum Press. 2017.

Bessette, Lee Skallerup. *Adjunct Run*. https://adjunctrun.readywriting.org/ Accessed 27 November 2017.

Bretz, Andrew. "The New Itinerancy: Digital Pedagogy and the Adjunct Instructor in the Modern Academy." *Digital Humanities Quarterly* Vol. 11, No. 3 (2017). http://www.digitalhumanities.org/dhq/vol/11/3/000304/000304.html

Clement, Tanya [panel chair], Alison Booth, Ryan Cordell, Miriam Posner, Maria Sachiko Cecire. "Challenges for New Infrastructures and Paradigms in DH Curricular Program Development," panel at the 2017 Digital Humanities Conference in Montréal, Québec, Canada August 8-11, 2017. https://dh2017.adho.org/abstracts/176/176.pdf.

Cordell, Ryan. "Abundance and Usurpation While Building a DH Curriculum." http://ryancordell.org/research/abundance/ 23 August 2017.

Davis, Rebecca Frost, Matthew K. Gold, Katherine D. Harris and Jentery Sayers, eds. *Digital Pedagogy in the Humanities*, Digital Edition (peer editing version) https://digitalpedagogy.mla.hcommons.org/. Accessed 27 November 2017.

English, James F. *The Economy of Prestige: Prizes, Awards, and the Circulation of Cultural Value.* Cambridge: Harvard University Press, 2008.

Finley, Ashley. "Women as Contingent Faculty: The Glass Wall." *On Campus With Women* featured article of the Association of American Colleges and Universities. Vol. 37, No.3 (Winter 2009). http://archive.aacu.org/ocww/volume37_3/feature.cfm?section=1 Accessed 27 November 2017.

Gonzales, Andrea and Sophie Houser. *Tampon Run*. http://tamponrun.com/ Accessed 27 November 2017.

Higgen, Parker. Tweet dated 6 January 2015. https://twitter.com/xor/status/552456370629672960. Accessed 27 November 2017.

Honn, Joshua. "Never Neutral: Critical Approaches to Digital Tools Culture in the Humanities." https://figshare.com/articles/Never_Neutral_Critical_Approaches_to_Digital_Tools_Culture_in_the_Humanities/1101385 Accessed 21 November 2017.

Jacobs, Ken, Ian Perry, and Jenifer MacGillvary. "The High Public Cost of Low Wages: Poverty-Level Wages Cost U.S. Taxpayers $152.8 Billion Each Year in Public Support for Working Families." UC Berkeley Center for Labor Research and Education. April 13, 2015 Report. http://laborcenter.berkeley.edu/the-high-public-cost-of-low-wages/

Jasnik, Scott. "Humanities Job Woes." *Insider Higher Ed.* January 4, 2016.
https://www.insidehighered.com/news/2016/01/04/job-market-tight-many-humanities-fields-healthy-economics Accessed 27 November 2017.

Knapp, Laura G., Janice E. Kelly-Reid and Scott A. Ginder. "Employees in Postsecondary Institutions, Fall 2009, and Salaries of Full-Time Instructional Staff, 2009-10," a Report published by the U.S. Department of Education. November 2010. https://nces.ed.gov/pubs2011/2011150.pdf Accessed 27 November 2017.

Koseff, Alexei. "Part-time community college instructors to get job protections" [*sic*]. *Sacramento Bee*. 30 September 2016. http://www.sacbee.com/news/politics-government/capitol-alert/article105301086.html Accessed 27 November 2017.

Losh, Elizabeth, ed. *MOOCs and Their Afterlives: Experiments in Scale and Access in Higher Education.* Chicago: University of Chicago, 2017.

Manyika, James, Susan Lund, Jacques Bughin, Kelsey Robinson, Jan Mischke, and Deepa Mahajan. "Independent Work: Choice, necessity, and the Gig Economy." *McKinsey Global Institute*. October 2016. https://www.mckinsey.com/global-themes/employment-and-growth/independent-work-choice-necessity-and-the-gig-economy.

McGrail, Anne. "Whole Game: Digital Humanities at Community Colleges." *Debates in Digital Humanities 2016*, eds. Lauren F. Klein and Matthew K. Gold. Minneapolis: University of Minnesota Press, 2016. http://dhdebates.gc.cuny.edu/debates/text/53

McPherson, Tara. "Theory/Practice: Lessons Learned from Feminist Film Studies," on the panel "Alternate Histories of the Digital Humanities: a Short Paper Panel Proposal," convened by Roger Whitson and featuring Whitson, Amy Earhart, Steven Jones and Padmini Ray Murray, at the Digital Humanities 2017 conference July 8-11, 2017 in Montréal, Québec, Canada. https://dh2017.adho.org/abstracts/115/115.pdf Accessed 27 November 2018.

Molloy College DH Adjunct Job Advertisement. https://main.hercjobs.org/jobs/10389448/new-media-and-digital-humanities-adjunct. Accessed 18 November 2017. [The link will expire; see screenshot in Appendix.]

Nazer, Daniel and Elliot Harmon. Electronic Frontier Foundation, "Stupid Patent of the Month: Elsevier Patents Online Peer Review." August 31, 2016. https://www.eff.org/deeplinks/2016/08/stupid-patent-month-elsevier-patents-online-peer-review

New Faculty Majority "Women and Contingency Project." http://www.newfacultymajority.info/women-and-contingency-project/ Accessed 27 November 2017.

Pierazzo, Elena. "The Disciplinary Impact of the Digital: DH and ‚The Others'." Keynote at the Digital Humanities Summer Institute 2017, Victoria, B.C., Canada 16 June 2017. Abstract viewable here: http://dhsi.org/schedule.php
Precaricorps "True Stories." https://precaricorps.org/about/true-stories/ Accessed 27 November 2017.
Ramsay, Stephen. "The Hot Thing." (2012) https://github.com/sramsay/sramsay.github.com/blob/master/_posts/2012-04-09-hot-thing.markdown. Accessed 27 November 2017.
Risam, Roopika and Susan Edwards. "Micro DH: Digital Humanities at Small Scale." Conference talk at Digital Humanities 2017 Conference in Montréal, Québec, Canada August 8-11, 2017. Abstract viewable here: https://dh2017.adho.org/abstracts/196/196.pdf
Silva, Liana. "How Many Women Are Adjuncts Out There?" *Chronicle of Higher Education.* 27 May 2015. https://chroniclevitae.com/news/1017-how-many-women-are-adjuncts-out-there Accessed 27 November 2017.
Stanley, Sara Catherine. "Why DH?" (2017) http://scatherinestanley.us/2017/06/why-is-dh Accessed 21 November 2017.
Varner, Stuart. "Digital Humanities or Just Humanities?" https://stewartvarner.com/2013/11/digital-humanities-or-just-humanities/ Accessed 21 November 2017.
Warford, Erin. "StoryTelling with Digital Maps," a workshop at the 2017 Summer Digital Humanities Workshop Series at Canisius College. https://blogs.canisius.edu/digital-humanities/gis2017/ Accessed 27 November 2017.

Transposição Didática e atuais Recursos Pedagógicos: convergências para o diálogo educativo

Ana Maria Bosse
anahboss@hotmail.com
Universidade Federal de Santa Catarina, Brazil

Juliana Bergmann
jcfbergmann@gmail.com
Universidade Federal de Santa Catarina, Brazil

Resumo: Esta pesquisa, desenvolvida com alunos do 3º ao 5º ano do Ensino Fundamental brasileiro, tem como objetivo analisar a importância da renovação dos recursos pedagógicos no contexto educacional, da sociedade contemporânea, e refletir sobre as possibilidades e potencialidades destes recursos no processo de repensar o papel da escola nesta cultura digital, para assim atender as necessidades educacionais e favorecer o diálogo educativo.

Introdução

Atualmente, em nossa sociedade, as Tecnologias Digitais de Informação e Comunicação, associadas à internet, têm proporcionado mudanças constantes na circulação dos saberes, na produção e apropriação dos conhecimentos, passando a **informação** a ser o **bem** de maior valor social, e como já apontado por Pérez Goméz (2015:15), nesta era "a atividade principal dos seres humanos tem a ver com a aquisição, o processamento, a análise, a recriação e a comunicação da informação". As constantes inovações tecnológicas, desta cultura digital, vêm influenciando e interferindo nas relações interpessoais, despertando novas formas de gerenciar socialmente o conhecimento, de ensinar e aprender.

Nesta perspectiva, podemos destacar duas situações recorrentes no contexto educacional desta sociedade: 1) os recursos pedagógicos oferecidos nas escolas muitas vezes não levam em conta o uso potencial das novas mídias pelos alunos, ignorando todas as experiências cotidianas que eles desenvolvem e adquirem com essas novas tecnologias; 2) muitas vezes a escola dispõe de inúmeros recursos tecnológicos (midiáticos) de última geração, mas estes são subutilizados, sem que o educador os inclua em seu planejamento, seja por desconhecimento, seja por não acreditar que fará qualquer diferença ao aluno. Assim, se faz necessário pensarmos em elos que favoreçam esta aproximação.

Uma proposta para fomentar um maior diálogo educativo, conforme a presente pesquisa – realizada em uma escola de Ensino Fundamental onde atuo como Coordenadora Pedagógica –, dá-se através da compreensão do uso e da definição dos recursos pedagógicos no processo de ensino e aprendizagem, considerando que estes precisam ser constantemente reavaliados, de forma a beneficiar principalmente a transposição didática dos saberes, acreditando, assim, que o caminho para a construção de um novo pensar e de um novo fazer se edifica no questionamento, na pesquisa, no revisitar e analisar os modelos existentes para então propor novos indicativos. Para exemplificar, apontamos que o uso da internet, dos sites e dos aplicativos, através dos computadores, dos *tablets* e dos celulares, utilizados como recursos pedagógicos, podem proporcionar novas práticas para aproximar o conteúdo didático com a práxis da sala de aula, estabelecendo uma conexão concreta com a cultura cotidiana do aluno.

Recursos Pedagógicos na era digital

Dentro de todo este entrelaçar de mudanças advindas das novas tecnologias, é notório que a informação está à disposição em qualquer momento, a todo tempo, nos mais diversos locais; as tecnologias de comunicação trazem consigo esta potencialidade, fenômeno intitulado "*ubiquidade*" (Santaella, 2013); as potencialidades da co-

municação, principalmente com os dispositivos móveis e digitais, são inúmeras.

As novas gerações já nascem imersas nesse contexto da cultura digital. Desde muito cedo os sujeitos interagem com as mais diversas tecnologias de informação e comunicação e o mundo do ciberespaço já é parte constituinte do seu cotidiano. Assim, se adaptam a ele muito rapidamente e trafegam por entre essas novidades tecnológicas com desenvoltura e habilidade. Diante desta realidade, precisamos refletir, analisar e repensar o papel da escola e do ensino de modo que compreenda o contexto da sociedade atual.

Rivoltella (2007, *apud* Didonê, 2007), propõe que a mídia pode e deve permear os processos de ensino e aprendizagem, como acontece com a escrita, destacando que o papel assumido pelo professor que usa as novas tecnologias midiáticas não se limita a falar, mas sim, a direcionar o uso dos meios de comunicação pelos alunos.

A partir destas reflexões, podemos destacar que no atual contexto educacional nos encontramos diante de *"escolas analógicas e cabeças digitais"* (Petarnella, 2008), sendo pertinente e necessário trazer o mundo vivencial do aluno – tecnológico e midiático desta cultura que já faz parte do nosso cotidiano, para o ambiente escolar, e assim favorecer um verdadeiro diálogo educativo em que todos se beneficiem.

Recursos Pedagógicos: caminho para o diálogo educativo

Acreditando nas potencialidades dos recursos pedagógicos e na contribuição destes para aproximar e envolver o aluno no processo de ensino e aprendizagem, ponderamos também a importância destes como elementos que fazem parte da cultura do homem, que o colocam em contato com o seu tempo, com a sua historicidade.

Ao considerarmos que os recursos pedagógicos comportam em si a missão e o potencial, de se bem utilizados, de aproximar o aprendiz da sua aprendizagem, possibilitando maior entendimento na relação com o currículo pedagógico, mais interação na relação dos sujeitos envolvidos neste processo educacional, compreendemos que eles abrem para uma nova linguagem do aprender. De acordo com Eiterer e Medeiros (2010: 1), definimos como recursos pedagógicos "o entendimento daqueles lugares, profissionais, processos e materiais que visem assegurar a adaptação recíproca dos conteúdos a serem conhecidos aos indivíduos que buscam conhecer", e atendendo o importante papel que estes ocupam e desempenham no universo pedagógico, ainda compete destacar que sua abrangência está além da materialidade dos recursos em si.

Atualmente estamos diante de outro pensar pedagógico, que leva em consideração a importância da transposição didática nas relações de aprendizagem, nas relações entre aluno, professor, conhecimentos científicos, currículo, escola, prática pedagógica e recursos pedagógicos, e Almeida (2011: 11), enfatiza que "as nossas discussões acerca da transposição didática têm de ser entendidas dentro de uma concepção multiforme e ininterrupta". Pois, se é ao fazer pedagógico que compete tornar esta cultura transmissível e assimilável, ainda de acordo com o autor (Almeida, 2011), de algum modo é necessário transcender as diferenças e, através da interdisciplinaridade, rompermos com uma técnica homogeneizadora e homogeneizante de currículo, que engessa os conhecimentos, e que não compreende o valor da contextualização na prática educativa. Faz-se necessário pesarmos o fazer pedagógico através da prática reflexiva, e conforme Perrenoud (2002: 65), "a prática reflexiva é uma relação com o mundo: ativa, crítica e autônoma. Por isso, depende mais da postura do que de uma estrita competência metodológica". Nesse sentido, diante de todo o contexto apresentado sobre as condições da escola contemporânea e do aluno nesta sociedade da informação - da cultura digital, esta pesquisa, prima por investigar as possibilidades do uso de recursos pedagógicos e tecnológicos digitais (*tablets*, celulares, internet, sites, aplicativos), promover o diálogo educacional entre professor e aluno bem como favorecer a transposição didática, estimular no aluno o hábito da pesquisa e tornar mais significativo ao aprendiz o processo de ensino e aprendizagem, e analisar se estes recursos podem proporcionar uma nova relação no diálogo entre currículo, metodologia, professor e aluno.

References

Almeida, G P. (2011). *Transposição didática por onde começar?* São Paulo: Cortez Editora.

Didonê, D. Pier Cesare Rivoltella: F*alta cultura digital na sala de aula. Nova Escola*. Disponível em: <http://novaescola.org.br/formacao/formacao-continuada/pier-cesarerivoltella-falta-cultura-digital-sala-aula-609981.shtml>. Acesso em: 14 maio 2016.

Eiterer y Medeiros, C. L. *Recursos Pedagógicos*. Disponível em: http://www.gestrado.net.br/pdf/155.pdf. Acesso em 09/11/2017.

Freire, P. (2007). *Educação como Prática da Liberdade*. 29. ed. Rio de Janeiro: Paz e Terra.

Gentile, P. Antonio Nóvoa: *Professor se forma na escola. Nova Escola*. Disponível em: <https://novaescola.org.br/conteudo/179/entrevista-formacao-antonio-novoa>. Acesso em: 20 novembro 2017.

Moran, J. M., Masetto, M. T. y Behrens, M. A. (2003) *Novas Tecnologias e Mediação Pedagógica*. 7. ed. São Paulo: Papirus.

Pérez Gómez, Á. I. (2015). *Educação na era digital: a escola educativa*. Porto Alegre: Penso.

Perrenoud, P. (2002). *A Prática reflexiva no Ofício do Professor*. Porto Alegre: Artmed.

Petarnella, L. (2008). *Escola analógica: Cabeças digitais: o cotidiano escolar frente às tecnologias midiáticas e digitais de informação e comunicação*. Campinas, SP: Alínea.

Sacristán, J. G. (2000). *O currículo: uma reflexão sobre a prática*. Porto Alegre: Artmed.
Santaella, L. (2013). *Comunicação ubíqua: Repercussões na cultura e na educação*. São Paulo: Paulus.

Hurricane Memorial: The United States' Racialized Response to Disaster Relief

Christina Boyles
christina.boyles@trincoll.edu
Trinity College, United States of America

On September 20, 2017, Hurricane Maria made landfall in Puerto Rico. As the strongest hurricane to hit the island since 1928, the storm has caused significant damage—especially to infrastructure including roads, dams, communications networks, the electrical grid and the water supply. With much of the island still without power, and with limited aid coming from the United States, Puerto Rico is being left to deal with a humanitarian crisis on its own. The slow nature of the United States' response, coupled with Donald Trump's barrage of tweets, highlight the ways in which colonial narratives are feeding into disaster response efforts. For example, when San Juan Mayor Carmen Yulin Cruz requested an increase in federal aid, Trump replied, "Such poor leadership ability by the Mayor of San Juan, and others in Puerto Rico, who are not able to get their workers to help. They want everything to be done for them when it should be a community effort" (@realDonaldTrump). He later went on to claim that Puerto Rico's need for aid was hurting the federal budget and to claim that Hurricane Maria was not "a real catastrophe" for the island ("Trump compares Puerto Rico to Katrina"). Trump's victim-blaming behavior highlights both his lack of empathy for the citizens of Puerto Rico and the racial prejudice that undergirds the U.S. colonial enterprise. Although rarely so blatant, such behavior is not new; rather, the United States has an ongoing legacy of racialized disaster relief that is grounded in its colonial endeavors, particularly in Puerto Rico.

According to *El Nuevo Día*, the most widely-circulated newspaper in Puerto Rico, "El huracán María no superó a San Felipe II según un informe preliminar", or "Hurricane Maria did not surpass the strength of the San Felipe II Hurricane" (Ortega Marrero). Nevertheless, the two storms bear striking similarities. Both hit the island of Puerto Rico as category 5 hurricane, both crossed the island from the southeast corner and moved through the center of the island to the northwest corner, and both had significant long-term effects on the island.

While coverage of the 1928 storm's devastation in Florida is prominently displayed in novels and journalistic reports, coverage of the damage in Puerto Rico is almost non-existent in the mainland United States. I argue that the vulnerabilities created by the hurricane of 1928 were pivotal to the United States colonial agenda in Puerto Rico, resulting in a land grab by corporations and government entities that would impede the island's agricultural industry and economy for decades. This is made evident by the fact that the U.S. downplayed effects of the storm, the U.S. implemented policies to hurt small farmers & agricultural workers, and the U.S. denied that their actions caused economic and environmental harm to Puerto Rican citizens.

To make these connections clearer and to bring the stories of the storm's underrepresented victims back into our cultural memory, I have launched a digital work called the Hurricane Memorial project. This site includes my preliminary research, visualizations of my findings, and interviews with survivors and their family members.

As Florida and the Caribbean start to recover from Hurricane Maria, it is important to note that those living in economically disadvantaged communities will suffer the greatest from the storm's damage—just as they did in 1928. Aid quickly was rushed to Florida, while the federal government is "killing [Puerto Rico] with inefficiency" ("I Am Mad As Hell"). Such a response demonstrates the ways in which United States' racialized response to natural disasters is deeply rooted in its colonial enterprise. Failing to address these issues risks reinforcing harmful colonial narratives and causing irreparable harm to communities throughout the Caribbean and the world.

References

@realDonaldTrump. "Such poor leadership ability by the Mayor of San Juan, and others in Puerto Rico, who are not able to get their workers to help. They want everything to be done for them when it should be a community effort." Twitter, 30 September 2017, 5:26 A.M. https://twitter.com/realDonaldTrump/status/914089003745468417
Hurston, Zora Neale. *Their Eyes Were Watching God*. 1937. HarperPerennial, 2006.
"'I Am Mad As Hell': San Juan Mayor Carmen Yulin Cruz Criticizes Maria Response." *YouTube*, uploaded by NBC Nightly News, 29 September 2017. https://www.youtube.com/watch?v=41hl5RwfOVc
Ortega Marrero, Melisa. "El huracán María no superó a San Felipe II según un informe preliminar." *El Nuevo Dia* [Guaynabo, Puerto Rico], 29 September 2017.
Sharp, Deborah. "Storm's path remains scarred after 75 years."*USA Today*, 4 September 2003.
Sterghos Brochu, Nicole. "Florida's Forgotten Storm: the Hurricane of 1928." *South Florida Sun-Sentinel,* 2003.
"Trump compares Puerto Rico to Katrina, 'a real catastrophe." *YouTube*, uploaded by USA Today, 3 October 2017. https://www.youtube.com/watch?v=-J18rugiTxoU

Backoff Lemmatization as a Philological Method

Patrick J. Burns
pjb311@nyu.edu
Institute for the Study of the Ancient World, United States of America

Automated lemmatization, that is the retrieval of dictionary headwords, is an active area of research in Latin text analysis. Latinists have available web-based applications like Collatinus (Ouvard and Verkerk, 2014) and LemLat (Bozzi et al., 1992) and web services like Morpheus (Almas, 2015). LatMor (Springmann, 2016) and TreeTagger (Schmid, 1994) offer lemmatization as a byproduct of their primary tasks as morphological taggers. Recent work, to name a few developments, has seen lexicon-assisted tagging and rule induction (Eger et al., 2015; cf. Juršic, 2010) as well as neural networks (Kestemont and De Gussem, 2017) used as strategies for improving Latin lemmatization.

In this short paper, I describe the implementation of the Backoff Lemmatizer (https://github.com/cltk/cltk/blob/master/cltk/lemmatize/latin/backoff.py) for the Classical Language Toolkit, an open-source Python platform dedicated to developing natural language processing tools for historical languages (Johnson, 2017). The Backoff Lemmatizer is in fact not a single lemmatizer but rather a customizable suite of sub-lemmatizers, based on the Natural Language Toolkit's SequentialBackoffTagger. The SequentialBackoffTagger allows the user to "chain taggers together so that if one tagger doesn't know how to tag a word, it can pass the word on to the next backoff tagger" (Perkins, 2014, 92). While the backoff process was originally designed to handle part-of-speech tagging, and so, a task with a limited tagset, it works well for lemmatization (~90.34% accuracy compared to the 93.49% to 95.30% range reported in Eger et al., 2015).

A default class for sequential lemmatization, BackoffLatinLemmatizer, is available through the CLTK "Lemmatize" module using the following backoff sequence: 1. a dictionary-based lemmatizer for high-frequency, indeclinable vocabulary; 2. a unigram-model lemmatizer based on training data; 3. a rules-based lemmatizer based on regular expression patterns; 4. a variation on the previous regular-expression-based lemmatizer that factors in principal-part information; 5. another dictionary-based lemmatizer using the Morpheus lemma dictionary; and finally 6. an identity lemmatizer that returns the token as lemma.

Although currently available and tested only for Latin, the Backoff Lemmatizer is in theory language agnostic, since the sub-lemmatizers can be passed language-specific training data and models. So, for example, the UnigramLemmatizer requires training data in the form of a Python list of tuples of the form *[('token1', 'lemma1'), ('token2', 'lemma2'), ...]*. A Latin model with data in this form based on The Ancient Greek and Latin Dependency Treebank (Celano, Crane, and Almas, 2017) is available in the CLTK Latin corpora, but a similar model could be built for any language. Similarly, the RegexLemmatizer relies on a custom dictionary of regular expression patterns extracted from Latin morphological patterns. But again, a list of patterns could be written for any language and worked into this sub-lemmatizer. Furthermore, the sub-lemmatizers can be added or removed as necessary, and can be reordered based to optimize accuracy for a given language or language domain. Accordingly, the BackoffLemmatizer is particularly well-suited to less-resourced languages (Piotrowski, 2012, 85): a language without sufficient training data could build a backoff chain that ignores the UnigramLemmatizer and rely only on dictionary- and rules-based sub-lemmatizers.

Because of its multipass combination of probabilistic tagging based on existing Latin text, Latin lexical data, and a ruleset based on Latin morphology, the Backoff Lemmatizer can be described as following a philological method. By this, I mean that the process reflects the reading, decoding, and disambiguating strategies of the modern Latin reader (McCaffrey, 2006). For example, the process echoes the classroom process of Paul Diederich, who describes groups of students reading together and analyzing their text first through a combination of previous knowledge and dictionary lookups, but then "if no member of the group can clear up the difficulty, they resort to a formal analysis of the endings" (Hampel, 2014, 95).

One limitation of the current Backoff Lemmatizer setup is its binary sequential decision making; that is, a token is assigned a lemma based on the first match encountered in the backoff chain. By way of conclusion, I will discuss work in progress on a progressively scored Backoff Lemmatizer, or one that returns the lemma with the highest likelihood found after a token passes through and is assigned a score by every sub-lemmatizer in the chain.

References

Almas, B. (2013). *Morpheus-Wrapper*. https://github.com/PerseusDL/morpheus-wrapper (accessed 21 November 2017).

Bozzi, A., G. Cappelli, M. Passarotti, E. Pulcinelli, and P. Ruffolo. (1992). *LemLat*. http://www.ilc.cnr.it/lemlat/ (accessed 21 November 2017).

Celano, G. G. A., G. Crane, and B. Almas. (2017). *The Ancient Greek and Latin Dependency Treebank*. https://perseusdl.github.io/treebank_data/ (accessed 21 November 2017).

Eger, S., T. vor der Brück, and A. Mehler. (2015). Lexicon-Assisted Tagging and Lemmatization in Latin: A Comparison of Six Taggers and Two Lemmatization Methods, in Proceedings of the 9th SIGHUM Work-

shop on Language Technology for Cultural Heritage, *Social Sciences, and Humanities*: 105–13.
McCaffrey, D. (2006). Reading Latin Efficiently and the Need for Cognitive Strategies, in *When Dead Tongues Speak: Teaching Beginning Greek and Latin*, ed. J. Gruber-Miller. New York: Oxford University Press.
Hampel, R. L. (2014). *Paul Diederich and the Progressive American High School*. Charlotte, NC: Info Age.
Juršič, M., I. Mozetic, T. Erjavec, and N. Lavrac. (2010). LemmaGen: Multilingual Lemmatisation with Induced Ripple-Down Rules. *Journal of Universal Computer Science*: 1190–1214. https://doi.org/10.3217/jucs-016-09-1190.
Johnson, K. P. (2017). *CLTK: The Classical Language Toolkit*. https://github.com/cltk/cltk. (accessed 21 November 2017).
Kestemont, M., and J. De Gussem. (2017). Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages. https://arxiv.org/abs/1603.01597v2.
Loper, E., S. Bird, and T. Tresoldi. (2017). *NLTK 3.2.5 Documentation: nltk.tag.sequential*. http://www.nltk.org/_modules/nltk/tag/sequential.html (accessed 21 November 2017).
Ouvard, Y., and P. Verkerk. (2014). *Collatinus Web*. http://outils.biblissima.fr/en/collatinus-web/index.php (accessed 21 November 2017).
Perkins, J. (2014). *Python 3 Text Processing with NLTK 3 Cookbook*. Birmingham, UK: Packt Publishing.
Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. San Rafael, CA: Morgan & Claypool Publishers
Schmid, H. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*, In Proceedings of the Conference on New Methods in Language Processing, Manchester, UK.
Springmann, U., H. Schmid, and D. Najock. (2016). LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity. *Open Linguistics* 2(1). https://doi.org/10.1515/opli-2016-0019. (accessed 21 November 2017).

Las humanidades digitales y el patrimonio arqueológico maya: resultados preliminares de un esfuerzo interinstitucional de documentación y difusión

Arianna Campiani
acampiani@ucmerced.edu
University of California Merced, United States of America

Rodrigo Liendo
rodrigo@liendo.net
Universidad Nacional Autónoma de México, Mexico

Nicola Lercari
nlercari@ucmerced.edu), University of California Merced, United States of America

El uso de tecnologías digitales para el registro y conservación del patrimonio ha demostrado ser de gran utilidad ya que permite contar con una documentación exacta que puede constituir la base para proyectos de restauración, pero también de investigación y difusión (De Reu et al., 2013; Forte et al., 2015). En las últimas dos décadas, el cambio climático, la creciente inestabilidad política y el saqueo han llevado al deterioro de numerosos sitios arqueológicos mesoamericanos (Juárez Cossío, 2000; Lario Villalta, 2000; Noriega y Quintana, 2002). En este escenario, la documentación del patrimonio digital y la difusión de datos en línea se convierten en recursos invaluables para registrar, monitorear y preservar el patrimonio cultural maya del sur de México. (Forte et al., 2015)

La Coordinación Nacional de Monumentos Históricos del Instituto Nacional de Antropología e Historia (INAH) ha implementado el Laboratorio de Imagen y Análisis Dimensional para integrar un acervo tridimensional del patrimonio arquitectónico, pero, en cuanto al patrimonio arqueológico la documentación digital se ha limitado a edificios específicos de pocos sitios arqueológicos. En la última década, universidades de los Estados Unidos y Canadienses que conducen investigaciones en la península de Yucatán han empleado tecnología LiDAR y otras herramientas digitales para la documentación de sitios arqueológicos, no obstante estas iniciativas raramente contemplan la participación de universidades mexicanas o estudiantes locales (Golden et al., 2016; Hare, 2014; Hutson 2015; Hutson et al., 2016; Magnoni et al., 2016; Reese-Taylor et al., 2016).

En 2018, gracias a una colaboración entre la Universidad Nacional Autónoma de México y la Universidad de la California- Merced hemos empezado los trabajos de levantamiento digital en el sitio arqueológico de Palenque, Chiapas, patrimonio de la UNESCO desde 1980. En paralelo con las actividades de excavación en el Grupo IV, al noroeste del núcleo cívico-ceremonial, hemos empleado un escáner láser terrestre (TLS) y dos drones con cámaras de alta resolución para producir mapas y modelos 3D de los edificios y de sus espacios asociados, con una precisión al centímetro. En un lugar de la importancia de Palenque, donde los edificios necesitan de constante mantenimiento, esta labor nos parece relevante y necesaria.

En cuanto al centro del asentamiento y a los edificios monumentales con ello asociados, los vuelos con drones permiten no solo tener un registro cuidadoso sino complementar el levantamiento hecho manualmente a través de los años. Además, la fotogrametría consiente situar los trabajos de restauración llevados a cabo y reflexionar

sobre la manera en que estos complementan y a la vez modifican la percepción de las construcciones, puestos que dejan a la vista una sobreposición de diferentes etapas constructivas.

En acuerdo con los arqueólogos y conservadores del INAH, se escanearon la Casa E y C del Palacio y el Templo de las Inscripciones con énfasis en la Tumba de Pakal, ya que a corto plazo el Instituto empezará un proyecto de investigación y restauración de dichos edificios. La documentación producida servirá para planear las excavaciones en el Palacio y a la vez constituye la base para el monitoreo de los edificios y de sus decoraciones en piedra y estuco, y para evaluar la eficacia de las técnicas empleadas para su conservación.

A mediano plazo esperamos contar con un dron con cámara LiDAR para hacer prospección más detallada, perfeccionar el mapa de la ciudad y planear las excavaciones de acuerdo a las preguntas de investigación de los diferentes investigadores y estudiantes involucrados.

Estas técnicas digitales de documentación arqueológica y de monitoreo del patrimonio que empezamos a emplear en Palenque han sido adoptadas por el equipo de UC Merced en otros proyectos. Por ejemplo, en el sitio patrimonio mundial de Çatalhöyük, Turquía, el registro se ha complementado de modelos predictivos para la conservación gracias a la comparación de los datos 3D (con el uso del software open source Cloud Compare) y su implementación en una plataforma GIS (ESRI) (Campiani, Lercari y Lingle, 2018).

A parte de contar con el equipo para el mapeo digital, y paralelamente a la documentación, el objetivo de las dos instituciones es formar estudiantes gracias a la experiencia en campo, la organización de talleres y el intercambio de estudiantes y profesores. A través de estas estrategias, los datos recolectados por el equipo interinstitucional pueden ser analizados por todos los usuarios mediante software abiertos. A la fecha se ha empezado con la formación de arqueólogos en la temporada 2018.

A la vez, con el programa Unity, tanto para Çatalhöyük como para el sitio histórico de Bodie, California, en UC Merced se han desarrollado tres apps con fines diferentes: una para la simulación de las excavaciones y la interpretación de la estratigrafía (Lercari et al., 2017), una para los restauradores para la comparación de los elementos arquitectónicos y su estado de conservación (Lingle y Seifert, 2017) y otra app para guiar al público en el parque de Bodie (Lercari et. al, 2018). Los códigos generados constituyen la base para los trabajos a implementar en Palenque en cuanto a estudio y difusión.

Con fundamento en estas premisas pensamos que nuestra colaboración interinstitucional pueda sentar las bases metodológicas para el estudio y monitoreo del patrimonio arqueológico maya, gracias a la participación interdisciplinaria, el intercambio y formación de estudiantes y profesores, el desarrollo de nuevos métodos para el estudio arqueológico, la conservación y la difusión.

En esta ponencia breve queremos presentar los resultados de la primera temporada de campo con el empleo de estas tecnologías y reflexionar sobre objetivos a futuro y buenas prácticas en cuanto a documentación, difusión y divulgación de conocimiento para un público especializado y el público en general, para que el uso de la tecnología digital aplicada a la documentación del patrimonio arqueológico maya se vuelva un puente entre investigación y sociedad.

Referencias

Campiani, A., Lercari, N. y Lingle A. (2018). Analytical models for at-risk heritage conservation and 3D GIS. *Society for American Archaeology Conference: Abstracts of the 83rd annual meeting*. Washington DC, p.83, http://www.saa.org/Portals/0/SAA/MEETINGS/2018%20Abstracts/Individual%20Level%20Abstracts_C_D_2018.pdf (consultado el 1 de mayo de 2018)

De Reu, J., Plets, G., Verhoeven, G., De Smedt, P., Bats, M., Cherretté, B. y De Maeyer, W. (2013). Towards a Three-Dimensional Cost-Effective Registration of the Archaeological Heritage. *Journal of Archaeological Science*, 40 (2): 1108–21.

Forte, M., Dell'Unto, N., K. Jonsson K. y Lercari, N. (2015). Interpretation process at Çatalhöyük using 3D. In Hodder I. y Marciniak, A. (eds), *Assembling* Çatalhöyük. Maney Publishing, pp. 43-57.

Golden, C., Murtha, T., Cook, B., Shaffer, D.S., Schroder, W., J. Hermitt, E., Alcover Firpi, O. y Scherer, A. K. (2016). Reanalyzing environmental lidar data for archaeology: Mesoamerican applications and implications. *Journal of Archaeological Science: Reports*, 9: 293-308.

Hare, T., Masson, M. y Russell, B. (2014). High-density LiDAR mapping of the ancient city of Mayapan. *Remote Sensing* 6 (9): 9064–85.

Hutson, S. R., Kidder, B., Lamb, C., Vallejo-Cáliz, D. y Welch, J. (2016). Small Buildings and Small Budgets. Making Lidar Work in Northern Yucatan, Mexico. *Advances in Archaeological Practice* 4(3): 268-83.

Hutson, S. (2015). Adapting LiDAR data for regional variation in the tropics: A case study from the Northern Maya Lowlands. *Journal of Archaeological Science: Reports*, 4: 252–63.

Juárez Cossío, D. (2000). El Proyecto Yaxchilán y las alternativas de conservación en la década de los setenta. *XXII Simposio de Investigaciones Arqueológicas en Guatemala: Sitios arqueológicos en el área Maya: un reto para la conservación*. The Getty Conservation Institute, pp. 27-37.

Lario Villalta, C.R. (2000). El reto de conservación Tikal, Guatemala. *XXII Simposio de Investigaciones Arqueológicas en Guatemala: Sitios arqueológicos en el área Maya: un reto para la conservación*. The Getty Conservation Institute, pp. 59-69.

Lercari, N., Jaffke, D., Aboulhosn, J., Baird, G. y Guillem, A. (2018). Citizen Science Archaeology at Bodie

State Historic Park. *Society for American Archaeology Conference: Abstracts of the 83rd annual meeting*. Washington DC, p. 283, http://www.saa.org/Portals/0/SAA/MEETINGS/2018%20Abstracts/Individual%20Level%20Abstracts_I_L_2018.pdf (consultado el 1 de mayo de 2018)

Lercari, N., Shiferaw, E., Forte M. y Kopper R. (2017). Immersive Visualization and Curation of Archaeological Heritage Data: Çatalhöyük and the Dig@IT App. Journal of Archaeologial Method and Theory: 1-25.

Lercari, N., Lingle, A. y Umurhan O. (2016). Catalhöyük Digital Preservation Project. Çatalhöyük *2016 Archive Report*. http://www.catalhoyuk.com/sites/default/files/media/pdf/Archive_Report_2016.pdf (consultado el 2 de febrero de 2017).

Lingle, A. y Seifert, J. (2017). Update on the Çatalhöyük Digital Preservation Project. Çatalhöyük *2017 Archive Report*. http://www.catalhoyuk.com/sites/default/files/Archive_Report_2017.pdf (acceso 1 Mayo 2018)

Magnoni, A., Stanton T., Barth, N., Fernandez-Diaz, J. C., Osorio León, J. F., Pérez Ruíz, F. y Wheeler, J. A. (2016). Detection Thresholds of Archaeological Features in Airborne Lidar Data from Central Yucatán. *Advances in Archaeological Practice* 4(3): 232-248.

Noriega, R. y Quintana, O. (2002). Programa de restauración: Proyecto Protección de Sitios Arqueológicos en Petén. In Laporte, J.P., Escobedo, H. y Arroyo B. (eds), *XV Simposio de Investigaciones Arqueológicas en Guatemala, 2001*. Museo Nacional de Arqueología y Etnología, pp. 228-238

Reese-Taylor, C., Anaya Hernández, A., Flores Esquivel, F. C. A., Monteleone, K., Uriarte, A., Carr, C., Geovannini Acuña, H., Fernandez-Diaz, J. C., Peuramaki-Brown M. y Dunning, N. (2016). Boots on the Ground at Yaxnohcah: Ground-Truthing Lidar in a Complex Tropical Landscape. *Advances in Archaeological Practice* 4(3): 314-338.

Cartonera Publishers Database, documenting grassroots publishing initiatives

Paloma Celis Carbajal
pceliscarbaj@wisc.edu
University of Wisconsin-Madison, United States of America

Starting in Buenos Aires with Eloísa Cartonera in 2003, Cartonera publishers emerged as a reaction to the over commercialization of the book industry and its ever-growing conglomerates. With their unique hand embellished covers and their peculiar aesthetics, these publishers have challenged how books and literature are produced and distributed. Their collective manual process is equal to the intellectual one, resulting in a more democratic mode of production.

For thirteen years, the Cartonera Publishers Database has been documenting and preserving the diverse initiatives that stem from these grassroots projects which use recycled cardboard as book covers. The database is comprised of more than 1,200 entries which include Dublin Core metadata, scanned images of the back and front covers, copyright pages, and title pages, and audio files of interviews of several members of Cartonera publishing houses. An electronic crosswalk connects these entries to local cataloging of the Cartonera Book Collection. The audio files and an online full-text book "Akademia Cartonera: A primer of Latin American Cartonera Publishers" are additionally indexed and marked using TEI. This database is the only digital reference tool on these multi-pronged publishing initiatives. The ultimate goal is to connect this locally focused digital humanities project with cartonera books held at other institutions around the world in an interinstitutional Cartonera Catalog.

In the past year, I have been studying the possibility of using crowd sourcing and/or folksonomies to supplement the current content with the goal of providing a deeper understanding of the variety of contexts in which these books are created while also offering a space for the Cartonera publishers to contribute other content created directly by them. My proposed papers addresses the database and initial efforts to expand our work.

References

Cartonera Publishers Database, http://digital.library.wisc.edu/1711.dl/Arts.EloisaCart (accessed 20 November 2017).

Integrating Latent Dirichlet Allocation and Poisson Graphical Model: A Deep Dive into the Writings of Chen Duxiu, Co-Founder of the Chinese Communist Party

Anne Shen Chao
mrsannechao@gmail.com
Rice University, United States of America

Qiwei Li
liqiwei2000@gmail.com
University of Texas Southwestern Medical Center, United States of America

Zhandong Liu
zhandonl@bcm.edu
Baylor College of Medicine, United States of America

Chen Duxiu (1879-1942) co-founded the Chinese Communist Party in 1920, and served as its secretary general

from 1921 to 1927. He was a prolific author and a cultural rebel whose writings transformed the intellectual and social landscape of 20th century China. Yet from 1904 to about 1919, Chen advocated Western democracy and Social Darwinism as solutions to save China. His turn to communism was an abrupt transition, and many historians credited this to the influence of his colleague, and co-founder of the CCP, Li Dazhao (1888-1927). Both Li and Chen had studied in Japan, and through their interaction with Japanese Socialists and fellow students, became acquainted with literature on socialism and anarchism. Some say that Li was the theoretician who understood Bolshevism and Marxism in depth, while Chen did not become well-versed in Marxism until he founded the CCP (Yoshihiro, 2013).

In this paper, we applied topic modeling (Blei et al., 2012) to a select number of Chen's and Li's published articles, in an attempt to detect the difference, if any, in their interpretation on the subject of socialism, Marxism, communism and Bolshevism. We integrated two well-developed statistical methodologies, the Latent Dirichlet Allocation (LDA) and the Poisson Graphical Model (PGM), to probe in finer detail the broad themes in the 892 pieces of Chen's essays, correspondences, and occasional poetry, comprising a total of 1,347,699 Chinese characters. Based on the word counts per topic, we then implemented the PGM method to study the association among different topics. The use of PGM minimizes any misleading inference caused by confounding variables, and it also leads to a more concise structure of the network of topics.

Specifically, we chose 263 articles written by Chen Duxiu and 53 written by Li Dazhao, containing words related to Marxism, socialism, Bolshevism, and communism (Ren, 2018; Li ,1984). (Both selections covered the length of the men's publishing career; Chen passed way at age 63, while Li was executed at age 39). A document-term matrix (bag-of-words data) was generated from the preprocessed text. Next, we carefully selected a set of seed words for each of *K* topics of interest. We then applied the topic modeling method LDA to the bag-of-words data to find the remaining mixtures of words associated with each topic. Consequently, we could interpret each estimated topic by abstracting the top ranking terms within that topic. We then generated a new document-topic matrix from the document-term matrix by calculating the counts of those top words from the same topic. Finally, we applied the Poisson Graphical Model to the document-topic matrix to infer the conditional independence between each pair of topics. The resulting graph is a network visualization where each node represents a topic, and each edge indicates the conditional dependencies among the topics, meaning the two topics that are linked by an edge are correlated even after adjusting for all the other topics in the corpus.

The results yield several initial observations: Chen used a smaller set of vocabulary words over and over again to emphasize a point, while Li adopted a more discursive style with fewer repeats of the same word. Chen used many more verbs (such as: "agitate," "struggle," "unite," "lead," "develop," "carry out"), thereby exhorting his readers to action, while Li tended to use descriptive words. Chen focused on the present by analyzing different political groups: "Guomindang," "warlords," "proletariat," "bourgeoisie," "military," "students," "masses" and "imperialists." Li painted a larger scenario by using words such as "world," "humanity," "philosophy," "phenomenon," "relationship," "history" and "religion." The general conclusion at this early stage of analysis is that Chen urged his readers to put into action his plans to bring China under communism, while Li tended to explain to his readers the nature of Bolshevism and Marxism.

More interestingly, these calculations yielded "orbits" of vocabulary for each man's important ideas. For instance, Chen's use of the word "revolution" appeared three times in the 8 topics that we studied. In the first sub-topic, "revolution" appeared with words such as "class," "bourgeoisie," "proletariat," "develop," "strength," and "movement." In the second sub-topic, "revolution" again appeared alongside "peasants," "bourgeoisie," "proletariat," "lead," "China," "masses," "movement," and "action." In the third sub-topic, "revolution" appeared with "bourgeoisie," "proletariat," "struggle," "China," "Guomindang," "movement." Li, when he discussed "revolution," which appeared twice in the four topics we studied, he often used words such as "people," "Russia," "movement," "government," "masses," "future," and "China." While the general trend of these two men's writing is clear by a casual browsing of all of these articles, but this method of calculation demonstrates in a quantitative manner the qualitative interdependence of topics, and diagrams in an easy to read manner the network configuration of the vocabulary of each man.

References

Blei, David M. (2012) *Probabilistic topic models: Communications of the ACM*, 55(4): 77-84.

Li D. (1984). *Li Dazhao Wenji* [A literary collection of Li Dazhao]. Beijing: Renmin chubanshe.

Ren J. ed. (2008), *Chen Duxiu zhuzuo xuanbian* [A selected collection of Chen Duxiu's writing]. Shanghai: Shanghai renmin chubanshe. 6 vols.

Yoshihiro, I., tr. by Fogel J. (2013) *Formation of the Chinese Communist Party.* New York: Columbia University Press.

Sensory Ethnography and Storytelling with the Sounds of Voices: Methods, Ethics and Accessibility

Kelsey Marie Chatlosh
kchatlosh@gradcenter.cuny.edu
The Graduate Center, CUNY, United States of America

In contemporary anthropology, nearly all of us work with sound – usually oral interviews – but its quality as such is often taken for granted. Audio files of interviews are often quickly transcribed or qualitatively coded into text, then analyzed and written into books. And the soundscapes of our fieldwork sites are often taken for granted as well. Their meanings and textures as sounds are thus erased. The small sounds of voices and places often invoke an intimacy that anthropologists may attempt to render in text through "thick description" (Geertz 1973) and hopefully also "sincerity" (Jackson 2005), drawing from hermeneutic and poetic approaches in literary studies (Clifford and Marcus 1986, Behar and Gordon 1995).

Meanwhile, a growing body of interdisciplinary scholarship on sound studies foregrounds sound as "a modality of knowing and being in the world," of creating a sense of place or a narrative (Feld 2000). Performance studies scholars have also provided many contributions towards thinking about "the hegemony of textuality" (Conquergood 2002: 147) and, conversely, the "repertoire" of manifestations of knowledge and memory that exist outside the written, institutionalized archive (Taylor 2003). Needless to say, ontologies, storytelling, memories, and place- and identity- making are canonical topics of study in anthropology. As Steven Feld, an anthropologist and one of the leading theorists of sound studies, has discussed, there are many possibilities in "doing ethnography through sound—listening, recording, editing, and representation" that will hopefully one day be more than just "mostly about words" (Feld and Brenneis 2004: 461, 471). Further, as anthropologist and sound studies theorist, Roshanak Kheshti, has argued: "considering sound through the critical genealogy of feminist or race theory forces you to move beyond sound as an object and think of sound instead as an analytic or a hermeneutical tool for understanding inequality…" and the "social worlds" that scholars study (Brooks and Kheshti 2011: 330).

I am interested in approaches to methods, ethics and accessibility when working with the sounds of voices that cross-cut anthropology – specifically sensory ethnography, or ethnographic methods that foreground the senses – sound studies (and sound arts), and digital humanities. Anthropologists are not that common in the realm of digital humanities. However, many of us, one could argue, do projects that could be construed as "digital humanities," that is: "digital methods of research that engage humanities topics in their materials and/or interpret the results of digital tools with a humanities lens" (Lexicon of DH Workshop, The Graduate Center Digital Initiatives, tinyurl.com/lexicondh). And the thing with DH is, once we (scholars) start paying more acute attention to the ways in which our research is digital this can open up new questions and also new methods for doing what it is that we do, in terms of both research and pedagogy. This is particularly true, I suggest, for sound studies – given the importance of digital tools and platforms for recording, mixing, sharing and listening to audio.

Yet, new methods, digital tools and projects emerging through DH and internet research in general open up an array of rather new ethical and accessibility concerns (see e.g. Barnes 2006, Markham and Buchanon 2012). What constitutes personhood or "human subjects" on the internet? What data is or should be "public"? When should consent protocols be required? Can images or audio files of people and their voices bely anonymity? Who has access to make digital projects or to engage them, particularly in relation to differences of class, ability, and language fluency? How is the internet – its structure, its users, its algorithms – racialized and gendered (e.g. McPherson 2012, Noble 2018)? In what ways may some DH projects follow a practice of extraction without reciprocity? Indeed, anthropologists wrestle a lot with that last question in particular when extracting stories of individuals that then advance our careers, while many DH-ers may be, e.g., web scraping.

This short paper presentation will examine the possibilities of cross-cutting methodological approaches to anthropology, sound studies and arts, and digital humanities, specifically when recording and sharing the sounds of peoples' as a mode of storytelling. I will focus on oral interviews in particular. Driven by the aforementioned anthropological and interdisciplinary concerns, this paper will discuss the interplays of method and theory when cross-cutting these approaches, and issues of ethics and accessibility when recording and sharing sound. This includes being wary of institutional compliance with Institutional Review Boards but also following a feminist ethics beyond compliance, that, for example, foregrounds consent as not a one-time signature but reiterated, negotiated and subject to change (see Davis and Craven 2016). I will also consider various levels of intrusiveness and impact that the recording and sharing of the sounds – especially the sounds of peoples' voices – may have, and the potential roles of shared sounds within larger networks of listeners and what their availability may foreclose (e.g. Sugarman 1997, Brooks and Kheshti 2011, Kunreuther 2014, Kheshti 2015). Lastly, I will discuss digital modalities for sharing research with sound and their (limited) possibilities for storytelling, specifically for doing and sharing anthropological and other research in a more accessible form – with the exceptions structured by access to technology, limited hearing ability and translatability across languages and contexts. I will highlight free and open-source resources, such as sound archives and editing and hosting technologies, as well as low-cost Do-It-Yourself (DIY) microphones and speakers.

While websites are often great platforms for sharing oral history projects and other sounds, I will also discuss examples of other modalities for sharing sounds, such as exhibits and events, as well as digital platforms for scholarly publishing (e.g. Manifold). I will include a brief survey of various free online platforms that seem to have high potential for use in scholarship and pedagogy. These in-

clude: the SoundCloud online streaming platform, the Oral History Metadata Synchronizer in coordination with Omeka, podcasting via iTunes, StoryMaps for sharing audio on a map, and Chirbit for sharing audio on social media or embedding audio on a website. I will also discuss examples for the in-person sharing of sounds during, for example, an exhibit or class, including a brief survey of different kinds of speakers and headphones and different ways of transferring pre-recorded or live sounds to them, as well as spatial considerations for sharing sound. For example, placing numerous speakers inside an enclosed space, such as a tent, may allow for a focused listening space that is still shared and not as individuated as when using headphones (an idea I learned from sound artist Grant Smith of Reveil Radio in London). While I do not plan to conduct a full comparative analysis of these platforms, I will briefly discuss what I find to be some of openings and limitations of each.

In sum, this presentation aims to bridge together a number of themes: sound studies, oral histories, ethics, accessibility, and modalities for sharing sounds. I emphasize the intention that motivates my attempt to bridge these various themes: In my opinion, when recording and sharing human voices the researcher must always be vigilant in their ethical considerations (beyond IRB approval) at every step of the research design and practice, and then the sharing of these sounds is what makes their collection most worthwhile and to do so requires considerations of accessibility and modalities and ethics for such sharing.

References

Barnes, S. (2006). "A Privacy Paradox: Social Networking in the United States," *First Monday* 11(9). http://firstmonday.org/article/view/1394/1312 (accessed 26 April 2018).

Behar, R. and Gordon, D. A. (1995). *Women Writing Culture*. Berkeley, CA: University of California Press.

Brooks, D. and Kheshti, R. (2011). The Social Space of Sound, Theatre Survey 52: 329-334.

Clifford, J. and Marcus, G. ed.s. (1986). *Writing Culture: The Poetics and Politics of Ethnography*. Berkeley: University of California Press.

Conquergood, D. (2002). Performance Studies: Interventions and Radical Research, *The Drama Review* 46: 145-156.

Davis, D.-A. and Craven, C. (2016). *Feminist Ethnography: Thinking through Methodologies, Challenges, and Possibilities*. Lanham, MD: Rowman and Littlefield.

Feld, S. (2000). Sound Worlds. In Sound, Kruth, P. and Stobart, H. (eds). Cambridge, England: Cambridge University Press.

Feld, S. and Brenneis, D. (2004). Doing Anthropology in Sound, *American Ethnologist* 31(4): 461-474.

Geertz, C. (1973). *The Interpretation of Cultures*. New York: Basic Books.

Jackson, J. Jr. (2005). *Real Black: Adventures in Racial Sincerity*. Chicago, IL: Chicago University Press.

Kheshti, R. (2015). *Modernity's Ear: Listening to Race and Gender in World Music*. New York: New York University Press.

Kunreuther, L. (2014). *Voicing Subjects: Public Intimacy and Mediation in Kathmandu*. Berkeley, CA: University of California Press.

Markham, A. and Buchanan, E. (2012). Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). Association of Internet Researchers (AoIR). http://aoir.org/reports/ethics2.pdf (accessed 26 April 2018).

McPherson, T. (2012). Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation. In *Debates in the Digital Humanities*, Gold, M. (ed). Minneapolis, MN: University of Minnesota Press with Manifold. http://dhdebates.gc.cuny.edu/debates/text/29 (accessed 26 April 2018).

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Sugarman, J. (1997). *Engendering Song: Singing and Subjectivity at Prespa Albanian Weddings*. Chicago, IL: University of Chicago Press.

Taylor, D. (2003). *The Archive and the Repertoire*. Durham, NC: Duke University Press

Seinfeld at The Nexus of the Universe: Using IMDb Data and Social Network Theory to Create a Digital Humanities Project

Cindy Conaway
cindy.conaway@esc.edu
SUNY Empire State College, United States of America

Diane Shichtman
diane.shichtman@esc.edu
SUNY Empire State College, United States of America

This Digital Humanities project is an interdisciplinary project effort that uses the lens of, and data from, the U.S. TV show*Seinfeld*to explore questions about television and other media.*Seinfeld* has significant cultural influence over other media, but what is its **reach**, meaning the many other media items cast and crew worked on, also known as the **overlap**? We are starting with data from the Internet Movie Database (IMDb). This makes this project somewhat different from other Digital Humanities projects as we're using an existing database rather than primary sources. An associate professor of media studies, accustomed to conducting critical analysis of television shows, and an associate professor of information systems, more used to working with non-media studies data, are wor-

king to populate a relational database, to use quantitative analysis, and a social science theory--social network theory, particularly "Small Worlds" theory--to explain trends in media industries, including questions of genre, gender, race, and age in entertainment businesses.

Seinfeld (NBC 1989-1998) was a US-based half-hour, multi-camera, situation comedy, one of several that featured stand-up comics in stories similar to their own lives. Although it ended nearly 20 years ago, it heavily influences TV shows of today, including "hangout" sitcoms, one-camera comedies featuring conversation and digression, and antihero dramas. Journalist Jennifer Keishen Armstrong writes in the bestselling *Seinfeldia* that the show "snuck through the network system to become a hit that changed TV's most cherished rules; from then on, antiheroes would rise to prominence, unique voices would invade the airwaves, and the creative forces behind shows would often gain as much power and fame as the faces in front of the cameras" (Armstrong, 2016). It's a singularly important show for a variety of reasons.

Clearly, *Seinfeld* hassignificant cultural impact on other shows and movies, but what we wanted to know is, what is its 'reach'? Reach is defined as other media that texts cast and crew from *Seinfeld* worked on before, during, and after their appearance(s) on the show. Such texts exist in every media type (movies, video games, web-based media). When two media items share cast/crew, we look for overlap.

Dr. Conaway worked on the project for two years, using cut and paste and Excel spreadsheets for items and people, before involving Dr. Shichtman, who has created a relational database that may be searched. We first used MySql and an Amazon Web Services server, have recently shifted to the college's virtual machine and the Oracle database management system. We involved two students in a grant funded practicum in the Fall term as well.

Our research revealed that the 1551 cast/crewhad worked on over 32,500 other discrete media texts, starting in 1936, and with many texts still on the air today, often with an overlap of more than one. Nearly every televison series, TV movie, and TV special we could think of included overlap. Only recently, in "peak TV"—in which there are over 500 scripted TV shows in production this year alone, in addition to reality, sports, and news shows (many of which also have overlap)—are we seeing well-known US TV series with no overlap. Our research found that although most were US-based, there were media items from over 60 countries.

Social network theory would help us answer some questions. As Duncan Watts writes in *Six Degrees: the Science of a Connected Age*, „Affiliation networks . . . are . . . networks of overlapping cliques, locked together via the comembership of individuals in multiple groups" (Watts, 2004). Small worlds theory discusses how networks of people influence each other, and each others' connections.

Questions include, what genres did the cast/crew, presumably chosen for a common comic sensibility, work on other than comedy? What genres included the most cast/crew? What genres have less overlap, none at all, and what might be some reasons for that? What is the importance of gender, race, and age?

We looked for other, similar projects that used IMDb and found that there were few that did. Some computer scientists had used IMDb to trace the overlaps among actors involved in 'adult' films in the database as an example of a 'small world' environment. Media History scholars had traced 'race films' that ended before our database started, and Digital Humanities scholars used it to look at patterns of exhibition of films or specifically how Australians worked together, but not to examine how cast circulated among media.

IMDb, it turns, out, is a challenging tool for this purpose. Deb Verhoeven, Associate Dean of Engagement and Innovation of the University of Technology Sidney, who has done a lot of Digital Humanities work on Australian films explained in 2012 that IMDb consists of "elaborated sets of lists" created by fans, writing:

Accordingly, the primary users of filmographic catalogues are not cinema historians, information managers, analytical filmographers, or cinema scholars, but members of the public, film buffs, students and so on who are content to navigate these databases using the small number of structured search fields provided. (Verhoeven, 2012)

IMDb, which started in the early 1990s, is very robust, and provides information for free download using Python, but is not usable 'as is.' Entries may be misleading, incomplete, or unclear, with genres in particular organized in unhelpful ways. The Downloadable information includes the full cast and some types of crew members, but not others. In addition, the fields of the two faculty members made shared vocabulary difficult, and getting complete and clean data that could be turned into tables and graphs meant conducting additional research outside of IMDb, and reorganizing the data significantly from the way Dr. Conaway initially tagged it. SUNY Empire State College also lacks the structures that many institutions have for conducting Digital Humanities work.

However, we have been able to create some early data visualizations that will show a microcosm of how the US entertainment industry works for various types of actors and crew members, using specifically the data from television programs. We've compared *Seinfeld*'s numbers of actors and crew to that of other shows, analyzed how the media items break down by genre, and visualized how women's careers wax and wane in different patterns from men's careers. In the future we will do the same for subgenres, actors of color, and actors of various age groups.

References

Armstrong, J.K., 2016. *Seinfeldia: How a Show about Nothing Changed Everything*. Simon and Schuster.

Bajak, A. 2017. Seinfeld, big data and measuring the Internet's emotional landscape. *Mediashift.*

Gold, M.K. 2012. *Debates in the Digital Humanities*. University of Minnesota Press.

Gold, M.K. and Klein, L.F., 2014. *Debates in the Digital Humanities*. University of Minnesota Press.

Lavery, D. And Dunne, S.L. 2006. *Seinfeld, Master Of its Domain*. New York: Continuum.

Verhoeven, D. New cinema history and the computational turn. Beyond Art, Beyond Humanities, Beyond Technology: A New Creativity, Proceedings Of the World Congress Of Communication and the Arts Conference, University Of Minho, Portugal. 2012

Watts, D.J. 2004. *Six Degrees: The Science Of a Connected Age*. WW Norton & Company.

Exploring Big and Boutique Data through Laboring-Class Poets Online

Cole Daniel Crawford
cole_crawford@fas.harvard.edu
Harvard University, United States of America

Though quantitative methods are becoming increasingly common within the humanities, few researchers readily describe their primary texts as data. Most prefer to see their objects of study as contextually situated and socially constructed entities with independent value that resist complete digital representation. Miriam Posner argues that for many humanities researchers, describing an artifact as data implies "that it exists in discrete, fungible units; that it is computationally tractable; that its meaningful qualities can be enumerated in a finite list; and that someone else performing the same operations on the same data will come up with the same results." Defined this way, digital artifacts and metadata seem to simultaneously insist on particular interpretations and to be bereft of deeper meaning outside of an aggregate state, thereby resisting the hermeneutic methodologies which form the core of humanistic inquiry.

This position stems from understanding data primarily through a big data mindset. As corporations, governments, and universities have increasingly addressed business problems by embracing data analytics, the essential qualities of big data (large volume, high velocity, and heterogenous variety) have created the illusion among many that such datasets can perfectly model an imperfect and unpredictable world, gaining credibility simply by increasing in volume. The computational authority of big data is persuasive because it presents a seemingly objective, number-driven way of knowing reality – an epistemology of the database, predicated on scale, comprehensiveness, and reproducibility.

While an immense and complete archive possesses an undeniable allure (Manovich, 2012; Kaplan, 2015), there is still value in examining individual records and investigating the intangible stories and datapoints that hide in database gaps or reside outside of databases entirely. I use Cheryl Ball et al's term "boutique data" to emphasize the ongoing importance of small, localized, partial, and qualitative datasets to the humanities research process. I frame boutique data as both a thing (a boutique dataset) and a theoretical approach to data-intensive work in the humanities. While big data are often automatically generated, boutique data are manually curated – subjective, created *capta* as opposed to given data (Drucker, 2011). Big data hides the work and decisions that drive data processing, while boutique data foregrounds the hidden labor and assumptions that shape data. Big data fits information into a predetermined mold, while boutique data models are built from the bottom up. Where a big data mindset treats gaps in data coverage as a corrupting null to be fixed, a boutique approach to data sees these gaps not as empty voids but as evocative absences worth further investigation. In this presentation, I will examine both the successes and failures of a boutique approach to data through a case study of *Laboring-Class Poets Online* and speculate about possible future improvements to the project.

The texts and histories studied by scholars of laboring-class culture are riddled with gaps. Since the publication of E. P. Thompson's *The Making of the English Working Class* over fifty years ago, researchers have increasingly viewed laboring-class poets and their writing as subjects worthy of scholarly inquiry. Rather than portraying proletarian writers as isolated anomalies or novelties, such as how George Thomson characterized Robert Burns as a "heav'n taught ploughman" in his famous obituary for the Scottish bard, modern critics acknowledge that working-class writing was a significant, widespread phenomenon. However, while some British laboring-class poets such as Burns or John Clare have achieved near-canonical status, most of these writers are still obscure figures. Information on their lives and access to their writing remains scarce and scattered, hindering research on both their personal histories and their poetry.

Laboring-Class Poets Online (*LCPO*) addresses this gap by aggregating biographical and bibliographical information about the more than 2,000 British laboring-class poets who published between 1700 and 1900 and the texts they produced. *LCPO* draws on collaborative research initially collected by an international distributed team of researchers over several decades and presented as biographical entries in *A Database of British and Irish Labouring-Class Poets and Poetry*. *LCPO* transforms these freeform biographical snippets into structured, web-accessible records. This structure facilitates a pro-

sopographic approach to British working-class literary studies. Lawrence Stone defines prosopography as "the investigation of the common background characteristics of a group of actors in history by means of a collective study of their lives." This methodological shift from the study of individual biographies to collective biographical and bibliographic patterns enables a more comprehensive understanding of laboring-class literary production at a time of great social and economic change. Users can ask questions about laboring-class literature holistically and map trends and themes, including the impact of industrialization; the role of religion as a vehicle for literacy and a source of aesthetic influence; the tension between increased urbanization and a celebration of regional identity, often demonstrated through writing in dialect; the transformation of the publishing industry and the role of patronage and subscription publishing; the growth of literary miscellanies and magazine publishing; and the influence of organized labor movements (e.g., Chartism or Christian Socialism) on laboring-class artistic expression. Scholars can investigate emigration patterns, education level, labor engagement, health outcomes, poet occupations, and interactions with the criminal justice and social relief systems. Publications can similarly be filtered and searched by typical facets such as publication date, author, or location, but also by subscription lists, patronage, cost, or print run size.

Users can interact with aggregate data through numerous data visualizations including geographic maps that show poet and publication locations; timelines of individual lives or major events which shaped the working classes; and network graphs that display connections between writers based on correspondence, personal relationships, or literary influence. Each of these visual forms encourages users to shuttle back and forth between individual records and aggregate analysis. Users can also create collections of content for further interpretation and analysis, correct mistakes in poet entries, or contribute new data to the website. All data presented through *Laboring-Class Poets Online* are freely available for download or access via a REST API.

This information is vital for scholars of working-class writing and culture, but it is also an instance of boutique humanities data (capta): a collaboratively and manually created and curated small dataset of several thousand entities extracted during ongoing research. While scholars often use context to interpret data points in historical documents, databases and computational methods typically lack this capability. Uncertainty is embedded in historical sources, but databases often strip away ambiguity to perform the computational functions that make their use worthwhile. By taking a boutique approach to historical and literary information, *LCPO* retains much of this ambiguity and offers insight into how humanities researchers can accommodate a complex understanding of space and time as continuously unfolding events.

References

Ball, C., Graban, T. S. and Sidler, M. (Forthcoming). The Boutique is Open: Data for Writing Studies. In Rice, J. and McNely, B. (eds), *Networked Humanities: Within and Without the University*. Parlor Press.

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, 5(1) http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Kaplan, F. (2015). A Map for Big Data Research in Digital Humanities. *Frontiers in Digital Humanities*, 2 doi:10.3389/fdigh.2015.00001. http://journal.frontiersin.org/article/10.3389/fdigh.2015.00001/abstract

Goodridge, J. (ed). (2017) *A Database of British and Irish Labouring-Class Poets and Poetry, 1700-1900.*

Manovich, L. (2012). Trending: The Promises and Challenges of Big Social Data. In Gold, M. (ed), *Debates in the Digital Humanities*. University of Minnesota Press.

Posner, M. (2015). Humanities Data: A Necessary Contradiction *Miriam Posner's Blog* http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/

Stone, L. (1972). Prosopography. In Gilbert, F. and Graubard, S. (eds), *Historical Studies Today*. New York.

Organizing communities of practice for shared standards for 3D data preservation

Lynn Cunningham
lynncunningham@berkeley.edu
University of California Berkeley, United States of America

Hannah Scates-Kettler
hannah-s-kettler@uiowa.edu
University of Iowa, United States of America

Scholars are producing and using 3D content more than ever due the advancement and availability of 3D technology. How is this 3D content and its metadata being captured, disseminated, and preserved? How is this digital scholarship being made available and discoverable for pedagogical and research purposes?

Although there is great interest in 3D applications in research, there is currently little available guidance regarding the preservation of digital objects and associated information in perpetuity. The preservation and sharing of research data is a necessary, invaluable responsibility of libraries, museums, and other cultural heritage institutions, and although standards and best practices have been developed for many kinds of digital data to ensure assets can be accessed and reused in perpetuity, the applicability of these standards to 3D data is limited.

Building off the discoveries made during the 2015/2016 NEH Advanced Challenges in Theory and Practice in 3D Modeling of Cultural Heritage Sites, this paper explores one of the main threads of discussion throughout the NEH Summer Institute: research longevity and publication. Underpinning the issue was concerns of the preservation of 3D data and their overall discoverability and (re)use beyond their creation.

This paper investigates the current state of existing standards and schemas for 3D data and explores what more needs to be done (and is being done) by practitioners, librarians and curators to ensure that this digital content is preserved and disseminated, enabling further humanistic inquiry and advancing scholarship of our shared cultural heritage.

In 2017 the Institute for Museum and Library Services received several proposals regarding the advancement of 3D research and support. Two of these grants were funded which are working in tandem to discuss issues related to 3D and virtual reality, and preservation and best practices for 3D data curation. This paper will focus on the developments regarding the latter IMLS grant - the Community Standards for 3D Data Preservation (CS3DP). According to the CS3DP grant proposal (Moore et.al., 2017):

> The project team surveyed an international community including individuals involved in digital curation and 3D data acquisition and research, primarily at universities and museums. Of 104 respondents 70% said that they did not use best practices or standards for preservation, documentation, and dissemination of 3D data. Of those not using standards/best practices, 69% said that they did not use them because they were unaware of such standards.

In order to respond to the lack of consensus around 3D data standards, the grant team will develop "a community-developed plan to move 3D preservation forward [and] recommendations for standards and best practices" for data creators and preservation specialists alike (Moore et. al., 2017). By the time of the 2018 DH conference, the CS3DP grant will have convened around 70 data creators and professionals to address the issues of 3D data preservation. This paper will report on initial findings and ongoing discussions and areas of work, as well as solicit feedback from the DH conference goers about other areas of concern, development and needs.

References

3D-ICONS Guidelines and Case Studies. https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/3D-ICONS/Deliverables/3D-ICONS%20Guidelines%20and%20Case%20Studies.pdf (accessed 27 April 2018).

Advanced Challenges in Theory and Practice in 3D Modeling of Cultural Heritage Sites. https://advanced-challenges.com/ (accessed 27 April 2018).

Alliez, P., Bergerot, L., Bernard, J.-F., Boust, C., Bruseker, G., Carboni, N., Chayani, M., et al. (2017). *Digital 3D Objects in Art and Humanities: Challenges of Creation, Interoperability and Preservation. White Paper : A Result of the PARTHENOS Workshop Held in Bordeaux at Maison Des Sciences de l'Homme d'Aquitaine and at Archeovision Lab. (France), November 30th - December 2nd, 2016.*

Cook, M., Hall, N., Laherty, J. (2017). Developing Library Strategy for 3D and Virtual Reality Collection Development and Reuse. IMLS grant proposal: https://www.imls.gov/sites/default/files/grants/lg-73-17-0141-17/proposals/lg-73-17-0141-17-full-proposal-documents.pdf (accessed 27 April 2018).

D'Andrea, A. and Fernie, K., Addison, A. c., De Luca, L., Guidi, G. and Pescarin, S.(2013). CARARE 2.0: A metadata schema for 3D cultural objects. *2013 Digital Heritage International Congress (DigitalHeritage)*, vol. 2. pp. 137–43 doi:10.1109/DigitalHeritage.2013.6744745.

Moore, J., Rountrey, A., Scates Kettler, H. (2017). Community Standards for 3D Data Preservation (CS3DP). IMLS grant proposal: https://www.imls.gov/sites/default/files/grants/lg-88-17-0171-17/proposals/lg-88-17-0171-17-full-proposal-documents.pdf (accessed 27 April 2018).

Guidi, G., Micoli, L. L., Gonizzi, S., Navarro, P. R. and Russo, M. (2013). 3D digitizing a whole museum: A metadata centered workflow. *2013 Digital Heritage International Congress (DigitalHeritage)*, vol. 2. pp. 307–10 doi:10.1109/DigitalHeritage.2013.6744768.

Legacy No Longer: Designing Sustainable Systems for Website Development

Karin Dalziel
kdalziel@unl.edu
University of Nebraska–Lincoln, United States of America

Jessica Dussault
jdussault@unl.edu
University of Nebraska–Lincoln, United States of America

Gregory Tunink
techgique@unl.edu
University of Nebraska–Lincoln, United States of America

Introduction

The Center for Digital Research in the Humanities (CDRH) at the University of Nebraska–Lincoln is home to digital collections such as *The Walt Whitman Archive*, *The Willa Cather Archive*, *The Journals of Lewis and Clark*, and *O Say Can You See*. These projects contain overlap between subjects, individuals, and locations, yet are siloed, and many

are built in aging, unsupported technologies with no interoperability or common search. In order to address this, the Center has developed an API ("Henbit") as part of a modular software stack to index and display data and content.

Challenge

Over the past twenty years, the Center has created over 30,000 TEI files in addition to other data sets such as VRACore documents, spreadsheets, and databases. Sites showcase the content and metadata of these files using a variety of technologies, many of which are no longer maintained. In addition, some sites used commercial software which became unsustainable when costs went up, cementing a commitment to open source. This experience informed and reinforced our adopted design philosophy, which can be summed up as:

- Keep it simple, stable, and sustainable
- Embrace modularity by writing software for one purpose
- Avoid over-engineering solutions (i.e. graphical interfaces where command-line will do)
- Provide comprehensive documentation

The Center has been inspired to think bigger about what can be accomplished by including existing data in a new framework. An exciting next step is creating a site to search all Center data, find commonalities between projects, and read materials across sites for comprehensive research. This approach will also help solve accessibility issues of older project sites which do not meet modern requirements. As projects become unsustainable, the Center may retire them while keeping all content available.

While having one place to view and search the Center's data is important, it's also critical to allow the creation of independent sites which utilize unique organization and include special features requested by principal investigators for new and evolving projects. Quickly creating bare bones sites to view in-progress TEI is essential, as it allows metadata experts and PIs to refine their data and arguments. Such sites should be written for ease of maintenance, freeing future developer time to work on new projects rather than sustaining old ones.

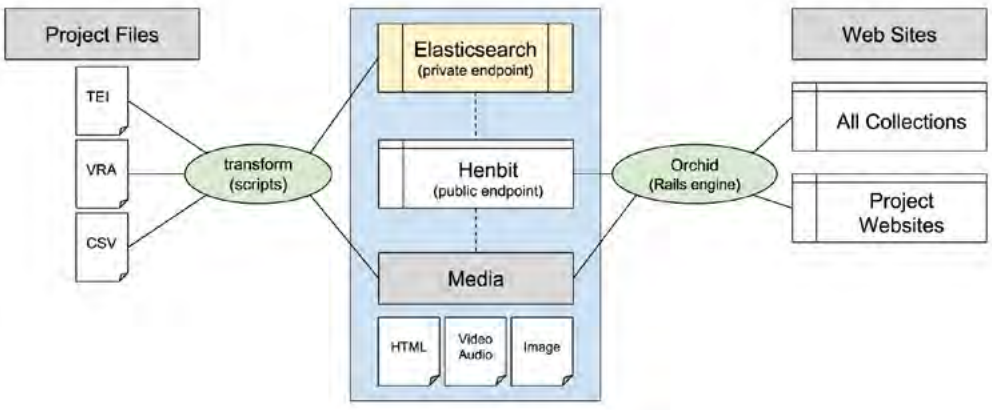
Solution

The Center explored the possibility of using existing software to address these challenges, such as XTF, Blacklight, and Fedora. These packages did not fit the Center's needs; though comprehensive, they were not flexible enough to accommodate the variety of document types and project site requirements. Additionally, many solutions would lock the API into using Solr instead of allowing an interchangeable search engine (Blacklight, 2017; DuraSpace, 2017).

Instead of heavily customizing existing software, The Center decided to create a modular solution. The system consists of several components:

- data repository for project files and scripts for transformation
- document datastore and search engine (Elasticsearch)
- Ruby on Rails (Rails) API to serve data (Henbit)
- media retrieval system for associated images, audio, and video
- template generator for rapid website creation (Orchid)

With a modular software stack, future changes in technology and project needs can be accommodated with independent upgrades rather than massive redesigns and rewrites.



Project Files and Scripts

The data repository houses original files for projects, such as TEI-XML, VRACore, CSV, and Dublin Core. The repository also contains CLI scripts which create HTML and populate search indexes with document content and metadata (CDRH, 2017a). New projects use generalized scripts, which are organized to allow overriding functionality in individual projects. Older websites may continue to use existing XSLT and populate legacy Solr indexes while their existing sites are supported, as well as populate Elasticsearch using the standardized script. Static HTML files derived from this process are used to create a document which can be viewed in a browser, regardless of the original data format.

Henbit (Public Endpoint)

Henbit is a Rails powered API (application program interface) which creates appropriate requests for the backend index, and returns JSON. Currently, Henbit uses Elasticsearch as a backend, but most of its features (sorting, filtering, aggregating on ranges, etc) could be ported to a different backend. The OpenAPI specification was used during Henbit's creation to fit current design practices (CDRH, 2017b).

Media Retrieval

In legacy sites, associated media lived inside the website directory. The Center has created a standard URL path for media files. It will be easier to optimize serving specific file types with this common retrieval structure. In the near future, the CDRH will be implementing a IIIF image server to serve images of varying sizes and resolutions.

Orchid (Rails Engine)

Orchid is a Rails engine which connects Rails 5 applications and Henbit. Orchid and a supporting gem, api_bridge, provide a template website that allows users to browse, search, filter, and view documents. This template is highly customizable, and can be altered to allow different URLs, search behavior, and anything possible in Rails (CDRH, 2017c).

Current Implementation and Future Plans

Beta versions of all components were released in 2017. In late 2017 the framework was used to build *The Complete Letters of Willa Cather* (launched January 2018). *The Complete Letters* demonstrates the customization which can be accomplished with this modular system. The CDRH is currently developing another project, *Family Letters*, which will also take advantage of the data repositories, scripts, Henbit, and Orchid template.

In the meantime, older websites are being converted for the new system. Updated documents and original XSLT have been reorganized into the structure required by the data repository scripts and are being posted to the Elasticsearch index. Once a site for Centerwide projects has been created, older sites can be retired as needed, replaced by content now available through the new API and supporting website.

The decision to use custom built software rather than an existing, out of the box solution, was not easy. Though at times it felt like reinventing the wheel, our highly customizable and flexible implementation prepares for future technological developments and enables flexibility in meeting project requirements.

Notes

https://cdrh.unl.edu
http://whitmanarchive.org, http://cather.unl.edu, https://lewisandclarkjournals.unl.edu, and http://earlywashingtondc.org
https://xtf.cdlib.org, http://projectblacklight.org, and http://fedorarepository.org
https://github.com/CDRH/data
https://github.com/CDRH/api
https://github.com/OAI/OpenAPI-Specification
http://iiif.io
https://github.com/CDRH/orchid
https://github.com/CDRH/api_bridge
http://cather.unl.edu/letters

References

Blacklight (2017). "Project Blacklight." http://projectblacklight.org.

CDRH (2017a). "CDRH Data Repository." *GitHub*. https://github.com/CDRH/data.

CDRH (2017b). "Henbit." *GitHub*. https://github.com/CDRH/api.

CDRH (2017c). "Orchid." *GitHub*. https://github.com/CDRH/orchid.

DuraSpace (2017). "Fedora Repository." http://fedorarepository.org

Histonets, Turning Historical Maps into Digital Networks

Javier de la Rosa Pérez
versae@stanford.edu
Center for Interdisciplinary Digital Research, Stanford University, United States of America

Scott Bailey
scottbailey@stanford.edu
Center for Interdisciplinary Digital Research, Stanford University, United States of America

Clayton Nall
nall@stanford.edu
Department of Political Science, Stanford University, United States of America

Ashley Jester
ajester@stanford.edu
Center for Interdisciplinary Digital Research, Stanford University, United States of America

Jack Reed
pjreed@stanford.edu
Digital Library Systems and Services, Stanford University, United States of America

Drew Winget
awinget@stanford.edu
Digital Library Systems and Services, Stanford University, United States of America

Introduction

The study of communication networks, specifically road networks, is a topic of broad interest to the scholarly community. It allows researchers to draw conclusions that range from historical events (Antrop, 2004; Trombold, 1991) to transit and traffic (Bash et al., 2017; Yang and Yagar, 1995), while adding a tangible and understandable dimension to their work. The appearance of Geographical Information Systems (GIS) made it possible to perform such analysis efficiently and accurately. It is just recently that the study of topological and growth properties of road networks are giving us the chance of understanding the bigger picture of cities (Antrop, 2005; Kasanko et al., 2016).

In the American landscape, network analysis of road networks has shown evidence that the construction of interstate highways affected the political and geographic polarization of cities, undermining representation and posing a threat to democracy itself (Nall, 2015; Ejdemyr et al., 2005). Most of these studies, however, rely on "the only rigorous year-to-year record of the construction of interstate highways and the incorporation of existing freeways into the system" (Nall, 2018), the Federal Highway Administration PR-511 database (FHWA PR-11). While the FHWA PR-11 is the most complete database available, it is based on highway construction records, which oftentimes misrepresent the complexity of turning political promises into reality, and does not include data on the development of road networks before the interstates. One way to approach this lack of data is to resort to roadmap collections, which might be a better proxy to understand the reality of transportations networks. Unfortunately, despite the number of digitized and scanned map collections, the lack of their availability in standard network data formats still represents a burden for the study of historical road networks. Although network analysis tools exist, we are not able to fully leverage their potential regarding historical datasets without a huge amount of manual work to generate network data.

As an alternative, modern approaches of road extraction from maps promise fully automated methods that rarely generalize well (Mena, 2003, Sharma et al., 2013), or rely on good quality labeled data (Isola et al., 2016), which is non existent or very difficult and costly to gather. We are then left to semi-automated methods where the researcher is guided to enter some crucial information needed for the automated process to start. However, these methods are usually conceived for satellite imagery or raster images of maps, lacking proper support for the variety of style and format found when dealing with collections of historical maps, and producing vector information not in network format. In order to fill this gap, we are presenting Histonets, a web-based platform to assist in the conversion of historical maps into digital networks, turning intersections into nodes and roads into edges.

Methodology

The platform begins with a login screen, after which each researcher can create a number of collections of images of maps by linking them from IIIF-compliant repositories. Furthermore, researchers are able to create settings for similar images (according to their criteria). Once images are selected, the pipeline for the Histonets platform is comprised of 4 steps: image preparation and cleaning, pattern matching, pathfinding, and graph correction. Cleaning can be fully automated or fine-tuned by adjusting the parameters of several actions to be applied. Once clean, image color depth is reduced by an automatic color clustering algorithm that only needs the final number of colors (defaults to 8).

With the image clean and posterized, the pattern matching step begins. In order to identify intersections and corners that will eventually become the nodes of the graph, researchers must circle around them, and, with a couple of samples, Histonets will try to find other instances in the images, taking into account rotation and orientation of the templates. Identifying roads is done by selecting their colors and a threshold. Areas under a certain threshold are removed as well. A final preview of the resulting graph is shown for the whole image. If the graph complies with the expectations the researcher can start a batch process to apply the same parameters to the whole collection. The tasks can be monitored and canceled. The final result of the process for each image map is a downloadable file in a compatible graph format, including Gephi and GraphML (sees Figure 1).

Figure 1. Sample of image input (upper left), internal output (upper right), and final graph as produced by Histonets (lower)

Discussion

Although in early stages, Histonets has already proved to reduce substantially the amount of hours of manual labour, cutting down the time needed to process an entire collection. Moreover, the easy parallelization built-in in Histonets is only limited by the computational resources available, making it easier for cloud or high performance computing center deployments to further boost its performance. However, without a proper benchmarking framework it is still difficult to assess its accuracy and completeness. One of our goals moving forward is to test and measure these factors, and adjust the platform for greater reliability.

While Histonets, as a whole pipeline, is focused specifically on extracting road networks from historical maps, collaborators have already identified uses outside of Political Science or History. As a general low-barrier and user friendly computer vision application, we have shown it to be useful for identifying capital letters in Medieval manuscripts, counting glyphs in Egyptian hieroglyphs, or even identifying architectural features. With its balance between meeting specific research needs and generalizable applicability, Histonets has a bright future as an adaptable tool in the Digital Humanities.

References

Antrop, M. (2004) Landscape change and the urbanization process in Europe. *Landscape and urban planning* 67.1, pp. 9-26.

Antrop, M. (2005) Why landscapes of the past are important for the future. *Landscape and urban planning* 70.1, pp. 21-34.

Bast, H., et al. (2017) Fast routing in road networks with transit nodes. *Science* 316.5824, pp. 566-566.

Champion, T. (2001) Urbanization, suburbanization, counterurbanization and reurbanization. *Handbook of urban studies* 160: 1.

Ejdemyr, S., Nall, C., and O'Keefe, Z. (2015) Building Inequality: The Permanence of Infrastructure and the Limits of Democratic Representation.

Isola, P., et al. (2016) Image-to-image translation with conditional adversarial networks. *arXiv preprint* arXiv:1611.07004.

Mena, J. B. (2003) State of the art on automatic road extraction for GIS update: a novel classification. *Pattern Recognition Letters* 24.16, pp. 3037-3058.

Nall, C. (2015) The political consequences of spatial policies: How interstate highways facilitated geographic polarization. *The Journal of Politics* 77.2, pp. 394-406.

Nall, C. (2018) The Road to Inequality: How the Federal Highway Program Polarized America and Undermined Cities. Cambridge University Press.
Kasanko, M., et al. (2016) Are European cities becoming dispersed?: A comparative analysis of 15 European urban areas. *Landscape and urban planning* 77.1, pp. 111-130.
Sharma, N, Bedi, R., and Dogra, A. K. (2013) A Survey on Road Extraction from Color Image using Vectorizaton. *IJRET: International Journal of Research in Engineering and Technology* 2.10.
Trombold, C. D. (1991) ed. *Ancient road networks and settlement hierarchies in the New World.* Cambridge University Press.
Yang, H., and Yagar, S. (1995) Traffic assignment and signal control in saturated road networks. *Transportation Research Part A: Policy and Practice* 29.2, pp. 125-139.

Alfabetización digital, prácticas y posibilidades de las humanidades digitales en América Latina y el Caribe

Gimena del Rio Riande
gdelrio.riande@gmail.com
CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Argentina

Paola Ricaurte Quijano
ricaurte.paola@gmail.com
Tecnológico de Monterrey, México

Virginia Brussa
virbrussa@gmail.com
Universidad Nacional de Rosario, Argentina

Atraviesan al concepto étnico-geográfico definido como Latinoamérica distintos procesos regionales en los que se observan políticas para impulsar estrategias de acceso a internet, incorporación de las tecnologías digitales al sistema educativo y/o implementación de programas de alfabetización digital. Desde la Cumbre Iberoamericana de San Salvador en el año 2008 se viene sosteniendo, por ejemplo, la necesidad de "impulsar políticas, que incluyan el marco de la colaboración público-privada, encaminadas a facilitar la integración plena de las y los jóvenes en la Sociedad de la Información y del Conocimiento a través del acceso universal a las Tecnologías de la Información y de la Comunicación (TIC) y el desarrollo de contenidos digitales, mediante programas de alfabetización digital que reduzcan la brecha existente y con la mira puesta en facilitar el acceso al empleo, el emprendimiento y la realización personal" (INTEF, 2013).

Pero ¿qué es la alfabetización digital en el marco de un campo científico como el de las Humanidades Digitales, un nuevo espacio de producción académica nacido bajo el amparo de las Digital Humanities del norte global? Como es sabido, las Humanidades Digitales (HD) se han consolidado como un campo académico en franca expansión, principalmente en países de habla anglosajona. Así y todo, su recepción ha sido diferente para nuestra región y, al día de hoy, no se han absorbido en el currículo universitario o actividades de investigación del mismo modo. Las HD dan cuenta de un diálogo entre las humanidades y la informática entendida como digitalidad, y también de la posibilidad de crear nuevos objetos de estudio y líneas de investigación mixta, aunque, tal vez la apuesta más interesante y menos apreciada de las HD sea los puentes interdisciplinares que tienden y ofrece a las distintas disciplinas humanísticas (del Rio Riande, 2016).

Si bien la alfabetización digital y el desarrollo de competencias supone mucho más que infraestructuras, la posibilidad de acceso físico, real y efectivo a las tecnologías, así como el desarrollo de políticas institucionales relativas a su impulso siguen siendo un desafío para el crecimiento de las HD como campo científico. Algunos de los elementos que es necesario considerar tienen que ver con la implantación de una cultura digital que no sea únicamente instrumental sino que implique una reflexión crítica acerca de la relación entre tecnología, humanidades y producción de conocimiento.

Por otra parte, las investigaciones recientes y las políticas educativas nos alertan sobre la importancia de lo que podríamos denominar "multiliteracies" o multialfabetizaciones (Cope & Kalantzis, 2000), incluyendo, en ese sentido, no sólo aspectos de manejo de herramientas, sino del impacto en cómo "leer", "traducir" aquello computacional desde un aspecto crítico. Clave, por ejemplo, es el proceso que contiene a los datos de investigación u objetos intensivos en datos digitales.

Con el fin de indagar acerca del estado de las prácticas digitales en la región, diseñamos una encuesta abierta, orientada a estudiantes, profesores, investigadores, bibliotecarios y documentalistas en América Latina en el marco del proyecto *Prácticas digitales en América Latina y el Caribe* (http://openlabs.limequery.com/954661?lang=es-MX). La encuesta buscó medir el conocimiento y las prácticas de estos agentes de producción en el ámbito académico sobre recursos para la investigación, la publicación científica y la preservación (desde los procesadores de texto, pasando por los repositorios, hasta las infraestructuras digitales). El proyecto se desarrolla en conjunto con Humanidades Digitales CAICYT (Centro Argentino de Información Científica y Tecnológica del CONICET-Argentina), +Datalab del Centro de Investigación en Mediatizaciones (Facultad de Ciencia Política y Relaciones Internacionales) de la Universidad Nacional de Rosario (Argentina) y Openlabs de la Escuela de Humanidades y Educación del Tecnológico de Monterrey (México). Se

recogieron, hasta el momento, más de 300 respuestas de diversos países de América Latina. Una primera versión de esta encuesta se realizó en Argentina en 2015-2016 en el marco del convenio entre CIM-CAICYT de CONICET.

Presentaremos en esta ocasión los resultados obtenidos respondiendo a estos imperativos, discutiendo los hallazgos clave sobre las interacciones entre la investigación, el acceso a la tecnología entre estudiantes cultural y lingüísticamente diversos, como parte del estado de la cuestión en espacios académicos y su incidencia sobre el desarrollo del campo científico de las Humanidades Digitales en América Latina así como de políticas y currículos universitarios más reales y democráticos.

References

Arellano, A. (2007). "De la epistemología de la ecología política latouriana a una epistemología de sustento antropológico". *Convergencia*. *Revista de Ciencias Sociales*, 44 (mayo-agosto).

Cope, B. & Kalantzis, M. (2000). Introduction. In Cope, B. & Kalantzis, M. (eds.), *Multiliteracies: Literacy learning and the design of social futures*. South Yarra, VIC: MacMillan.

Kreimer, P., Vessuri, H., Velho, L. & Arellano, A. (2014). Perspectivas Latinoamericanas en el estudio social de la Ciencia y la Tecnología. México, Siglo XXI.

del Rio Riande, G. (2016). Humanidades Digitales: estándares para su consolidación en el campo científico argentino. *SEDICI*. Repositorio Institucional de la UNLP. http://sedici.unlp.edu.ar/handle/10915/62008

Instituto Nacional y del Profesorado (INTEF) (2013). Declaración sobre innovación y TIC del Foro de Ministros de Educación de las Américas. [Blogpost] *Educalab.* http://blog.educalab.es/intef/2013/06/26/declaracion-sobre-innovacion-y-tic-del-foro-de-ministros-de-educacion-de-las-americas/

Listening for Religion on a Digital Platform

Amy DeRogatis
derogat1@msu.edu
Michigan State University, United States of America

What does religion in the United States sound like, and where should one go to listen for it? What are the different ways that religious individuals and communities make themselves heard--to each other, to their gods, and to others? How is religious pluralism reshaping the sounds and spaces of North American religious life? How might we begin to reconceptualize religion and its place in North American life if we begin by using our auditory perception as a source of knowledge? And how might this knowledge be represented and transformed through the use of new digital media?

I co-direct "The American Religious Sounds Project," a collaborative initiative of Ohio State and Michigan State Universities to leverage opportunities afforded by the new digital environment to consider what religion sounds like in the United States. The project centers on (1) the construction of a unique sonic archive, documenting the diversity of everyday American religious life through newly produced field recordings, interviews, oral histories, and related materials; and (2) the development of a new digital platform and website, which draws on materials in our archive to engage users in telling new stories about religious diversity in the U.S. This multi-modal platform includes a searchable archive, database-driven visualizations, which invite users to explore, discover, and listen for surprising connections among our materials, and a curated gallery of multimedia exhibits, which allow for greater interpretation and contextualization. Future phases include plans for museum installations, traveling exhibits, and community-based workshops.

It has become commonplace (if arguably inaccurate) to describe the United States as the most religiously diverse country in the world. Scholars of North American religions have recognized the pressing need for new approaches to documenting and making sense of this diversity. Our approach stems from our particular interests in the material and sensory cultures of American religions and in the varied ways that religion has become newly visible and audible in American life, confounding once dominant assumptions about secularization and privatization. Rather than retreating quietly into an interiorized or immaterial realm of personal belief, religion has remained an integral feature of the modern world, and religious communities have inscribed themselves on urban landscapes and soundscapes in a variety of ways.

The working we are doing through the American Religious Sounds Project also has been stimulated by a "sensory turn" in scholarship across the humanities and social sciences. Historians, anthropologists, geographers, and others have been attending to the cultural values and social ideologies expressed through different ways of sensing the world and to the multi-sensorial modes through which modern culture was constituted. The nascent field of sound studies, defined broadly as the cultural study of sound and listening, has proven particularly generative, giving rise to new ways of thinking about critical questions that have long animated humanistic inquiry, including the legacies of industrialization and urbanization, the role of technological production and mediation, and the construction of ethnic, racial, religious, sexual, gendered, and class-based differences. Research on sound and through sound provides a rich medium for understanding religious groups, people, events, and conflicts.

Religious studies scholars, however, have paid far more attention to visual and material culture than to audi-

tory culture. In part, this can be attributed to the limitations of the textual media through which scholars have traditionally presented their research, including published monographs and journal articles. Such media have not readily lent themselves to engagement with sonic materials, for sound can be difficult to represent in such formats. Acutely sensitive to this problem, many ethnomusicologists and sound artists have begun experimenting with digital tools and platforms, like soundmapping, but such approaches have not yet made their way into the discipline of religious studies. Scholars of religion should take greater advantage of the opportunities afforded by the new digital environment, while also reflecting critically on its limitations. The American Religious Sounds Project is designed to do both.

Our sound selections are robustly multi-religious, including a wide range of Christian and non-Christian traditions. We include the formal sounds of religious institutions, such as prayer, chanting, and hymns, as well as the informal, and often unintentional, sounds that arise during relaxed coffee hours and spontaneous conversations, ambient and incidental noises like laughter and crying, clapping and shouting, and the shuffling and movement of lived community during worship. We record regular weekly and daily services, as well as seasonal festivals and other special events. We move outside of formal religious institutions to capture the sounds of devotion in homes and schools, public parks and interfaith chapels, coffee shops and workplaces, as well as at ostensibly "secular" gatherings such as a school graduation, public arts festival, or college football games. For example, our researchers recently recorded the sounds of a public Christmas tree lighting, an interfaith prayer vigil against violence, a neo-Pagan brewing mead in his home kitchen, an anti-Islam protest rally, a (secular) Sunday Assembly meeting in a coffee shop, and a Bhutanese Nepali Hindu festival. By casting our net widely, we aim to build a resource that is broadly comprehensive, comparative, and even a bit provocative. We do not intend to answer definitively the question of what counts as religious, but to invite critical reflection on what is at stake in that designation and to consider the role that auditory perception plays in its constitution.

In this paper, I will introduce the project and present our website, which we expect to launch in March 2018. I will solicit critical feedback and offer reflections of my own on the capabilities and limits of new digital methods for enhancing our research of the varied sonic cultures of North American religious life. One of the goals of the American Religious Sounds Project is to provide a bridge between our academic settings and our local communities. That work must be done carefully and respectfully in the present political and religious climate of the United States. I will end with some thoughts on the precarious work of the public presentation of religious sounds and communities on an open accessible digital platform.

Words that Have Made History, or Modeling the Dynamics of Linguistic Changes

Maciej Eder
maciejeder@gmail.com
Institute of Polish Language (Polish Academy of Sciences), Poland; Pedagogical University in Kraków, Poland

Introduction

In the last decades, quantitative linguistics (following exact and social sciences) has developed a considerable number of statistic methods providing an insight into measurable phenomena of natural language. Although to a lesser extent, it also applies to the analysis of diachronic changes. The basic tool used to assess the chronology of linguistic changes is a rather effective yet simple method of trend search: the examined features are analyzed by mapping the frequency of the described phenomenon on a timeline (Ellegård, 1953). This timeline-centric visualization has become a standard in several studies and corpus tools. The most spectacular example is the corpus of several dozens of million of documents (mainly in English) accompained by the service Google Books Ngram Viewer http://books.google.com/ngrams, which, according to its authors, enables to examine changes taking place not only in the language, but also in culture (Michel et al., 2011).

A significant drawback of simple graphic representation of the trend, and hence of mapping the frequency of the examined phenomenon on a timeline, is a tacit assumption that the researcher knows in advance which elements of the language are subject to change. In other words, the method of plotting and inspecting the trend may be applied only to verify hypotheses stipulated earlier by traditional diachronic linguistics. For example, knowing in advance that Polish underwent the gradual replacement of the inflected ending *-bychmy* with *-byśmy*, one might draw the trendline and capture the dynamics of that change. Although many prominent diachronic works were based upon such an approach (Biber, 1988; Hilpert and Gries, 2009; Hu et al., 2007; Reppen et al., 2002; Smith and Kelly, 2002; Can and Patton, 2004), one might be interested in trend search without any *a priori* selection of the analyzed linguistic changes to be traced.

Needles to say, *some* selection of potential language change predictors (e.g.a predefined set of words, certain collocates, etc.) will always be the case. The strategy followed in this study was to analyze a considerably large set of 1,000 most frequent words without any further filters, with the assumption that some of them will turn out stronger than others. Arguably, in such a big set one should find a few dozen of function words, and a vast majority of content words. Another remark that has to be formulated here is that the language change cannot be reliably separated

from the stylistic drift (e.g.in literary taste of the epoch). This fact is well known in stylometric approaches to style ("stylochronometry"), where the actual changes in the system and stylistic signals of, say, the predominant genres are usually difficult to be told apart.

Supervised classification and the timeline

The most natural strategy to assess the discriminative power of numerous features at a time is to apply one of the multivariate methods. Since none of the out-of-the-box techniques is suitable to analyze temporal datasets, some tailored approaches have been proposed, e.g.using a variant of hierarchical clustering (Hilpert and Gries, 2009; Hulle and Kestemont, 2016). These methods, however, share a common drawback, namely their results are by no means stable. Also, no cross-validation can be considered a downside.

To assess these issues, an iterative procedure of automatic text classification was applied (Eder and Górski, 2016). Its underlying idea is fairly simple: first, we formulate a working hypothesis that a certain year – be it 1835 – marks a major linguistic break. The procedure randomly picks *n* text samples written before and after the assumed break; the samples then go into the *ante* and *post* subsets. In this study, a period of 20 years before and after the assumed break was covered (with an additional gap of 10 years), 500 text samples of 1,000 tokens were harvested into each of the subsets. To give an example: for the year 1835, 500 random samples covering the time span 1810–1830 were picked into the first subset, and another 500 samples from the years 1840–1860 into the second subset. Next, the both subsets are randomly divided into two halves, so that the training set and the test contain 500 samples representing two classes (*ante* and *post*). Then we train a supervised classifier – in this case, Nearest Shrunken Centroids – and record the cross-validated accuracy rates. Then we *dismiss* the original hypothesis, in order to test new ones: we iterate over the timeline, testing the years 1836, 1837, 1838, 1839, … for their discriminating power. The assumption is simple here: any acceleration of linguistic change will be reflected by higher accuracy scores.

Data and results

The above procedure has been applied to the Corpus of Historical American English (COHA), containing ca. 400 million tokens and covering the years 1810–2009 (Davies, 2010). The corpus provides the original word forms, part-of-speech tags, and the base word forms (lemmata). The results reported below were obtained using the lemmatized version of the corpus.

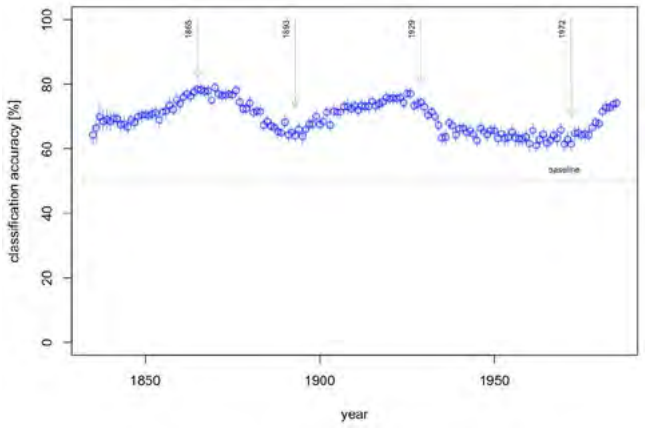


Fig. 1: Language change accelleration in the American English corpus: classification accuracy over the years 1835–1985.

In Fig. 1, the classification accuracy rates for the COHA corpus were shown (1,000 most frequent lemmata, NSC classifier). As one can observe, the scores obtained for each period are higher than the baseline, suggesting the existence of a temporal signal. Obviously, the higher the scores the faster the evolution of language, since the distinction between the period before and after the tested breakpoint is simpler for the classifier. More important, however, is the fact that the scores are not even: the signal becomes stronger in some periods, clearly indicating an acceleration of the language change. One of the stylistic breaks takes place in the 1870s (i.e.after the Civil War), the other in the 1920s (in the period of prosperity before the Great Depression); the third peak is not fully formed yet, even if one can observe an acceleration of language change at the end of the 20th century. Needless to say, any attempts at finding direct correlations between historical events and stylistic breaks are subject to human prejudices, and therefore might introduce substantial bias to the results. Even though, the coincidence of the three observed peaks and a few major changes in the American culture is rather striking.

Distinctive features

The results obtained in the above experiment seem to be rather promising. However, from the perspective of historical linguistic even more interesting is the question which features (words) were responsible for an given change observed in the dataset. It has been reported in several stylometric studies that attributing authorship relies, in most cases, on many features of individually very weak discriminative power. In the context of language change, a similar question can be asked: is it but a few characteristic words that trigger the change, or, alternatively, is the stylistic drift spread across dozens of tiny changes in word frequencies?

To answer the above question, one has to extract the features that played a prominent role in telling apart the *ante* and *post* periods as described above. The features exhibiting the biggest variance (that is, the overall impact on the results) are shown in Fig. 2. An important caveat needs to be formulated here: the plot shows the outputted weights from the classifier, rather than direct word frequencies. The underlying assumption is that the features' weights (to be precise: the *a posteriori* probabilities returned by the classifier) reflect the changes in actual word frequencies as combined with all the other frequencies being analyzed.

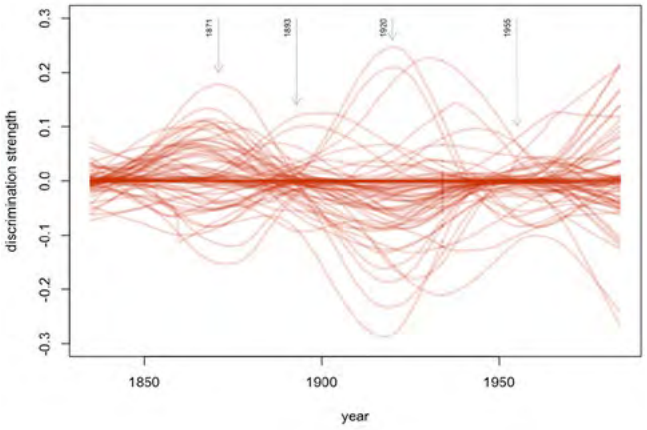


Fig. 2: Seventy-six linguistic features (words) that contributed considerably to the stylistic drift.

The main stylistic breaks form, again, three peaks that culminate roughly in the same years as presented in Fig. 1. What is counterintuitive, however, it is the fact that the features tend to form sinusoidal waves of their periodical discrimination power. Interestingly, these high impact features turned to be very frequent words that usually occupy the top positions on the frequency list. The 25 words of the highest discrimination strength are as follows:

the, *and*, *week*, *that*, *'s*, *last*, *is*, *be*, *of*, *it*, *we*, *i*, *to*, *was*, *mr.*, *our*, *my*, *been*, *not*, *u.s.*, *you*, *new*, *upon*, *there*, *has*

Even more interesting are individual trajectories of the high-impact words. In Fig. 3, one can observe a collinearity of function words: *the*, *and*, *that*, *is*, *been*, as opposed to the possessive *'s*. These function words seem to have impacted the language change at the turn of the 19th century. The possessive, in turn, contributed to the evolution of language roughly at the times of the Prohibition. (Again, this is not to say that any direct links between function words and actual events in history should be drawn).

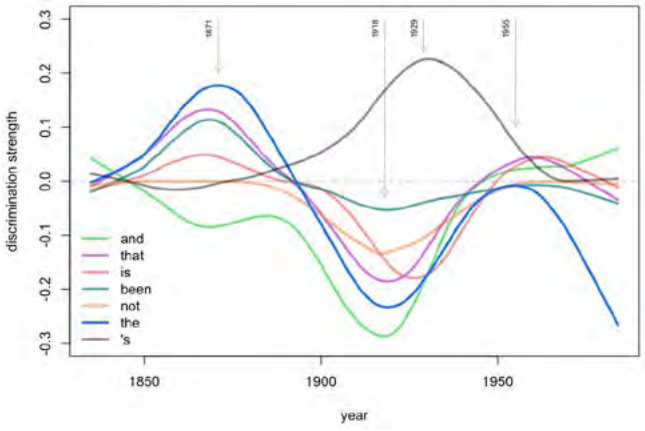


Fig. 3: Function words of the highest impact on the stylistic drift.

A different pattern is revealed by the "social" words, especially personal pronouns. It has been shown that these words, e.g. *I*, play prominent role in betraying someone's personality (Pennebaker, 2011). Certainly, traces of such individual profiles will hardly be noticeable at the

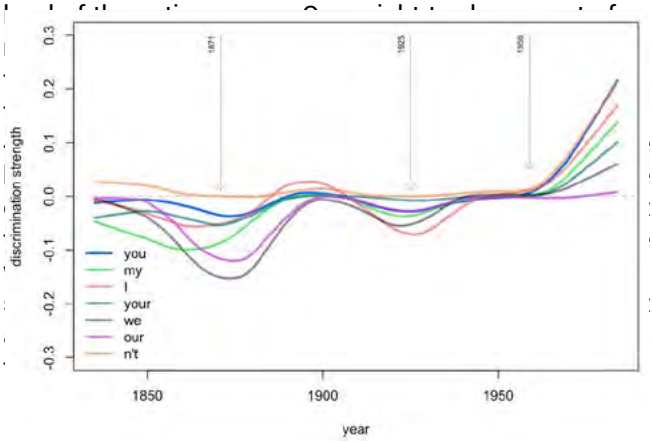


Fig. 4: High impact personal pronouns and contractions.

Conclusions

In this paper, we used a tailored stylometric method to assess the question of language change over time. Our chosen technique proved to be useful indeed, especially when one focuses on tracing the very linguistic features that were responsible for the observed change. The results were counterintuitive, since the set of strongly discriminative features contained common function words, which formed sinusoidal trajectories of their impact over time. One of the most interesting aspects of language development – overlooked in numerous existing studies – is the question of the dynamics of linguistic changes. Our study corroborated the hypothesis that epochs of substantial stylistic drift are followed by periods of stagnation, rather than forming purely linear trends.

Acknowledgements

This research is part of project UMO-2013/11/B/HS2/02795, supported by Poland's National Science Centre.

References

iber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Can, F. and Patton, J. M. (2004). Change of writing style with time. *Computers and the Humanities*, 38(1): 61–82.

Davies, M. (2010). The Corpus of Historical American English (COHA): 400 million words, 1810-2009 https://corpus.byu.edu/coha/.

Eder, M. and Górski, R. L. (2016). Historical linguistics' new toys, or stylometry applied to the study of language change. In, *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 182–84 http://dh2016.adho.org/abstracts/398.

Ellegård, A. (1953). *The Auxiliary Do: The Establishment and Regulation of Its Use in English*. Stockholm: Almquist & Wiksell.

Hilpert, M. and Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4): 385–401.

Hu, X., McLaughlin, J. and Williamson, N. (2007). Syntactic Positions of Prepositional Phrases in the History of Chinese: Using the Developing Sheffield Corpus of Chinese for Diachronic Linguistic Studies. *Literary and Linguistic Computing*, 22(4): 419–34.

Hulle, D. van and Kestemont, M. (2016). Stylochronometry and the Periodization of Samuel Beckett's Prose. In, *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 393–95 http://dh2016.adho.org/abstracts/70.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014): 176–82.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. New York: Bloomsbury Press.

Reppen, R., Fitzmaurice, S. M. and Biber, D. (eds). (2002). *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Smith, J. A. and Kelly, C. (2002). Stylistic Constancy and Change Across Literary Corpora: Using Measures of Lexical Richness to Date Works. *Computers and the Humanities*, 36(4): 411–30.

The Moral Geography of Milton's Paradise Lost

Randa El Khatib
elkhatib.randa@gmail.com
University of Victoria, Canada

John Milton's *Paradise Lost* creates an extraordinarily rich and complex sense of space. The epic poem elegantly captures the cartographical leap of the sixteenth and seventeenth century that owes to advancements in navigation techniques and rapid colonial expansion. The world image was rapidly changing and gaining a more distinct contour as newly colonized lands were becoming better described and known. Maps in this time could often be considered prototypes since cartographers were still experimenting to find a more accurate mimesis of the world. At the same time, the strong foundation of *Paradise Lost* and many other retellings of the Genesis captures the saturation of the seventeenth century in religious tradition and references to sacred places. In this way, *Paradise Lost* can be seen as a prototype of its own that brings together spatial traditions, new and old, real and imaginary, into a single medium. To date, Milton's spatial allusions – spanning biblical, classical, and contemporary temporalities – have predominantly been studied in relation to the textual sources that had influenced them. However, *Paradise Lost* was written at a time when the visual tradition of mapping places of the bible with cartographic exactitude had reached its peak, seen in the King James Bible, which was also Milton's family Bible – a tradition that, in retrospect, is an early example of a geospatial, text-to-map project. Milton construed his spatiality on the existing framework of this visual tradition, and consolidated the geographies of classical antiquity and of his contemporary world. These temporalities were conceived to have progressed on a linear spectrum of geographical continuity, according to the prevalent notion of historical sequence of a seventeenth-century audience. By superimposing these layers, Milton uses textual sources to assign moral valence to geographical points; these inform the readers' understandings of the epic and of the space of human history that it encompasses. The GIS-based digital project, "A Map of the Moralized Geography of Paradise Lost," explores the multi-temporal complexity of Milton's spatial allusions through an open access map depicting the moralized geography of *Paradise Lost*. These multiple temporalities are delineated by various layers of georectified historical maps, including the map that supplies the visual paratext of the King James Bible, as well as John Speeds map of "The Turkish Empire" (1626). The interactive dimensions of the map permit users to recover and evaluate nuance (by resituating geographical names in their poetic contexts) even as they seek to apprehend and deduce larger patterns.

The most powerfully apparent pattern is the concentration of Milton's spatial allusions on the Mediterranean

world, forming a thick chain around the Mediterranean basin. Sites of biblical or classical significance were, in the seventeenth century, in territories almost entirely controlled by the Ottoman Empire; this superimposition creates a polarized dynamic of moral valence. Additionally, Milton's map is coordinated with a map based on place names extracted from the Book of Genesis in order to investigate the scope of influences of the biblical book itself on the epic poem. The extraction of geo-coordinates from both works was carried out manually for the sake of accuracy, since the limitations of present geoparsing techniques with variant and historical place names remain a methodological sticking-point. The Genesis map is less complex than the initial one, making it clear that it was literary and exegetical writings, and religious culture more broadly, that built thick association. This condition reinforces the status of geographical references in Milton's epic as references, as vectors that import or apply associations established through cultural tradition or poetic technique. In this way, *Paradise Lost* functions like an early modern chorography that contextualizes place names at use. The fruit of this project is a navigable visual network that invites users to trace contextualized recurring patterns in multiple temporalities.

References

Galey, A. and Ruecker, S. (2010). How a prototype argues. *Literary and Linguistic Computing*, 25(4), 405–24.

Gillies, J. (1994). *Shakespeare and the Geography of Difference*. Cambridge: Cambridge University Press.

Gregory, I. and Murrieta-Flores, P. (2016). Geographical information systems as a tool for exploring the spatial humanities. In Crompton C., Lane, R.J., and Siemens, R. (eds), *Doing Digital Humanities: Practice, Training, Research*, pp. 177–92.

Hill, L. *Georeferencing*. (2014) Cambridge, MA: The MIT Press.

Jacobson, M. (2014). *Barbarous Antiquity: Reorienting the Past in the Poetry of Early Modern England*. Philadelphia: University of Pennsylvania Press.

Jessop, M. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing*, 23(3), 281–93.

Lim, W. (2010). John Milton, orientalism, and the empires of the east in Paradise Lost. In Johanyak D. and Lim W. (eds), *The English Renaissance, Orientalism, and the Idea of Asia*. Palgrave Macmillan, pp. 203–235.

McLeod, B. (1999). The 'Lordly eye': Milton and the strategic geography of empire. In Rajan B. and Sauer E.(eds.), *Milton and the Imperial Vision*. Duquesne University Press, pp. 48–66.

Milton, J. (2007). *The Complete Poetry and Essential Prose of John Milton*, Kerrigan W., Rumrich, J. and Fallon S. (eds). The Modern Library.

Ng, M. (2014). Milton's maps. *Word and Image*, 29(4), pp. 428–442.

Locative Media for Queer Histories: Scaling up "Go Queer"

Maureen Engel
mengel@ualberta.ca
University of Alberta, Canada

This paper reports on the completion and launch of the locative media app "Go Queer." Taking the theorization, iteration, and development of "Go Queer" as a model and case study, the paper argues that locative media is uniquely suited to re/mediating queerness. It then proposes that these findings can be used as a framework and set of best practices for developing a variety of queer history applications.

Go Queer is a ludic, locative media experience that occurs on location, in the city, on the playful border between game and story, the present and the past, the queer and the straight, the normative and the *slant*. The app takes the city of Edmonton's queer history as its text, and produces a locative, spatialized narrative of that history by displaying text, images, video and audio in place at the actual locations where they occurred, thus creating what Richardson and Hjorth (2014, 256) call "the hybrid experience of place and presence." The app invites its users to drift queerly through the city, discovering the hidden histories that always surround us, yet somehow remain just beyond our apprehension. It compiles these traces into a media layer that augments quotidian city space, juxtaposing the past onto the present, creating a deep, queer narrative of place. By bringing together the physical navigation of the contemporary city with the imaginative navigation of its queer past, the app enacts a praxis that I characterize as a *queer ludic traversal*, one that renders the navigation itself as queer as the content that it presents. In so doing, the app produces the experience of *place*, in Lucy Lippard's (1997) formulation that

> Place is latitudinal and longitudinal within the map of a person's life. It is temporal and spatial, personal and political. A layered location replete with human histories and memories, place has width as well as depth. It is about connections, what surrounds it, what formed it, what happened there, what will happen there. (7)

The app proposes that a productive and underrepresented setting for queer play is the space of the city itself, and that the hybrid reality of locative media provides specific affordances to enable particularly queer navigations, occupations, and constructions of urban space.

The app arises from, and takes shape in relation to, a range of theoretical inspirations. First are the contributions queer theories of space, the urban, and community, such as David Bell's (2001) observation of "the special relationship between the city and the deviant" (84) and Theories recognizing the very public-ness of the formation, circu-

lation, and inhabiting of queer identities (D'Emilio, 1983; Berlant and Warner, 1998); central here is Sara Ahmed's theorization of "orientation" and her contention that "orientations are about the directions we take that put some things and not others in our reach" (552). New theorizations of space and place that have come to be called *the spatial turn* have similarly mobilized our thinking, challenging us to imagine space as a complex social production (Lefebvre, 1992) and asking us to think through how we move in space as either *tactical* or *strategic* (deCerteau, 2011). Praxis-based interactivity, which I draw principally from the field of Game Studies, has introduced concepts like the fidelity context (Galloway 2004) and ambient experience (Flanagan 2009). Deep mapping offers new possibilities for modeling space, particularly historical space, by bringing together the explanatory and critical capacities of both narrative and mapmaking (Bodenhamer 2007). These theoretical methods intersect in locative media itself, the vehicle for "Go Queer" and a platform, I argue, that holds significant promise for queer scholarship and expression.

By exploring how each of these theoretical arenas is literalized in the app itself, this paper aims to provide a framework and method for other practitioners interested in deploying locative media technologies to engage queer subjects, histories, and cultural productions.

References

Anthropy, Anna. *Rise of the Videogame Zinesters: How Freaks, Normals, Amateurs, Artists, Dreamers, Drop-Outs, Queers, Housewives, and People Like You Are Taking Back an Art Form*. New York: Seven Stories Press, 2012.

Bell, David, and Jon Binnie. "Authenticating Queer Space: Citizenship, Urbanism and Governance." *Urban Studies* 41.9 (2004): 1807–1820. usj.sagepub.com. Web. 12 Feb. 2015.

Bell, David, and Gill Valentine. *Mapping Desire: Geographies of Sexualities*. 1 edition. London ; New York: Routledge, 1995.

Berlant, Lauren, and Michael Warner. "Sex in Public." *Critical Inquiry* 24.2 (1998): 547–566.

Binnie, Jon et al. *Pleasure Zones: Bodies, Cities, Spaces*. Syracuse: Syracuse Univ Pr, 2001.

Bodenhamer, David J. "Creating a Landscape of Memory: The Potential of Humanities GIS." *Journal of Humanities & Arts Computing: A Journal of Digital Humanities* 1.2 (2007): 97–110.

Cabiria, Jonathan. "Virtual World and Real World Permeability: Transference of Positive Benefits for Marginalized Gay and Lesbian Populations." *Journal of Virtual Worlds Research* 1.1 (2008): n. pag. Web.

Certeau, Michel de. *The Practice of Everyday Life. Trans*. Steven F. Rendall. 3rd Revised edition edition. University of California Press, 2011.

Chisholm, Dianne. *Queer Constellations: Subcultural Space In The Wake Of The City*. 1 edition. Minneapolis: Univ Of Minnesota Press, 2004.

Chrisman, Nicholas R. "Design of Geographic Information Systems Based on Social and Cultural Goals." *Photogrammetric Engineering & Remote Sensing* 53.10 (1987): 1367.

Craig, William J., and Sarah A. Elwood. "How and Why, Community Groups Use Maps and Geographic Information." *Cartography & Geographic Information Systems* 25.2 (1998): 95.

Crampton, J.W. "Maps as Social Constructions: Power, Communication and Visualization." *Progress in Human Geography* 25.2 (2001): 235–252.

Crang, Mike. "Public Space, Urban Space and Electronic Space: Would the Real City Please Stand Up?" *Urban Studies* (Routledge) 37.2 (2000): 301–317.

Danielson, Laura. "An Exploration of Deep Maps." N.p., thepoliscenter.iupui.edu.

Désert, Jean-Ulrick. "Queer Space." *Queers in Space: Claiming the Urban Landscape*. Ed. Gorden Brent Ingram, Gordon B. Ingram, and Yolanda Retter. Seattle, Wash: Bay Pr, 1997. 17–26.

Dodge, Martin, Rob Kitchin, and Chris Perkins, eds. *The Map Reader: Theories of Mapping Practice and Cartographic Representation*. Chichester, West Sussex, UK; Hoboken, NJ: Wiley, 2011.

Flanagan, Mary. *Critical Play: Radical Game Design*. The MIT Press, 2013.

Gieseking, Jen Jack et al., eds. *The People, Place, and Space Reader. People, Place, and Space: A Reader*: Routledge, 2014.

Goodchild, Michael F., and Donald G. Janelle. "Toward Critical Spatial Thinking in the Social Sciences and Humanities." *GeoJournal* 2010: 3.

Gregory, Derek. *Geographical Imaginations*. Cambridge, MA : Blackwell, 1994.

Gregory, Ian, and Paul S. Ell. Historical GIS [electronic Resource] : *Technologies, Methodologies and Scholarship* / Ian N. Gregory, Paul S. Ell. Cambridge, UK ; New York : Cambridge University Press, 2007. Cambridge Studies in Historical Geography: 39.

Halberstam, Judith. "What's That Smell? Queer Temporalities and Subcultural Lives." International Journal of Cultural Studies 6.3 (2003): 313–333.

Hall, Stuart. *"Encoding/Decoding." Media and Cultural Studies: KeyWorks*. Ed. Meenakshi Gigi Durham and Douglas M. Kellner. Revised Edition. Malden MA: Blackwell, 2006. 163–173. KeyWorks in Cultural Studies.

Harris, Trevor M., John Corrigan, and David J. Bodenhamer. *The Spatial Humanities : GIS and the Future of Humanities Scholarship*. Bloomington: Indiana University Press, 2010.

Harris, Trevor, and Daniel Weiner. "Empowerment, Marginalization, and 'Community-Integrated' GIS." *Cartography and Geographic Information Systems* 25.2 (1998): 67–76.

Hjorth, Larissa, and Sun Sun Lim. "Mobile Intimacy in an Age of Affective Mobile Media." *Feminist Media Studies* 12.4 (2012): 477–484.

Johnston, Lynda, and Robyn Longhurst. *Space, Place, and Sex: Geographies of Sexualities*. Lanham: Rowman & Littlefield Publishers, 2009.

Juliano, Linzi. "Digital: A Love Story; Bully; Grand Theft Auto IV; Portal; Dys4ia (review)." *Theatre Journal* 64.4 (2012): 595–598.
Kirkpatrick, Graeme. Computer Games and the Social Imaginary. *Polity*, 2013.
Knowles, Anne Kelly. *Past Time, Past Place : GIS for History* / Edited by Anne Kelly Knowles. Redlands, Calif. : ESRI Press, 2002.
Lefebvre, Henri. *The Production of Space*. 1 edition. Wiley-Blackwell, 1992.
Lippard, Lucy. *Lure of the Local: Senses of Place in a Multicentered Society*. 1 edition. New York: The New Press, 1998.
Lynch, Kevin. *The Image of the City*. Cambridge, Mass.: The MIT Press, 1960. Mattern, Shannon. Deep Mapping the Media City. Univ Of Minnesota Press, 2015.
McLafferty, Sara. "Mapping Women's Worlds: Knowledge, Power and the Bounds of GIS." *Gender, Place and Culture* 9.3 (2002): 263–269.
Murphy, Kevin. "Walking the Queer City." *Radical History Review* 62 (1995): 195–201.
Paglen, Trevor, and John Emerson. *An Atlas of Radical Cartography*. Ed. Lize Mogel and Alexis Bhagat. Slp edition. Los Angeles: Journal of Aesthetics and Protest Press, 2008.
Pavlovskaya, Marianna. "Theorizing with GIS: A Tool for Critical Geographies?" *Environment and Planning A* 38.11 (2006): 2003–2020.
Presner, Todd, David Shepard, and Yoh Kawano. *HyperCities: Thick Mapping in the Digital Humanities*. Cambridge, Massachusetts: Harvard University Press, 2014.
Retter, Yolanda, Anne-Marie Bouthillette, and Gordon Brent Ingram, eds. *Queers in Space: Communities, Public Places, Sites of Resistance*. Seattle, Wash: Bay Press, 1997.
Ridge, Mia, Don Lafreniere, and Scott Nesbit. "Creating Deep Maps and Spatial Narratives through Design." *Journal of Humanities & Arts Computing: A Journal of Digital Humanities* 7.1/2 (2013): 176–189.
Rundstrom, Robert A. "GIS, Indigenous Peoples, and Epistemological Diversity." *Cartography and Geographic Information Systems* 22.1 (1995): 45.
Shaw, Adrienne. "Putting the Gay in Games: Cultural Production and GLBT Content in Video Games." *Conference Papers -- International Communication Association* (2008): 1–29.
Skeggs, Beverley et al. "Queer as Folk: Producing the Real of Urban Space." *Urban Studies* 41.9 (2004): 1839–1856.
Soja, Edward W. *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. 2nd edition. London; New York: Verso, 2011.
Warf, Barney, and Santa Arias. The Spatial Turn : *Interdisciplinary Perspectives* / Edited by Barney Warf and Santa Arias. London : Routledge, 2009
Wood, Denis, John Fels, and John Krygier. *Rethinking the Power of Maps*. New York: The Guilford Press, 2010.

Analyzing Social Networks of XML Plays: Exploring Shakespeare's Genres

Lawrence Evalyn
lawrenceevalyn@gmail.com
University of Toronto, Canada

Susan Gauch
segauch@gmail.com
University of Arkansas, United States of America

Manisha Shukla
mshukla@email.uark.edu
University of Arkansas, United States of America

Introduction

Our inquiry considers the speech interactions of characters within plays as a proxy for broad narrative structures. We analyze computationally-generated social networks of 37 plays by Shakespeare to see whether, and how, they can be used to distinguish between Shakespeare's comedies, tragedies, and histories.

Because dramatic performances enact social encounters, social network analysis translates surprisingly well to fictional societies. Stiller et al. have shown that social networks in Shakespeare's plays mirror those of real human interactions, particularly in size, clustering, and maximum degrees of separation (2003). However, as fictions, these networks are shaped not only by sociological principles, but also by narrative structures. Moretti uses social networks to examine the plots of three Shakespearean tragedies, and to contrast the structure of chapters in English and Chinese novels (2011). Alberich et al. (2002) and Sparavigna (2013) also discuss the interplay between social and narrative constraints on networks. We emphasize this distinction to look for specifically literary features of our networks.

Recent papers presented at DH2017 sought ways to richly quantify the details of one or two plays (Fischer et al., 2017; Tonra et al., 2017). At another scale, Algee-Hewitt examined 3,439 plays by looking only at the Gini Coefficient of each play's eigenvector centrality (2017). With our three dozen plays, we attempt to strike a fruitful middle ground in the inevitable balancing act between detail and scale. Each play is considered individually, but at a level of abstraction which allows rapid and direct comparisons.

Creation of social network graphs

Our parser tracks characters present on stage during speech. This approach is highly extensible: it can parse any play that follows TEI P5 guidelines for performance texts. Each speaking character is connected to all cha-

racters currently present on stage. These connections are recorded in a network graph, with characters as nodes and shared speech as edges. Edges are directional, and weighted based on the number of lines spoken. In future, we plan to extend our parser to identify the specific addressees of a character's speech, allowing us to model more detailed interactions.

To verify that our parser is accurate, we compare our generated network of Hamlet to Moretti's well-known handmade model of that play (2011). Despite some minor differences in peripheral characters like "Servant", and our less-minor difference of including the play-within-the-play, the two networks are highly similar. Our network graph supports Moretti's reading. Our tool also improves on Moretti's model by adding direction and weight to each connection. Although this level of detail turned out not to be necessary for the basic task of using network graphs to distinguish between Shakespeare's genres, it may be useful in future work examining a less homogenous corpus of plays, or in work asking different questions about this corpus.

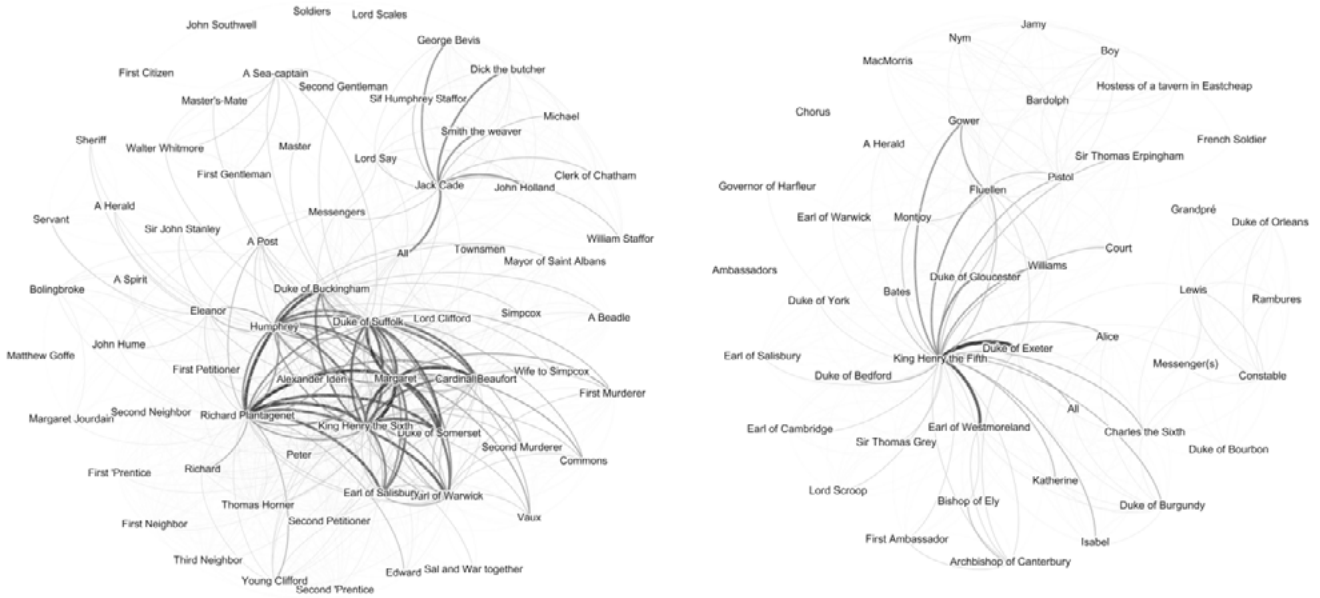
Using networks to identify genre

We then use our generated network graphs to test our central question: whether the social network enacted by a play's characters can be used as a proxy for features of the play's narrative content. More specifically, we ask whether social networks can be used to distinguish between the dramatic genres of tragedy, comedy, and history. Using a support vector machine with fivefold validation, we tested 17 different mathematical features of the networks. No single feature was independently sufficient to identify the genre, though graph density came closest (83% accuracy). However, if features are used in combination, the network graphs can indeed achieve full accuracy. One combination of features which does achieve 100% accuracy is edges, words, and degree. We are currently exploring other combinations that might also be capable of accurately identifying genres.

Discussion

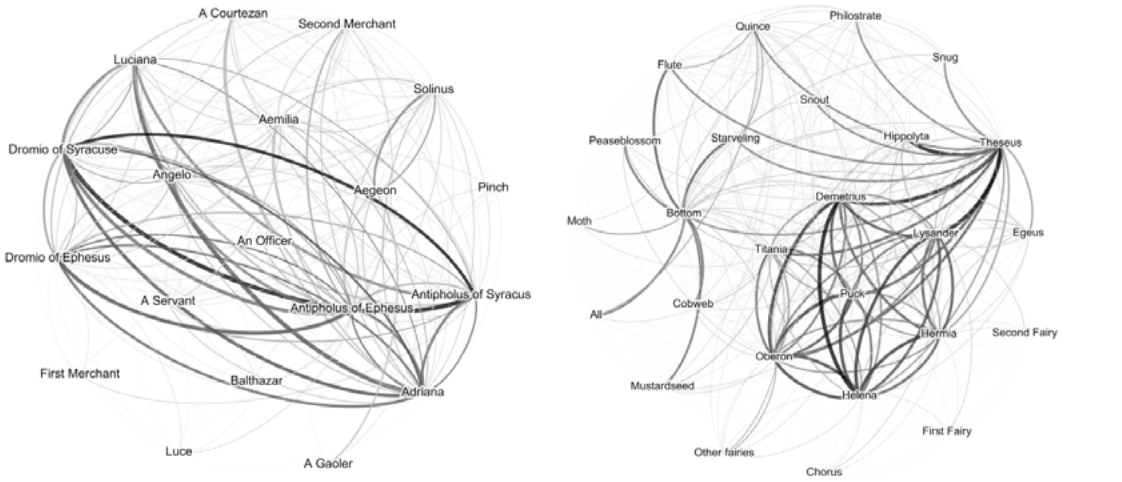
History, comedy, tragedy

The potential utility of graph density in distinguishing genres is visually obvious when individual comedy and history networks are compared. Histories feature highly dispersed networks, with large numbers of very minor characters, such as "First," "Second," and "Third" members of groups like soldiers and ambassadors, who each interject briefly in a single scene. Connections form chains of acquaintance with little overlap, so even the monarchs have low eigenvector centrality.

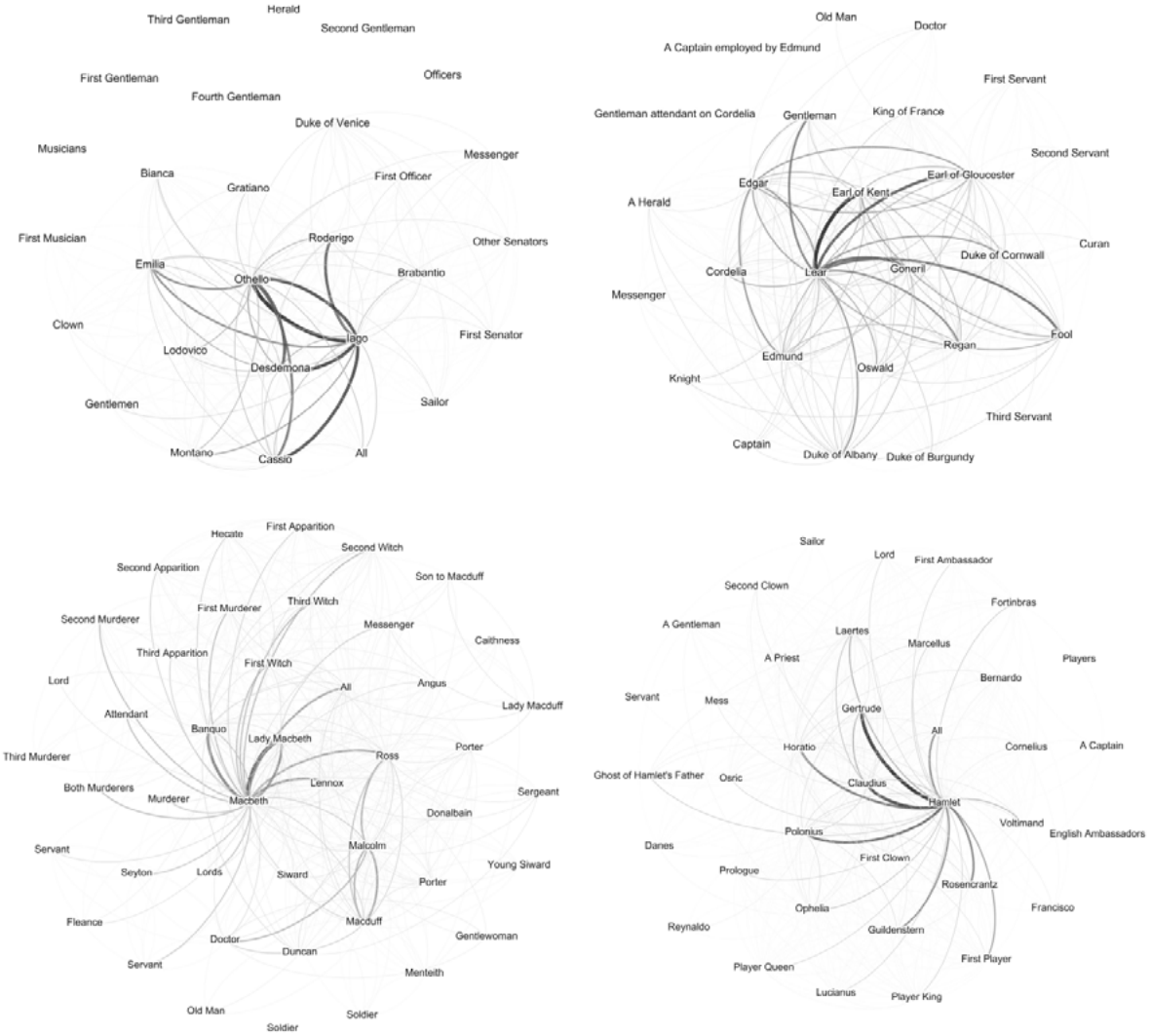


Social network graphs of the histories *Henry VI, Part 2* and *Henry V*.

Comedies, in contrast, feature networks with far fewer characters, in which nearly everybody speaks to nearly everybody else at some point. Although comedies often have multiple subplots, these separate stories do not result in highly-separated networks. We theorize that comedic networks are strongly shaped by the plays' final "resolution" scenes, which bring together the full cast. The average eigenvector centrality of the characters in comedies is much higher than in tragedies or histories; this suggests that many more of the characters in a comedy are "important," reflecting a focus on ensemble stories.



Social network graphs of the comedies *The Comedy of Errors* and *A Midsummer Night's Dream*.



Social network graphs of the tragedies *Othello*, *King Lear*, *Macbeth*, and *Hamlet*.

Graph density is insufficient, however, to fully distinguish the tragedies, which feature networks somewhere between history and comedy in their density. They often have a dense core with a secondary ring of more peripheral characters. What seems to distinguish them is the existence of the central tragic hero, whose influence directly touches more of the network than the protagonists of histories, but whose connections are less interconnected than the ensembles of comedies. These subtleties are better captured, it seems, by the combined metric of "edges, degree, and words."

The "problem plays"

We then use our preliminary identification of each genre's features to examine Shakespeare's various contested genres. Training our model only on the plays for which there is strong consensus, we applied it to the "Roman plays," the "problem plays," and the "romances" in turn. Of the Roman plays, all but *Antony and Cleopatra* are identified as tragedies by every metric; *Antony and Cleopatra* is identified by "edges, words, and degree" as a history and by "degree, modularity, and density" as a comedy. Of the problem plays, *All's Well that Ends Well* is always identified as a comedy; *Troilus and Cressida* and *Measure for Measure* are both identified as a comedy by all metrics except for "edges, criticality, and degree", which identify them as tragedies. The four romances, despite visually unusual networks which support literary arguments that Shakespeare's writing had grown more experimental at the end of his career, are identified as comedies by every mathematical metric. We treat none of these identifications as definitive declaration of the plays' "real" genres, but use them to distinguish between plays whose generic ambiguity lies in their subject matter, and plays whose ambiguity lies in their structure.

Conclusion

Our parser successfully and rapidly produces sophisticated social network graphs of TEI plays that can be used to computationally identify theatrical genre in Shakespeare's plays. Thirty-seven plays is a small scale for this approach: since the parser is highly extensible and can be used with any plays encoded in TEI, future work need not be restricted to the Early Modern period. It need not even be restricted to works written in English. Our networks of the well-studied works of Shakespeare can provide a baseline against which to contextualize analysis of these elements in works for which there is far less critical consensus.

References

Alberich, R.,Miro-Julia, J., and Rosselló., F. (2002). Marvel Universe Looks Almost Like a Real Social Network. arXiv:cond-mat/0202174v1

Algee-Hewitt, M. (2017). Distributed Character: Quantitative Models of the English Stage, 1500-1920. *Digital Humanities 2017: Book of Abstracts*. Montreal: McGill University and Université de Montréal, pp. 119–21.

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., and Trilcke, P. (2017). Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts. *Digital Humanities 2017: Book of Abstracts*. Montreal: McGill University and Université de Montréal, pp. 437–41.

Moretti, F. (2011). Network Theory, Plot Analysis. *New Left Review,* 68: 80–102.

Sparavigna, A. C. (2013). On Social Networks in Plays and Novels. *International Journal of Sciences*, 2: 20–25.

Stiller, J., Nettle, D., and Dunbar, R. I. M. (2003). The Small World of Shakespeare's Plays. *Human Nature*, 14(4): 397–408.

Tonra, J., Kelly, D., and Reid, L. (2017). Personæ: A Character-Visualisation Tool for Dramatic Texts. *Digital Humanities 2017: Book of Abstracts*. Montreal: McGill University and Université de Montréal, pp. 627–30.

Resolving the Polynymy of Place: or How to Create a Gazetteer of Colonized Landscapes

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Diane Katherine Jakacki
diane.jakacki@bucknell.edu
Bucknell University, United States of America

In working with British colonial records and German church manuscripts of colonized and missionized landscapes in the North American mid-Atlantic, the authors have grappled with the problem of polynymy in their attempt to create a gazetteer of places. As Presner and Shepard (2016) have argued, unlike conventional positivistic approaches to mapping, DH and geohumanities have developed a rich vocabulary with which to describe and analyze the human perception of place. Whether through "deep maps" that recount the stories of place and experience or through the multiple layers of temporally inflected information, the spatial turn has revealed the need to see the practice of mapping as "arguments or propositions that betray a state of knowledge." (Presner and Shepard 2016, 207). However, whereas there are sophisticated models of temporal-spatial mapping now available to DHers working with historical materials, to date little critical attention has been paid to the place/person variable. The work of Ann Knowles (and her students) has paved the way for sophisticated representations of the experien-

ce of place (Knowles 2008; 2015). In her arguments for a nonpositivistic geo-practice within the humanities, Knowles has opened up the field to the "fuzzy data" of critical humanistic inquiry. Privileging design over data, Knowles' prize-winning visualizations of the Holocaust challenge us to reconsider in sophisticated ways the experience of landscapes. (Knowles 2014) On a similar path, as Presner and Shepard conclude, virtual reality and gaming allow for an experiential and avatar-based investigation of dynamic, embodied, albeit presentist, multiple perspectives of place. Students at Bucknell have already produced sophisticated critical cartographical visualizations of the Susquehanna river in the Colonial period that draw in part on Knowles' perspectives. This paper will explore the problem of creating a gazetteer of colonized landscapes, specifically those of the mid-Atlantic in the 18th century, in which the name of a place (toponym) changes depending on the person or political entity who is describing that place. In colonized landscapes, there can be multiple names for one place. Maps of this period are veritable palimpsests of conquests and defeats; and travel diaries, mission records and letters contain accounts of human experience of places that are multiply identified. The task is made more complicated still when one factors time into the equation: when competing spatial identities persist across generations. Using the case study of the research project "Moravian Lives" we will ask how we can create a gazetteer of places using authority IDs, when that very authority is itself the product of apolitical-historical struggle. "Moravian Lives" is an international collaborative DH project that aims to make available to the scholarly and lay community the vast corpus of life writings of members of the Moravian Church from the mid-18th century to today (http://moravianlives.org). Facing the simultaneity of multiple names for a place, can we create a system of "triples" that satisfactorily reflects the multiple perspectives and presence or absence of agency of those who name place? Drawing on the substantial cultural-historical GIS of the Susquehanna river produced by Faull and a team of Bucknell staff and students that supported the Department of the Interior designation of the Susquehanna River as a National Historic Water Trail in 2012, the Moravian Lives gazetteer aims to provide the most comprehensive place-name resource for researchers in many fields. The construction of an historical gazetteer for Moravian Lives involves complexities that arise from not only the naming of places but also how their spatial identities reflect respective, concurrent relationships to those places by Native American peoples, Moravian missionaries,and colonial representatives. There are multiple names for a single place as well as multiple understandings of place names, and these differences depend on who it was who did the naming. An example of this challenge is 18th-century Shamokin in Pennsylvania. Shamokin was at that point an Iroquois settlement at the confluence of the north and west branches of the Susquehanna River, encompassing the shores of both branches and an island at the river's fork. To Shikellamy, an Oneida emissary of the Six Nations of the Iroquois or Haudenosaunee, who oversaw the Algonquin-speaking nations of the Lenni Lenape, Shawnee, and Mahican in Iroquoia (present-day Pennsylvania and New York), and who lived in the town in the 1740s, "Shamokin" would have constituted the whole area of the rivers' confluence. To Count Nikolaus von Zinzendorf, the founder of the Moravian Church who visited Shikellamy in 1742, "Shamokin" represented an opportunity for Moravian missionaries offered to them by Shikellamy in the form of space for a blacksmith's shop and mission. While the location of that mission was small, it loomed large in Zinzendorf's interest in founding "Heiden-Collegia", or colleges of the "heathen", in Pennsylvania. To Conrad Weiser, a German settler and negotiator between the colonial government in Philadelphia and the Indian nations, and who worked with Shikellamy on several treaties between the Iroquois and the Colonial government, "Shamokin" would have represented a strategic and ultimately military outpost that would become the site of Fort Augusta during the French and Indian War. These "Shamokins" co-existed, with Native American, Moravian, and Colonial inhabitants and visitors relating to it in discrete yet overlapping ways. One byproduct of our work on the gazetteer could thus be the proposition of authority lists to the OCLC's VIAF council, thereby introducing and linking our information where there is currently no match. In compiling a gazetteer we realize that there is already a VIAF authority ID for Shamokin that is recognized by the Library of Congress/NACO but refers to another (modern) place called Shamokin some 18 miles to the east. (Shamokin, PA VIAF ID: 146606881 (Geographic). We cannot therefore "re-mint" an authority name for these Shamokins. Furthermore, a part of the 18th century Shamokin is now Sunbury (the site of Fort Augusta and Shikellamy's grave) also has its own VIAF ID, (3 Sunbury, PA VIAF ID: 123181256 (Geographic) but, for the historical and cultural studies scholar, it might be inaccurate, misleading, and in some ways irresponsible to equate Sunbury with or consider it as a variant for the historic Shamokin. How can we recognize spatial multivalence (or "polynymy") in the Moravian Lives gazetteer? How does the scholar act responsibly while acknowledging their own potential complicity in political-historical renegotiations and multiple cultural understandings of place? In effect, must we not push back at the idea of *an* authority, and work toward a system that recognizes and synchronizes multiple authorities? We propose a two-phased approach to developing the Moravian Lives gazetteer, which will expand geographically to places beyond North America and will need to resolve polynymic complexities in Central Europe, the Arctic areas of Greenland and Newfoundland, the Caribbean, South Africa and Australia. The first phase involves "stabilizing" all of the place names without giving primacy to any one of them. Each would be assigned a unique HTTP URI offering information about each toponym pertinent to its

own cultural relationships and link to its siblings. In this way we can push back against the need to choose one authority (whether it be restoring an indigenous name or opting among European ones) and demonstrate that these names are not "same as" or "variants" of the others. This, in turn, allows us to reflect upon colonial places in a much more nuanced way that takes into account geographical features and proximity (viz. 'Peace huts on the Susquehanna', 'an der Höhle bei Bethel'). It also enriches the companion personography under development for Moravian Lives. In the visualization already available through Moravian Lives, each person is associated with place using a single-point Google location (see Figs. 2 and 3); but by integrating the cultural historical mapping already completed for the Susquehanna river project, we can now connect these people with better suited vector data referencing each unique place's footprint or range at the same time acknowledging that our identification involves a consideration of certainty (or "fuzziness") by the editor. Through this process, we will strengthen the interlinking of temporo-spatial data within the Moravian Lives project, weaving together the text-based gazetteer with the mapped data. The second phase is to submit our set of authority files to the OCLC and its VIAF council through a member advocate (such as the Moravian Archives). Our work will then be reviewed, assessed against existing identified geographic places in the VIAF database, and where appropriate we hope that new VIAF IDs will be minted. In this way we will make these places discoverable to other researchers considering similarly complex cultural landscapes.

References

Faull, Katherine. ""Digital Lives: Reading Moravian Memoirs in the Age of the Internet" Short paper, *DH 2017*. Montreal, Canada.

---. "Charting the Colonial Backcountry: Joseph Shippen's Map of the Susquehanna River." *The Pennsylvania Magazine of History and Biography*, vol. 136, No. 4, 2012, 461-465.

---. "Writing a Moravian Memoir: The Intersection of History and Autobiography" in *Life Writing and Lebenslauf: Pillars of an Invisible Church*, eds. Christer Ahlberger and Per van Wachtenfeld, Artos Publishers, 2017.

Grumet, Robert. *Manhattan to Minisink. American Indian Place Names in Greater New York and Vicinity*. Norman, OK: University of Oklahoma Press, 2013.

Horsman, Stuart. "The Politics of Toponyms in the Pamir Mountains." *Area*, vol. 38, no. 3, 2006, pp. 279–291. JSTOR, www.jstor.org/stable/20004545.

Jakacki, Diane and Janelle Jenstad. "Mapping Toponyms in Early Modern Plays with the Map of Early Modern London and Internet Shakespeare Editions Projects." *Early Modern Studies and the Digital Turn*. Laura Estill, Diane Jakacki, Michael Ullyot, eds. Malden, MA: ITER. 2016

Meredyk, Steffany, Bethany Dunn, under the supervision of Katherine Faull. "A Corridor of Fear: Stories along the Susquehanna River, 1754-1768". *Stories of the Susquehanna Valley project*. 2013. Stable URL: http://bit.ly/2iXbJzA

Knowles, Ann. "Inductive Visualization: A Humanistic Alternative to GIS" (2015). *GeoHumanities* 1(2), pp. 233-265. DOI 10.1080/2373566X.2015.1108831.

---. *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (2008), edited by Knowles, digital supplement edited by Amy Hillier. Redlands, Cal.: ESRI Press Knowles, Ann, Tim Cole, and Alberto Giordano, eds. Geographies of the Holocaust. Bloomington, IN: U of Indiana Press. 2014.

Moravian Lives project website: http://moravianlives.org/

Oetelaar, Gerald A., and David Meyer. "Movement and Native American Landscapes: A Comparative Approach." *Plains Anthropologist*, vol. 51, no. 199, 2006, pp. 355–374. OCLC Virtual International Authority File webpage: http://www.oclc.org/en/viaf.html

Presner, Todd and David Shepard, "Mapping the Geospatial Turn" in *A New Companion to Digital Humanities*, eds. Susan Schreibman, Ray Siemens, and John Unsworth, (Oxford: Wiley Blackwell, 2016) 201-212.

Radding, Lisa, and John Western. "What's In A Name? Linguistics, Geography, And Toponyms." *Geographical Review*, vol. 100, no. 3, 2010, pp. 394–412. JSTOR, www.jstor.org/stable/25741159.

Audiences, Evidence, and Living Documents: Motivating Factors in Digital Humanities Monograph Publishing

Katrina Fenlon
kfenlon2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Megan Senseney
mfsense2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Maria Bonn
mbonn@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Janet Swatscheno
jswatsc2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Christopher R. Maden
crism@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Introduction

How humanities scholars communicate their research - with one another, with interdisciplinary communities, and with diverse publics - continues to shift with the emergence of new publishing models. We do not understand enough about why scholars choose to publish in different modalities, or what the implications of their choices are for the use, evaluation, and sustainability of research. Thus, publishing systems and services lag behind the advance of digital methods and modes of communication.

This paper presents selected results of a multimodal study of humanities scholars' digital publishing needs. Building on national survey of humanities scholars in the United States, initially reported at DH2017 (Senseney et al., 2017), this paper describes preliminary outcomes of a series of interviews with humanities scholars who have a manifest interest in experimental digital publishing. This study seeks to deepen our understanding of scholarly goals for digital publication.

Outcomes of this study are guiding the development of a service model for library-based humanities publishing, as part of the Publishing Without Walls (PWW) project (http://publishingwithoutwalls.illinois.edu/). Funded by the Andrew W. Mellon Foundation, the University of Illinois Library is leading the PWW initiative in partnership with the Graduate School of Library and Information Science, the Illinois Program for Research in the Humanities, and the African American Studies Department at the University of Illinois. PWW aims to develop a scalable, shareable model for monograph publishing within libraries, with the goal of bridging gaps in current publishing systems, such as gaps between the complex materials scholars want to publish and what existing systems can accommodate, between scholarly practices and existing publishing tools, and between publishing opportunities at resource-rich and under-resourced institutions.

This paper focuses on humanities scholars' motivations for publishing digital, open access, and multimedia monographs. We explore three central motivations for digital publishing: (1) the desire to reach diverse audiences; (2) the desire to integrate interactive, multimedia, and linked evidence; and (3) the desire to publish "living" documents. These factors have implications for digital humanities scholars in understanding the impact of different modes of sharing, for libraries seeking to support digital scholarship, for data models underlying enhanced publications, and for publishing service models.

Methods

This study comprised a set of semi-structured interviews with humanities scholars. Interview participants were self-selected from among scholars who had already participated in the PWW initiative in some way, whether by attending publishing workshops or submitting to the new series. Nineteen interviews have been conducted to date; more are planned for summer 2018. All interviews are recorded and transcribed, and a formal analysis of resulting transcripts is underway. Participants are all affiliated with academic institutions. They include faculty, postdoctoral research associates, and academic professionals with backgrounds in humanities disciplines, information science, and communications.

Three motivations for enhanced digital publishing

Multiple audiences

Scholars turn to open access (OA) monograph publishing to increase impact by reaching more readers, not only within their disciplines but also cross-disciplinary peers and the general public. Visibility and broad dissemination are established motivations for OA book publishing; evidence suggests that these motivations are rewarded, as OA books receive significantly more usage and citation than non-OA counterparts (Emery et al., 2017). Yet, our study indicates that humanities scholars want more than to reach large audiences. They want to reach diverse audiences, ranging from peers in other disciplines to practitioners, policymakers, and the public. Despite potential impact, participants acknowledged that certain prevalent models of OA monograph publishing suffer from a lack of "institutional weight" and "automatic audiences." However, participants described leveraging their own social and research networks to promote their work directly.

Interactive, multimedia, and linked evidence

Authors pursue opportunities for representing new kinds of evidence in new contexts. The potential benefits of multimedia publishing are largely unrealized in publishing practice due to the challenges of managing complex digital publications (Jankowski et al., 2012). Scholars want to integrate or actionably link to more kinds of evidence, including multimedia sources, interactive visualizations, data sets, and curated collections. They also want to make their sources interactive, to allow readers opportunities to visualize, explore, and assess bodies of evidence while anchoring them to narrative descriptions and interpretations. One participant described his primary goal for multimedia publishing as making evidence "come alive in a narrative history."

Living documents

Some humanities scholars want to publish what participants call "living," evolving documents —works-in-progress that are subject to indefinite change. Participants value immediacy of entrance into ongoing scholarly dialogue, both for obtaining rapid feedback from peers and for flag-planting. Some participants see self-publication as a route toward obtaining high-quality peer review more quickly than through the conventional publication; the complexity of peer review in interdisciplinary settings — like the digital humanities — can lead to dilatory, frustrating review processes, which one participant compared to "the phenomenon of too many cooks in the kitchen," and which may yield "diluted" end work. The ultimate manifestation of a "living" document is a publication that facilitates ongoing co-authorship, annotation, interlinking, and revision. One participant described an ideal publication as an online document that "people can comment on, that can directly link to its sources and other people can link to it, that has an attached data set of results that other people can make use of and check," and which is subject to versioning. He described this as an evolving or living document and noted that, "at the moment, most of our research papers are dead documents."

Future work

While openness is a core value of digital humanities scholarship (albeit with qualifications see, e.g., Spiro, 2012), it is not clear how different modes of publication can most effectively open humanities research: to the stratified audiences identified in this study, to deep interaction with sources, and to ongoing evolution. This paper describes outcomes of our study on what humanities scholars need from the next generation of publishing systems and services, and how this study is guiding development of a new model for library-based publishing that can support and sustain highly diverse and broadly impactful research products.

References

Emery, C., Lucraft, M., Morka, A., and Pyne, R. (2017). *The OA effect: How does open access affect the usage of scholarly books?* Springer Nature.

Jankowski, N., Scharnhorst, A., Tatum, C., and Tatum, Z. (2012). Enhancing Scholarly Publications: Developing hybrid monographs in the humanities and social sciences. *Scholarly and Research Communication*, 4(1).

Senseney, M. F., Velez, L., Maden, C. R., Swatscheno, J., Bonn, M., Green, H., and Fenlon, K. (2017). Informing library-based digital publishing: A survey of scholars' needs in a contemporary publishing environment. Presented at the Digital Humanities (DH2017), Montréal, Canada.

Spiro, L. (2012). "This Is Why We Fight": Defining the values of the digital humanities. *Debates in the Digital Humanities*, 16.

Mitologias do Fascínio Tecnológico

Andre Azevedo da Fonseca
azevedodafonseca@gmail.com
Universidade Estadual de Londrina (UEL), Brazil

A cultura digital do século XXI tem sido marcada pela ascensão de um imaginário mágico em relação ao poder das tecnologias. Por meio de uma produção monumental de símbolos, as indústrias culturais e a publicidade das mais diversas empresas de tecnologia têm veiculado mensagens a fim de relacionar o consumo tecnológico à conquista progressiva da autonomia, da liberdade, da felicidade e, em última instância, da transcendência. Este imaginário que induz à devoção das tecnologias parece seduzir as novas gerações com a promessa da elevação dos seres humanos à condição de semidivindades a partir do consumo físico e simbólico de produtos e marcas.

No entanto, sob o brilho deste deslumbre, o Estado e as corporações têm se movimentado no sentido de empregar recursos tecnológicos de forma sistemática para aprofundar o controle social de natureza tecnocrática, de modo que cidadãos e consumidores são observados e analisados em sua intimidade. Ofuscados pelo brilho mágico das tecnologias, usuários entregam voluntariamente informações detalhadas de suas personalidades e experiências pessoais para delegar aos algoritmos de inteligência artificial decisões cada vez mais importantes de suas experiências humanas, tornando-se mais vulneráveis a estímulos publicitários e propagandas ideológicas cada vez mais personalizadas e eficientes.

Entre os vários elementos para que o capitalismo informacional lograsse legitimar essa sociedade de controle tecnocrático, observamos uma intensa produção simbólica nas indústrias culturais no sentido de instrumentar a cultura digital com um fabuloso repertório iconográfico para, primeiramente, exorcizar os temores apocalípticos que as tecnologias sem limites haviam inspirado na humanidade – sobretudo após o advento da bomba atômica e da chamada crise da razão – e, em seguida, substituir os antigos temores por uma nova devoção aos mitos tecnológicos. Nesse contexto, mitologias ancestrais que expressavam as maldições divinas decorrentes da desobediência de homens e mulheres que ousaram ultrapassar os limites do conhecimento foram esvaziadas e invertidas, de modo que os consumidores contemporâneos, mais do que apenas perder o medo, passaram a cultuar esses mitos: da maçã proibida do Éden à mação mordida da Apple, do terrível Big Brother de George Orwell ao sedutor Big Brother da Endemol, da maldição do monstro de

Frankenstein à celebração do gênio do cientista impetuoso no imaginário do Vale do Silício.

O objetivo desta pesquisa é compreender essa dinâmica de subversão de mitologias empregadas para superar os temores, atribuir uma conotação religiosa às experiências com tecnologias e, enfim, ofuscar o controle tecnocrático do ecossistema digital. Para isso, sob a perspectiva da Comunicação, da História Cultural e dos estudos de mitologia e imaginação social, analisamos um conjunto de símbolos evocados na imprensa, no cinema e na publicidade de empresas de tecnologia contemporâneas, situando-as no contexto histórico da utilização de arquétipos e mitologias na publicidade a partir do final do século XX. Como resultado, identificamos um conjunto de mitos e imagens arquetípicas manipuladas nas mídias para associar o consumo tecnológico ao imaginário sagrado da superação do pecado original, da reconquista do paraíso e da transcendência da condição humana.

Referencias

Barthes, R. (2009). *Mitologias*. 4 ed. Rio de Janeiro: Difel.

Chartier, R. (1985). *A história cultural: entre práticas e representações*. Rio de Janeiro: Difel/Bertrand Brasil.

ELLUL, J. (1964). *The technological society*. New York: Vintage Books.

JUNG, C. G. (2000). *Os arquétipos e o inconsciente coletivo*. 2 ed. Petrópolis: Vozes.

ROSZAK, T. (1972), *A contracultura: reflexões sobre a sociedade tecnocrática e a oposição juvenil*. 2 ed. Petrópolis: Vozes.

TURNER, F. (2006). *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago: The University of Chicago Press.

Latin@ voices in the Midwest: Ohio Habla Podcast

Elena Foulis
foulis.5@osu.edu
OSU, United States of America

In recent years, the use and development of Podcasts has significantly grown. Podcasts allow us to listen to topics we are interested in and learn more about an issue or community. Podcasts like *This American Life*, *Radio Ambulante* and *Latino USA*, put at the center of their stories experiences of people and places. Indeed, using audio as a medium to tell the larger stories of our community has proven successful as signaled by all top 10 iTunes podcasts—5 of which are documentary style. Creating university based podcast like, *Ohio Habla*, will allow us to connect and learn more the Latin@/Hispanic experiences locally, while amplifying the voices of the community everywhere. Language and cultural studies are in a unique position to utilize this medium to advance the understanding of how culture and language is both transmitted and analyzed.

The *Ohio Habla* podcast is primarily produced by students in advanced Spanish language and Latin@ studies classes together with their professor. Each student plans, researches, secures a podcast guest and carries out the interview. Students are able to continue to develop their written, reading, speaking and listening skills and they are responsible to produce one whole 30-45 minute podcast.

Ohio Habla is an extension of the digital oral history project, ONLO (oral narratives of Latinos/as in Ohio), however, it focuses topics, rather than life history. On the other hand, in the case of Latin@ students, they collect family stories, instead of interviewing a member of the community. Podcasting can help document issues that are of interest to our community and potentially be able to share it more widely. Finding new and real ways to use language and storytelling is of great benefit to our students, and podcast in the foreign language classroom can accomplish this. Our own teaching methodology here at Ohio State encourages second language learners to use the language communicatively and in real situations that are as authentic as possible. Podcasting is a great way to use the language in real and creative ways, and most importantly, in community—an element that is often left out the foreign language classroom for various reasons (mainly, time).

Teaching methodologies

As a pilot project, this use of podcasting in the classroom may pave the way for further research opportunities about the benefit of podcasting in advanced language courses, service-learning and heritage language learners. This course enhancement will also provide students with a structured opportunity to make deeper connections with Latino/a campus community, to reflect on that experience, and to gain interviewing skills that will serve them in the future. Additionally, students will be instructed in (1) Language and intercultural skills 2) Organizational and professional skills through the research of a topic, securing a guest that can speak about the topic, preparing the guest with agreed up points of conversation, and practicing before recording the interview (3) Technical skills through using recording equipment and editing software.

Spotting the Character: How to Collect Elements of Characterisation in Literary Texts?

Ioana Galleron
ioana.galleron@univ-paris3.fr
U. Sorbonne Nouvelle Paris 3, France

Fatiha Idmhand
fatihaidmhand@yahoo.es
U. de Poitiers, France

Cécile Meynard
cecile.meynard@gmail.com
U. d'Angers, France

Pierre-Yves Buard
pierre-yves.buard@unicaen.fr
U. de Caen

Julia Roger
julia.roger@gmail.com
U. de Caen, France

Anne Goloubkoff
anne.goloubkoff@unicaen.fr
U. de Caen, France

What is a literary character made of? To this question, a pragmatic answer is to say that it exists as a result of a chain of different linguistic elements, scattered throughout the text. The aim of this paper is to propose a digital method for collecting these elements, so as to analyse their nature, to observe their repartition in texts, and, ultimately, to contribute to a deeper understanding of the functions the literary device called "character" assumes in a text.

Projects dedicated to named-entity recognition put a great deal of effort into using Natural Language Processing (NLP) techniques for identifying names of people, places and organisations mentioned in various types of discourses, especially political ones, as well as the co-referential chains built on the basis of these names. However, in spite of important advances in the field, much remains to be done in order to train the computer to link correctly various phrases referring to the same entity, as well the pronouns pointing to it (see Schnedeker and Landragin, 2014). In our case, we are interested in such elements of a co-referential chain that bear characterization features, and this is, inevitably, a supplementary complication. In addition, we are interested in certain elements (eg. "his brother" in the phrase "John is his brother") that are often left aside in named entity recognition, as performing another functions than strictly pointing towards an entity. Therefore, NLP techniques did not appear adapted to our needs.

We will therefore resort to "crowd-reading", as another means, offered by the explosion of the digital sphere, to make sense from texts. Very similar to the crowdsourcing, the crowd-reading asks to benevolent contributors to annotate a document, bringing in their own view and understanding, instead of transcribing, or adding in information based on a (sometimes external) form of authority. Considering the nature of the work to be done, the crowd-reading appeared as a valid technique in our case.

In a first stage, we submitted a short text (Julio Cortazar "Continuidad de las parques") to the manual annotation of a hundred students from our universities. This brought to the fore the sheer variety of elements considered to be participating to the characterization of a literary "person" (nouns and adjectives, of course, but verbs and adverbs too), as well as the need to dispose of a controlled vocabulary allowing to understand what kind of characterization each respondent attached to the various strings of characters selected as participating to this function.

In a second phase, we have decided to build an interface, offering a more ergonomic experience to our respondents, and allowing us to extract automatically the linguistic elements selected, as well as to group them by categories. Built with XML Mind, this interface is in fact based on a text lightly encoded with TEI tags, in which our respondents add, every time they select a string of characters, an <rs> tag, bearing in addition two attributes:

a @key attribute, defined by each respondent every time he or she encounters a new character. The keys are subsequently available for reuse in the rest of the text. We expect the number of keys to vary considerably from a reader to another.

an @ana attribute, with a set of constrained values. Based on another project of character analysis, these values have been defined in Galleron, 2017, and cover aspects such as the ontological type of a character, its sex, age, family position, nationality, occupation, and so on.

The text submitted to annotation has been changed for this second experience: it concerns now the "Jardin aux sentiers qui bifurquent" ("Jardín de los senderos que se bifurcan") by Jorge Luis Borges. At the date of this proposal, the second campaign of crowd-reading has not started yet. We'll have a significant number of answers before the DH conference. Our respondents will be recruited again amongst the students enrolled in literary studies in our universities: while they have a certain level of training in linguistics, literature and poetics, so as to be able to recognise the type of linguistic elements we look for, their reading still remains close of the "non-informed", "amateur" reading of the "man in the street" (see Dufays, 205).

The results will be analysed so as to observe what kind of linguistic units have been identified most often, and what kind of values of the @ana attribute have been mobilised most often. We will further discuss the divergences between the selected elements, and those we were expecting to be selected. This will allow us, on the one hand, to suggest a possible use of our interface as a remediation tool in literary studies, for students with difficulties in extracting pertinent information from a text, so as to respond a specific task. On the other hand, we will advance an hypothesis about the observed distribution of the most frequent elements of characterization, that are far to appear where, intuitively, one would expect them to

be grouped together (so as to "introduce" the character) as shown by our first campaign of crowd-reading, and by our own annotation endeavours.

References

Dufays, Jean-Louis, Gemenne, Louis, et Ledur, Dominique (2005). *Pour une lecture littéraire. Histoire, théories, pistes pour la classe*, Bruxelles: De Boeck – Duculot.

Galleron, Ioana (2017). Conceptualisation of theatrical characters in the digital paradigm: needs, problems and foreseen solutions. *Human and Social studies*, De Gruyter. 6: 1 (Published Online: 2017-04-18 | DOI:https://doi.org/10.1515/hssr-2017-0007).

Schenedeker, Catherine; Landragin, Frédéric (2014). Les chaînes de référence: présentation. *Langages*, 3:145, 3-22.

Archivos Abiertos y Públicos para el Postconflicto Colombiano

Stefania Gallini
sgallini@unal.edu.co
Universidad Nacional de Colombia, Colombia

El 26 de noviembre de 2016 el Congreso colombiano votó su aprobación al "Acuerdo final para la terminación del conflicto y la construcción de una paz estable y duradera", firmados por el presidente de la República Santos y el comandante del Estado Mayor central de la guerrilla de las FARC-EP Jiménez. El fin legal del largo conflicto armado interno puso en marcha la creación de una nueva institucionalidad estatal para transitar hacia los necesarios procesos de, que lentamente y entre muchas resistencias ha ido tomando cuerpo en el año que siguió a la aprobación parlamentaria.

Dos de los institutos más significativos creados por los Acuerdos de paz son la Jurisdicción Especial para la Paz y la Comisión para el Esclarecimiento de la Verdad, la Convivencia y la no Repetición. Estas instancias se suman a los esfuerzos de la Comisión Nacional para la Memoria Histórica (creada en 2011) y de distintas iniciativas (regionales y locales, públicas y privadas, académicas, cívicas y gremiales) por recopilar, organizar, preservar y a menudo hacer público un complejo acervo de información acerca de la historia y la memoria del conflicto colombiano.

Estos archivos y repositorios – los que ya existen y los que la implementación del Acuerdo creerá a partir de los hallazgos de la justicia transicional, la Comisión de la Verdad y las iniciativas de la sociedad civil y organizaciones de derechos humanos – son las fuentes con las cuales el país apuesta reconstruir las bases de justicia, reparación y no repetición que deberán sostener el nuevo pacto social de la nación.

La situación no es nueva en el escenario global. Durante el siglo XX y lo que va corrido del XXI, muchas veces se han constituido archivos de derechos humanos, de comisiones de la verdad, de memoria de las víctimas al finalizar un conflicto armado interno o una dictadura. Los ejemplos van desde el Cono Sur latinoamericano á Irlanda, Suráfrica y Guatemala, para citar algunos.

Sin embargo, a diferencia de los casos anteriores, los archivos del conflicto armado interno de Colombia se construyeron, consolidarán e interrogarán en pleno auge de la era digital. Esta circunstancia influye de manera radical en cuestiones de adquisición, preservación, seguridad y acceso a la información, pero también implica dos consecuencias importantes: la oportunidad que la dimensión colaborativa y abierta de los archivos en la era digital brinda para alcanzar los objetivos de esclarecimiento de la verdad histórica y judicial, y el protagonismo que la adopción de técnicas y herramientas digitales y de la informática humanística puede jugar para permitir la efectiva apropiación social de los datos.

Los archivos de la era digital son intrínsecamente distintos a sus antepasados. Estos son repositorios participativos, de-institucionalizados, de acceso abierto, de contenidos digitales, de-localizados, que funcionan en red con otros archivos y repositorios documentales, capaces de generar y actualizar continuamente sus formas de hacerse accesible y apropiable por parte de un público heterogéneo.

El conflicto armado interno dejó detrás suyo un enorme volumen de datos que, junto con la complejidad de la gestión de esta información sensible, requiere pensar en metodologías y herramientas tanto archivísticas como informáticas que aseguren la interoperabilidad de los datos, la seguridad de la preservación y no obsolescencia tecnológica de la información, el procesamiento automatizado (incluyendo la georeferenciación) de metadatos, el tratamiento, la transcripción automatizada y codificación de fuentes orales, entre otros aspectos.

El Pensamiento Archivístico crítico y las Humanidades Digitales ofrecen una matriz epistemológica y técnica para pensar los archivos del conflicto y gestionar su información propiciando su visibilidad ante la opinión pública, visualización adaptada a las necesidades de distintos actores, interoperabilidad, traducibilidad en evidencia judicial, entre otros. Se trata de pensar de qué manera tanto las herramientas como la perspectiva cultural de las DH pueden contribuir a esta tarea colectiva.

La ponencia presentará los avances del proyecto de investigación que, con asentamiento en el Laboratorio de Cartografía Histórica e Historia Digital de la Universidad Nacional de Colombia en Bogotá, un grupo de docentes y estudiantes está desarrollando sobre las temáticas descritas, que tiene además el propósito de ofrecer lineamientos de políticas públicas en el tema de los archivos de la historia y la memoria del conflicto armado interno colombiano en la era digital.

Respuestas a evaluaciones – enero 2018

Dos de las evaluaciones sugieren mayor precisión en la propuesta. Agradezco esta oportunidad para poder aclarar que la ponencia presentará los primeros avances de un proyecto de investigación que está apenas empezando y que discute una materia – la nueva institucionalidad de Verdad, Justicia y No Repetición del conflicto colombiano - que también ha sido formalizada hace tres semanas (enero 2018). La participación en DH2018 durante la fase inicial del proyecto justamente apunta a encontrar en el congreso aquella retroalimentación de pares que no es posible siempre encontrar en el ámbito nacional, donde las HD se encuentran en un estadio todavía embrionario, aunque acelerado y entusiasta.

El objetivo principal de la ponencia es por ende presentar críticamente el caso colombiano como ocasión de construcción (a veces) y organización (a veces, cuando los repositorios ya existan) de archivos para la reconstrucción de la memoria, la historia y a menudo la verdad judicial del conflicto colombiano, en un momento histórico en el cual la revolución digital y la expansión de sus consumidores/actores abre el escenario a posibilidades, pero también a desafíos no antes conocidos.

Se tendrán a la vista archivos que ya existen (i.e. Centro Nacional de Memoria. "Archivo Virtual de Los Derechos Humanos Y Memoria Histórica." http: //www.archivodelosddhh.gov.co/, los de matriz periodística como Verdabierta, los de ongs y asociaciones de víctimas, los de historia oral de los movimientos, ver por ej. Suárez Pinzón, Ivonne. "El Archivo Oral de Memoria de Las Víctimas AMOVI-UIS: Un Archivo de Derechos Humanos." UIS y Corporación Compromiso, 2014. https://www.uis.edu.co/webUIS/es/amoviUIS/documentos/presentacionAMOVI-UIS.pdf.), pero también archivos los que se van a levantar a partir de las nuevas indagaciones e instituciones (p. ej. las instancias de la JEP y de la Comisión de la Verdad). En la ponencia será posible referirse a archivos o datasets más en detalle, pero sería apresurado indicarlos en esta fase que es todavía exploratoria. Igualmente, aunque me interesan especialmente algunos problemas "técnicos" (la interoperabilidad y la georeferenciación), la intención de la ponencia es presentar problemas y discutir desafíos, a la luz de una discusión que es álgida en Colombia (Centro Nacional de Memoria Histórica. "Política Pública de Archivos de Graves Violaciones a Los Derechos Humanos, Infracciones a Los Derechos Humanos, Infracciones Al DIH, Memoria Histórica Y Conflicto," February 2015. http://www.centrodememoriahistorica.gov.co/descargas/mesasRegionalesArchivos/Politica-publica-archivos-integrada-20-2-1.pdf).

Los referentes teóricos los he encontrado – como se indica en el Abstract – en el Pensamiento Archivístico crítico (MacNeil, Heather, and Terry Eastwood, eds. *Currents of Archival Thinking, 2nd Edition*. Santa Barbara, California: Libraries Unlimited, 2017; Schwartz, Joan M., and Terry Cook. "Archives, Records, and Power: The Making of Modern Memory." *Archival Science*, no. 2 (2002): 1–19; Weld, Kirsten. *Paper Cadavers: The Archives of Dictatorship in Guatemala*. American Encounters/Global Interactions. Durham: Duke University Press, 2014; Centro Nacional de Memoria Histórica. "Seminario Internacional Archivos Para La Paz." Centro Nacional de Memoria Histórica, 2014. http://www.centrodememoriahistorica.gov.co/centro-audiovisual/videos/seminario-internacional-archivos-para-la-paz) y las Humanidades Digitales.

References

Sanmiguel, Lahdy Diana del Pilar Novoa, and Diego Andrés Escamilla Márquez. "Archivos orales y memoria del conflicto armado interno colombiano: retos y posibilidades." *Advocatus* 14, no. 27 (March 1, 2017): 153–73. http://www.unilibrebaq.edu.co/ojsinvestigacion/index.php/advocatus/article/view/732.

Centro Nacional de Memoria Histórica. "Política Pública de Archivos de Graves Violaciones a Los Derechos Humanos, Infracciones a Los Derechos Humanos, Infracciones Al DIH, Memoria Histórica Y Conflicto," February 2015. http://www.centrodememoriahistorica.gov.co/descargas/mesasRegionalesArchivos/Politica-publica-archivos-integrada-20-2-1.pdf.

Centro Nacional de Memoria Histórica. "Seminario Internacional Archivos Para La Paz." Centro Nacional de Memoria Histórica, 2014. http://www.centrodememoriahistorica.gov.co/centro-audiovisual/videos/seminario-internacional-archivos-para-la-paz.

"Algunas Notas Sobre Los Repositorios Institucionales (Parte I) -." *Infotecarios* (blog), August 22, 2017. http://www.infotecarios.com/algunas-notas-los-repositorios-institucionales-ri-parte-i/.

Colectivo de Historia Oral. "Colectivo de Historia Oral (Colombia)." *Colectivo de Historia Oral* (blog). Accessed November 28, 2017. https://colectivohistoriaoral.wordpress.com/category/historia-oral/.

Jelin, Elizabeth. *Los Trabajos de La Memoria*. Memorias de La Represión 1. Madrid: Siglo XXI, 2002.

Suárez Pinzón, Ivonne. "El Archivo Oral de Memoria de Las Víctimas AMOVI-UIS: Un Archivo de Derechos Humanos." Universidad Industrial de Santader y Corporación Compromiso, 2014. https://www.uis.edu.co/webUIS/es/amoviUIS/documentos/presentacionAMOVI-UIS.pdf.

Brodsky, Marcelo. "Buena Memoria," 1997. http://v1.zonezero.com/exposiciones/fotografos/brodsky/defaultsp.html.

Histórica, Centro Nacional de Memoria. "Archivo Virtual de Los Derechos Humanos Y Memoria Histórica." Accessed November 28, 2017. http://www.archivodelosddhh.gov.co/.

Humanidades Digitales en Cuba: Avances y Perspectivas

Maytee García Vázquez
maytee.garcia.vazquez@gmail.com
Cubaliteraria, Cuba

Sulema Rodriguez Roche
sulema1985@gmail.com
Universidad de La Habana, Cuba

Ania Hernández Quintana
aniahdez@fcom.uh.cu
Universidad de La Habana, Cuba

Hasta hace pocos años, la Web proporcionaba información unilateralmente. Por un lado, estaban las grandes empresas e instituciones, que eran las que poseían espacio en la red, y por el otro, los usuarios, en actitud receptora y pasiva. Esa tendencia está siendo modificada por el movimiento denominado Web 2.0 que propugna que todos somos potenciales surtidores de contenidos y creadores de los registros del conocimiento. La evolución natural en la sociedad de la información se expresa en la metáfora del paso del ciudadano 1.0, consumidor de recursos, al ciudadano 2.0, creador de recursos, evidenciando la horizontalización y democratización de las fuerzas que rigen la red. Las Humanidades Digitales son un resultado de esas transfiguraciones digitales y como un ámbito disciplinar de convergencia cultural e investigativa, dejó de ser una moda para convertirse en una urgencia para cultura y memoria del mundo.

En Cuba, ya se notan de forma clara las comprensiones sobre la necesidad de fomentar este ámbito de teoría y práctica; aprovechando un contexto de crecimiento tecnológico en el país, en el que empieza a notarse la presencia digital en casi todos los sectores de la sociedad, con demandas infocomunicacionales y culturales crecientes y multilaterales. El presente trabajo tiene la finalidad de compartir los avances más visibles, así como las proyecciones hacia lo profesional, lo académico, lo investigativo y lo institucional, como parte de la agenda de las Ciencias Sociales y Humanísticas en Cuba.

Se abordan los conceptos de partida de las Humanidades Digitales para Cuba, como un campo interdisciplinar dispuesto para dar espacio a las reflexiones y prácticas suscitadas por los cambios que produce la introducción de las tecnologías digitales en el universo de la cultura y la información; con énfasis en el desafío epistemológico y metodológico para la articulación de conocimientos y prácticas profesionales y de investigación que enfrentan las ciencias humanas en el cibermundo. Se abordan como una oportunidad de transformación sinergética del consumo cultural, cada vez más urgente, en tanto se demanda mayor conocimiento de investigadores y usuarios, que a su vez demandan información, de forma activa, en espacios colaborativos.

La manera en que se puede trabajar en las Humanidades Digitales, partiendo de los límites casi precarios del desarrollo tecnológico en Cuba, ha creado proyectos sui géneris. Estos son inimaginables en Europa o en Estados Unidos. Varios ejemplos: la circulación de USB, tan común en el paso de los archivos, crea un sistema de distribución de conocimiento muy diferente. Dentro de esos sistemas de distribución, hay una relación política diferente hacia los derechos de autor que cambia la manera en que el conocimiento fluye. Se crean también proyectos digitales que pueden pasarse de máquina en máquina por USB, muy diferentes a aquellos que se colocan en un servidor. Se pueden desarrollar nuevas pistas para el análisis textual, por ejemplo.

En Cuba esto tiene un aspecto político que no siempre se verbaliza, pero son reconfiguraciones de trabajo en equipo que transforma la manera en que la investigación se ha hecho hasta ahora. Eso cambia la relación hacia la investigación a nivel social y su papel en la formación de grupos sociales para el trabajo cultural, ahora más equitativas. De la misma forma, existen jerarquías que vienen de la organización tradicional del trabajo de investigación, en la cual el personal técnico se encuentra separado del investigador y ambos de los bibliotecarios; transitando hacia un modelo colaborativo e interdisciplinar.

Los primeros pasos en Cuba proceden de los años 90 del siglo XX, marcado por un período social complejo en el país, con la publicación del primer libro digital. De forma aislada, varias instituciones académicas y de investigación han realizado proyectos de Humanidades Digitales y finalmente en mayo de 2017, se avanzó hacia una estrategia de articulación, con el primer curso de Humanidades Digitales impartido por profesoras del Laboratorio de Innovación de Humanidades Digitales (LINHD). Uno de los resultados de ese encuentro fue la disposición de crear iniciativa profesional que articule y visibilice el trabajo en Humanidades Digitales que se ha venido realizando en el país.

Se han identificado algunos focos muy visibles, en especial el de la carrera Ciencias de la Información, de la Universidad de La Habana, cuyo propósito es transversalizar las Humanidades Digitales en el campo de las Ciencias de la Información en Cuba. Ese proyecto, con una vocación claramente pedagógica, comenzó a trabajar en noviembre de 2016. Las investigaciones resultantes del primer año tienen como principio que las Humanidades Digitales se interesan por el estudio, preservación y acceso a la información registrada, objetivos que disciplinariamente enfrenta también la comunidad científica, académica y profesional del campo de esta carrera, y que para ello las Humanidades Digitales se distinguen por el uso intensivo de métodos de procesamiento automático y semiautomático, expresados científicamente a través de contribuciones en congresos, experiencias en laboratorios de I+D+I y en programas de formación universitaria.

Asimismo, el equipo se preocupa por las colecciones digitales, y en consecuencia por el requerimiento de protocolos de preservación de sus contenidos, determinados por funciones y estructuras más sofisticadas que modifican los procesos de gestión de esos recursos electrónicos a través de métodos globales como open data, linked data, linguistic link data y TEI.

Es un hecho que las bibliotecas digitales clasifican como uno de los sistemas de información más complejos por la multidisciplinariedad que implican; además, por la convergencia de conocimientos que supone organizar, difundir y usar información en este tipo de repositorios; y especialmente por las complicadas e interdependientes multirrelaciones que activa, que llegan a la autotransformación y a la construcción de contrahegemonías emancipatorias.

Las primeras siete investigaciones del grupo exploraron los conceptos de las humanidades digitales en su multiplicidad y complejidad, las redes profesionales y los currículos de humanidades digitales. Además, se realizaron indagaciones más enfocadas a la solución de problemas como el procesamiento de una revista infantil con carácter patrimonial con el método linked data y la creación de un espacio de aprendizaje colaborativo para estudiantes de Ciencias de la Información.

En Feria del Libro de la Habana, a realizada en el mes de febrero de 2018, se desarrolló un programa especial denominado "Cuba Digital". Libros digitales, aplicaciones móviles, conferencias de investigadores nacionales y extranjeros y proyectos cubanos, entre otros, integraron las propuestas de ese espacio, que contó con la coordinación de la Editorial Cubaliteraria.

La lista de proyectos e instituciones cubanas involucradas en proyectos que apuntan a las humanidades digitales, crece. Un levantamiento preliminar en la capital destaca los siguientes: Instituto de Historia de Cuba; la Fundación Fernando Ortiz con su proyecto Archivo de la palabra; el Instituto de Literatura y Lingüística, dedicado al estudio y descripción del español de Cuba; el proyecto www.postdata.club, del Centro Martin Luther King; la Biblioteca Nacional de Cuba con su catálogo digital, y el proyecto Mirador, en colaboración con Infomed, la red nacional de información en salud en Cuba, enfocado en el rescate de colecciones patrimoniales en Cuba y también aliado del Grupo de Investigación de Humanidades Digitales para las Ciencias de la Información, de la Universidad de La Habana. Convocados por la editorial digital Cubaliteraria, del Instituto Cubano del Libro, que lidera la producción de ebooks y multimedias sobre literatura cubana.

En el futuro cercano, se proyecta una postura sinérgica que aproveche los aprendizajes de las Humanidades Digitales del Sur, con un enfoque parecido a la realidad cubana y se articule en una iniciativa nacional o proyecto de Asociación de Humanistas Digitales.

Corpus Jurídico Hispano Indiano Digital: Análisis de una Cultura Jurisdiccional

Víctor Gayol
vgayol@colmich.edu.mx
El Colegio de Michoacán, A.C., Mexico

El proyecto *Corpus de derecho castellano-indiano / digital* es una propuesta colectiva e interdisciplinaria que abarca la compilación, digitalización, procesamiento, macroanálisis y publicación anotada en línea del conjunto de los textos jurídicos vigentes en el marco de la monarquía castellana entre el siglo XIII y principios del XIX. El núcleo principal del proyecto es la construcción de un modelo para el macroanálisis de estos textos jurídicos y, en consecuencia, la generación de herramientas analíticas y de consulta del corpus que permitan comprender la interrelación entre sus distintos elementos semánticos y conceptuales y su transformación a través de los siglos y así proponer una interpretación de cómo es que posiblemente funcionaban en el contexto del discurso y la práctica en el orden jurídico tradicional de la cultura jurisdiccional, tanto en el ámbito de la doctrina, del ejercicio de la potestad normativa como en el del actuar cotidiano del aparato de gobierno e impartición de justicia.

El proyecto implica diversas conexiones y diálogos en diversos ámbitos. En el ámbito interdisciplinario, entre los historiadores de la corriente crítica (cultural) del derecho, lingüistas, humanistas digitales y programadores; en el ámbito teórico y metodológico, entre dos posturas acaso antagónicas en apariencia: la lectura densa y cercana de los textos jurídicos hecha por la historia cultural del derecho a lo largo de varias décadas y la lectura distante. Lo anterior nos obliga a discutir ciertos principios teóricos, como lectura densa, tomada por la historia cultural del derecho de la idea de descripción densa (Geertz, 1973), como sistema capaz de ser leído como texto en relaciones contextuales, o un nivel más complejo (Genette, 1992) y su noción de transtextualidad. Varios historiadores del derecho han aplicado incluso algo parecido a la lectura cercana del criticismo literario (Clavero, 1991). Esto interesa al estudiar el derecho de antiguo régimen frente a la posibilidad de aplicación de metodologías computacionales enfocadas, generalmente, a una lectura distante (Moretti, 2013) en la búsqueda de estructuras formales mediante el análisis de grandes cantidades de texto/data. Es justamente necesario pensar en la posibilidad de ensayar no sólo una minería de texto cuantitativa sino en aspectos más cualitativos, modelando campos semánticos que se transforman históricamente.

Cabe aclarar que el criterio de selección de fuentes para la conformación del corpus es complejo y presenta muchos problemas. Responde a una historiografía que ha definido el campo de lo jurídico en el antiguo régimen

hispánico como algo más allá del texto jurídico normativo (entendido como ley). Incluye la doctrina de los juristas y de los teólogos por considerarse que la cultura jurídica tiene una estrecha relación con la doctrina católica. El corpus completo abarcaría tanto normas como doctrina y costumbre y se consideran textos jurídicos producidos tanto en Castilla como en los territorios americanos de la monarquía. Por lo tanto, no se trata de un corpus reunido de antemano en su propia época, sino de un corpus compuesto por el conjunto de la comunidad de historiadores dado que se ha analizado su utilización práctica a lo largo de los siglos y en un contexto cultural determinado (Castilla y sus dominios ultramarinos entre los siglos XIII y XIX). Tener claro cómo suponemos que se definía un texto jurídico en el antiguo régimen es de suma importancia ya que el interés del proyecto es generar una comunidad colaborativa de investigación interdisciplinario que determine sus elementos semánticos necesarios para poder caracterizar digitalmente este tipo de textos. Esto es primordial puesto que son textos completamente distintos de los literarios o de otra índole que se han considerado, por ejemplo, en la iniciativa TEI. Dicho de otra forma, el nodo fundamental del problema es cómo se construye un corpus histórico jurídico particular para que sea útil en las humanidades digitales.

Como la reunión del corpus completo es un proyecto a muy largo plazo, en una etapa piloto consideramos que trabajar con los textos normativos puede ser suficiente para ensayar la propuesta de un modelo flexible y escalable. Además, para el caso de los textos normativos ya existe un ordenamiento y un proceso de digitalización previo de esa parte del corpus. De unas 35,355 normas referenciadas se han puesto en línea, de manera digital básica, 26,831 por un grupo de académicos españoles que viene trabajando al respecto desde la década de 1970 y en el que se han ya recogido la mayor parte de las normas legisladas entre el año 1020 y 1868.

Por tanto, el objetivo de esta ponencia es discutir los diferentes ejes de nuestra propuesta teórica: 1) el aspecto de su realidad digital, es decir, cuáles son los requisitos para una digitalización óptima de fuentes jurídicas que se presentan en la realidad de maneras diversas –manuscritas, impresas, cuyos contenidos varían ortográfica y semánticamente a lo largo de los siglos-, 2) el problema de qué se concibe como texto propio de la cultura jurisdiccional en el orden jurídico tradicional –no sólo los obviamente jurídicos en apariencia-, y, en consecuencia, 3) los retos que implica el diseño de herramientas digitales propias que permitan el macroanálisis de los textos como datos masivos. Esto, a su vez, implica un problema mayor y de fondo que es el de la conexión entre un necesario abordaje hermenéutico de los textos jurídicos (lectura densa) en una perspectiva de larga duración –desde la baja edad media hasta el fin de la edad moderna– para entender su contexto cultural de sentido, y el reto de procesar dichos textos entendidos como corpus y en forma de datos masivos mediante computadora (lectura distante), no sólo en procesos de segmentación del corpus para su visualización (nubes de palabras, frecuencias relativas y absolutas, KWIC), sino la posibilidad de ensayar, sobre todo, un modelado tópico semántico con objeto de reflexionar sobre cuál sería un modelo de macroanálisis adecuado para este tipo de corpus. Finalmente, proponer un modelo particular para la edición digital del corpus de los textos jurídicos propios de la cultura jurisdiccional del orden jurídico tradicional.

Referencias

Clavero, B. (1991). *Antidora: antropología católica de la economía moderna*. Milano: Giuffrè.

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Genette, G. (1992). *The architext : an introduction*. Berkeley: University of California Press.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Designing writing: Educational technology as a site for fostering participatory, techno-rhetorical consciousness

Erin Rose Glass
erglass@ucsd.edu
UC San Diego, United States of America

In the past ten years, advancements in computing technology have lent themselves to diverse applications in teaching and learning such as seen with MOOCS, learning managements systems, networked collaborative pedagogy, virtual/augmented reality course modules, and algorithmic-driven approaches to personalized learning. While these engagements represent a variety of exciting (though often controversial) new directions for educational technology, the changing socio-technological conditions of our information landscape call for new critical approaches towards its development and use. Information communication technology (ICT) in educational settings should not only be evaluated according to the way it supports intended learning goals, but also according to the type of technological consciousness it produces in students. In this paper I will draw from methods and values in participatory design (Simonsen), critical pedagogy (Freire; Shor), and the digital humanities (Drucker & Svensson; Rockwell & Sinclair) to outline a way that academic technological practices and infrastructure might be re-engineered to foster more critical and participatory relationships to digital technology within higher education. I will focus specifically on how this approach has particular value for the teaching and use of writing in un-

dergraduate and graduate education in that it enables a praxis-oriented approach to analyzing and designing digitally-mediated rhetorical situations within and beyond academia. I will then describe KNIT, a digital commons at UC San Diego that aims to develop a participatory model of educational technology, and describe the challenges and opportunities experienced in its development.

Participatory approaches to ICT

The general user has little expectation or ability of being able to understand or modify the code of ICTs that mediate their everyday communicative activities, such as email, social media, Internet searching, or text editing. While this lack of critical user participation in software oversight and production may appear as natural, inevitable, and relatively inconsequential, I will argue that it has been normalized through corporate technical policies, cultural myths regarding programming, and the use of technology in educational settings. To demonstrate the range of alternatives to passive relationships to software, I will point to a number of software cultures, projects, and visions in which the everyday user has greater opportunity to democratically participate in shaping the technical functionality and policy of their digital tools. I will argue that examples such as the Free Software community, the Platform Coop movement, and Alan Kay's 1968 vision for Dynabook represent promising alternative software models that foster participatory design consciousness in the general user that could be fruitfully applied in educational settings. By implementing tools in the classroom that allow for participatory design and oversight, students would have the opportunity to experience greater forms of creative and critical control over ICTs that might lead them to question the lack of similar rights with regard to ICTs in everyday life. In this way, fostering participatory design approaches to digital technology stands as one promising approach to fostering critical and practical resistance in students to exploitative practices inherent in everyday ICTs such as dataveillance and algorithmic influence and manipulation. It also offers the possibility of turning educational technology into a site for producing open source ICT alternatives for general public use.

Techno-rhetorical consciousness

Participatory design approaches towards educational technology also have direct application for writing-intensive courses in the humanities in that they can help foster "techno-rhetorical consciousness," or a sensitive understanding of the way digital technology mediates rhetorical situations. By providing students with the perspective and control over ICTs normally only afforded by corporate or administrative entities, students have the opportunity to study more directly the way ICTs mediate their intellectual activities and communities, and explore how technical modifications might help support personal and collective intellectual goals and values. For example, access to data produced and transmitted through ICTs would enable students to use text analysis techniques from the digital humanities to study patterns in their individual and collective intellectual activities for the purpose of understanding the social dynamics of knowledge production and transmission. It would allow them to gain basic familiarity with algorithmic techniques that have increasing power in everyday life. And it would also provide students with the opportunity to experiment with how different aesthetic and algorithmic design features might better support individual cognitive activities related to writing process or productive intellectual exchange among students. These opportunities would not only have rich potential for the use and development of educational technology itself, but would also help students consider the way digital technology mediates the production and transmission of knowledge and power in everyday life.

KNIT, a digital learning commons

To explore some of these ideas in practice, we have launched KNIT, a digital commons for UC San Diego and institutions of higher education in the San Diego area. KNIT uses the free and open source software package Commons in A Box and thus, unlike many forms of proprietary software in educational settings, remains open to critical study and modification by the user community. In the final portion of my talk, I will discuss how we are using KNIT to test-drive participatory design practices for educational technology at UC San Diego and how we envision using it to give students a leadership position in its development and governance. I will also discuss the institutional, technical, and educational challenges of this approach and provide recommendations and resources for those interested in experimenting with this method at their home institution.

References

Drucker, Johanna, and Patrik BO Svensson. *The Why and How of Middleware*. Vol. 10, no. 2, 2016. *Digital Humanities Quarterly*.

Freire, Paulo. *Pedagogy of the Oppressed*. Continuum, 1993.

Rockwell, Geoffrey, and Stéfan Sinclair. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press, 2016.

Shor, Ira. *Critical Teaching and Everyday Life*. University of Chicago Press, 1980.

Simonsen, Jesper, and Toni Robertson. *Routledge International Handbook of Participatory Design*. Routledge, 2012.

Expanding the Research Environment for Ancient Documents (READ) to Any Writing System

Andrew Glass
asg@uw.edu
Microsoft Corp., University of Washington, United States of America

The Research Environment for Ancient Documents (READ) is an integrated Open Source web platform for epigraphical and manuscript research. The original goal of the READ platform was to support scholars in researching and presenting studies of handwritten documents and inscriptions preserved in Gāndhārī language using the Kharo hī script. Since many of the workflows supported by READ are common to epigraphic and manuscript studies in other textual traditions we wanted to investigate how READ could be generalized to support other writing systems. This presentation will share the results of that investigation with examples from English, Aramaic, Chinese, and Mayan.

Three core components of the READ data model depend on the writing system used by the source material:

1. The link between physical and textual data
2. The constraint mechanism that allows a user to edit text without disrupting links
3. The sort weight API that allows data in the model to be displayed in an expected sort order

Part One. The database model underlying READ was designed to reflect the separate components and layers of interpretation which manuscript scholars and epigraphers typically use in their research (letter forms => paleography; graphemic units => phonology; inflectional forms => morphology, etc.). Furthermore, the model recognizes a continuum of factual confidence beginning from statements of fact (e.g., the name of a collection in which an item is kept), to data which may have multiple or variant interpretations (e.g., the transcription of a sample of writing). Such variant data is linked back through the model to original facts. At the crux of this system of links are the references between segments on an image each containing an orthographic unit in the writing system and the transcription of that unit. Because READ was originally developed for Kharo hī, an alphasyllabary or Abugida-type writing system, this link maps image segments to syllable clusters. Other writing systems can be supported by mapping the syllable cluster to the appropriate orthographic units. This has been tested by mapping syllable clusters as follows: English letters, Aramaic syllables, Chinese logographs, and Mayan syllables and logographs.

Part Two. READ is intended to be a working environment for born-digital text editions. A critical feature of the model is that links created within the system must be preserved during repeated editing. The editing interface allows users to modify linked syllable clusters. By constraining edits to valid transcriptions of a syllable cluster defined for the language, READ can keep track of user edits and prevent links from being broken. Other writing systems can be supported by defining the valid transcription forms for the orthographic units. In most cases this is less complex than for ak ara-based writing systems. This has been tested by defining valid orthographic units as follows: English – Consonants, Vowels; Aramaic syllables - Consonants, Vowels, Consonant with modifier; Chinese – Logograph; Mayan – Logograph, CV syllable. All systems also permit orthographic units to be Digits and Punctuation signs.

Part Three. READ uses custom sort tables to weight the orthographic units and subunits used by the model. Having custom sort tables allows correct sorting of Romanized transcription when the expected sort order is not equal to standard 'ABC' order. Other writing systems represented in Romanized transcription with non-standard sorts require dedicated sort tables. Alternatively, writing systems represented in native script via Unicode may be sorted via their Unicode weights. This has been tested using standard ABC weights for English, custom weights for Mayan transcription, Unicode weights for Hebrew transcription of Aramaic, and Pinyin sort weights for Chinese logographs.

The outcome of these investigations has been that the READ architecture is generalizable, and that the READ platform could be employed by projects with a focus on documents in any writing system.

The Latin American Comics Archive: An Online Platform For The Research And Teaching Of Digitized And Encoded Spanish-Language Comic Books Through Scholar/Student Collaboration

Felipe Gomez
fgomez@andrew.cmu.edu
Carnegie Mellon University, United States of America

Scott Weingart
scottbot@cmu.edu
Carnegie Mellon University, United States of America

Daniel Evans
djevans@andrew.cmu.edu
Carnegie Mellon University, United States of America

Rikk Mulligan
rikk@cmu.edu
Carnegie Mellon University, United States of America

Overview

This short paper looks into the process of developing the Latin American Comics Archive (LACA), a project created by our team at Carnegie Mellon University. LACA combines ongoing research in the Humanities with digital technologies as a tool for enhancing access and analysis capabilities for both scholars and students of these materials.The curated digital archive includes representative samples of Latin American comics digitally encoded in Comic Book Markup Language (CBML), while a technical foundation combining the open source content management system Omeka with TEI Boilerplate offers a customizable front-end for public or restricted access to the individual items and curated collections of the comics. This allows students and researchers access to source materials and possibilities to collaborate in their exploration, definition, tagging, and annotation for the analysis of visual and verbal language, cultural and linguistic characteristics or themes, and a variety of formal categories.

Statement of the problem

Despite the overdue growing recognition of the genre of comics in academia, the study of foreign/second language comics within the United States has encountered specific obstacles. Primary-source research of Spanish-language comics has often proved to be challenging. Among other difficulties, collections are most often housed in the source countries, and a desired piece of documentation may sometimes be in libraries hundreds or thousands of miles away. Items may be both in public and private hands, and access to certain items is often highly restricted due to their fragility, rarity, and value. Oftentimes, specific documents aren't cataloged in the archives' container lists, making the identification, location, and access of relevant materials problematic. When using traditional research methods, these challenges have to be confronted and resolved by the researcher, who works in isolation with the source documents. Many of these issues also generate constraints in the realm of teaching, where the limitations to the access of sources restricts course conceptualization and implementation, and where students don't usually have much agency or opportunities to engage in larger debates and conversations with other students or scholars of Latin American comics.

Digital tools have the potential to facilitate or solve many of these issues for research and teaching of this important cultural and literary medium. Indeed, they have the ability to address precisely the core values that Spiro (2012) associates with work in the Digital Humanities -- openness, collaboration, diversity, experimentation, collegiality, and connectedness. These tools can, for instance, create optimal opportunities to view and use some of these sources online, thus granting access to an audience who may never have had the chance to see them in the "analog" era, and opening and expanding the possibilities for a richer and deeper type of collaborative research. Our goal is to expand the possibilities of using Spanish-language comics by identifying and piloting the use of digital tools with which digital copies of representative Latin American comics can be made, accessed, and annotated in collaboration with students and scholars. Our focus is on developing an archive of sources that scholars and students can use for analysis, interpretation, and research employing digital tools.

Critical Context

LACA seeks to insert itself in the broad scholarly landscape created at the intersection of comics scholarship (e.g. Priego, 2016; Walsh, 2012), visual ontologies and comics (e.g. Bateman et. al., 2017; Turton, 2017), work done to encode comics elements (e.g. Dunst et. al., 2016; Haidar and Ganascia, 2016; Kuboi, 2014), and work on the value of comics as a pedagogical tool (e.g. Brooks, 2017).

Methodology

Given the team's expertise in Digital Humanities (DH) and Digital Scholarship, and with the support of an institutional Mellon DH seed grant, the project was initiated in the summer 2016. LACA was modeled after existing specialized collections such as MIT's Comics and Popular Culture archive, UNAM's specialized online resource http://www.pepines.unam.mx/, and the Grand Comics Database (GCD) with the purpose of combining the PI's ongoing research and teaching experience on Spanish-language Latin American comics with the use of DH tools to create an environment enabling students and scholars to have access to and collaborate in the analysis of the digital materials. At the current stage, LACA includes a small digital sample of Latin American comics produced throughout the last century, provided through a combination of previously digitized materials, materials we scanned, and those provided by authors themselves.

The presentation will detail three parts of the project:

1. Curating the comics and creating the archive;
2. Creating the online Metamedia platform to house digitized sources for the research and teaching of Spanish-language comics through student/scholar collaboration;
3. Piloting and implementing the digital archive for research and teaching.

Insights/Results

LACA was piloted at CMU over the past year as an instrument in courses for undergraduate students of Spanish

language and culture. Students and faculty collaborate in the analysis and CBML coding of the comics. In the process, students learn the basics of TEI and CBML, as well as critical approaches to Spanish-language comics, and their work contributes to the availability of comics on the site. Students are also able to develop integrated textual and visual competence, knowledge, and skills. The pilot courses provide initial evidence that coding the comics facilitates students' attention to details, notice of patterns, and, in general, collaborative advancement in the analysis and understanding of the linguistic and cultural elements contained in the comics. At the same time, it also helps students keep in mind communication to a wide public audience. The PI has benefitted from the additional opportunities afforded to glean information about students' progress toward cultural, linguistic, visual, and digital literacy. Thus, it is suggested that LACA could be of use and applicable to other courses in Hispanic studies, Modern Languages, and the Humanities.

We intend to make LACA publicly available for use as a hub where students and scholars interested in experimenting with the inquiry of Latin American comics can interact. This would help transform and expand the scale of traditional research methods used, and could open new modes and possibilities for text analysis that can be employed into the realm of student agency and learning. However, as we advance in the process to attain this goal, we acknowledge that IP/copyright permissions remain a challenge. Some creators have granted permission to distribute their works; others will only be used as part of course materials. Despite this, we think it is important to keep in mind Walsh's (2012) point that "nothing prevents a scholar from applying CBML markup to any text as part of a strategy for reading, interpretation, and analysis. The end goal of markup is not and should not always be publication of a digital surrogate. The encoding of a text may be a rigorous intellectual activity that has great value as process, not just as product."

References

Bateman, J. A., Veloso, F. O. D., Wildfeuer, J., Cheung, F. H. and Guo, N. S. (2017). An open multilevel classification scheme for the visual layout of comics and graphic novels: Motivation and design. *Digital Scholarship in the Humanities*, *32*(3), 476–510. https://doi.org/10.1093/llc/fqw024

Brooks, M. (2017). Teaching TEI to undergraduates: A case study in a digital humanities curriculum. *College and Undergraduate Libraries*, *0*(0), 1–15. https://doi.org/10.1080/10691316.2017.1326331

Dunst, A., Hartel, R., Hohenstein, S. and Laubrock, J. (2016). Corpus Analyses of Multimodal Narrative: The Example of Graphic Novels. *Digital Humanities 2016: Conference Abstracts*. Krakow, Poland. Retrieved from http://dh2016.adho.org/static/data-copy/387.html

Haidar, A. and Ganascia, J. (2016). Automatic Detection of Characters in Case Insensitive Text in Comics. In *Digital Humanities 2016: Conference Abstracts* (pp. 425–426). Jagiellonian University & Pedagogical University, Kraków.

Kuboi, T. (2014). Element Detection in Japanese Comic Book Panels. *Master's Theses and Project Reports*. https://doi.org/10.15368/theses.2014.141

Priego, E. (2016). Comics as Research, Comics for Impact: The Case of *Higher Fees, Higher Debts*. The Comics Grid: Journal of Comics Scholarship. 6, p.16. DOI: http://doi.org/10.16995/cg.101

Spiro, L. (2012). "This Is Why We Fight": Defining the Values of the Digital Humanities. In M. K. Gold (Ed.), *Debates in the Digital Humanities*. Minnesota: University of Minnesota Press. Retrieved from http://dhdebates.gc.cuny.edu/debates/text/13

Turton, A. (2017). Towards Feminist Data Production: A Case Study from Comics. In *Digital Humanities 2017: Conference Abstracts*. Montreal, Canada. Retrieved from https://dh2017.adho.org/abstracts/493/493.pdf

Verba Volant, Scripta Manent: An Open Source Platform for Collecting Data to Train OCR Models for Manuscript Studies

Samuel Grieggs
sgrieggs@nd.edu
University of Notre Dame, United States of America

Bingyu Shen
bshen@nd.edu
University of Notre Dame, United States of America

Hildegund Muller
hmuller@nd.edu
University of Notre Dame, United States of America

Christine Ascik
cascik@nd.edu
University of Notre Dame, United States of America

Erik Ellis
erik.z.ellis.67@nd.edu
University of Notre Dame, United States of America

Mihow McKenny
mihow.p.mcKenny.5@nd.edu
University of Notre Dame, United States of America

Nikolas Churik
nchurik@nd.edu
University of Notre Dame, United States of America

Emily Mahan
emahan@nd.edu
University of Notre Dame, United States of America

Walter Scheirer
walter.scheirer@nd.edu
University of Notre Dame, United States of America

Introduction

The transcription of handwritten historical documents into machine-encoded text has always been a difficult and time-consuming task. Much work has been done to alleviate some of that burden via software packages aimed at making this task less tedious and more accessible to non-experts. Nonetheless, an automated solution would be a worthwhile pursuit to vastly increase the number of digitized documents. As part of a continuing effort to expand the footprint of digital humanities research at our institution, we have embarked on a project to automatically transcribe and perform automated analysis of Medieval Latin manuscripts of literary and liturgical significance. Optical Character Recognition (OCR) is the process of converting images containing text into a machine encoded document. Recent advances in artificial neural networks have led to software that can transcribe printed documents with near human accuracy (LeCun et al., 2015). However, this level of accuracy breaks down when working with handwritten, and especially cursive, documents except when applied to restrictively specific domains.

Neural networks that are trained for this task require thousands of labeled examples so that their millions of parameters can be optimized. While there are thousands of high-quality scans of manuscripts available on the Internet, very few of these documents have been annotated for OCR tasks, and there is only a limited selection of ground-truth data which is annotated and segmented at the word-level (Fischer et al., 2011; Fischer et al., 2012) . There is no data available that provides annotations at the character-level. Normally, machine learning researchers would outsource the production of this ground-truth data to a platform such as Amazon's Mechanical Turk service, which allows crowdsourcing of human intelligence tasks. This is not an option for transcribing Medieval manuscripts, because it requires domain specific expertise. We put together a team of expert Medievalists and Classicists to generate the ground-truth data, and we have been developing a software platform that breaks the tedious task of producing pixel-level training data into more tractable jobs. The goals of this software go beyond just Latin manuscripts: it can be used to generate source data for any machine learning task involving document analysis. We are releasing it publicly, as free and open source software, in hopes that others can also use it to generate data, and help bring further advances in machine learning for handwritten text recognition.

Related work

State-of-the-art solutions to handwritten digit recognition on the MNIST dataset have achieved accuracies greater than 99% and have led some to declare handwritten OCR a solved problem (Wan et al., 2013). However, Cohen et al. have shown that adding the English alphabet to the dataset drops accuracy by more than 20% even when using the same methodology (Cohen et al., 2017). Some of the difference can be attributed to the fact that characters like "I", "l", and "1" are often ambiguous without context --- especially when handwritten. To combat this, many handwritten text recognition algorithms will often use recurrent neural networks that look at the whole word and utilize a language model to overcome ambiguities (Fischer et al., 2009; Sánchez et al., 2016). Additionally, Convolutional Neural Networks (CNN) have been shown to have promise in segmenting biomedical images, which are also difficult to ground truth (Ronneberger et al., 2015). A similar approach could be used to segment individual letters in manuscripts. Incorporating human performance information into the machine learning process has been shown to improve the accuracy of tasks like face detection (Scheirer et al., 2014). We hypothesize that incorporating a human weighted loss function will lead to similar improvements in this task as well.

Workflow

Currently the software runs in Google App Engine using high-resolution source images. We are in the process of setting up the software to be run in a vagrant environment to make it available for local environments. The vagrant script will provision a Virtual Machine, either locally or to the cloud to serve the software and configure it to work with a user-provided library of documents. In either case, transcribers can access the software via a web browser. The user then proceeds to segment the document by lines and words by drawing bounding boxes, and characters by drawing over them. It also collects text annotations of the text at the word- and character-level. It stores all the information in a MySQL database.

Line and word level

Our process starts by having experts segment the document into lines. Transcribers use a modified version of the Image Citation Tool from the Homer Multitext Project to quickly break the document down into CITE URNs representing each line by drawing boxes around them (Blackwell and Smith, 2014). After all the lines are selected the process is repeated for each word. A screenshot of these processes is shown in Fig. 1.

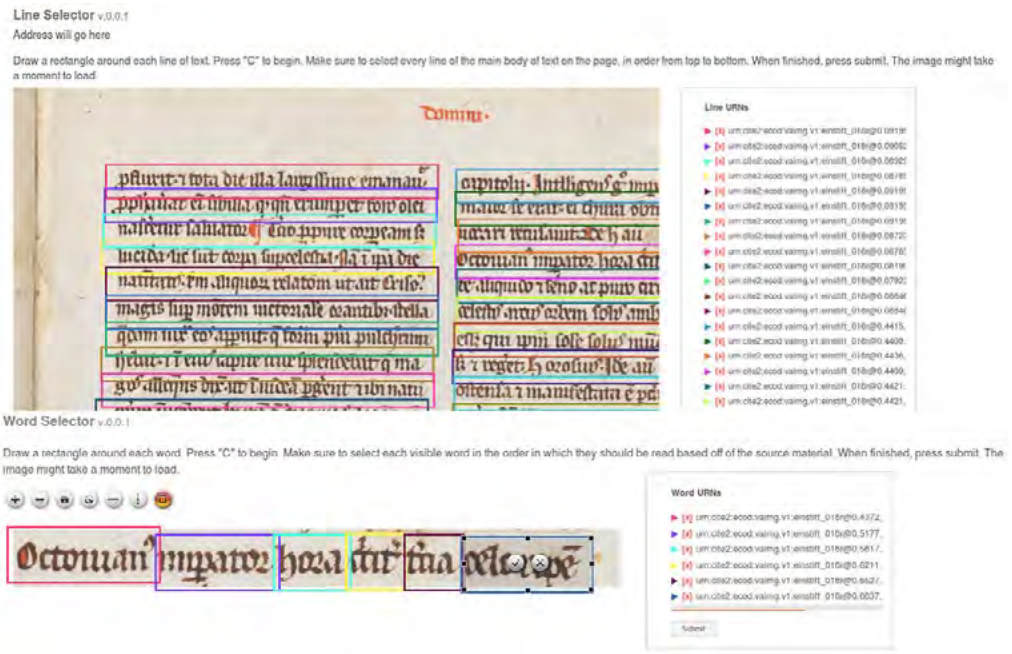


Figure 1: An example of the interface for selecting lines and words. Manuscript: Einsiedeln, Stiftsbibliothek, Codex 629(258), f. 4r – [Jacobus de Voragine] Legenda aurea sive lombardica (http://www.e-codices.unifr.ch/en/list/one/sbe/0629)

Pixel level annotation

After segmenting the document into words, our software prompts the expert to segment and annotate each word letter by letter. Instead of using a bounding box, we have the user trace over each character in the word using a pen tool. This gives us a pixel-by-pixel segmentation of the image that can be used to train a CNN to segment the characters automatically, much in the same way segmentation models are trained for other computer vision tasks (Ronneberger et al., 2015). At this stage the expert will also select which letter best represents each character from an array of buttons, as shown in Fig. 2.

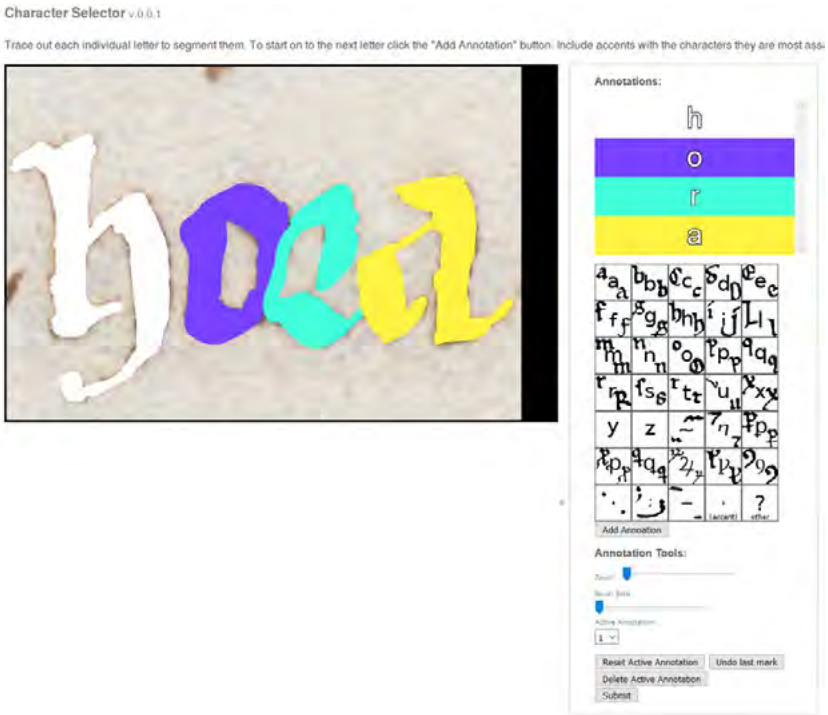


Figure 2: An example of the tool used to collect pixel level ground-truth at the character level.

Psychophysical measurements

The final stage collects psychophysical measurements of the human process of reading. The software brings up individual characters, as shown in Fig. 3, and asks the transcriber to pick an annotation for a character without word context. They will also be asked to select the difficulty of each character. The software also records how long it takes for the user to submit an answer and compares whether the user selected the same character that was selected during the word-level annotation.

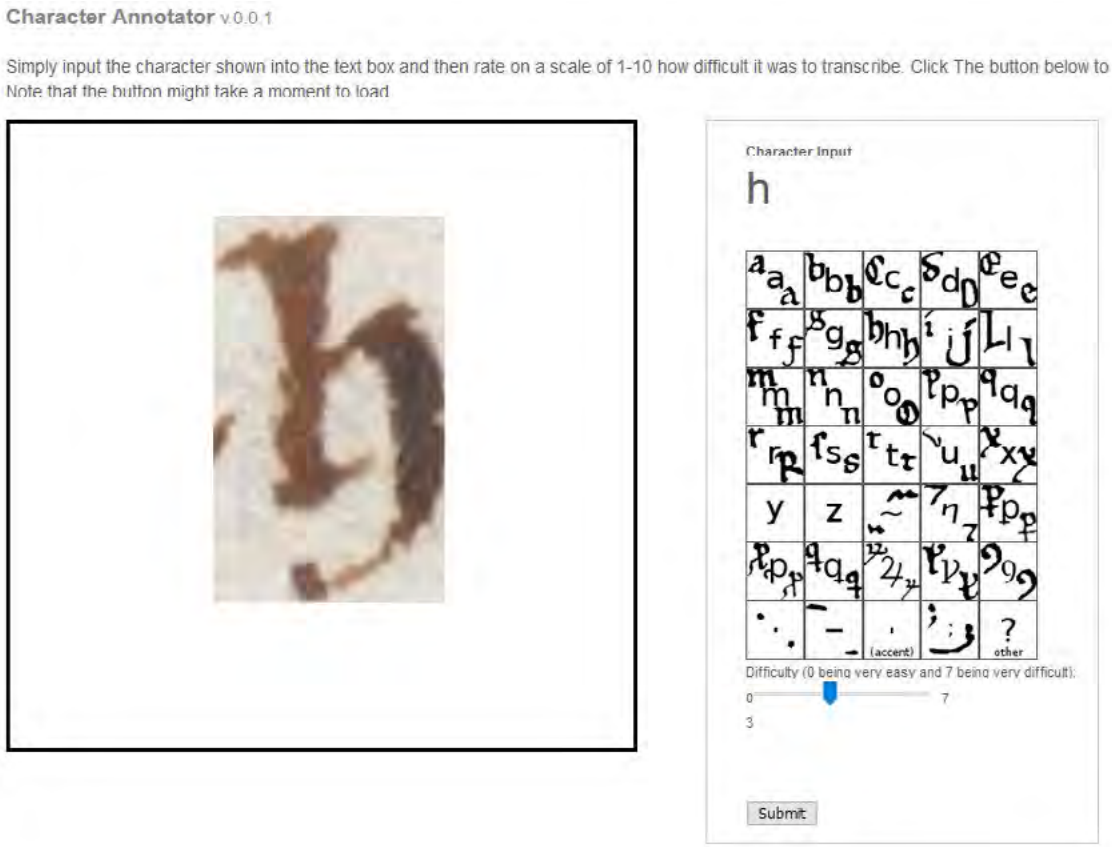


Figure 3: A screenshot of the psychometric data collection stage.

Outcomes

The software produces a segmented image for each document that can be used as training data for machine learning-based segmentation. Furthermore, it provides the pyschophysical measurements on the reading difficulty of each character. We also designed it to produce word-level segmented data in a similar format to the IAM Historical Document Database (Fischer et al., 2012; Fischer et al., 2011). Finally, the user will be able to export the transcribed document into a standard markup language such as TEI.

References

Blackwell, C. W. and Smith, D. N. (2014). The Homer Multitext and RDF-Based Integration. *Papers of the Institute for the Study of the Ancient World*, 7.

Cohen, G., Afshar, S., Tapson, J. and Schaik, A. van (2017). EMNIST: an Extension of MNIST to Handwritten Letters. *CoRR*, abs/1702.05373.

Fischer, A., Frinken, V., Fornés, A. and Bunke, H. (2011). Transcription Alignment of Latin Manuscripts Using Hidden Markov Models. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. *ACM*, pp. 29–36.

Fischer, A., Keller, A., Frinken, V. and Bunke, H. (2012). Lexicon-free Handwritten Word Spotting Using Character HMMs. *Pattern Recognition Letters*, 33(7): 934–942.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G. and Stolz, M. (2009). Automatic Transcription of Handwritten Medieval Documents. Virtual Systems and Multimedia, 2009. VSMM'09. *15th International Conference on. IEEE*, pp. 137–142.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep Learning. *Nature*, 521(7553): 436–444.

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.

Sánchez, J. A., Romero, V., Toselli, A. H. and Vidal, E. (2016). ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset. Frontiers in Handwriting Recognition (ICFHR), 2016 *15th International Conference on*. IEEE, pp. 630–635.

Scheirer, W. J., Anthony, S. E., Nakayama, K. and Cox, D. D. (2014). Perceptual Annotation: Measuring Human Vision to Improve Computer Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8): 1679–1686.

Wan, L., Zeiler, M., Zhang, S., Cun, Y. L. and Fergus, R. (2013). Regularization of Neural Networks Using Dropconnect. Proceedings of the *30th International Conference on Machine Learning* (ICML-13). pp. 1058–1066.

Indagando la cultura impresa del siglo XVIII Novohispano: una base de datos inédita

Víctor Julián Cid Carmona
vjcid@colmex.mx
El Colegio de México, México

Silvia Eunice Gutiérrez De la Torre
segutierrez@colmex.mx
El Colegio de México, México

Guadelupe Elisa Cihuaxty Acosta Samperio
cihuaxtysamperio@gmail.co
Universidad Nacional Autónoma de México

El objetivo del sistema es facilitar el estudio sistemático de los libros anunciados en la *Gazeta de México*, primera publicación periódica de América. Esta publicación, que imprimió su primer fascículo en 1722, ofrecía en algunos números información sobre las novedades bibliográficas de la época.

Estas notas incluían menciones de títulos, nombres de autores e impresores, ubicación de las imprentas y lugares de venta de los libros, entre otros (imagen 1).

La información de las novedades fue enriquecida con un proceso de investigación documental en el que se identificaron 32 campos distintos tales como: catálogos en los que se encuentra registrada la obra anunciada, disponibilidad en pdf, cargo o profesión de los autores, archivo de autoridad virtual internacional de los mismos (VIAF), página exacta del anuncio, precio de la obra, idioma, asignación temática sistematizada, entre otros.

Libros nuevos.

VNO en octavo cuio titulo es: *Razones para convencer à el Pecador para que salga de pecado, y se ponga en gracia de Dios*; su Autor, el R. P. M. Ignacio Fiol, de la Compañia de Jesvs, impresso en la Imprenta de D. Joseph Bernardo de Hogal, en la Calle nueva de la Monterilla.

Otro en lo mismo, intitulado: *Epitome de la vida de la V. Virgen Mariana de JESUS, Flores, y Paredes, Beata Jesuita, nombrada la Azucena de Quito, ilustre en Santidad, Profecias, y Milagros*, sacado de la Vida, q̃ latamente escriviò el P. Jacintho Morân de Butrõ, de la misma Compañia. Vendese en la Libreria de los Herederos de la Viuda de Miguel de Rivera Calderon, en el Empedradillo.

Otro en doze, intitulado: *Sol del Alma, cuias luces alumbran la razõ, y encienden los afectos. Meditaciones Santas del Señor San Bernardino de Sena*; escritas por el R. P. Fr. Francisco de la Concepcion Barbosa, Hijo de la Provincia del Santo Evangelio de Mexico, y Predicador Mayor en su Convento Grande. Hallarâse en la Sacristia del V. Orden Tercero de Penitencia de N. S. P. San Francisco, de esta Corte.

Otro en diez y seis, intitulado: *Devocion à Señor San Joseph, Esposo de la Virgen Maria N. Señora*, por el P. Francisco Garcia de la Compañia de JESVS.

Vn Sermon predicado en la Iglesia del Convẽto de Señoras Religiosas de la Seraphica Virgen Santa Catharina de Sena, de esta Corte, el dia primero de Abril del año de 1729. en la nueva Festividad de las Llagas de dicha Seraphica Virgen; por el M. R. P. M. Fr. Juan de Alvarado, del Sagrado Orden de Predicadores, Hijo de esta Provincia de Santiago de Nueva-España.

☞ *Impressos donde esta Gazeta.*

Imagen 1. Anuncio de libros en la *Gazeta de México*, 16 de septiembre de 1807

La propuesta tiene dos propósitos principales. Por un lado, representará uno de los catálogos digitales más completos de las obras novohispanas del siglo XVIII. El *Catálogo de impresos Novohispanos (1563-1766)* coordinado por Guadalupe Rodríguez, por ejemplo, únicamente contiene 505 registros. Por otro lado, el estudio de las gacetas de México se ha abordado, hasta ahora, desde dos perspectivas: tratarlas en su generalidad (Drwall, 1980; Ruíz Castañada, 1969, 1970, 1971); o bien, referirse a la representación de algún tema específico en sus páginas (ver Guedea, 1989, 1991). En este sentido, el proyecto **Ciudad letrada: la *Gazeta de México* y la difusión de la cultura impresa durante el siglo XVIII**, permite un nuevo acercamiento a las gacetas de México como fuente de la cultura impresa de la época.

A esto se aúna el hecho de que ha sido diseñada como una herramienta que apoye en las tareas del investigador interesado en los impresos de aquel siglo. El modelo de la base de datos, es del tipo Entidad-Relación. Para acercar a las personas a esta información se utilizó la plataforma Omeka 2.0, donde los vínculos se construyeron con el complemento llamado ItemRelations y la prueba de concepto puede ser consultada en: http://sandbox.colmex.mx/~silvia/omeka25/.

El procedimiento para integrar los datos que contiene la base, implicó la revisión de cada uno de los 1370 fascículos durante los 42 años de edición de la *Gazeta* con el fin de identificar los anuncios de libros nuevos. La información de cada uno de los anuncios se complementó con datos bibliográficos obtenidos de bibliografías especializadas y catálogos de bibliotecas con el propósito de enriquecer la información original y hacerla más útil. Esto dio como resultado la identificación de 1872 anuncios de libros y folletos, publicados entre los años de 1657 y 1809; es decir, libros en un rango de centuria y media que

jamás habían sido identificados sistemáticamente, por lo cual hablamos de una base de datos inédita.

Entre las características especiales de este desarrollo caben destacar las múltiples formas de explorar y acceder a los registros. Entre ellas: la exploración por etiquetas, por índices, por mapa de ubicación de imprentas o lugares de venta, navegación hipervinculada de los resultados y de cada registro.

La búsqueda por etiquetas ofrece una vista de pájaro sobre los temas más frecuentes, los cuales fueron desagregados de su forma clásica (es decir en triadas) para permitir exploraciones más granulares.

Imagen 2. Fragmento de la nube de etiquetas

Por otro lado, los índices fueron generados con el complemento 'Reference' desarrollada para Omeka por Daniel Bertherau. Estos índices permiten una navegación exploratoria de los temas, autores, impresores, años de publicación, lugares de impresión y de venta, etc., ordenados alfabéticamente junto con sus ocurrencias (imágenes 3 y 4).

Imagen 3. Índices disponibles para búsqueda sistemática

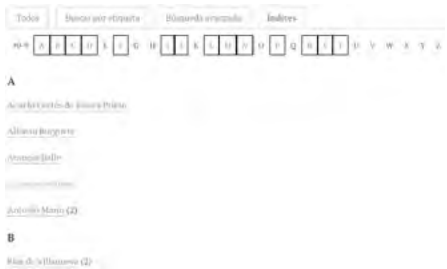


Imagen 4. Ejemplo de un índice (impresores)

Los mapas se crearon utilizando Carta, un complemento de AcuGis. En ellos se puede observar la distribución y concentración de imprentas (imagen 5) y lugares de venta (imagen 6) y de esta forma identificar los espacios clave de la Ciudad Letrada.

Imagen 5. Ubicación de imprentas en la Ciudad de México, siglo XVIII

Imagen 6. Librerías y otros lugares de venta

Por último, con navegación hipervinculada nos referimos a que sobre cada metadato se puede pulsar (imagen 7) para desplegar otros registros con esa misma característica (imagen 8).

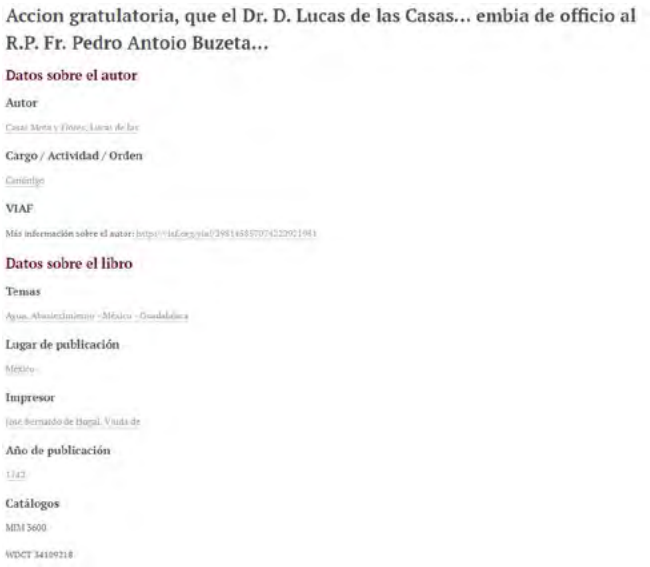


Imagen 7. Despliegue de registro con datos hipervinculados

Imagen 8. Despliegue de registros coincidentes (mismo autor)

Por mencionar un ejemplo de uso, imaginemos el siguiente escenario: digamos que el usuario explora el mapa de lugares de venta y abre la ubicación del Colegio de San Ildefonso (como en la imagen 6), al pulsar sobre el hipervínculo que dice 'Libros impresos en esta ubicación', el sistema despliega la lista completa de registros de la base que se vendían en ese lugar; en el despliegue de estos datos (imagen 9), se observa que todos son libros seculares y relacionados con las ciencias y la educación. Esto es interesante, si se considera que la mayoría de los registros son de contenido religioso y casi todos los puntos de venta ofrecían en mayor cantidad obras de esta naturaleza y podría guiar al estudioso del tema hacia nuevas preguntas como ¿todos los colegios vendían libros seculares?, ¿qué otros puntos distribuían obras de esta índole?, ¿por qué es más difícil encontrar el nombre de los autores de estas publicaciones?, etc.

Imagen 9. Libros a la venta en el Colegio de San Ildefonso

Cabe mencionar que se identificó un conjunto considerable de obras de carácter técnico o científico, algunos diccionarios generales y especializados, varios textos literarios y, principalmente, obras de contenido religioso, histórico, biográfico, dogmático y devocional.

Para concluir, consideramos que este desarrollo posibilitará a los interesados en impresos del siglo XVIII nuevas vetas de investigación a partir de la información sistemática que incluye. En particular, resultará útil para tratar asuntos relacionados con autores, impresores y comerciantes del libro en México durante el siglo XVIII. Además, ofrece la posibilidad de indagar sobre los mecanismos de propaganda de este bien cultural, así como saber en qué lugares se conservan ejemplares de estos documentos actualmente o, tener acceso a versiones electrónicas de ellos, en varios casos.

Referencias

AcuGis (s.f.). *"Carta 2.1.1". Github*. https://github.com/AcuGIS/Carta.

Adank, P. A. D. (1980). Accommodation and innovation: the Gazeta de México, 1784 to 1810 Arizona: Arizona State University Doctorado.

Bertherau, D. (s.f.). *"Reference 2.4.2". Github*. https://github.com/Daniel-KM/Reference.

Castera, I. (1785). Plano Geométrico de la Imperial, Noble y Leal Ciudad de México, teniendo por extremo la Zanxa y Garitas del Resguardo de la Real Aduana

Madrid, en la Calle de Atocha, frente de la Aduana vieja, Manzana 159, N.o 3.
Domínguez Rodríguez, G. (2012). Introducción. *Repertorio de impresos novohispanos (1563-1766)*, vol. 12. Xalapa, Veracruz: Biblioteca Digital de Humanidades. Universidad Veracruzana.
Guedea, V. (1989). La medicina en las gacetas de México. *Mexican Studies/Estudios Mexicanos*, 5: 175–99.
Guedea, V. (1991). *Las gacetas de México y la medicina: un índice*. México: Universidad Nacional Autónoma de México, Instituto de Investigaciones Históricas.
Ruíz Castañeda, M. del C. (1969). La Gaceta de México de 1722 primer periódico de la Nueva España. *Boletín del Instituto de Investigaciones Bibliográficas*, 1(1): 39–59.
Ruíz Castañeda, M. del C. (1970). La segunda Gazeta de México (1728-1739, 1742). *Boletín del Instituto de Investigaciones Bibliográficas*, 3(2): 23–42.
Ruíz Castañeda, M. del C. (1971). La tercera Gaceta de la Nueva España. Gaceta de México (1784-1809). *Boletín del Instituto de Investigaciones Bibliográficas*, 3(6): 137–150.

Puesta en mapa: la literatura de México a través de sus traducciones

Silvia Eunice Gutiérrez De la Torre
segutierrez@colmex.mx
El Colegio de México, Mexico

Jorge Mendoza Romero
enciclopedia.flm@gmail.com
Fundación para las Letras Mexicanas, Mexico

Amaury Gutiérrez Acosta
agutierrez@conabio.gob.mx
CONABIO, Mexico

Para responder cuáles han sido las tendencias de la circulación de la literatura de México hacia otros espacios lingüísticos, se partió de los datos disponibles en la *Enciclopedia de la literatura en México* (ELEM, www.elem.mx) para realizar un estudio de las traducciones de obras de escritores mexicanos, escritas en español y traducidas a 33 lenguas (incluidos los 7 idiomas indígenas del país de los que hubo al menos un registro). En esta presentación breve, daremos cuenta de los resultados de una investigación en curso sobre el modelado y puesta en mapa de estos datos.

En México, el estudio cuantitativo de las traducciones de la literatura nacional tiene un antecedente emblemático en la obra pionera de José Ignacio Mantecón, Índice de las traducciones impresas en México, de 1959. En este trabajo se recopilaron 544 traducciones hechas en México en ese año, y se registraron aspectos tales como el género al que pertenecían, lo que permitió derivar conclusiones como el hecho de que el grupo más representativo de traducciones lo constituían las obras literarias (35%), de las cuales un 13% eran libros infantiles (Mantecón, 1959: 14, 18). Sin embargo, además de que este trabajo no ha sido replicado, este estudio sólo da cuenta de las traducciones al español como lengua meta.

Otra referencia, en la que se perfila el objetivo de nuestra investigación –el estado de la traducción de la literatura de México– se encuentra en la introducción que hace Rosenzweig del intercambio epistolar entre Alfonso Reyes y el traductor al checo Zdeněk Šmíd. En ésta se lee lo siguiente:

> Salvo excepciones, la literatura mexicana, al igual que la latinoamericana, se comenzó a traducir a comienzos de los años treinta del siglo XX. Inicialmente se hicieron traducciones al inglés y al francés; en un segundo momento, impulsadas por el francés, a otras lenguas europeas como el alemán, neerlandés, checo e italiano. Las primeras novelas mexicanas que se tradujeron fueron Los de abajo y Mala yerba de Mariano Azuela; *El águila y la serpiente* y *La sombra del Caudillo*, de Martín Luis Guzmán; El indio, de Gregorio López y Fuentes; y *¡Vámanos con Pancho Villa!* de Rafael F. Muñoz. (Rosenzweig, 2014: 13)

No obstante, este extracto carece de referencias numéricas exactas y tampoco responde quiénes fueron esos primeros traductores al inglés, cuándo comenzaron exactamente las traducciones al francés o cuándo a otras lenguas europeas. Y es que, a excepción de algunas listas de idiomas específicos –como los 327 registros de obras de la literatura mexicana traducidas al inglés en Estados Unidos (Boyd, 2012); el catálogo análogo de 99 registros de obras traducidas al alemán (Küpper, s.f.); o la lista de las traducciones al italiano (Tedeschi, s.f.) – no existe ningún compendio que ofrezca el panorama completo de la proyección de la literatura mexicana en un sentido global. Por tal razón, la bibliografía de más de 1500 traducciones de la ELEM es una base de datos única en su tipo de la que es necesario expandir sus posibilidades heurísticas. Pero antes, algunas palabras sobre esta enciclopedia.

La ELEM comenzó a organizar el conocimiento en torno a la cultura literaria de México (oral y escrita) desde 2011, cuando fue creada. Cuenta con los registros de 13,040 personas (autores, traductores, investigadores literarios) y más de 40,000 obras impresas (primeras ediciones), que conforman una bibliografía general de la literatura en México, la cual abarca casi V siglos de cultura literaria. Entre sus prioridades se encuentra el registro de las obras traducidas a otros espacios lingüísticos con el propósito de observar, a través de las lenguas meta y los países del mundo en que son impresas, el grado de recepción de la literatura del país.

Por esto, emprendimos un trabajo colaborativo y transdisciplinario en el que se planteó un modelado de los datos disponibles en la enciclopedia (ver Imagen 1)

bajo el concepto de puesta en mapa (en analogía de la puesta en página del mundo editorial) y en consonancia con la línea de las Humanidades Digitales denominada spatial humanities. En este caso específico, designa al desarrollo de una interfaz que permite captar geopolíticamente la circulación de la literatura de los autores mexicanos que escriben en español (con algunas traducciones indirectas) hacia 19 lenguas indoeuropeas, 7 lenguas indígenas de México, además de estonio, euskera, finés, hebreo, húngaro, japonés y turco. El corpus del que partimos contempla un universo de 1658 primeras ediciones que se desdobla, a partir de las reimpresiones y reediciones de muchos títulos, en un total de 2088 objetos.

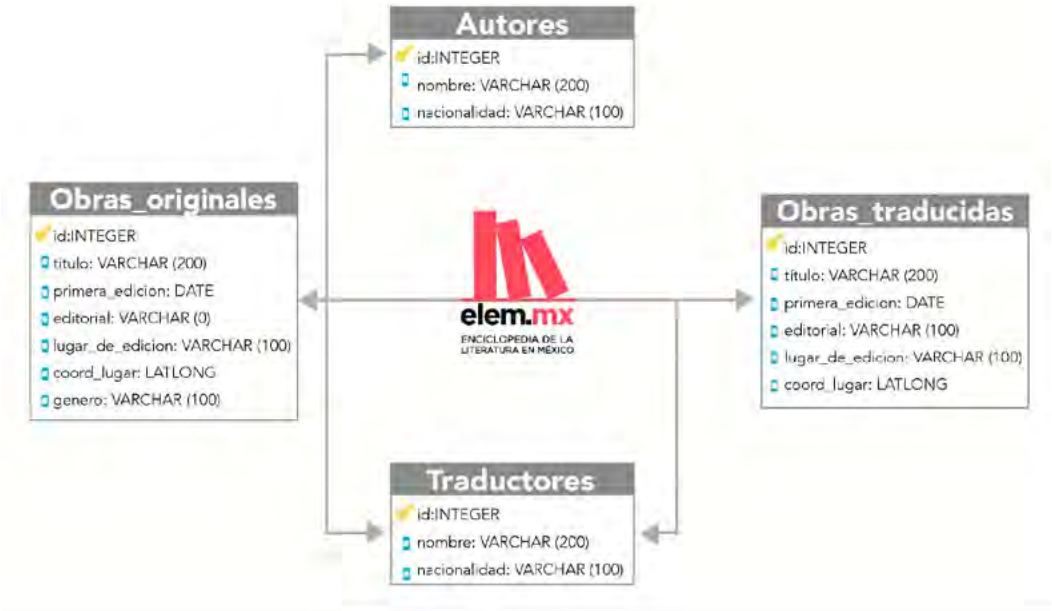


Imagen 1. Estructura de la base de datos

En un primer acercamiento, nos interesó indagar las relaciones espacio-temporales de las obras traducidas para responder las siguientes preguntas:

- ¿Qué autores o géneros han sido los más traducidos?
- ¿En qué años?
- ¿En qué geografías?
- ¿A qué idiomas?

Imagen 2. Perspectiva general de la puesta en mapa

Para facilitar la exploración de estas relaciones, se creó un prototipo de interfaz interactiva que permite iniciar investigaciones a partir de la puesta en mapa de los datos. El código en desarrollo de este prototipo se encuentra disponible en GitHub (Gutiérrez, 2017) y su versión para consulta estará en: http://elem.mx/estgrp/datos/1335.

Se describen las etapas de desarrollo a continuación. A partir de una consulta SQL a la base de la ELEM se creó un archivo separado por comas (csv) usando un script de Python (parser.py en el repositorio de GitHub). Estos insumos fueron transformados para obtener un formato adecuado para el consumo en Javascript: JSON. Para la arquitectura de la aplicación web se usó una herramienta para hacer empaques o bundles llamada Webpack (https://webpack.js.org/). La biblioteca usada para la creación del mapa es una herramienta de código abierto llamada Leaflet en su última versión 1.2.0 (http://leafletjs.com/). El desarrollo de la aplicación se puso en marcha en Javascript para la interfaz ya que la información, por el momento, existe de manera estática. En el futuro, cuando se integre con la base de datos con la dorsal final o backend, será deseable que las consultas de datos se realicen desde este punto y se exponga un end-point adecuado para el consumo.

La interfaz pretende facilitar la visualización e interacción con los datos de la base, así como el análisis exploratorio de los mismos (Behrens, 1997). Los usuarios podrán elegir filtros tales como: lengua meta, género literario, año de la traducción y, explorar los registros por ubicación geográfica. Además se provee de la siguiente información sobre los objetos: título de la traducción, autor/a, traductor/a, editorial de la traducción y título original de la obra.

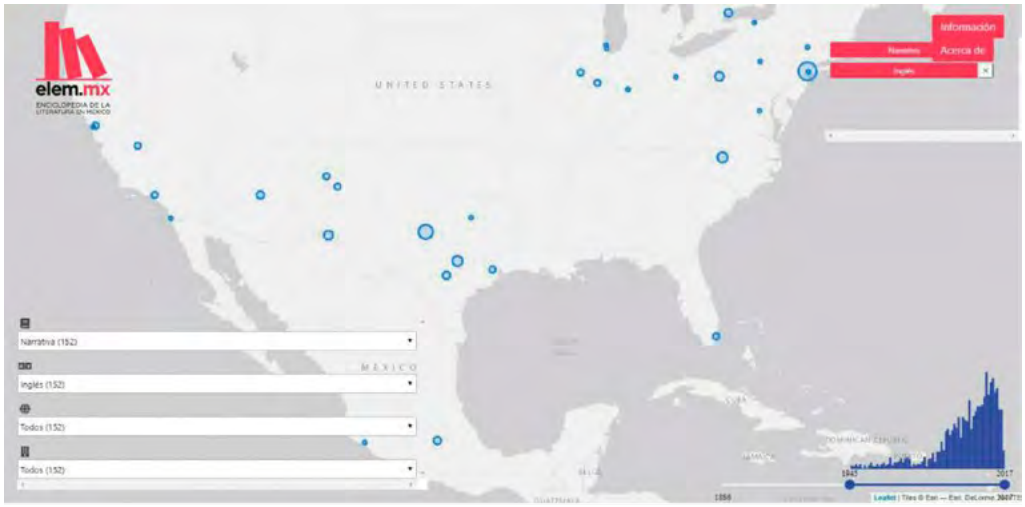


Imagen 3. Perspectiva del filtro: narrativa/inglés/1945-2017

Uno de los potenciales usos de esta herramienta puede ilustrarse a partir del siguiente ejemplo en el cual se usó el filtro de idioma (inglés), el de género literario (narrativa) y el rango de años de edición (1945-2017). La vista de los datos nos permitió observar un comportamiento no previsible. El título *Kill de Lion!* fue editado en México, D. F., en inglés. Es decir, el espacio geográfico no corresponde necesariamente con el espacio lingüístico, como se hubiera podido suponer en un principio.

Los especialistas e interesados en la cultura de la traducción literaria podrán contar con una visión de conjunto para realizar análisis e interpretaciones más minuciosas sobre la circulación de la cultura literaria de México a través de sus traducciones. Además, la puesta en mapa se irá actualizando conforme a las actividades de catalogación de la enciclopedia, lo que permitirá un acercamiento a las traducciones hacia otras lenguas aún no contempladas hasta ahora. Asimismo, los interesados en los contactos entre lenguas contarán con los insumos para poner en perspectiva las relaciones diglósicas, trasladadas a la cultura impresa, entre lenguas hegemónicas y lenguas minorizadas a partir de la traducción.

Referencias

Behrens, J. (1997), Principles and procedures of exploratory data analysis. *Psychological Methods* 2 (2): 131.

Boyd, M. (2012). *A Conflict of Narratives: The Influence of US Ideological Constructions of Mexican Identity in the Translation of Mexican Literature into English*. Universidad de York, Toronto, Canadá.

Burns, P. et al. (2017). Mapping Linked Data Subject Headings in the Library Catalog. *DH2017*. Montreal, Canadá.

Gutiérrez, A. (2017). Literature Translation. *GitHub*. https://github.com/amaurs/literature-translation

Küpper, K. (s.f.) Mexiko / Mittelamerika. *Archiv für übersetzte Literatur aus Lateinamerika und der Karibik*.http://www.lateinamerikaarchiv.de/antiquariat/mittelamerika-mexiko.html

Flexibility and Feedback in Digital Standards-Making: Unicode and the Rise of Emojis

S. E. Hackney
s.hackney@pitt.edu
University of Pittsburgh, United States of America

Background

The infrastructures that we use to navigate the world often become invisible as they become indispensable (Bowker and Star, 2000). However, critical examination of information systems is necessary to understand their implicit biases, and the ways that they invite some types of engagement and restrict others. Structures of power continue to be replicated in the ways that technologies are deployed in our lives (Noble, 2016; Tufekci, 2016), and the inability to access and assess the standards which make digital communication possible risks the uncritical perpetuation of those power structures (Drabinski, 2013). The moments of rupture, when an established system takes on a new facet with unintended consequences, can be an important moment of visibility, where we are able to reveal its ideological foundations, and the ways that its users adapt their own behaviors to it, or push back against its uncomfortable constraints (Raley, 2006; Marino, 2007). The introduction of emojis to the Unicode Standard, and their widespread adoption over the decade from 2006-2017 is one such moment of transition.

Scholars of standards and standardization argue that the input of users is necessary for a standard to meet the needs of those users (Foray, 1994), and while the process of adding content to the Unicode Standard remains rigid, the unicode.org website provides an explicit record of the development and evolution of the face that Unicode presents to its users, and is able to be read as a text which reveals the contemporary state of Unicode and the cultural ideologies which shape it.

Methodology

While major language- and script-based additions are made with each update to the Unicode Standard, my analysis focuses on changes to the unicode.org website, and its role as an intermediary document between the Consortium, the Standard itself, and everyday users. The introduction of emojis in various updates to the Standard has resulted in changes to the content and structure of the unicode.org website that reflect an increased engagement with end users, which I argue is the result of increased semantic value of emoji characters for the user[1], as compared to an individual character in a language's written script. It is my intention, through this analysis, to describe the types of changes that happen to the governing body and public documents of Unicode as major changes happen to the Standard itself.

A timeline was created of the dates of major updates to the Unicode Standard since its introduction in 1991, using the official release dates for updates to the Unicode Standard as maintained by the Unicode Consortium. I cross-reference this document with the rollout of each new version by the major platforms[2], with a particular emphasis on updates featuring new emoji characters, beginning with Unicode 6.0 in 2010[3].

With this timeline in mind, I scraped the unicode.org domain using Python and the Beautiful Soup[4] library to collect the URLS of all the unique pages under the parent domain, as well as a table of links between those pages. This serves as a source-target list for the creation of a network visualization of the unicode.org domain, using the network visualization software Gephi.[5] This process is repeated using archived versions of the unicode.org site, available from the Internet Archive's Wayback Machine[6], resulting in several structural snapshots of the unicode.org website over time, which can then be overlaid and compared to one another to note particular areas of change within the site.

Additionally, using points of change within the site structure as a guide, I also collect and code page content data to reflect the type of changes made to those pages during each major update. This coding is done on two axes: The first labels each change as being content- or structure-based (eg. adding text or links to a page, respectively), and the second designates which aspect of the Standard and/or Consortium is being addressed by the change. Examples of this second type of labelling would be "Emoji," "Membership," "Meta-Documentation," or "Language Scripts." This coding is done in two phases— an initial survey of this data in order to formally create labelling categories, and then a closer examination of the updates to apply those labels.

1 A notable exception to this semantic shift is written Chinese, which is already a semantic-character-based language, as opposed to syllable- or alphabet-based, as are the rest of the world's major languages. Thomas S. Mullaney gives a thorough historical analysis of the implication of this on text-encoding technologies in *The Chinese Typewriter* (MIT Press, 2017).

2 https://unicode.org/emoji/format.html#col-vendor lists the major "vendors" of emojis, or platforms with proprietary visual displays of emojis. These vendors are Apple, Google, Twitter, Facebook, Facebook Messenger, Windows, and Samsung.

3 While the first major batch of emojis were incorporated into Unicode in 2010, and the first official "Emoji 1.0" release was in 2015, work has been done within the standard since late 2006 to consider the addition and management of emoji-like characters within Unicode— hence the specific 2006-2017 emphasis of this research. (https://www.unicode.org/reports/tr51/#Introduction)

4 https://www.crummy.com/software/BeautifulSoup/

5 http://gephi.io

6 https://web.archive.org/

Discussion and next steps

This research project addresses issues of digital infrastructure from a unique angle: one that considers the socially-constructed nature of technology, as well as the meta-narrative of maintenance and upkeep of a system that has become crucial to our ability to communicate in a digital world. Through analysis of the secondary documents relating to the Unicode Standard, it is possible to gain invaluable insights into the ways that knowledge is organized collectively and continuously, as well as the embedded values that shape who can access and influence that knowledge.

This case study will provide a foundation for more expansive examination of systems of digital infrastructure. It is a beginning point both for further analysis of the adoption and adaptation of Unicode (and emojis in particular), but also as a framework for examining other forms of scaffolding which uphold the content of digital spaces.

References

Bowker, G. C., and Star, S. L. (2000). *Sorting things out: Classification and its consequences*. Cambridge: MIT Press.

Drabinski, E. (2013). Queering the catalog: queer theory and the politics of correction. *The Library Quarterly* 83(2): 94-111. doi:10.1086/669547

Foray, D. (1994). Users, standards and the economics of coalitions and committees. *Information Economics and Policy*, *6*(3): 269-293.

Marino, M. C. (2007, December 4). Critical code studies. *Electronic Book Review.* Retrieved from http://electronicbookreview.com/thread/electropoetics/codology

Noble, S.U. (2016). A future for intersectional black feminist technology studies. *The Scholar & Feminist Online.* 13.3 - 14.1. Retrieved from: http://sfonline.barnard.edu/traversing-technologies/safiya-umoja-noble-a-future-for-intersectional-black-feminist-technology-studies/0/

Raley, R. (2006). Code.surface || Code.depth, *Dichtung Digital*. Retreived from http://www.dichtung-digital.org/2006/01/Raley/index.htm

Tufekci, Z. (2016, June). *Machine intelligence makes human morals more important.* [Video file]. Retrieved from https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important

The Digital Ghost Hunt: A New Approach to Coding Education Through Immersive Theatre

Elliott Hall
elliott.hall@kcl.ac.uk
King's College London, United Kingdom

Figure 1 Heather Agyepong, disrupting an ordinary school day in KIT Theatre's Alfred the Great Time Travel Adventure

Introduction

The Digital Ghost Hunt combines coding education, Augmented Reality and live performance into an immersive storytelling experience. Students ten to eleven years old (Key Stage 2 in the UK) will explore the haunted Battersea Arts Centre with devices they've learned to program themselves. The key objective of The Digital Ghost Hunt is to present technology to students as an empowering tool, where coding emerges as – and fuses with – different forms of storytelling. It seeks to shift the context in which students see coding and engage groups who may be uninterested in or feel excluded by digital technology, opening up an imaginative space through play for them to discover the creative potential of technology on their own terms.

The Digital Ghost Hunt has been awarded funding through the UK Arts and Humanities Research Council (AHRC) New Generation of Immersive Experiences call, as part of an application led by Mary Krell, Senior Lecturer in Media Practice at the Centre for Material Digital Culture in the University of Sussex. A 'scratch' – a prototype of the experience – will be developed by Elliott Hall of King's Digital Lab and Tom Bowtell of Kit Theatre. It will be performed at the historic Battersea Arts Centre with a two-form entry of students from local schools.

Structure of the experience

The Digital Ghost Hunt is split into two parts. The first part begins with a regular coding class that suddenly goes haywire. While the teacher is trying to restore order, the lesson will be interrupted by Ms. Quill, Deputy Undersecretary of Paranormal Hygiene (Ghost Removal Section). She will enlist their help in the Ministry's work as apprentice ghost hunters. Students will use a simplified Python library to program their ghost hunting devices, which are based on two microcomputers: the Raspberry Pi and the BBC Micro:bit.

The coding in the project will focus in particular on two learning goals of the UK's National Curriculum: "Design, write and debug programs that accomplish specific

goals, including controlling or simulating physical systems; solve problems by decomposing them into smaller parts," and "Use sequence, selection, and repetition in programs; work with variables and various forms of input and output."It will teach students to take the overall goal of their devices – detecting 'paranormal' phenomena – and break it down into the discrete input, analysis and output tasks required, aided by the project's abstracted libraries.How they combine the functions of these libraries will be up to them, and will rely on their understanding of the fundamental logical structuresof programming to analyse sensor data, apply it to an algorithm, and debug when things go wrong.The project will also introduce students to embedded computing through the devices themselves. The emphasis will be on students taking ownership of their devices, deciding which of the ghost detectors they want to build and how it will work.

The second part is a ghost hunt, an immersive experience combining Augmented Reality (AR) and live theatre. Students will work together in small teams, using their devices to find objects and areas touched by the ghost. These traces will be both virtual objects in Augmented Reality, and actual physical phenomena such as radio waves, ultraviolet paint, and high-frequency sound. Each device will have different capabilities, forcing the students to work together to get all the clues. The ghost will in turn communicate with them, given life by actors, practical effects and the poltergeist potential of the Internet of Things. Only by using the devices they have programmed and working together can students unravel the mystery of why the ghost is haunting the building and set it free.

Coding, play and performance

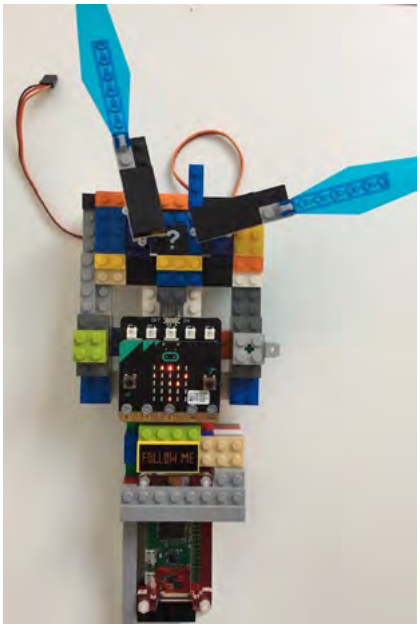


Figure 2 A proof of concept ghost hunting device using Lego and the Micro:Bit

Young people's familiarity with digital products are increasing, but their interest in learning the technology behind it is not, as evidenced in the UK by the low take up of the new GCSE in Computer Science(BCS, 2017). Teaching coding in schools is promoted by the UK Department of Education (DOE, 2014), but students often experience coding education as a classroom assignment, divorced from their intuitive and creative experiences with commercial digital applications.

There are several applications now using AR as a teaching tool (for example, The Battle of Mount Street Bridge (Schreibmen et al., 2017) and Virtual Roman Frontiers (Wilson at al.)) and initiatives to teach children coding, from commercial apps to coding clubs and the work of the Raspberry Pi and Micro:Bit foundations. These applications all seek the increase in engagement and experimentation that can occur when 'work' is reframed as 'play.' (Pellegrini, 2009)

However, these applications all take place within a screen, an approach that creates its own problems. A screen can shift a user's attention to the digital environment to the exclusion of the physical one. (Chrysanthi, 2012) The Ghost Hunt's approach is to bring AR interaction fully into the physical space without the mediating influence of a screen, reconnecting audiences to the world around them.

The addition of immersive theatre reframes the experience again, from 'play' to 'performance.' This second shift is important to reach groups not engaging with existing digital resources. In 2016, girls made up just 20% of entrants for the computer science exam, while pupils on free school meals made up just 19% of GCSE entrants even though they are 27% of the population (Cellan-Jones, 2016). Performance may draw in groups who would otherwise be uninterested in or feel excluded from traditional Computer Science education.

However, the performance should not be seen as secondary in any way to the coding elements of the project. The aim of the project is to expand the imaginative possibilities of digital technology through play; the coding elements are the means to that end, not the other way around. The Ghost Hunt seeks to shift how the context of computer science is perceived, from a skill intended only for a narrow group to a tool of creativity and play available to all.

As part of its evaluation, the project will use the student's code and feedback from educators on how the software libraries are used, as well as video, audio and device logging during the experience. It will be direct engagement with participants through formal and informal methods such as interviews, questionnaires and the creative material they create as part of the experience that will provide the crucial method of evaluation. The only way to assess the pedagogical value of the project in terms of creating a new and sustained interest in the possibilities of digital technology will be if students create new things on their own initiative, independent from the project's se-

tting and materials. This metric is beyond the scope of the pilot project but is something the project team are eager to explore in subsequent phases in collaboration with the educational partners.

Beyond the hunt

The lessons of the Digital Ghost Hunt scratch funded by the AHRC will direct refinement of the existing tools towards developing a technical and conceptual framework that can be adapted and implemented for different locations, stories and audiences. This short paper aims to present the practice-based collaborative framework of the Digital Ghost Hunt as conceived by its creators in its first funded iteration to elicit feedback from the Digital Humanities 2018 participants and integrate it into future development.

References

British Chartered Institute for IT (BCS). (2017) [online] *BCS deeply concerned over stagnation of number of Computer Science GCSE applicants*. Available at: http://www.bcs.org/content/conWebDoc/57904 *[Accessed 23/11/2017]*

Cellan-Jones, Rory. *(2017).* Computing in schools - alarm bells over England's classes. *BBC News*. [online.] Available at: *http://www.bbc.co.uk/news/technology-40322796 [Accessed 23/11/2017]*

Department of Education (DOE). (2014.) Teaching children to code. In: *D5: London*. London. Available at: https://www.gov.uk/government/publications/d5-london-summit-themes/d5-london-teaching-children-to-code [Accessed 23/11/2017]

Pellegrini, A. (2009) *The role of play in human development*. Oxford: Oxford University Press.

Chrysanthi, A., Papadopoulos, C., Frankland, T., and Earl, G. (2013). 'Tangible Pasts': User-centred Design of a Mixed Reality Application for Cultural Heritage. In: *Conference of Computer Applications and Quantitative Methods in Archaeology*. Southampton: Amsterdam University Press, pp. 31-41.

Schreibman, S, Papadopoulos, C., Hughes, B., Rooney, N., Brennan, C., Fionntann, M., Healy, H. Phygital Augmentations for Enhancing History Teaching and Learning at School. In: *Digital Humanities 2017*. Montreal. [online] Available at: https://dh2017.adho.org/abstracts/401/401.pdf [Accessed 23/11/2017]

Wilson, L., Weeks, P, Rawlinson A., Dobat, E., Fluegel, C., Hermann, C. (2017). Virtual Roman Frontiers: 3D Visualisation and Innovative Technology Applications for the Antonine Wall. In: *3D Imaging in Cultural Heritage*. London: The British Museum. Available at: https://www.3dimaginginculturalheritage.org/resources/3D_Imaging_in_Cultural_Heritage_Abstracts.pdf [Accessed 23/11/2017]

Exploration of Sentiments and Genre in Spanish American Novels

Ulrike Edith Gerda Henny-Krahmer
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Germany

Background, aims, and hypotheses

In 19th century Spanish American novels, the expression of emotionality is an essential characteristic of the texts belonging to different subgenres.[1] Especially during the Romantic period in the first half of the century, many sentimental novels have been written (Zó, 2015). But emotions also play an important role in other types of novels: a love story is often a basic plot element for example in historical or costumbrista novels. Also, there are novels characterized more by negative emotions, like Cuban anti-slavery novels (Rivas, 1990), Argentine anti-Rosas novels (Molina, 2011: 285-312, García Ardeo, 2006), or sociopolitical novels in general.

In text mining, a common method to analyze emotions is Sentiment Analysis (Pang and Lee, 2008). Sentiment Analysis is the computational treatment of sentiment, opinion, or emotion in text. Sentiments are usually modelled in terms of polarity values (positive, negative, neutral) or emotion values (such as trust, fear, joy, etc.).

The aim of this proposal is to test several hypotheses about sentiments in subgenres with an explorative analysis of a corpus of Spanish American novels. To this end, sentiment values are used as features in a text classification task. A secondary objective of this contribution is to compare the results of two different sentiment lexica for Spanish to see how well they perform.

The first hypothesis of this proposal is that the degree and kind of emotionality in the novels differs for different subgenres. The second hypothesis here is that not just emotions in general matter, but also whether they are expressed in the direct speech of the characters of the novels or in narrated text.[2]

State of the Art

Two recent examples for the usage of Sentiment Analysis with literary texts are Zehe et al., 2016 for the prediction of happy endings in German novels and Kim et al., 2017 for the analysis of prototypical emotion developments in literary genres with English texts. Sentiment Analysis has been used with Spanish texts, as well, mainly for the analysis of reviews and tweets (see Henríquez Miranda and Guz-

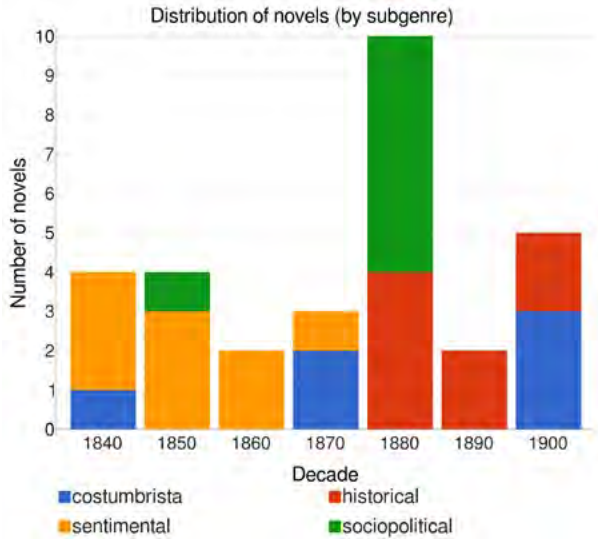
1 This contribution is concerned with the linguistic manifestation of emotions in literary texts on the textual surface. See Winko, 2003 for a discussion of how emotional meaning and literary texts are related

2 The anti-slavery novel, for example, has been defined in terms of its atmosphere of fear, but also by vigorous interferences of the narrator. Cf. Rivas, 1990.

mán, 2017 for an overview). To the best of my knowledge, there are no applications of Sentiment Analysis on Spanish novels yet, and the distinction of direct speech and narrated text has not previously been used in combination with the analysis of sentiments in literary texts.

Data

For this analysis, a corpus of 30 Spanish American novels has been selected. The collection has the following characteristics: The novels have been published between 1840 and 1910 (13 before 1880 and 17 after 1880), are from three countries (Argentina: 16, Cuba: 9, Mexico: 5), and have been written by 16 different authors.[3] Fig. 1 shows the distribution of novels per decade and subgenre:



Distribution of novels per decade and subgenre

As the texts at hand are not easily distinguishable genre fiction but more general literary fiction, the assignment of subgenre labels is a non-trivial task. For the assignment of subgenre labels to the novels, the subgenres as given in titles and subtitles of the novels were collected and subgenre assignments made in secondary literature were considered. Both types of information were used to derive four kinds of interpretive[4] subgenre labels corresponding to four broad types of novels: costumbrista (6 novels),[5] historical (8), sentimental (9), and sociopolitical (7) novels.[6]

Methods

In general, Sentiment Analysis can be done with a machine learning approach and a lexicon-based approach. Here, two sentiment lexica were used: (1) SentiWordNet 3.0, an adaptation of WordNet 3.0 for sentiment analysis (Miller, 1995, Baccianella et al., 2010) and (2) the NRC Emotion Lexicon (Saif and Turney, 2013). The two lexica differ in how sentiments are modelled and also in their volume. SentiWordNet has polarity values (positivity, negativity, neutrality) for WordNet synsets which range between 0 and 1 and sum up to 1. The NRC lexicon, in contrast, has only binary values (0 or 1), but those are provided for positivity and negativity as well as eight basic emotions (Trust, Fear, Joy, Sadness, Anger, Disgust, Anticipation, Surprise). SentiWordNet contains 117,653 entries, the NRC lexicon just 14,182.[7]

In order to use the sentiment lexica, the texts had to be lemmatized (for NRC) and annotated with WordNet synsets (for SentiWordNet) which was done with the NLP library FreeLing (Padró and Stanislovsky, 2012). To be able to use the distinction between direct speech and narrated text as a feature, the texts were annotated semi-automatically in their TEI master files (see Fig. 2):

```
<p><said>—Parece que duerme</said>, dijo examinando
atentamente las facciones de la viejecita, <said>¡quiera Dios que
este sueño alivie sus dolencias y reponga en un tanto sus ya
gastadas fuerzas!</said></p>
```

Example of a paragraph with annotated direct speech, from „Camila o la virtud triunfante“ (1856) by Estanislao del Campo

Each paragraph was split into sentences. Each sentence was annotated with FreeLing and the words with sentiment values were determined using the lexicons. The sentiment values for the words were summed up for each sentence.[8] For the eight basic emotions of the NRC (Trust, Fear, etc.), a sentence is assigned the emotion with a highest value in the sentence. Besides the sentiment features that come directly from the lexicons, the following features were determined for each sentence:

A Decision Tree classifier was used for the classification of the novels by subgenre, using the above-mentioned features (see Manning and Schütze, 1999: 578-589 on this method). The advantage of Decision Trees is that

3 This is a subcollection of a larger corpus of Spanish American novels being prepared in the context of the junior research group Computation Literary Genre Stylistics (CLiGS), see https://cligs.hypotheses.org/sprachen/english.
4 Because the many variations found had to be normalized for this computational analysis, an interpretive step was unavoidable.
5 Novels of manners in the context of the Costumbrismo movement.
6 The distribution of novels shows that there is a tendency for sentimental novels to belong to the first half and for non-sentimental novels to the second half of the century. This observation may be relevant for future tasks with a bigger corpus and interested in the development of genres over time. More detailed metadata for the novels can be found at https://github.com/cligs/projects2018/blob/master/sentgenre-dh/metadata.csv.
7 SentiWordNet can be used for Spanish because the synset IDs can be mapped to the Spanish version of WordNet. The NRC lexicon has been translated into Spanish automatically. See Baccianella et al., 2010 for evaluation reports for SentiWordNet. The authors of the NRC lexicon state that the translated versions may contain errors. An orthographic check on the NRC lexicon returned 409 entries that were not recognized as Spanish words. A further evaluation and improvement of the translated lexica is desirable.
8 The Sentiment Analysis could be refined further by considering the sentence structure (and negation), which is a future task.

they can be interpreted. This is desirable in an explorative analysis interested in the kind of sentiment-based features that are relevant to differentiate novels of different subgenres. When compared to other types of classifiers, Decision Trees do not necessarily yield the best results in terms of accuracy, but their interpretability is valued higher here in order to gain insight into how sentiments, the opposition of direct speech vs. narrated text, and subgenres are related.

Feature name	Description
emotional	Proportion of emotional sentences in the text. To determine emotionality, a threshold of 1 was set: all sentences with a positive value > 1 or a negative value < -1 were considered emotional.
neutral	Proportion of neutral sentences in the text, with a sentiment value between -1 and 1.
positive	Proportion of positive sentences in the text, with a sentiment value > 1.
negative	Proportion of negative sentences in the text, with a sentiment value < 1.

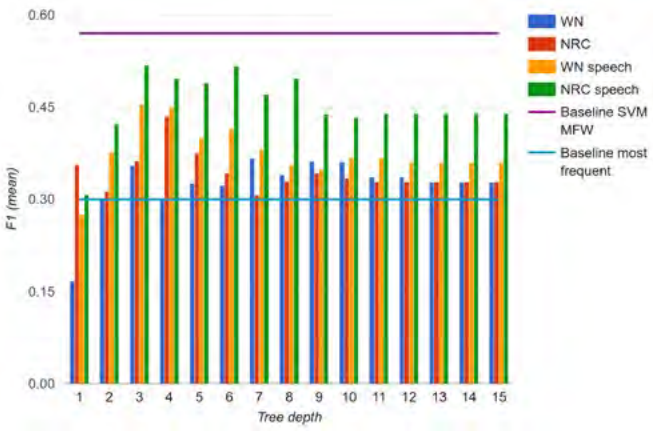
Additional features for the Sentiment Analysis

To generate data for the machine learning task, the values of the single sentences were aggregated into five sections and divided by the section length (number of sentences contained in the section), resulting in 150 data points for the 30 novels. 60 different experiments were run, varying the sentiment features and lexicon used, and the depth of the decision tree. A 5-fold cross-validation was applied.

Results and Discussion

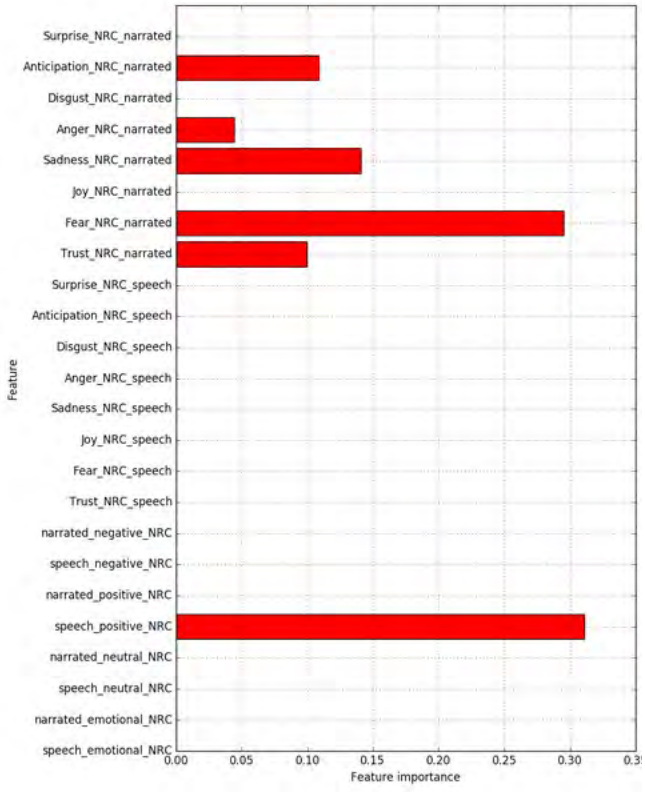
The results of the experiments are summarized in Fig. 4 below. The depth of the Decision Tree was varied between 1 and 15.[9] The accuracy is given as the mean F1 score obtained from the cross-validation. Four different sets of sentiment features were used: Features from the SentiWordNet lexicon (WN) and from the NRC lexicon (NRC), both without differentiating between direct speech and narrated text, as well as WN- and NRC-features with separate sentiment values for direct speech and narrated text (WN speech and NRC speech). The results of all experiments are compared to the "most frequent"-baseline and to a baseline obtained with an SVM classifier, using the 5,000 most frequent words.

Although the F1 scores are not very high (the highest mean value being at 0.52), almost all of them outperform the "most frequent"-baseline (0.3) which confirms that sentiment features are relevant for subgenre classification. Still, the results do not reach the best mean score of the MFW classification (0.57).[10] In terms of classification accuracy, a next step will be to combine both sentiment features and MFW to see if the sentiment features can contribute to improve the overall results.



Results for subgenre classification with sentiment features

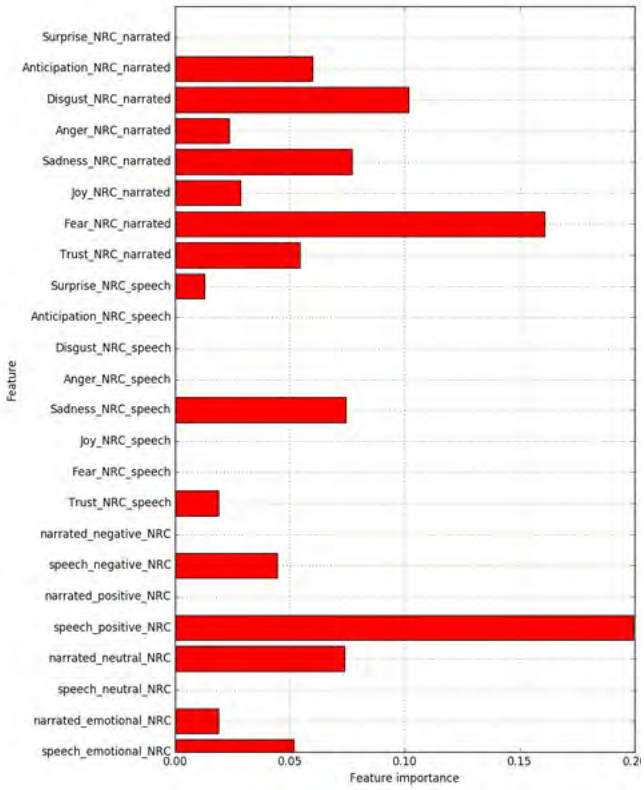
When comparing the results for the two different sentiment lexica, the NRC lexicon performs better than SentiWordNet, although the latter covers almost ten times as many words as the first one. A look into the feature importance shows that the eight basic emotions, which are only present in the NRC lexicon, are crucial (see Fig. 5 and 6).



Feature importance for a tree with depth 3, using NRC and speech vs. narrated text

9 Restricting the tree depth helps to prevent overfitting and usually leads to a better performance of the classifier on the test set.

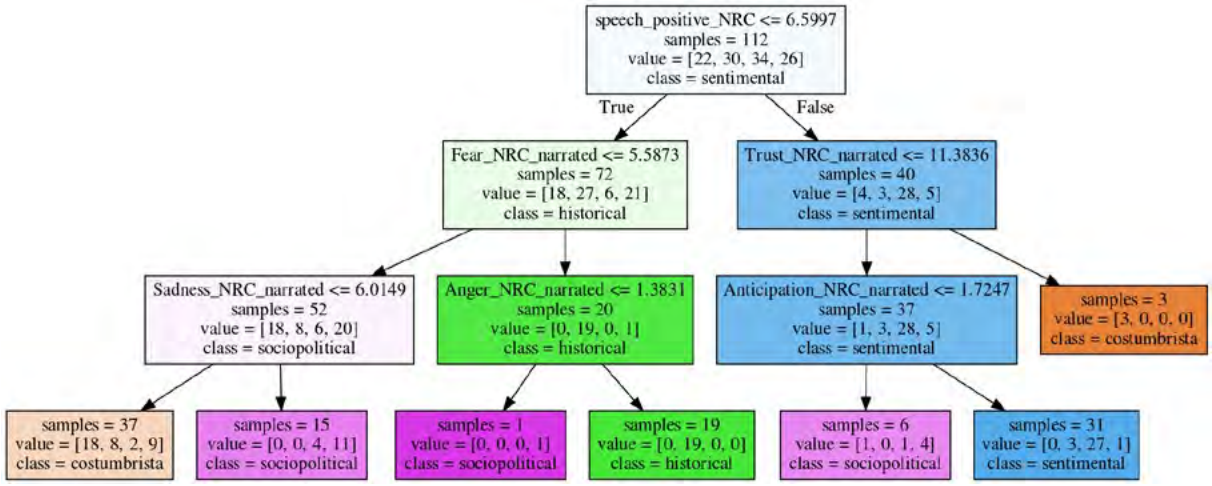
10 See Hettinger et al., 2016 for a discussion of various types of features (MFW, topics, networks) for subgenre classification, stating that genre classification in general works best with most frequent words, all words, and the like.



A Decision Tree for the classification of subgenres, based on the best parameters

Regarding the difference between speech and narration, Fig. 4 above also shows that the highest values for both WN and NRC are reached when the sentiment values are calculated separately for direct speech and narrated text. The best scores are obtained for the feature set "NRC speech". The most important feature in both example trees is positive speech, followed by narrated fear. Fig. 7 shows the Decision Tree corresponding to the feature importance in Fig. 5 above.

The tree shows that novels with higher values of positive speech are more likely to be sentimental novels. Other features that contribute to the distinction of sentimental novels are lower values of trust and higher values of anticipation in narrated text. The path for historical novels includes less positive speech and more fear and anger in narrated text. Costumbrista novels are characterized by less sadness in narrated text than sociopolitical novels and by more trust in narrated text than sentimental novels. Sociopolitical novels differ from historical novels in that they have a lower value of fear and anger in narrated text.[11]



Feature importance for a tree with depth 6, using NRC and speech vs. narrated text

11 The results of all experiments can be found at https://github.com/cligs/projects2018/tree/master/sentgenre-dh/.

Conclusion and Future Work

This exploration of sentiments in Spanish American Novels showed that Sentiment Analysis can be used as a basis for subgenre classification tasks. It has been shown that the distinction between emotions in direct speech and emotions in narrated improves the classification results considerably. Regarding the two sentiment lexica that were tested, the NRC Emotion Lexicon performs better than SentiWordNet.

The Decision Trees resulting from the classification give much insight into how sentiments in general, in direct speech and in narrated text are related to different types of novels. That the features can be interpreted easily contributes to a better understanding of what textual features are connected to the subgenres, but the classification results themselves can still be improved. Other classifiers, for example Random Forest trees or an SVM, might yield better results but will also be less interpretable. Another important next step is to increase the corpus size to make the results more stable.

References

Baccianella, S., Esuli, A. and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of LREC 2010.* Valletta, Malta: ELRA: 2200-2204. http://www.lrec-conf.org/proceedings/lrec2010/summaries/769.html (accessed April 27 2018).

García Ardeo, J. M. (2006). Eduardo Gutiérrez y sus dramas del terror. *Letras* 54: 77-94.

Henríquez Miranda, C. and Guzmán, J. (2017). A Review of Sentiment Analysis in Spanish. Una Revisión Sobre el Análisis de Sentimientos en Español. *TECCIENCIA* 12 (22): 35-48. doi: 10.18180/tecciencia.2017.22.5.

Hettinger, L., Jannidis, F., Reger, I. and Hotho, A. (2016). Classification of Literary Subgenres. *DHd2016.* Leipzig: Universität Leipzig: 154-158. http://dhd2016.de/boa.pdf (accessed April 27 2018).

Kim, E., Padó, S. and Klinger, R. (2017). Prototypical Emotion Developments in Literary Genres. *Digital Humanities 2017. Conference Abstracts.* Montréal: McGill University. https://dh2017.adho.org/abstracts/203/203.pdf (accessed April 27 2018).

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Mass.: The MIT Press.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11), 39-41.

Molina, H. B. (2011). *Como crecen los hongos. La novela argentina entre 1838 y 1872.* Buenos Aires: Teseo.

Padró, L. and Stanislovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012).* Istanbul, Turkey: ELRA: 2473-2479. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf (accessed April 27 2018).

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2 (1-2): 1-135.

Rivas, M. (1990). *Literatura y esclavitud en la novela cubana del siglo IXI.* Sevilla: Escuela de Estudios Hispano-Americanos.

Saif, M. and Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29 (3), 436-465.

Zehe, A., Becker, M., Hettinger, L., Hotho, A., Reger, I., and Jannidis, F. (2016): Prediction of Happy Endings in German Novels based on Sentiment Information. *Proceedings of DMNLP, Workshop at ECML/PKDD.* Riva del Garda, Italy. http://ceur-ws.org/Vol-1646/paper2.pdf (accessed April 27 2018).

Zó, R. E. (2015). *Emociones escriturales. La novela sentimental latinoamericana.* Saarbrücken: Editorial Académica Española.

Winko, S. (2003). Über Regeln emotionaler Bedeutung in und von literarischen Texten. In Jannidis, F., Lauer, G., Martínez, M., Winko, S. (eds.), *Regeln der Bedeutung.* Berlin: de Gruyter, pp. 329-348.

Digitizing Paratexts

Kate Holterhoff
kate.holterhoff@gmail.com
Georgia Institute of Technology, United States of America

Digital archivists tend to disagree about the place of paratexts. Whereas *Google Books* often scans texts at such a low resolution that anything but printed words are difficult to discern, Andrew Stauffer's *Book Traces* project and Steven Olsen-Smith and Peter Norberg's The Melville Marginalia Project aims to identify individual copies of nineteenth- and early-twentieth-century books in libraries by highlighting their unique marginalia and inserts. Illustrations, advertisements, marginalia, boards, and decorative initials—the effluvium of the print form—does not digitize easily. Moreover, in terms of library and information science, paratexts resist standard means of categorization. Paratexts are problematic because they offer an exception rather than a type. To scholars, they often seem extraneous or even detrimental to the written texts they accompany. Marginalia, for instance, simultaneously defaces and compliments a text. Advertisements are a distracting and commercial accretion to an artwork. And yet, all paratexts provide necessary context for understanding the complexity and fullness of print history. The question I will address in this paper is how archivists ought broadly to understand paratexts, and how specifically should they treat nineteenth-century illustrations.

Numerous digital archives have taken on the task of scanning, categorizing, and tagging illustrations (*e.g.* the *William Blake Archive,* the *Cervantes Project*, Cardiff University's *Illustration Archive*), and yet the purpose and constraints of this task remain unfixed. In fact, Julia Tho-

mas notes in her recent *Nineteenth-Century Illustration and the Digital* (2016), that owing to the uniquely important role of context for these paratexts—usually the book or periodical—"the digital might appear an alien environment for historic illustrations." While the role of the digital image archive concerned with illustrations remains unsettled, recent scholars have used the affordances of the digital archive to open up new avenues for curation and exploration. Using as a case study a digital archive that I direct and edit titled *Visual Haggard*, a *NINES* indexed and peer reviewed archive that contextualizes and improves access to the illustrations of Victorian novelist H. Rider Haggard (1856 - 1925), I argue that digitizing illustrations must be inclusive.

I will consider the problem of inclusion and exclusion in digital archive curation. As paratexts, illustrations are lumped together with a number of visual objects that initially accompanied fictions. For this reason I explain the necessity of using metadata to differentiate illustration types. The large decorative initials which appear in many nineteenth-century texts, but originated in medieval manuscripts, are less illustrations of the text than embellishments. However, their ideological function is significant and multifold. Similarly, advertisements were often in conversation with serialized fictions—whether thematically or stylistically. In this paper I discuss strategies to enable digital image archivists committed to creating an authentic encounter with the history of print to avoid ignoring or marginalizing these types of unique and difficult paratexts.

A Corpus Approach to Manuscript Abbreviations (CAMA)

Alpo Honkapohja
alpo.honkapohja@ed.ac.uk
University of Edinburgh, United Kingdom

As anyone, who has worked with medieval manuscripts, will know, sometimes more than half of the words are abbreviated. For example, in a forthcoming paper on Middle English and Latin manuscripts of the *Polychronicon*, we found that in the most heavily abbreviated Latin sections as many as 59 percent of the words could be abbreviated, while the number for Middle English was 21 per cent (Honkapohja and Liira, in preparation). Studies comparining Latin and Romance have met with similar results (Hasenohr, 1997; Careri et al., 2011). Nevertheless, in digital scholarship, abbreviations are typically seen as something to get rid of rather than useful data to mine.

A major reason for lack of attention given to manuscript abbreviations can be found in editorial practices inherited from printed editions. It is a standard practice for editors to expand abbreviations as "a service to the reader" (cf. Driscoll, 2009). Twentieth-century editorial theory often treats abbreviations as scribal variation as "accidentals" (see e.g. W. W. Greg, 1950), not relevant for the authorial "work" contained in the manuscripts, as much scholarship focuses either directly on the work or uncovering the work under layers of scribal copying and errors. The outcome is an editorial tradition in which silently expanding abbreviations is very much the norm.

Digital approaches for making use of abbreviations as data are available, but are often not used. TEI P5 guidelines introduced the possibility of encoding both the abbreviations and their expansions using the <choice> elements with <abbr> and <expan> (cf. Driscoll, 2006, 2009; Honkapohja, 2013). Still, many digital resources continue the practice of silently expanding abbreviations. Reasons may range from considering encoding abbreviations to be too labour intensive to basing the digital resources on printed editions which expand the abbreviations (cf. Honkapohja et al., 2009). Moreover, text retrieval systems are typically unable to recognize different forms of the same word and the problem is usually solved by normalisation (cf. Kestemont, 2015: 160). Furthermore, some research questions, including investigations into syntax or stemmatology, also require normalisation. However, while normalisation may be necessary for some research questions, it also discards large amount of potentially useful data, which makes other types of research impossible.

The fairly few scholars who do work with abbreviations have identified a number of potentially interesting lines of enquiry. Abbreviations can be used, for example, for identifying change of scribe in the text (cf. Kestemont, 2015) or in historical dialectology for identifying regional characteristics in scribal language (see e.g. Smith, 2016), or studying the effect of right-margin justification on scribal spelling (Shute, 2017), or hiding endings in multilingual business writing (Wright, 2011). Consequently, the practice of expanding abbreviations is discussed and criticised by a number of scholars (Driscoll, 2006; Kytö et al., 2011; Rogos 2011, 2012; Stutzman, 2014, Lass, 2004).

Even though the problems related to the prevailing practice of silently expanding are well known, and some resources such as the *Medieval Nordic Text Archive* (MENOTA) do encode them, there have been relatively few studies which would have attempted to use them as data (e.g. Camps, 2016; Honkapohja, 2018; Kestemont, 2015; Rogos, 2012; Smith, 2016; Shute, 2017), especially in comparison to fields such as stemmatology and stylometry. My proposal for short paper presents project plan and early results for a project, called *Corpus Approaches to Manuscript Abbreviations* (CAMA), funded for September 2017- February 2020.

The current project focuses on applying methodologies developed for corpus linguistics on abbreviations in the spelling system of Early Middle English, 1150-1350. The period is of interest as it was a formative one for the writing systems of English. Linguistic situation in England changed dramatically after the Norman Conquest of 1066, which introduced a new ruling class and relegated Engli-

sh to a tertiary role after Latin and Anglo-Norman French. When Middle English texts become more numerous in the 13th century, we find a very diverse dialect landscape in which the lack of a prestigious vernacular has led to the proliferation of local varieties, with almost every text appearing to represent a separate linguistic system.

Within the Early Middle English period, my project focuses on four research questions:

- (Q1) Does each scribe have an individual scribal profile of abbreviations?
- (Q2) Are some abbreviation usages connected to certain geographic areas?
- (Q3) How are Latin and Old English abbreviations distributed in Germanic and Romance vocabulary?
- (Q4) What is the function of abbreviations in the spelling system(s) of Middle English?

The data comes from the *Linguistic Atlas of Early Middle English* (LAEME), a corpus of ca. 650,000 divided into scribal samples of localised Middle English. Each text in LAEME is based on a diplomatic transcription from manuscript facsimiles, not editions, and using a mark-up system that encodes the expansions of abbreviations, but in a way which makes identifying the abbreviation easy and workable (LAEME: 3.3.1). Consequently, it can be used to compile a dataset, which can be analysed quantitatively.

The methodology is based on corpus linguistics, statistical analysis and historical dialectology. I will use corpus enquiries to compile a dataset of the findings, then subject the dataset to statistical analysis using R and tried and tested techniques such as linear regression, linear correlation, principal component analysis, chi square test and cluster analysis which have yielded results in previous studies of abbreviations and spelling variation (cf. Kestemont, 2015; Smith, 2016).

Compiling the dataset consists of three steps:

1. Corpus enquiries, using the web interface and scripts of LAEME.
2. Corpus enquiries for unabbreviated forms of the abbreviated words found in stage 1 in each text, in which a particular abbreviation is used. These can be localised, using the lemmas tagged in the LAEME (see 2.3.2: E).
3. Compiling a dataset the results, which will include **a) results of the corpus enquiries**, i.e. the abbreviation type, the abbreviated word, non-abbreviated variant(s), frequencies, **b) information included in the LAEME metadata**, i.e. text, lemma, grammatical tag, manuscript, date, script, place, co-ordinates in the LAEME localisation grid, and **c) additional variables needed for research questions Q1 and Q3**, i.e. word origin: Germanic/Romance/Latin (12), content vs. function word (13).

The dataset will be subjected to further analysis, using:

A) The inbuilt mapping function in LAEME, which allows dynamically creating feature maps, based on the distribution of any form, its lemma, or grammatical tag.
B) Statistical analysis,
   a) linear correlation and linear regression, using the form of the abbreviation as the dependant variable, and the results encoded in the dataset (2.3.3: 3) as independent variables, calculating which of them interact with the type of the abbreviation in a certain specimen (cf. Smith, 2016),
   b) Principal component analysis common in stylometry (cf. Kestemont, 2015: 168-70).

As I am giving the presentation fairly early in the funding period, I hope to receive valuable feedback on the methodology and also to build a bridge between corpus linguistics and stylometry, creating discussion on the value and potential of scribal 'accidentals' as data.

References

Camps, J-B. (2016). *La `Chanson d'Otinel':* édition *complète du corpus manuscrit et prolégomènes à l'édition critique*. Paris-Sorbonne.<https://doi.org/10.5281/zenodo.1116735>

Careri, M., de Saint-Pol Ruby, C. and Short, I. (2011). *Livres et* écritures *en français et en occitan au XIIe siècle: catalogue illustré*, *Scritture e libri del Medioevo 8*, Viella.

Driscoll, M. (2006). Levels of transcription. In Burnard, L., O'Brien O'Keeffe, K. and Unsworth, J. (eds), *Electronic Textual Editing*, Modern Language Association, pp. 254–261.

Driscoll, M. (2009). Marking up abbreviations in Old Norse-Icelandic manuscripts. In M.G. Saibene, M. and Buzzoni, M. (eds), *Medieval Texts – Contemporary Media: The Art and Science of Editing in the Digital Age*. Ibis, pp. 13–34.

Greg, W. W. (1951/1951). The Rationale of Copy-Text. *Studies in Bibliography* Vol. 3, pp. 19-36.

Hasenohr, G. (1997). Écrire en latin, écrire en roman: réflexions sur la pratique des abréviations dans les manuscrits français des XIIe et XIIIe siècles. In Banniard, M. (ed.), *Langages et peuples d'Europe: cristallisation des identités romanes et germaniques (VIIe-XIe siècle*). Toulouse-Conques, pp. 79-110.

Honkapohja, A. (2018). "Latin in Recipes?" A corpus approach to scribal abbreviations in 15th-century medical manuscripts. In Pahta, P, Skaffari, J. and Wright, L. (eds), *Multilingual Practices in Language History: English and beyond*. De Gruyter, pp. 243-271.

Honkapohja, A. (2013). "Manuscript abbreviations in Latin and English: History, typologies and how to tackle them in encoding." *Studies in Variation, Contacts and Change in English Volume 14: Principles*

and Practices for the Digital Editing and Annotation of Diachronic Data. <http://www.helsinki.fi/varieng/series/volumes/index.html>

Honkapohja, A., Kaislaniemi, S. and Marttila, V. (2009). Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora. In Jucker, A., Schreier, D. and Hundt, M. (eds), *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*. Ascona, Switzerland, 14–18 May 2008. Rodopi, pp. 451–474.

Honkapohja, A. and Liira, A. (in preparation). Abbreviations and Standardisation in the *Polychronicon*: Latin to English, and Manuscript to Print. In Wright, L (ed.), *The Multilingual Origins of Standard English (MOSTE)*. De Gruyter.

Kestemont, M. (2015). A Computational Analysis of the Scribal Profiles in Two of the Oldest Manuscripts of Hadewijch's Letters, *Scriptorium,* 69: 159-75.

Kytö, M., Grund, P. and Walker, T. (2011). *Testifying to Language and Life in Early Modern England: Including CD-ROM: An Electronic Text Edition of Depositions 1560-1760 (ETED)*. Benjamins.

LAEME = Laing, M. (2013). *A Linguistic Atlas of Early Middle English, 1150–1325*, Version 3.2. Edinburgh: © The University of Edinburgh.: <http://www.lel.ed.ac.uk/ihd/laeme1/laeme1.html>

Lass, R. (2004). Ut Custodiant Litteras: Editions, Corpora and Witnesshood. In Dossena, M. and Lass, R. (eds), *Methods and Data in English Historical Dialectology*. Peter Lang, pp. 21-48.

MELD = Stenroos, M., Thengs, K. and Bergstrøm, G. *A Corpus of Middle English Local Documents*, v. 2017.1. University of Stavanger. <http://www.uis.no/research/history-languages-and-literature/themest-programme/a-corpus-of-middle-english-local-documents-meld/>

MENOTA = *Medieval Nordic Text Archive*. <http://www.menota.org/EN_forside.xhtml>

Rogos, J. (2011). On the pitfalls of interpretation: Latin abbreviations in MSS of the Man of Law's Tale. In Fisiak, J. and Bator, M. (eds), *Foreign Influences on Medieval English*. Peter Lang, pp. 47–54.

Rogos, J. (2012). Isles of systemacity in the sea of prodigality? Non-alphabetic elements in manuscripts of Chaucer's 'Man of Law's Tale. <http://www.isle-linguistics.org/resources/rogos2012.pdf>

Shute, R. (2017). Pressed for Space: The Effects of Justification and the Printing Process on Fifteenth-Century Orthography. *English Studies* 98 (3): 262–82.

Smith, D. (2016). The predictability of {-S} abbreviation in Older Scots manuscripts according to stem-final littera. AMC Symposium. Conference paper.

Stutzmann, D. (2014). Conjuguer diplomatique, paléographie et édition électronique : les mutations du XIIe siècle et la datation des écritures par le profil scribal collectif. In Ambrosio, A., Barret, S. and Vogeler, G. (eds), *Digital Diplomatics. The computer as a tool for the diplomatist?*, Archiv für Diplomatik. Beiheft 14, 27190.

Wright, L. (2011). On Variation in Medieval Mixed-Language Business Writing. In Schendl, H. and Wright, L. (eds.), *Code-Switching in Early English*. De Gruyter, pp. 191-218.

On Natural Disasters In Chinese Standard Histories

Hong-Ting Su
r03944039@ntu.edu.tw
National Taiwan University, Taiwan

Jieh Hsiang
jhsiang@ntu.edu.tw
National Taiwan University, Taiwan

Nungyao Lin
nungyao@gmail.com
National Taiwan University, Taiwan

This paper describes a study which analyzes natural disasters described in the Chinese Standard Histories. We first define the scope and nature of disasters as presented in the Standard Histories. The records, in plain text but usually contain the dates, locations, type, and severity of the natural disasters, are then extracted. The extracted records are further annotated with metadata so as to meet the needs of the studies on the history of disasters. In order to ensure flexibility and extensibility, we have designed a markup language, WXML, to tag the information. A search/retrieval system with GIS is then developed to provide visualization of the distribution of time, space, and type of disasters of the search result.

We have made some preliminary observations. For instance, the number of disasters recorded during the Yuan Dynasty is significantly higher than the other dynasties (both in absolute number and on average). As another example, disasters seem to disproportionately concentrate around urban centers, in particular the capital of the time. This shows that the records in the Standard Histories may not accurately reflect the actual events, but rather how they were documented by the officials.

Natural Disasters described in the Chinese Standard Histories

Chinese Standard Histories (正史), 24 in total, are the official histories of the Chinese Dynasties. A Standard History is usually written during the succeeding dynasty, based on existing, often meticulously kept, records of the previous dynasty. These tomes start from *Shiji* (史記), written by Sima Qian (司馬遷) in the Han Dynasty around 90 BCE, and ends with *Ming Shi* (明史), the Standard His-

tory of Ming Dynasty (1368-1644). Together they cover about 2,500 years of China's written history. Fourteen of the standard Histories contain volumes of *Wuxingzhi* (Book of the Five Elements, 五行志), which record natural disasters and mysterious phenomena. Disasters are also documented in the *Benji* (Chronical of an Emperor, 本紀), another port of a Standard History. These records document the nature of disaster, time, location, and severity; thus serve as important source for modern studies of the history of disasters in China.

In this paper, we focus on the natural disasters recorded in the *Benji*'s and *Wuxingzhi*'s in the Standard Histories.[1] We exclude the human-caused and unexplainable phenomena described in the *Wuxingzhi*.[2] After analyzing the formation of the *wuxingzhi*'s and other studies of natural disasters, we classified the natural disasters into 14 categories: flood, rain, frost, hail, famine, drought, cold, snow, wind, locust, borer, plague, earthquake, and landslide.[3]

Processing the Records and Markup

We have designed an XML format (Wuxing Markup Language, or WXML) to tag the texts.

A *record* is a writing of natural disaster indicated in the text. A record contains the following elements: *event*, *time period*, *area*, *severity*, and *frequency*. A record may describe several *events*. For instance, a record of drought often also mentions famine. In this case, both events are tagged. *Time period* (written using dynasty, era, year, month, day) has three subtags: starting time, ending time, and duration. If only a date is indicated, that date is considered the starting date. If there's no mentioning of duration or ending date, then the ending date is the same as the starting date. If duration is vague (such as "it rained for some 30 days"), then the ending date tag will not be filled. The element *area* contains two subjects: location and range. Since one or several administrative regions, a river or a mountain range may be indicated in a disaster, the location tag may have multiple values. The range tag could also be an administrative region or a geographical entity. When a record describes the area as "capitol and its surrounding prefectures", the location will be the capitol of the time, and the range will be the "surrounding prefectures". *Severity* includes the effect, the damages, and the reactions that followed. For example, a flood may include the effect of the breaking of the embankment which results in flooding of the farms and houses (damages), which leads to the reduction in taxes in the following year (reaction). *Frequency* is less complicated, although not entirely trivial. A record may mention several earthquakes, without indicating the exact number. In this case, it will simply be tagged as "several".

Producing and Counting the events

We first use the 14 keywords of disasters to extract descriptions mentioning the disasters. The paragraphs are then parsed automatically to identify the records and their time, event, area, etc. We remark that each description may contain several events, several locations, or even several time periods. We then tag the events, time periods, and locations automatically from the descriptions. The dates are standardized using the Buddhist Studies Time Authority Databases developed at Dharma Drum College (http://authority.dila.edu.tw/time/). Geographic coordinates are provided using the Chinese Civilization in Time and Space developed at the Academia Sinica (http://ccts.ascc.net/). An expert is then asked to go through the result to correct manually.

Several ways have been used in the literature to count the number of events. A record may involve multiple locations, different years, and multiple disasters. The same disaster may also appear in different books. A simple way that counts only the appearance of a type of disaster was used in (Deng**,** 1973) (regardless of the frequency, locations and severity, it is counted as 1 if it appeared in China during that year at least once. Otherwise it is 0 for that year). This method was adopted later by other researchers (Luo, 2005). At the other extreme, each tuple of time, disaster, location is recorded as one event (Yuan**,** 2008). A third option is to specify a tuple of time and location as an event without consider the other attributes (Wang, 2005; Zhang, 2007). By using tags, our approach provides the flexibility of being able to adjust to any of these counting methods, without being forced to pre-select one, by simply turning on or off an attribute.

Using single time and type as the event unit (while counting multiple locations as one), we tabulated a total of 9,717 events of natural disasters mentioned in Chinese Standard Histories, after removing duplicates from 6,653 events mentioned in *Wuxingzhi* and 3,848 in *Benji*. (We also removed 489 duplicate events between *Yuanshi* and *New Yuanshi*, and 79 duplicate events between *Old Tangshu* and *New Tangshu*.) The time distribution is as follows:

1 We have also included the *Book of Signs* (靈徵志) of *Weishu* (魏書), which also contains a fair amount of natural disasters.
2 The name *Wuxingzhi* indicates a view of the world in which the five elements, metal, wood, water, fire, and earth interact with each other. Thus certain phenomena were interpreted as signals of the missing of balance. However, the portion of this type of writing diminished significantly after the 10th century (You**,** 2007).
3 *Fire* is not considered a natural disaster. Although some fire might be due to natural reasons such as forest fire caused by lightning, researchers of natural disasters usually regard fire, as a general category, a manmade disaster since it is often hard to identify the cause (Zhang, 2012)).

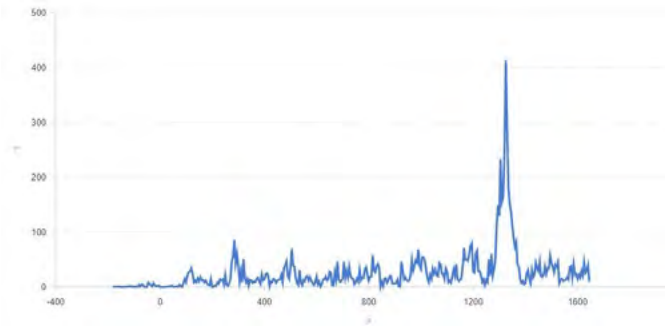


Figure 1 Distribution of natural disasters (X-axis: 5-year as a unit; Y-axis: frequency)

Note that the number of natural disasters recorded reached a peak during the Yuan Dynasty (1271-1368 BCE). (*Yuanshi*, 元史, only documented events occurred in China proper, not the Mongolian empire that ruled most part of the known world at the time.)

The system and some observations

We have built a system using the events of natural disasters mentioned above. Our interface allows one to specify one or several types of disasters, the era, and/or the areas and show the resulting data in number (or in graphs), on map, and also the texts of the events and their sources. The following is an example of disasters in the Guanzhong (關中) area.

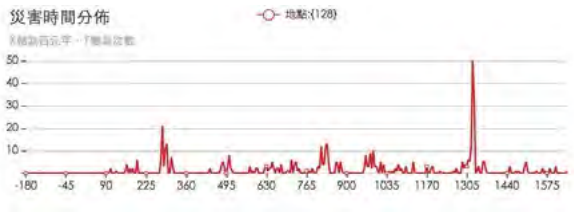


Figure 2 The number of disasters in Guanzhong area

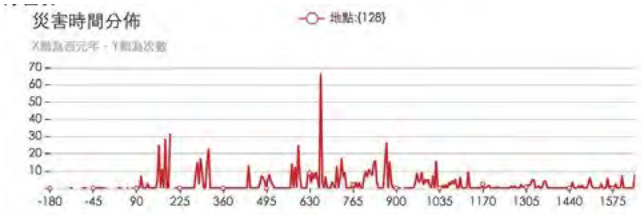


Figure 3 The percentage of disasters in Guanzhong area vs the country

The x-axis in both figures are years (in 5 years) in western calendar, while the y-axis of Figure 2 is the absolute number of disasters and the y-axis of Figure 3 is the *percentage* of *all* natural disasters recorded in the entire China during that time period. Note that although the number of disasters peaked around the year 1300, the percentage was dramatically high during the early Tang dynasty (618-907 BCE), when Changan (長安), a city in Guanzhong (關中), was the capital at the time. After the demise of Tang, the attention of latter empires gradually shifted to the northeast and south, and the percentage of disasters trailed off significantly, as Guanzhong gradually became irrelevant.

There are other interesting phenomena. For instance, there seemed to be more natural disasters during prosper periods. This may indicate that when the country was going through great turbulence such as foreign invasion or peasant revolt, the local officials simply did not bother to report natural disasters.

Concluding Remarks

In this paper we described a study on the natural disasters documented in the Chinese Standard Histories. We analyzed previous work on natural disasters and classified the events into 14 categories. We extracted texts of the records from *Wuxhingzhi* and *Benji*, and developed a markup language WXML to tag the events. We then build a system which is flexible in that one can use any of the measures mentioned above to show the results. Since the records are time-standardized and geo-referenced, our system also allows one to specify the type of disasters, time period, and locations and present the results either as charts or geographically. We are currently developing our system to allow full-text search to add flexibility.

We presented some preliminary observations. They seem to show that the natural disasters documented in the Standard Histories may not truthfully reflect the actual natural disasters that occurred. In other words, the records may reflect more on the circumstances under which the books were produced rather than the actual disasters that occurred. To more accurately capture natural disasters in Chinese history, one should at least also consult the local gazetteers (difangzhi, 地方志) (Chen, 2016). The WXML that we have designed is sufficiently flexible to incorporate those records as well.

References

You, Z. (2007). The Zheng (徵) and Ying (應) of Middle-age China. *Journal of Capital Normal University*, 2007.6: 10-16.

Zhang, G. (2012). *General Theory of Disasters*.

Deng, Y. (1973). *History of Disasters of China*. The Commercial Press.

Luo, C. (2005). Temporal-Spatial Distribution of East Han, MS Thesis, Zhengzhou Univ.

Yuan, Z. (2008). *Chinese Disaster History: Yuan Period*, Zhengzhoug University Press.

Wang, F. (2005). Disasters of two Jin, MS Thesis, Jiangxi Normal University.

Zhang, W. (2001). Preliminary Studies on the Natural Disasters of Han, PhD Thesis, Shaanxi Normal University.

Chen, S. (2016). Remapping Locust Temples of Historical China and the use of GIS, *Review of Religion and Chinese Society,* 149-163. Doi 10.1163/22143955-00302002.

REED London and the Promise of Critical Infrastructure

Diane Katherine Jakacki
diane.jakacki@bucknell.edu
Bucknell University, United States of America

Susan Irene Brown
sbrown@uguelph.ca
University of Guelph, Canada

James Cummings
james.cummings@newcastle.co.uk
Newcastle University, United Kingdom

Kimberly Martin
kmarti20@uguelph.ca
University of Guelph, Canada

Alan Liu has called upon digital humanists to think more critically about infrastructure - the "social cum technological milieu that at once enables the fulfillment of human experience and enforces constraints on that experience" (Liu, 2017). Liu's invitation comes at the moment when researchers involved in large-scale, long-term projects are shifting focus from remediation and the creation of digital incunabula to transmediation and the development of systems that support sustained discourse across ever-morphing digital networks, when we are recognizing the potential for "dynamism of the base or serialized form of the text—the state in which it is stored—as opposed to dynamic modes of presentation" (Brown, 2016: 288). REED London is one such project with a polyvalent dataset that spans over 500 years' worth of archival records, embracing from the start the need to establish a stable, responsive production and presentation environment primed for use by a wide range of scholarly audiences. Thus we find that we are immediately testing those infrastructural constraints. In this paper, members of the REED London project team will address the challenges we face as we develop and implement a framework that trains us to think about our collected data in relation to much larger networks of disparate resources and user needs.

REED London develops from a partnership between the Records of Early English Drama (REED) and the Canadian Writing Research Collaboratory (CWRC). Together we are establishing an openly accessible online scholarly and pedagogical resource of London-centric documentary, editorial, and bibliographic materials related to performance, theatre, and music spanning the period 1100-1642. With support from the Andrew W. Mellon Foundation and a CANARIE Research Software Program grant, a team of researchers in the digital humanities and performance history from the U.S., Canada, and the U.K. are building a stable, extensible editorial production and publication environment that will create new possibilities for scholarly presentation of archival materials gathered from legal, ecclesiastical, civic, political, and personal archival sources in and around London. The REED London project combines materials from three printed REED collections (*Inns of Court*, *Ecclesiastical London*, and *Civic London to 1558*), the prosopographical material from REED's *Patrons & Performances (P&P)*, the bibliographical materials of the *Early Modern LondonTheatres (EMLoT)* database, and in-progress and planned digital collections focusing on London area performance spaces, most notably the Globe, Rose, and Curtain theatres and Civic London 1559-1642.

REED is an internationally renowned scholarly project that has worked to locate, transcribe, and edit evidence of drama, secular music, and other communal entertainment in Britain from the Middle Ages until 1642. Since 1979 REED has published twenty-seven printed collections of transcribed records plus contextual materials. REED has long recognized the importance of online access to its resources,first with *P&P* and *EMLoT*, and more recently with the born-digital collection *Staffordshire*. REED has wrestled with the balance between what was once considered its "core" print publication activities and "adjunct" digital efforts, in the process migrating its data across a succession of programs and formats from Basic and dBASE to TEI P5 XML and MySQL (Hagen, MacLean, and Pasin, 2014). REED has developed its digital resources in ways that complicate integration (*P&P* exists in a Drupalinstance; *EMLoT* was built in a version of Django that is now out-of-date; *REED Staffordshire* was lightly tagged in TEI and relies on EATSML for entity management, an XML format used by the Entity Authority Tool Set (EATS) for serialisation of its data). The components of REED London must therefore first be made intra-operable before they can become interoperable (Jakacki, 2016). The partnership with CWRC supports broader adoption of standards for TEI text markup, RDF metadata specifications, and named entity aggregation, most immediately with the ingestion of *EMLoT* and the printed *Inns of Court* collection.

CWRC is an online infrastructure project designed to enable unprecedented avenues for studying the words that most move people in and about Canada. Built with funding from the Canada Foundation for Innovation, the CWRC platform supports best practices in the production of online collections, editions, born-digital essays, anthologies, collections, monographs, articles, or bibliographies, and supports the inclusion of visual, audio, and video sources (About CWRC/CSÉC). It supports collaboration through the use of interoperable data formats and

interlinking of materials, and for teams like REED London provides invaluable tools for communicating, tracking activity, and workflow. We envision that as the partnership develops and as REED London advances through production toward publication we will take full advantage of CWRC's functionality. From the start we have worked directly in CWRC's unique editor, CWRC-Writer, which allows us to edit REED London records, essays, and bibliographical material using more diplomatic and critical TEI P5 XML markup and at the same time creating semantic web annotations with RDF to identify, manage, and interlink entities contained within. The platform is also helping us to develop a better editorial workflow through management of access to data and editing by role, team communications, tracking and reporting of team activities.

To ensure REED London's stability and sustainability while extending its content and value to new generations of scholars the project is being built within the CWRC environment. The scope of REED London would not be possible without the sophisticated, integrated platform that CWRC provides. The focus of our first year is the design and construction of a collaborative online production and publication environment. Extending from CWRC's existing integrated content management and preservation system, the enhanced environment will accommodate the range of record texts, editorial and bibliographical content from the source materials, while a customized browser-based CWRC-Writer platform will support the team's goal of developing online editorial collaboration and review. The resulting streamlined production and publication environment will yield multi-faceted user-centered editions, meaning that agile component archival and editorial parts can cohere according to various criteria in response to scholars' research and teaching needs. In this way we are establishing a platform that produces new forms of "edition" that combine customized textual and contextual materials, exportable customized datasets and dynamic data visualizations. It also means that we will be able to realize the promise of extending the value of these materials to colleagues in fields beyond performance history, including political, religious, and cultural studies, and linguistics.

The partnership between CWRC and REED allows us to explore the potential for new research applications associated with prosopography, networks, and deep contextualization. REED London's wealth of references to very itinerant individuals across contemporaneous records means that we will be able to discern patterns through linking, analysis, and visualization. We will leverage REED's named entities for linking people, places, events, and organizations. Our team has healthy debates about the problematic present of linked data. Brown has stated that, "linking up with other data means connecting one ontology to another, and this brings with it a pressure toward generalization rather than specificity" (Brown, Simpson, et. al., 2015). Cummings has posited that "being able to seamlessly integrate highly complex and changing digital structures from a variety of heterogeneous sources through interoperable methods without either significant conditions or intermediary agents is a deluded fantasy" (Cummings 2014). Still, as a group we hope that by publishing our ontologies as a means of relating these entities as linked open data, we will be able to contribute to larger dialogues about class and society in Britain - certainly over the 500 years covered by REED London, but also about the development of Britain and Europe. CWRC content will be aggregated by the Advanced Research Consortium (ARC), and REED London will benefit from that aggregation, as we anticipate that people who figure in the REED London corpus, such as Elizabeth I, Francis Bacon, and Inigo Jones will be discoverable by scholars searching for these known figures across other linked resources. Perhaps more important, REED London records include extended references to thousands of Londoners who were in some way connected to performance, but who were not defined by that connection: civic officials, guild members, lawyers, clerks, priests, etc. The work of this project thus holds as yet unrealized value for a much broader understanding of British historical subjects.

Working within CWRC's platform and optimizing CWRC-Writer has allowed the core REED London team to move efficiently to an advanced planning phase. By the end of 2017 we will have designed templates for all record formats from *Inns of Court* and mapped database fields from *EMLoT* to align with the record parts from the print collections. We will have harvested a preliminary "white list" of named entities (people, places, organizations) from all three print collection indexes, P&P, and Staffordshire. Because of this efficient onramp we will be able to focus in the first half of 2018 on ingesting data, records, and contextual materials from Inns of Court and EMLoT. We will test the REED-specific entity list on ingested materials. We will also begin to user-test the editorial workflow system with the larger project team of REED editors and staff. By June 2018 we will have begun semantic tagging and experimentation with the CWRC HuViz semantic web visualization tool. At the DH 2018 conference we will report on further customization of the CWRC interface, our plans for data discovery and research collaboration, and present preliminary plans for user-responsive editions and data linkage.

References

Brown, S. (2016). Tensions and Tenets of Socialized Scholarship. *Digital Scholarship in the Humanities*, 31 (2): 283-300.

Brown, S., Simpson, J., CWRC Project Team, and Inke Project Team. (2015) An Entity By Any Other Name: Linked Open Data as a Basis for a Decentered, Dynamic Scholarly Publishing Ecology. *Scholarly and Research Communication* 6 (2). http://src-online.ca/index.php/src/article/view/212/409.

Canadian Writing Research Collaboratory project website. http://www.cwrc.ca/en/.
Cummings, J. (2014). The Compromises and Flexibility of TEI Customisation. In Mills, C., Pidd, M. and Ward, E. (eds), *Proceedings of the Digital Humanities Congress 2012*.
CWRC: About CWRC/CSÉC webpage. http://www.cwrc.ca/about/#whatis
CWRC Humanities Visualizer webpage. http://www.cwrc.ca/uncategorized/huviz-tool/
Early Modern London Theatres website. http://www.emlot.kcl.ac.uk
Entity Authority Tool Set (EATS) website. https://eats.readthedocs.io/en/latest/index.html
Hagen, T., MacLean, S., and Pasin, M. (2014). Moving Early Modern Theatre Online: the Records of Early English Drama introduces the Early Modern London Theatres. http://static.michelepasin.org/public_articles/2014-REED_McLean-Pasin.pdf
Jakacki, D. (2017) REED London: Humanistic Roots, Humanistic Futures. Paper given at MLA 2017. http://dx.doi.org/10.17613/M67794
Jakacki, D. (2016) REED and the Prospect of Networked Data. Paper given at the Conference of the Canadian Society for Renaissance Studies.http://dx.doi.org/10.17613/M6CK59
Liu, A. (2017) Toward Critical Infrastructure Studies", paper given at the University of Connecticut. https://www.youtube.com/watch?v=2ojrtVx7iCw
Records of Early English Drama project website. http://reed.utoronto.ca
REED Patrons and Performances website. https://reed.library.utoronto.ca
REED Staffordshire Collection website. https://ereed.library.utoronto.ca/collections/staff/

Large-Scale Accuracy Benchmark Results for Juola's Authorship Verification Protocols

Patrick Juola
juola@mathcs.duq.edu
Duquesne University, United States of America

Authorship attribution, the analysis of a document's contents to determine its author, is an important issue in the digital humanities. An accurate answer to this question is important, as not only do scholars rely on this type of analysis, but they are also used, for example, to help settle real disputes in the court system (Solan, 2012). It is thus important both to have analyses that are as accuracy, and to know what the expected accuracy levels are.

In keeping with good forensic practice, scholars such as Juola (2015) have proposed formal protocols for addressing authorship questions such as "were these two documents written by the same person?" Juola (2015) described a simple and understandable protocol based on a relatively small number of distractor authors, multiple independent analyses (e.g, separate analyses based on character n-grams, on word lengths, and on distributions of function words), and a data fusion step based on the assumption that the analyses were biased towards giving correct answers. Juola (2016) proposed minor revisions using Fisher's exact test to formalize the probability of a spurious match. The revised protocol has been formalized into a software-as-a-service product called Envelope to provide a standard (and low cost) authorship verification service.

We reimplemented Juola's (2016) protocol on a corpus of blog posts to determine whether, in fact, the protocol yields acceptable accuracy rates. Our reimplementation used the JGAAP open-source software package, an ad-hoc distractor set of ten authors (plus the author of interest), and the five analyses listed in Juola (2016): Vocabulary overlap, word lengths, character 4-grams, 50 MFW, and punctuation.

Blog data was taken from the Blog Authorship Corpus [Schler et al. (2006)] a collection of collected roughly 140 million words of blog text from 20,000 bloggers collected in August 2004. From this collection, we gathered 4000 examples of authors who had written 300 or more sentences. Ten of these authors were reserved, following Juola (2015;2016) as fixed distractor authors, while the others were randomly paired to create wrong-author test sets.

To test same-author accuracy, the first hundred sentences of each of the remaining 3990 blogs were used as "known documents" in the Envelope protocol, while the last hundred sentences of that author were used as "unknown documents." Perhaps obviously, the correct answer for these tests is that the documents should verify as the same author. To test different-author accuracy, the first hundred sentences of every author in the set was used as a "known document" and compared to the last hundred sentences of the other, paired, author. This procedure generated nearly four thousand test cases of both same and different authors. Each test case was analyzed five times and the rank sum of the known document within the eleven candidate authors calculated as an overall similarity measure from 5..55. This was converted to a *p*-value using Fisher's exact test.

Juola (2016) recommends a seven-point evaluative scale, as follows:

- $p < 0.05$ (Strong indications of same authorship)
- $p < 0.10$
- $p < 0.20$
- $p < 0.80$ (Inconclusive)
- $p < 0.90$
- $p < 0.95$
- $p >= 0.95$ (Strong indications of different authorship)

The results of these experiments are presented in table 1. The final column indicates the odds ratio; the likelihood that any particular finding at that level corresponds to an actual correct author.

p-value	Same Author	Different author	Odds
< 0.05	2948	748	3.941
< 0.10	246	359	0.686
< 0.20	195	396	0.492
< 0.80	409	1390	0.294
< 0.90	54	234	0.231
< 0.95	47	230	0.204
> 0.95	91	663	0.137

These results show that, in the same-author case, the proposed protocol is very good at identifying same-authors; roughly 3/4 of the actual same-author cases tested at the 0.05 level or better. Because of this, any result less stringent than "strong indications of same authorship" is actually evidence *against* same-authorship. The different-author case is more problematic; in theory, if there is no relationship between the known and questioned documents, the p-value should be uniformly distributed, representing a variety of chance relationships. However, the $0.20 < p < 0.80$ range ("inconclusive") contains 60% of the probability space, but only 1390/3990 = 35% of the different-author analyses. By contrast, the $0 < p < 0.05$ contains 19% of the analyses, while $0.95 < p < 1.00$ contains 17% of the different-author analyses. The observed distribution is thus highly weighted to the extremes of the probability space.

These results indicate that the underlying independence assumptions -- that (e.g.) similarity measured by analysis of word lengths is independent of similarity derived from the most common (function) words -- are not held generally. If a set of genuinely independent analyses could be found, the accuracy of this protocol would be greatly enhanced. Assuming the same distribution for the same author case, the odds ratio for the "strongly indications of same authorship" would be closer to 15:1 rather than 4:1.

Nevertheless, these results do show that, suitably interpreted, Juola's proposed protocol yields accurate results in a high proportion of test cases. We continue to work both on the development of a better analysis suite (with better independence properties) as well as continuing to replicate this experiment to obtain more accurate estimates.

References

Juola, Patrick. (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities. Volume* 30, Issue suppl_1, 1 December 2015, Pages i100–i113,https://doi.org/10.1093/llc/fqv040

Juola, Patrick. (2016). Did Aunt Prunella Really Write That Will? A Simple and Understandable Computational Assessment of Authorial Likelihood. *Workshop on Legal Text, Document, and Corpus Analytics - LTDCA 2016*, San Diego, California.

J. Schler, M. Koppel, S. Argamon and J. Pennebaker. (2006). Effects of Age and Gender on Blogging in- *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Solan, Lawrence M. "Intuition versus algorithm: The case of forensic authorship attribution."*JL & Pol'y* 21 (2012): 551.

Adapting a Spelling Normalization Tool Designed for English to 17th Century Dutch

Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, The Netherlands

Wijckmans Tessa
tessa_wijckmans@hotmail.com
Huygens ING/Nederlab, The Netherlands

Context

One of the bigger problems in comparing historic Dutch texts is wildly differing spelling of the same word. Seventeenth century Dutch did not have standardized spelling. Many spelling variants of the same word coexisted, making it very difficult to use any language processing tools on such texts because they depend on the same word being spelled the same way. So, for example basic algorithms like named entity recognition to recognize place or personal names, or even just part-of-speech tagging to find the grammatical context of words to analyze, for example, changing meanings of words of phrases work less well on older texts. Other languages, of course, have the same problem.

The Dutch digital research platform *Nederlab* aims to provide researchers with as many current and historic Dutch text and a toolset to do research on them. As such, spelling normalization would be an important addition to their tools. This project is a collaboration between the CREATE-project of the University of Amsterdam and *Nederlab* to tackle that problem. To deal with the problem, rather than developing a tool from scratch, we chose to adapt an existing tool to this situation: VARD2.

VARD2

VARD2[4][5] (an acronym of VARiant Detector) is a Java tool developed by Alistair Baron. It uses two lists (a normalized word list and a variant list) to suggest or replace variant words with their normalized counterparts. The normalization suggestions using a combination of four different methods: 1. known variant replacements; 2. character edit distance; 3. letter rules and 4. phonetic distance. Not all of these were useful for Dutch: the phonetic matching algorithm for example is based on English phonemes and hence did not work on these texts, but the re-spelling rules and the known word replacements worked very well.

VARD2 was designed to normalize Early Modern English, but is modifiable for other languages with a custom configuration. To create a configuration we used the modifiable parts of VARD2: the letter rules, the variant list and the normalized word list.

Corpus

We used the 1657 edition of the Dutch translation of the bible as a training set. Not only because there was a modernized version of it available that stuck rather closely to the original word order, but also because it would make it possible to later include another edition of the same book printed in 1637 to easily find more spelling variants for the words we had manually respelled or checked in the 1637 edition. We were able to make a golden standard of modernized spelling for the books Genesis and Exodus.

Choices

We chose to only do orthographic respelling, in order to preserve grammatical relevant elements of the texts as those may be relevant to research using natural language processing. One problem were words that did not follow Dutch re-spelling rules or did not have a clear Dutch respelling: foreign words, particularly place names and personal names, We chose to ignore such words as they would taint re-spelling rules for Dutch.

Problems & solutions

The first problem we encountered was the lack of any usable existing word list of all possible conjugations in modern Dutch. To get as many possible conjugations of every Dutch word that occurs in the *Woordenboek der Nederlandse Taal*[6] *(WNT)* a two-pronged approach was necessary. A set of algorithms, one per word class provided possible conjugations for each word in the *WNT*. First approach: for some word classes we were able to check the conjugations manually, but the large numbers of nomina and verbs made that impossible to do in this project. Second approach: for those the resulting word lists were checked automatically against the occurrences of those words in the *Corpus of Spoken Dutch*[1], *Dutch Wikipedia*[2] and *Verbix*[3].

Another problem, there was no set of respelling rules available that was effective for respelling Early Modern Dutch - the rule sets available did correct some spellings but caused mistakes in others. Extracting re-spelling rules from patterns in our golden standard provided an effective set of rules, especially when we generalized the rules where possible to catch similar instances.

Third, VARD2 could not handle word variations where two words should be re-spelled to a single word. Our solution was to pre-process texts with a script to remove spaces from such words.

The fourth problem was that some homonyms had overlapping spelling variations but needed to be re-spelled to different spellings in modern Dutch. An example is the word 'nog': spelling variations 'nog' and 'noch' were used interchangeably, but in modern spelling those two spellings denote differences in meaning. The only way to determine the correct modernization is to take the grammatical context of the word into account, which VARD2 does not do. This necessitated a second pre-processing step: we were only able to run a few tests, but part of speech tagging the original text and (manually) selecting a few patterns that marked one meaning or the other seemed to provide enough information to deduce the correct re-spelling.

Results

All in all, with a few additions and modifications a tool like VARD2 can be successfully converted to work on a Early Modern Dutch. Tests on other types of texts (a treatise on mathematics from 1605, the description of a beached whale from 1599, a description of the New World from 1770, a poetry book from 1637 etc) show promising results, indicating that a little extra training can make this configuration work well for different genres. Automatic respelling of the entire 1657 bible at a 95% confidence level resulted in automatic re-spelling of 62% of 340,000 variants. For the earlier edition (1637), automatically correcting at 95% confidence corrects 60% of just short of 350.000 unknown words, at 75% confidence 84% of the variants were corrected. The paper will show the results of automatically re-spelling 17th century texts using a VARD2 trained on just the first two chapters of the bible.

4 http://ucrel.lancs.ac.uk/vard

5 Baron, A. and Rayson, P. (2008). VARD 2: *A tool for dealing with spelling variation in historical corpora. Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008.

6 http://wnt.inl.nl

Differential Reading by Image-based Change Detection and Prospect for Human-Machine Collaboration for Differential Transcription

Asanobu Kitamoto
kitamoto@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems; National Institute of Informatics, Japan

Hiroshi Horii
a-horii@amane-project.jp
AMANE LLC, Japan

Misato Horii
ymisachi@amane-project.jp
AMANE LLC, Japan

Chikahiko Suzuki
ch_suzuki@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan

Kazuaki Yamamoto
yamamoto.kazuaki@nijl.ac.jp
National Institute of Japanese Literature, Japan

Kumiko Fujizane
zanezane@post.ndsu.ac.jp
Notre Dame Seishin University, Japan

Introduction

Japanese books in the Edo period (1603-1868) were mainly published by woodblock print. Their caligraphic writing style using differnt characters prevents native Japanese people to read and understand the content, and the knowledge of the past has been buried in libraries. To change this situation, NIJL-NW project started a ten-year mass digitization program to create the open dataset of 300,000 old Japanese books [7]. To take advantage of emerging big data of Japanese culture, we are working on the development of "deep access" technology to make the content of books accessible by structuring the content by either manually or automatically.

This paper focuses on a series of old Japanse books called "Bukan" [6]. Bukan offers the directory of families of the state king (Daimyo) and bureaucrats of the central government (Bakufu) in the Edo period. Bukan has a unique history. It had been a best seller book for as long as 100 to 200 years, had been updated and published frequently with a peak frequency of a few times in a month, and had been the battle field of two commercial publishers competing each other to improve the quality of their own Bukan editions. Because of good coverage and quality of Bukan, the comprehensive analysis of Bukan is expected to improve our understanding on the political, administrative, and cultural structure in the Edo period.

Comprehensive analysis cannot be achieved, however, without a solution to the problem of multiple versions. Bukan had been published for a long period with high frequency, and it is not known how many versions had been published, or how to decide the proper ordering of existing versions. Moreover, the complete transcription of Bukan is not realistic due to a large amount of text across multipe versions. In short, two major problems, management of versions and reduction of transcription, need to be solved for comprehensive analysis of Bukan.

Method

We first propose the concept of "differential reading," which refers to the mode of reading books, such as close reading and distant reading. It is a reading focusing only on changes between different versions with support from digital tools. Algorithms to detect changes in different versions are two-fold; namely text-based and image-based approaches.

Text-based change detection is effective for manuscripts. Many tools, such as CollateX [2] and ViTA [9], have been developed for text comparison, or Versioning Machine [8], for structured text or TEI (Text Encoding Initiative). In the case of woodblock print, however, image-based change detection has a number of advantages. In the terminology of old Japanese bibliography, versions can be further classified into "publication" and "correction," where the former refers to the complete re-creation of the woodblock, while the latter refers to the application of small patches to the woodblock. Change detection on publication is an easy problem for image processing, and change detection on correction is also feasible by image matching because only a small part is corrected and other parts remain the same. Other advantages of image-based change detection include transcription-less change detection and non-textual change detection.

By taking advantage of image-based change detection, we formulate differential reading as a two-step process; namely machines work first to detect changes, and humans work next to read changes.

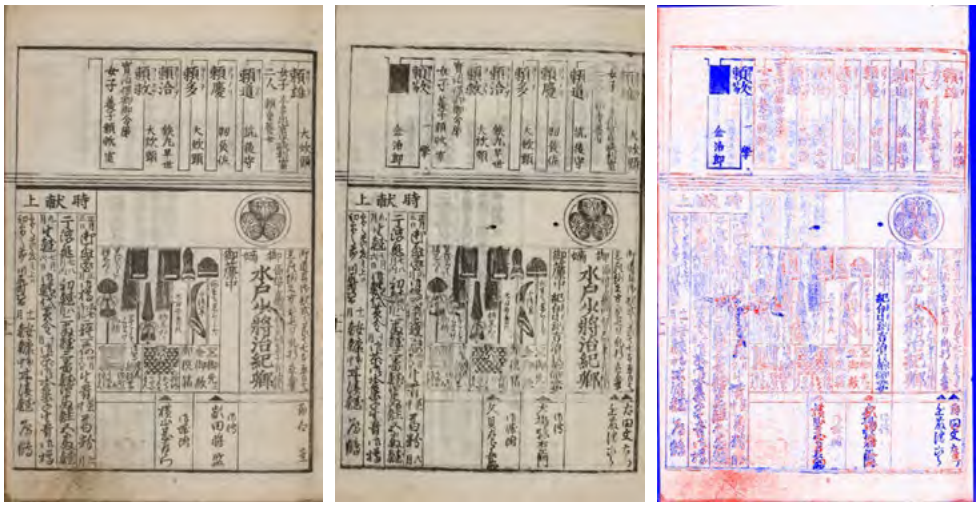


Figure 1: Comparison of two different versions of Bukan. Left: Kansei Bukan (1789); middle: Kansei Bukan (1791); right; the result of change detection, where red color represents regions present only on the 1789 version, and blue, the 1791 version.

Results

An image-based change detection algorithm was implemented on image processing library OpenCV 2.4 with a combination of algorithms such as FAST for feature detection, BRIEF for feature description, and Hamming distance for feature matching. In addition, RANSAC was used for estimating homography matrix for matching two images. Changes are then emphasized using a coloring scheme by assigning red and blue for large difference in pixel values and white for small difference in pixel values.

We compared two different versions of Bukan, Kansei Bukan (1789) [3] and Kansei Bukan (1791) [4] to check if the image-based change detection algorithm can identify changes between versions two years apart. Figure 1 shows the result of image-based change detection. It is clear that a part of the page, such as the genealogy of the family, has been changed from the 1789 version to the 1791 version. In the workflow of differential transcription, machine gernerated change infromation will be transferred to planned differential reading interface so that humans can focus only on a part of the image.

Differential transcription needs base transcription, on which transcription of subsequent versions depend. Initially the database of "Bukan Complete Collection" [1] uses Kansei Bukan (1789) as the base transcription. The database not only contains basic information about Daimyo, but also offers visualization about "Sankin Kotai," which is a required travel for Daimyo between their states and Edo city to meet Shogun (the national leader) in every two years or more often. Animated visualization in Figure 2 shows spatio-temporal and seasonal patterns of their trips coordinated by Bakufu. The database also offers the graphic design collection of Daimyo, such as family emblem, costumes, and tools they used for official activities.

We found one important missing element in creating the database; namely the standard ID system agreed within the community. Bukan is a collection of entities, such as people and political organization that changes over time. To uniquely identify entities appearing in different sources and to create a time-series database of linked entities, we need the standard ID system in the Edo period through collaboration with historians. With a proper ID system, this system may evolve into the information infrastructure of people and political entities for the historical studies of the Edo period.

Figure 2: Bukan Complete Collection website. Left: the list of Daimyo family emblems; right: animated visualization of spatio-temporal patterns of Daimyo trips. Only Japanese website is available at this moment

Discussion and Conclusion

The advantage of differential reading is two-fold. First, when reading two similar versions, differential reading has advantage over close reading by reducing the burden of human attention. A traditional approach of side-by-side comparison is error-prone, and machines can be optimized for pixel-level comparison without loss of attention by fatigue. For this type of task, human-machine collaboration should evolve into a combination that machines are in charge of low-level change detection while humans are in charge of high-level interpretation. Second, differential reading can be used as a component for differential transcription. The base transcription is required in any case, but the amount of transcription for subsequent versions is significantly reduced. A version management system may play an important role to optimize the transcription workflow, which is left for future work.

A proposed approach of differential transcription by human-machine collaboration is not only effective for Bukan, but also applicable to other woodblock print books with different versions. Our tools have been developed on IIIF (International Image Interoperability Framework), which allows us to apply our tools not only on NIJL-NW datasets but other datasets in the same manner. In the future, we plan to make a user interface on top of our IIIF Curation Viewer [5] and combine it with a workflow management tool to support efficient work of transcriptors.

References

Bukan Complete Collection, http://codh.rois.ac.jp/bukan/, (accessed April 27 2018).

CollateX, https://collatex.net/, (accessed April 27 2018).

Kansei Bukan (1789), https://doi.org/10.20730/200018823

Kansei Bukan (1791), https://doi.org/10.20730/200018825

IIIF Curation Viewer, http://codh.rois.ac.jp/software/iiif-curation-viewer/, (accessed April 27, 2018).

KITAMOTO, A., et.al. (2017) Structuring Time-Series Historical Sources by Human-Machine Specialization: Toward the Construction of Edo Information Platform Referring to "Bukan", *IPSJ SIG Computers and the Humanities Symposium 2017*, pp. 273-280 (in Japanese).

NIJL-NW project, http://www.nijl.ac.jp/pages/cijproject/index_e.html, (accessed April 27 2018).

Versioning Machine, http://v-machine.org/, (accessed April 27 2018).

ViTA (Visualization for Text Alignment), http://ovii.oerc.ox.ac.uk/vita/, (accessed April 27 2018).

The History and Context of the Digital Humanities in Russia

Inna Kizhner
inna.kizhner@gmail.com
Siberian Federal University, Russian Federation

Melissa Terras
m.terras@ed.ac.uk
University of Edinburgh, United Kingdom

Lev Manovich
manovich.lev@gmail.com
City University of New York, United States of America

Boris Orekhov
nevmenandr@gmail.com
National Research University Higher School of Economics, Russian Federation

Anastasia Bonch-Osmolovskaya
abonch@gmail.com
National Research University Higher School of Economics, Russian Federation

Maxim Rumyantsev
m-rumyantsev@yandex.ru
Siberian Federal University, Russian Federation

The history and context of the development of Digital Humanities in Russia as outlined in this paper shows that there are various influences at play which have led to the forming of the Russian DH field. We link the quantitative methods used to previous trends in scholarship, including mathematics, Russian editorial practices, and the development of museum computing in the country. By doing so we can consider the individual societal contexts which encourage a field to emerge, and although that field may look similar to outsiders, identify the lineage of intellectual approaches which still influence methods and cultures within the discipline.

The connection between Russian Formalism and the Digital Humanities (Allison et al., 2011; Moretti, 2013; Jockers, 2013; Stanford University, 2015) relates to the tradition that originated following the strengthening of Russian mathematics at the turn of the nineteenth century after the Moscow Mathematical Society was established in 1864. The influence of this school on literary studies can be traced through the twentieth century from Andrey Bely's experiments at the threshold of mathematics and poetry (Akimova, Shapir, 2006; Giansiracusa and Vasilieva, 2017) to the Moscow Linguistic Circle with Roman Jacobson as its chair (Akimova, Shapir, 2006; Pil'shchikov, 2015), to the Prague Linguistic Circle and further to the Tartu-Moscow School (Uspensky, 1998). Boris Jarkho's 'Research Methods for Literary Studies' written in 1936 anticipated the approach of Stanford Literary Lab not

only in its 'quantitative interpreting' (Underwood, 2017) but also in a skill of a scholar able to see wider contexts and make bridges across disciplines. The traditions are currently developed at the Centre for Digital Humanities at the Higher School of Economics in Moscow via digital tools (Skorinkin, 2017; Bonch-Osmolovskaya and Skorinkin, 2016; Orekhov and Tolstoy, 2017; Kuzmenko and Orekhov, 2016; Fischer et al, 2017).

Another tradition related to building the National Corpus of the Russian Language[1] can be traced back to Alexei Lyapunov (Sitchinawa, 2006), another famous Russian mathematician. The point here is not that mathematics sustained and influenced all the Russian humanities (Bakhtin's famous studies can provide an opposite example[2]) or that quantitative approach as a trendy international methodology was also present in this part of the world in the 1960s-1970s but that it provided the rigor and method to the field which was disconnected from the international research methods and standards. This disconnection resulted in a dramatic difference in academic cultures.

A recent paper (Underwood 2017) discusses distant reading as a part of the digital humanities project aimed at coping with confirmation bias. Underwood shows that (social) sciences provide the 'experimental structure' and help us build research design around hypothesis, samples and results. A specific Russian feature was that research methodology of this type was provided via mathematics, linguistics, and sciences. Social science and anthropology played a minor role in the interplay of influences (Gasparov, 2016).

A major part of the current Russian digital humanities project is connected to linguistics. However, linguistics did not only provide a set of formal features and a methodology to trace a formal technique in a literary work. It was an important initial influence, a novel method to do literary studies as a part of a new scientific perspective (Tynjanov, 1971; Jarkho, 2006) in the early twentieth century. The Moscow Linguistic Circle active from 1915 to 1924 held its meetings in Roman Jacobson's flat in Moscow and its members were over 60 linguists and scholars working in text analysis and literary studies[3]. Apart from its significant international influence, the society had an important impact on how Russian scholarship developed (Akimova, Shapir, 2006; Shapir, 1996; Pil'shchikov, 2015). Its traditions were continued in applying quantitative methods to studying poetry in the second half of the twentieth century (Akimova, Shapir, 2006; Bodrova, 2017). Its influence can be traced in a highly influential approach of applying structural linguistics to interdisciplinary cultural studies at Tartu University[4] also in the second half of the twentieth century (Gasparov, 2016).

A part of current projects in Russian digital humanities are connected to this tradition. The project of creating a semantic edition of Leo Tolstoy's complete works[5] (Bonch-Osmolovskaya, 2016) includes representative and interpretive components. The edition's interpretive part works with a humanistic data model of the characters' roles in *War and Peace* validated through the digital tools of natural language processing and extracting semantic roles (Bonch-Osmolovskaya and Skorinkin, 2016), this approach also includes a classification of characters using character networks (Skorinkin, 2017). The connection of digital approaches to the previous trends of scholarship (Russian Formalism and structural interdisciplinary studies initiated by scholars from Tartu and Moscow) is explicitly proposed and maintained through the Moscow-Tartu Summer School annually organized at the Higher School of Economics in Moscow.

Quantitative approaches to studying poetry has been a path traditionally pursued by Russian mathematicians or people related to mathematics. Andrey Bely who was closely related to Nikolai Bugaev[6], one of the first chairs of the Moscow Mathematical Society, developed a quantitative approach to studying poetic rhythm in 1910 and initiated a society where scholars were taught to use statistics to study poetry (Semyonov, 2009). Andrei Kolmogorov, a famous Russian mathematician, organized a seminar and published several papers in this field in the early 1960s (Semeyonov, 2009; Kolmogorov, 2015).

The tradition has been continued via digital tools where the authors show the limitations of digital analysis (Orekhov, 2014) or integrate mapping poetry in interdisciplinary cultural studies following the Tartu tradition (Kuzmenko and Orekhov, 2016).

Russian editorial practices in the second half of the twentieth century were focused on publishing complete works of the authors from the canon of the time. Thorough editorial work was limited by the editors' attempts to combine international standards of scholarly apparatus and the requirements of the moment. Twentieth century's attempts to create scholarly editions using interpretive practices of the time (Bonch-Osmolovskaya, 2016) resulted in a current need to build new epistemological foundations for contemporary scholarly editions. Digital methods and digital scholarly standards are probably the

1 The National Corpus of the Russian Language (http://www.ruscorpora.ru) includes over 600 million words. It was published online in 2004 and developed by the linguists from the Russian Academy of Sciences (Sitchinawa, 2006).

2 See, for example (Gasparov, 2002; Sedakova, 1992), for the discussion of the difference between Bakhtin and the Russian Formalism.

3 Tynaianov and Schklovsky, famous for their contribution to Russian Formalism, were members of the Moscow Linguistic Circle (Shapir, 1996).

4 Tartu University in Estonia, a part of Russia at that time, was home for the literary studies done in the tradition of the methodology looking at formal structural features.

5 A project that is currently developed at the Higher School of Economics and Leo Tolstoy museum. Apart from using a representational mark-up in TEI standards, the project includes experiments towards an interpretive component (Bonch-Osmolovskaya, 2016; Bonch-Osmolovskaya and Skorinkin, 2016).

6 Boris Bugaev's (Andrey Bely's) relations with his father and the influence of the academic environment on Bely's development have been widely discussed in literature (see, for example, Janecek, 2015 and Giansiracusa and Vasilieva, 2017).

best possible option to cope with epistemological difficulties in the field.

While editing textual materials was complicated by interpretive practices, visual editions in the 1970s, 1980s and early 1990s were introducing new standards of metadata and data models. Their editors made an important step towards digital practices and museum computing.

The editorial practices of printed visual editions of artworks related to the standards of publishing museum images (Kizhner et al, forthcoming), the quality of images and the scholarly apparatus accompanying visual editions in the 1970s and 1980s prepared the anticipations of standards for digital publishing and placing images in a wider context via digital tools (Polulyakh, 2009; Sher, 2006).

A specific Russian feature was looking for formal (structural) components to interpret a literary work, bringing a wide interdisciplinary context to interpretation. The tradition was sustained during the twentieth century before Russian scholars turned to digital humanities. The influence of social science, gender and race studies, enlarging or changing a canon did not leave significant traces even if (when) the ideas reached the community of scholars. A current exception are projects aimed at studying the nineteenth century literary canon and future developments seeking to compare it with contemporary canons (Vdovin and Leibov, 2013). The authors propose to build a canonical corpus and study the changes using a mark-up. The idea relates to Moretty's evolutionary theories (ibid) and the Russian traditions of observing the dynamics of a formal feature that can be traced back to Boris Jarcho's papers written in the 1930s.

The paper will demonstrate, using evidence from various sources that Russian traditions of quantitative interpreting, the influence of strong mathematics and a trend of placing cultural objects within a broader context were crucial for our understanding of how digital humanities, as a quantitative methodology, developed in the country, in a different way than it did elsewhere. Understanding these alternative histories will help us understand the range of activities taking place in Digital Humanities worldwide, by looking at the social, scholarly, and cultural contexts, helping the community to navigate and bridge differences.

References

Akimova, M. and Shapir, M. (2006) 'Boris Isaakovich Jarkho and the Strategy of Research Methods for Literary Studies', in Jarkho Boris, *Research Methods for Literary Studies (*ed. Maxim Shapir*),* Moscow: Phologica. In Russian.

Allison, S., Heuser R., Jockers, M., Moretti, F., Witmore, M. (2011) *Quantitative Formalism as Experiment,* Pamphlet 1. Stanford Literary Lab.

Bonch-Osmolovskaya, A., (2016) 'Digital Edition of Leo Tolstoy Works: Contributing to Advances in Russian Literary Scholarship'. *Journal of Siberian Federal University. Humanities and Social Sciences* 7 (9), pp. 1605-1614.

Bonch-Osmolovskaya, A. and Skorinkin D. (2017) 'Text Mining War and Piece: Automatic Extraction of Character Traits from Literary Pieces', *Digital Scholarship in the Humanities,* Vol. 32, Supplement 1.

Fischer, F., Orlova, T., Skorinkin, D., Palchikov, G., and Tyshkevich, N. 'Introducing RusDraCor - A TEI-Encoded Russian Drama Corpus for the Digital Literary Studies, in *International Conference Corpus Linguistics 2017: Book of Abstracts,* June 27-30, 2017, Saint Petersburg, pp. 28-32.

Gasparov, M., (2002) 'Michael Bakhtin in the Russian Culture of the 20th Century', in *Michael Bakhtin: Pro and Contra: Mikael Bakhtin's Heritage in the Context of the World Culture* (ed. Konstantin Isupov), Vol. 2, Saint Petersburg. In Russian.

Gasparov, B., (2016) 'Between Methodological Strictness and Moral Appeal: Questions of Language and Cultural Theory in Russia', History of Humanities, vol.1, number 2.

Giansiracusa, Noah and Anastasia Vasilyeva, 'Mathematical Symbolism in a Russian Literary Masterpiece', arXiv: 170902483v1 https://arxiv.org/pdf/1709.02483.pdf

Janecek, G., (2015) *Andrey Bely: A Critical Review,* Lexington: The University Press of Kentucky.

Jarkho B., (2006) *Research Methods for Literary Studies (*ed. Maxim Shapir*),* Moscow: Phologica. In Russian.

Jockers, M., 'Microanalysis: Digital Methods and Literary History. University of Illinois Press, Urbana, IL.

Kizhner, Inna, Melissa Terras, Maxim Rumyantsev and Kristina Sycheva, 'Accessing Russian Culture Online: The scope of digitization in museums across Russia', forthcoming.

Kolmogorov, A., (2015) *Studies in Poetry,* Moscow Centre for Mathematical Education Press, 2015. In Russian.

Kuzmenko, Elisaveta and Boris Orekhov, (2016) 'Geography of Russian Poetry: Countries and Cities Inside the Poetic World', in *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 830-832.

Moretti, F., (2013) *Distant Reading,* London: Verso.

Orekhov, Boris and Fekla Tolstoy, (2017) 'Textograf: A Web Application for Manuscript Digitization', in *Digital Humanities 2017: Conference Abstracts*. McGill University & Université de Montréal, Montréal.

Pil'shchikov, I., (2015) 'The Legacy of the Moscow Linguistic Circle and the Digital Humanities Today', in *Russian Formalism and the Digital Humanities: Abstracts*, Stanford University. https://digitalhumanities.stanford.edu/russian-formalism-digital-humanities-abstracts

Polulyakh, A., (2009), Photo capturing and digital technologies in museums: following traditions, In *Proceedings of 'ICT fro Regional Development' Conference*, 5-6 February 2009, Smolensk Regional Administration, Smolensk, pp. 217-222. In Russian

Sedakova, O., (1992) 'Michael Bakhtin: An Alternative Interpretation', http://www.olgasedakova.com/Moralia/267 In Russian.

Semyonov, V., (2009) 'Methods of Statistics for Studying Russian Poetry: Andrey Bely and Andrey Kolmogorov', *Journal of Moscow State University*, series 9, number 6.
Sher, J., (2006). 'Department of Museum Informatics at the Hermitage Museum (1975 - 1985), *Information Technology for Museums,* No 2, Saint Petersburg. http://kronk.spb.ru/library/sher-yaa-2006.htm In Russian.
Sitchinawa, D., (2005) 'The National Corpus of the Russian Language: a Brief Prehistory', in *The National Corpus of the Russian Language: 2003-2005,* Moscow: Indrik, pp. 21-30. In Russian.
Skorinkin, D., (2017) 'Extracting Character Networks to Explore Literary Plot Dynamics', in Proceedings of *Dialogue: Conference on Linguistic Computing*, Moscow, 31 May - 3 June, Issue 16 (23), Vol.1, pp. 257-270.
Shapir, M. (1996) 'An editorial note to Jacobson, Roman 'The Moscow Linguistic Circle', Philologica 3. In Russian.
Stanford University, (2015) *Russian Formalism and the Digital Humanities Conference: Book of Abstracts*, Stanford University. https://digitalhumanities.stanford.edu/russian-formalism-digital-humanities-abstracts
Tynjanov, J., (1971) 'On literary revolution', in *Readings in Russian Poetics (ed.* Ladislav Matejka and Kristina Pomorska), MIT Press.
Underwood, T., (2017) 'A Genealogy of Distant Reading', Digital Humanities Quarterly, Vol.11, No 2, http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html
Uspensky, B., (1998) 'On the origin of the Tartu-Moscow School', in 'The Tartu-Moscow School: History, Memoirs, Reviews' (ed. Sergey Neklyudov), Moscow: Languages of the Russian Culture, pp. 34-44. In Russian.
Vdovin, Alexey and Leibov, Roman (2013) 'Canonical texts: Russian poetry and teaching literature in the 19th century', in *Canonical Texts: Russian Pedagogical Practice in the 19th Century and Poetical Canon,* Tartu University Press. In Russian.

Urban Art in a Digital Context: A Computer-Based Evaluation of Street Art and Graffiti Writing

Sabine Lang
sabine.lang@iwr.uni-heidelberg.de
Heidelberg Collaboratory for Image Processing, Germany

Björn Ommer
ommer@uni-heidelberg.de
Heidelberg Collaboratory for Image Processing, Germany

Summary

The paper presents how digital photographs of street art and graffiti writing[1] are analyzed with computational methods by the Computer Vision group of Heidelberg University, where an interdisciplinary collaboration between art history and computer vision is embedded since 2009. The project on urban art started in November 2017 and has the following aims: It studies the effect of digital possibilities on street art and graffiti writing regarding access, dissemination and mobility. Per definition urban art is strongly attached to a street environment, which is canvas and frame at the same time. This resulting immobility of urban art is in contrast with traditional art, where the materiality simplifies a display at alternating locations. Eventually, the paper highlights why urban art can only endure within a digital context. An example of Bristol-born artist Banksy (*1974) illustrates this: In 2015, he put up a stencil on a wall in Calais, depicting *The Raft of the Medusa* (Fig.1); only two years later, workers painted over the wall and covered Banksy's work (Samuel, 2017). The project also establishes a data collection of urban art, consisting of reproductions from *Google Arts and Culture*, other image archives and a private collection by art historian and street art scholar Ulrich Blanché. Lastly, it demonstrates how computer-based tools are used to study images with regards to form and content. In this way, patterns over time and space or artistic networks are revealed and relations between artwork and urban environment can be evaluated. Therefore, the project team utilizes an interface, which was developed within the group and allows for a visual search based on multiple image regions in large image sets.

Evaluating street art and graffiti writing

In 2009, a collaboration between art history and computer vision was established within the Computer Vision group. Thus building a bridge between the two disciplines, which resulted in the realization of works, including the creation of an interface (Bell et al., 2014), reconstruction of drawing processes (Monroy et al., 2011) or the detection and analysis of gestures in medieval manuscripts (Yarlagadda et al., 2013), (Schlecht et al., 2011), (Yarlagadda et al., 2010). The group uses deep learning algorithms and unsupervised approaches to study visual similarities on image level (Bautista et al., 2017), (Bautista et al., 2016) and whole sequences (Milbich et al., 2017). The current project utilizes existing methods to study urban art. The presence of digital image collections of urban art and computational approaches enable both large-scale evaluations and detailed studies, which has not been done by scholars so far. Previous work mainly concentrated on terminology (Blanché, 2015), social aspects (Ross, 2016) or individual artists (Blanché, 2012), (Blanché, 2010), highlighted its mediality (Glaser, 2017) and generally justified its study in art history.

The presentation highlights the influence of digitization on urban art, describes the building of a suitable dataset and its evaluation through computational methods. (1) Digital possibilities have influenced all of humanities; for urban art, however, the effect is even more profound. Most traditional artworks are mobile; artists paint on canvas or paper, which allows for easy transportation and pu-

blic display at various places. In this way, styles, content, or individual motifs spread and art reveals itself to be less bound to a specific place. On the contrary, urban art is per definition tied to the street; its meaning only fully unfolds on site. The street not only provides a canvas, but also imposes form and additional meaning. As a result, urban art is greatly ephemeral: Works are being over-painted by authorities and artists (Samuel, 2017) – as the example of Banksy showed – or buildings are torn down. In reaction, works are increasingly documented and made available online. Since its start in November 2017, the project has studied the presence of urban art on the Internet. Its visibility on different websites has impacted the community and visuality of urban art: Communication between artists and fans has increased and is simpler, motifs are disseminated faster and wider, breaking national borders and indicating a tight network. It is only through digital possibilities that urban art can be preserved and disseminated – this distinguishes it from traditional art.

(2) In order to study form and content of images with computer-based tools, the project gathers a dataset of urban art, providing metadata if available. Images are taken from *Google Arts and Culture* or *Facebook's Global Street Art*. However, the project team also received a comprehensive set of photographs, capturing urban art in various cities worldwide between 2007 and 2017. All images were taken by art historian Ulrich Blanché; this unique data enables to address new questions regarding the capturing process: How did the photographic perspective and thematic focus change over time? Does it vary for different locations? Is there a correlation between alternating perspectives and Blanché's social role? Eventually, he captured urban art first as a simple admirer, then as a student and finally as a scholar – although the first role persisted throughout time. The final image collection, including metadata, will be published and can be used by other scholars. A large number of images contain large context regions and objects, including buildings or cars. To improve performance and detection, the data was pre-processed: Around 200 images from the Blanché-dataset were annotated with bounding boxes marking artwork or context.

(3) The project studies the visuality of urban art on the basis of this image collection using computer-based methods. It aims to find recurrences and variances of a motif, ultimately not only pointing to the same but different artists. On a smaller scale, the example of Cologne street artist 'kurznachzehn' illustrates this: She uses old family photographs to create paste-ups, which she attaches to walls in various German cities. Her most recognized motif is a young girl – the artist's mother as a child. The girl appears throughout her oeuvre in a similar pose but in varying scenarios: while picking up a dandelion (Fig.2), painting or feeding a little bird (Glaser, 2017), ('kurznachzehn', 2017). In order to study image collections, the project team utilizes unsupervised methods, which have been successfully applied on other tasks and do not rely on labeled data. This is valuable, since digital reproductions of urban art rarely have information regarding artist, title or creation date – this is mainly due to the anonymity of artists and legal reasons. Reproductions are evaluated on an interface, which not only allows to search for individual but also multiple regions and thus to consider geometrical relations between artworks and urban environment. The example of the dandelion-picking girl (Fig.2) by 'kurznachzehn' illustrates this: True to the nature of her gesture, she always appears close to the ground. Underlying algorithms use a SVM-classifier trained with one positive against many negative examples. While other retrieval systems require manual tagging, the algorithm purely operates on visual qualities. Currently state-of-the-art methods are being implemented, using CNN instead of HOG-features to train the classifier. Eventually, the interface not only detects identical motifs but also variations. First tests showed promising results; the project team studied images of artworks by Brazilian street artists OsGemeos. The user was interested in a figure seen from behind and a text region to its right; (Fig.3) shows the search results for the given queries after the second training round; the bottom row includes all correctly retrieved images as selected by the user. Results can now be analyzed regarding formal and semantic similarities or variances; also, it allows to evaluate the position of the motif in relation to the urban context. Future work should study the motif of the figure seen from behind also in the context of its general appearance in art history.

Applying computational methods to urban art data has emphasized chances and benefits, not only for art history but also for computer vision. Existing algorithms have been tested on challenging data and proofed their efficiency. However, working with urban art data has also highlighted some challenges: Collections are biased towards certain time periods, nationalities and dominated by works of popular artists. The new dataset, established within the project, is therefore extremely valuable. First tests, although overall successful, showed that algorithms are challenged by a dominating background, imaging mode (perspective) and the size of artworks. To remedy the latter, the project team decided to annotate part of the Blanché dataset with bounding boxes, which improved detection.

Conclusion

The presentation consists of two parts: A theoretical basis will be established in the first, discussing the influence of digital possibilities on aspects, such as mobility, access or dissemination, while the dataset will be introduced in the second half, which also includes presentations of search results on the interface. The project team aims to further establish urban art as a profound research topic in academia, point to new research questions and possible challenges when working with urban art data. Most importantly, the presentation emphasizes the chances offered by computer-based methods to study urban art in detail and on large-scale. (Words: 1485)

List of Illustrations

Fig.1: Banksy, *The Raft of the Medusa*, Calais, 2015

Fig.2: 'kurznachzehn', *Girl picking Dandelion*, Dusseldorf, 2013

Fig.3: Search results for figure and text to its right on interface; image collection of Brazilian street artists OsGemeos

References

Bautista, M., Sanakoyeu, A. and Ommer, B. (2017): Deep Unsupervised Similarity Learning Using Partially Ordered Sets. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. CVPR.

Bautista, M., Sanakoyeu, A., Sutter, E. and Ommer, B. (2016): CliqueCNN: Deep Unsupervised Exemplar Learning. *Proceedings of the Conference on Advances in Neural Information Processing Systems*. NIPS.

Bell, P., Ommer, B. and Takami, M. (2014): An Approach to Large Scale Interactive Retrieval of Cultural Heritage. *Eurographics Workshop on Graphics and Cultural Heritage*. The Eurographics Association.

Blanché, U. (2015): Street Art and related terms – discussion and attempt of a definition. *Street & Urban Creativity Scientific Journal. Methodologies for Research,* (1), pp. 32-40.

Blanché, U. (2012): *Konsumkunst. Kultur und Kommerz bei Banksy und Damien Hirst.* Bielefeld: Transcript Verlag.

Blanché, U. (2010): *Something to s(pr)ay: Der Street Artivist Banksy. Eine kunstwissenschaftliche Untersuchung.* Marburg: Tectum Verlag.

Glaser, K. (2017): *Street Art und neue Medien. Akteure, Praktiken,* Ästhetiken. Bielefeld: Transcript Verlag.

Milbich, T., Bautista, M., Sutter, E. and Ommer, B. (2017): Unsupervised Video Understanding by Reconciliation of Posture Similarities. *Proceedings of the IEEE International Conference on Computer Vision*. ICCV.

¿Metodologías en Crisis? Tesis 2.0 a través de la Etnografía de lo Digital

Domingo Manuel Lechón Gómez
domingo@sursiendo.com
Doctorado de Ecosur, Mexico; Sursiendo, Mexico)

La presente propuesta está basada en la experiencia y las reflexiones que han ido surgiendo en el transcurso del trabajo de investigación de doctorado todavía en proceso. Con el título "La disputa de Internet. Análisis de los marcos de acción colectiva del activismo tecnológico en México", este estudio busca describir, analizar e interpretar cuáles son los problemas sociotécnicos que diagnostican los propios actores sociales, quiénes son los causantes de esos problemas, quiénes componen un "nosotros" entre los movimientos ciberactivistas, quién es la audiencia a la que va dirigida la acción colectiva y cuál es la propuesta sociopolítica que plantean para resolver el problema. Todo ello desde la propuesta de los marcos de acción colectiva y la metodología basada en la etnografía de lo digital.

La participación en el Congreso tiene la intención de proponer preguntas para reflexionar e iniciar diálogos necesarios sobre los cambios epistemológicos y metodológicos que pueden darse con las investigaciones con/desde/en Internet, y las cuestiones éticas que subyacen en las ciencias sociales y las humanidades relacionadas con las redes digitales.

Partiendo de que Internet se inserta en un contexto histórico de profundos cambios sociales, a la vez que es uno de los dispositivos que potencia esas cambios en las sociedades actuales, cambios epistemológicos e incluso ontológicos. Como explicaba Priani (2012) desde hace años que se está dando un "desplazamiento del proyecto ilustrado", de la Modernidad, y con ello se ponen en cuestión el saber científico imperante y las formas de construir conocimiento adscritas a él. Esas transformaciones se vienen fraguando desde los años 60, y los llamados nuevos movimientos sociales dan cuenta de ello, impugnando al sistema desde el ecologismo, el feminismo, el antirracismo, el anticolonialismo, el antimilitarismo, etc. Los tecnoactivistas, que recogen enseñanzas de esos movimientos previos, de sus principios, acciones y propuestas, ahora actúan en el terreno de Internet incorporando cuestionamientos y evoluciones desde y hacia las ciencias.

Así, estos y otros cambios se están produciendo dentro mismo de las academias, como presentó Wallerstein (1996) en el Informe de la Comisión Gulbenkian, la hibridación de disciplinas es un hecho (necesario). Wallerstein y colaboradores invitan a explorar y dar palabra a lo que está ocurriendo en la actualidad en el campo de la ciencia y a idear las medidas institucionales que lo asienten y hagan operativo, para que las ciencias "sean más verdaderamente pluralistas y universales" (Wallerstein, 2006, 101). ¿Internet puede que haga más factibles esas transformaciones?

Uno de los aspectos que entran en debate ahí son las posiciones de objetividad y subjetividad, que por ejemplo Donna Haraway impugnó también por esas fechas con el "conocimiento situado" (1996). Con este concepto se pone en cuestión la construcción de conocimiento "desde afuera", problematiza aspectos tales como la influencia de la situación de "encuentro con el otro" en el investigador y los aspectos sensibles de la relación social que se plantea con los sujetos entrevistados u observados; y al abordar un hecho social prioriza la construcción conjunta de conocimiento entre el investigador y quienes devienen su objeto de estudio.

Con ello se puede vislumbrar que está en crisis el paradigma científico moderno, lo cual puede ser una buena oportunidad para debatir y trazar nuevos itinerarios.

En los estudios sociales sobre algún aspecto de las Tecnologías de la Información y la comunicación (TIC), o de Internet en concreto, también han ido cambiando los enfoques. Como por ejemplo es lo que Gálvez y colegas apuntaban: "El determinismo, ya sea tecnológico o social, ha marcado gran parte de las aproximaciones que se han hecho desde las ciencias sociales al estudio de la tecnología" (Gálvez y otros, 2003; p1). Ya cada vez más se mira desde una posición sociotécnica, tanto lo social como lo técnico se influyen mutuamente, y es necesario que cualquier investigación se aproxime desde ahí.

Por ello, entrando en temas metodológicos, por ejemplo, la etnografía de lo digital o virtual, que en un principio se asumió como el estudio de la práctica online, en la actualidad, lo que prevalece es un enfoque holístico en el que se superponen los campos online y offline (Hine, 2004). En definitiva, la etnografía virtual es un híbrido, en cuanto apunta a grupos en línea relacionados con situaciones fuera de línea.

Desde los movimientos conectados se da cuenta de otras formas de mirar Internet; por ejemplo, Carmona Jiménez (2011) apunta que junto a la noción de dispositivo sociotécnico que sume en cierta medida a Internet como un artefacto (socio-facto), el ciberespacio además permite considerarlo como un "lugar" en el que se gesta cultura (Hine, 2004) y proporciona una forma de "habitar", por lo que en verdad es un "espacio antropológico", pues hay una construcción simbólica del espacio. Estar en terreno exige que el investigador se convierta en usuario y su "observación participante" significa participar e interactuar (Carmona Jiménez, 2011).

Para Estalella y colaboradores, la etnografía de lo digital es "la adaptación de la metodología etnográfica a las propiedades de los fenómenos que se desarrollan a través de lo digital implica repensar muchos de sus conceptos básicos y planteamientos metodológicos" (Estalella y otros, 2006; p2).

Además, al tratarse de una investigación que se introduce en el mundo tecnoactivista en México, es impor-

tante considerar factores éticos, como el tratamiento de los anonimatos, el uso de programas de análisis de datos de código libre, las licencias de publicación, etc.

La participación en el Congreso puede aportar esas reflexiones que busquen nuevos itinerarios para iniciar diálogos sobre estas novedades, con sus dificultades y sus retos, para la reflexión sobre las ciencias sociales y humanísticas en la sociedad-red.

Referencias

Carmona Jiménez, J. (2011) Tensiones de la etnografía virtual: teoría, metodología y ética en el estudio de la comunicación mediada por computador. *Revista F@ro* No 13. Facultad de Ciencias Sociales, Universidad de Playa Ancha, Chile. En línea: http://web.upla.cl/revistafaro/n13/art03.htm

Estalella, A. (2007) *Etnografías de lo digital. borrador*. En línea: http://www.prototyping.es/wp-content/uploads/2014/05/Estalella_Etnografias-de-lo-Digital-borrador-parcial.pdf

Estalella, A.; Ardévol, E.; Domínguez, D.; y Gómez Cruz, E. (2006) Etnografías de lo digital, Actas del Grupo de trabajo, *III Congreso Online - Observatorio para la Cibersociedad* Del 20/11/2006 - 03/12/2006. En línea: http://mediaccions.net/wp-content/uploads/etnografias-digital-actas.pdf

Gálvez, A.M.; Ardèvol, E.; Nuñez, F. y González, I. (2003). "Los espacios de interacción virtual como dispositivos sociotécnicos". Comunicación presentada para el *VIII Congreso Nacional de Psicología Social*. Torremolinos, Málaga, Abril 2003.

Haraway, D. (1995) *Ciencia, cyborgs y mujeres. La reinvención de la naturaleza*. Madrid: Cátedra.

Hine, C.. (2000). Etnografía virtual. UOC, Barcelona.

Laraña, E. (1999). *La construcción de los movimientos sociales*. Madrid, Alianza, 1999.

Melucci, A. (1994) ¿Qué hay de nuevo en los nuevos movimientos sociales? En *Los nuevos movimientos sociales: de la ideología a la identidad* / coord. por Joseph Gusfield, Enrique Laraña Rodríguez-Cabello. págs. 119-150.

Mosquera Villegas, M. A. (2008) De la Etnografía antropológica a la Etnografía virtual. Estudio de las relaciones sociales mediadas por Internet. Fermentum. *Revista Venezolana de Sociología y Antropología*, vol. 18, núm. 53, septiembrediciembre, 2008, pp. 532-549 Universidad de los Andes Mérida, Venezuela.

Priani, E. (2012) Molinos o gigantes. Cambio y nuevas tecnologías en las humanidades. *Revista Virtualis* No. 5 27 de junio 2012. Publicado por Centro de Estudios sobre Internet y la Sociedad y el Tecnológico de Monterrey. p 9 a 12

Ruiz Méndez, M. R. y Aguirre Aguilar, G. (2015) Etnografía virtual, un acercamiento al método y a sus aplicaciones. En *Estudios sobre las Culturas Contemporáneas*. Época III. Vol. XXI. Número 41, Colima, verano 2015, pp. 67-96.

Tarrow, S. (1997). *El poder en movimiento. Los movimiento sociales, la acción colectiva y la política*. (H.b. Resines, Trad.) Madrid, España: Alianza.

Wallerstein, I. (ed.) (1996). *Abrir las ciencias sociales, Comisión Gulbenkian para la reestructuración de las ciencias sociales El Mundo del Siglo XXI*. México, ed. Siglo XXI.

Wallerstein, I. (2005). *Análisis de Sistema-Mundo Una introducción*. Madrid, España: Siglo XXI.

Hashtags contra el acoso: The dynamics of gender violence discourse on Twitter

Rhian Elizabeth Lewis
rhian.lewis@mail.mcgill.ca
McGill University, Canada

Introduction

The spring of 2016 has become known as the "Primavera Violeta" ("Purple Spring"), a period that saw the emergence of new digital activist networks tackling gendered and sexual violence in Latin America. Of the hashtags generated by these movements, few gained the public recognition and "celebrity status" of #MiPrimerAcoso ("My First Harassment" or "My First Abuse"), a hashtag that asked users to publically share their first experiences of sexual violence. On April 23, 2016, women in Mexico and across Latin America shared their stories via their personal Twitter accounts in response to a request tweeted by journalist Catalina Ruiz Navarro of the pop-feminism collective (e)stereotipas: "¿Cuándo y cómo fue tu primer acoso? Hoy a partir de las 2pmMX usando el hashtag #MiPrimerAcoso. Todas tenemos una historia, ¡levanta la voz!" *(When and how did your first acoso happen? Today from 2pm on, use the hashtag #MiPrimerAcoso. We all have a story, raise your voice!)*

Figure 1: A typical #MiPrimerAcoso tweet. In English: I was eleven years old and a man passed on a bicycle and grabbed my breast. A woman in the street blamed me for wearing that blouse.

After its initial launch, #MiPrimerAcoso spread rapidly throughout Mexico and quickly became a trending topic across Latin America. This analysis investigates the ways that Twitter users– activists, laypersons, public figures–

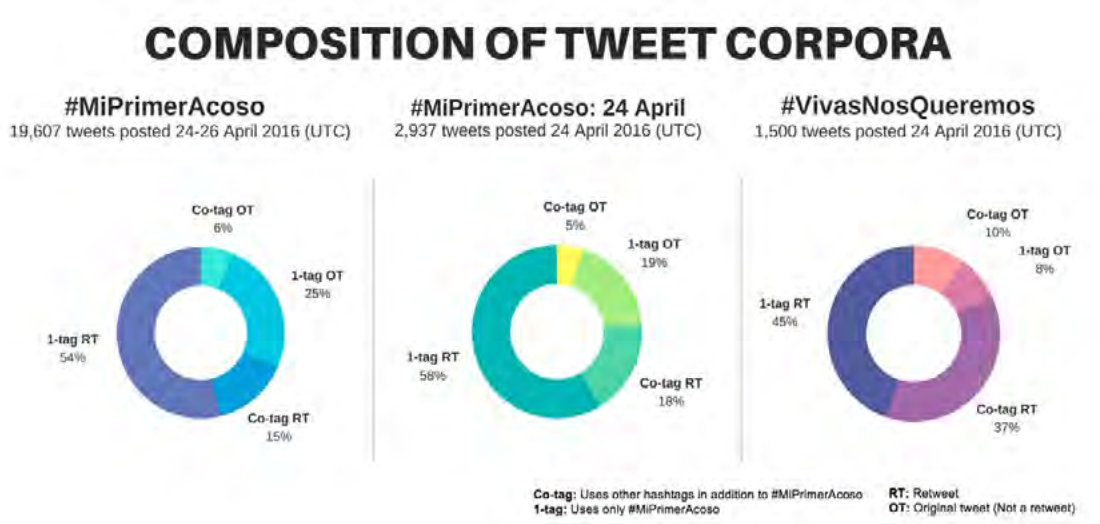
use hashtags to talk about trauma, paying special attention to the ways that quantifiable modes of Twitter engagement point to more complex affective experiences.

Methods

This project undertakes both qualitative and quantitative analyses of tweets posted using #MiPrimerAcoso in order to examine the key actors, contexts, and conditions that emerged from the hashtag's narrative premise. For the initial assessments, this analysis uses the #MiPrimerAcoso corpus collected and published by media company Lo Que Sigue. To provide a point of comparison, this project also analyzes a collection of tweets posted using another Primavera Violeta hashtag– #VivasNosQueremos ("We Want to Live")– whose corpus was collected and published by Lo Que Sigue at the same time as the #MiPrimerAcoso corpus.

Affective (and Effective) Tweeting

Hashtag dialogues serve to construct and re-construct bridges between different streams of dialogue within movements, between movement collaborators and stakeholders, and between activists, political powers, and the general public. To illustrate some of the preliminary findings of this exploration, I evaluate the prevalence of retweets and multiple-hashtag use (or "co-tagging") in the #MiPrimerAcoso corpus and another corpus published by #LoQueSigue of tweets posted using #VivasNosQueremos. Throughout this paper, I call upon Papacharissi's (2015) work on the affective properties of Twitter dialogues to further illustrate the forms of personal and political affect that drove the trans-national trajectory of #MiPrimerAcoso.



Although #MiPrimerAcoso is entangled with other Twitter dialogues on gender violence, it "stands alone" more often than one of its closest peers, and is less frequently retweeted and co-tagged. Here, I find that these concrete metrics summarize diverse modes of engagement: retweeting another user's personal story of violence is necessarily a different act than retweeting a popular news story about the hashtag. However, these metrics do demonstrate the ways in which use characteristics reflect the discursive mandate of a hashtag. Engagement with #MiPrimerAcoso might include reading, listening, creating original content, rebroadcasting, or responding to the content of other users within the affective public generated by the hashtag. This diverse set of practices allows Twitter users to "tune into an issue or a particular problem of the times but also to affectively attune with it, that is, to develop a sense for their own place within this particular structure of feeling" (Papacharissi 118). The Twitter users who tweeted their experiences of violence undertook a delegated task of content creation in response to the prompts posted by Ruíz Navarro. This guiding of the discussion allowed Twitter users to act and to *feel* using a pre-constructed response frame. By asking users to share the how and when of their first acoso, users tasked with personifying the political and *making it about themselves*. By focusing on a tweet structure that outlines an individually expressive personal action frame through the medium of shared experience, #MiPrimerAcoso allows its users to make "small and fitful contributions" (Bennett and Segerberg 2011) to a cause while feeling a profound sense of identification with the movement.

If we want to understand what it is people want from digital activism, #MiPrimerAcoso offers captivating insights regarding our need to see ourselves within online political movements. The secret of #MiPrimerAcoso's handling of collective and individual resonance lies in its personalization and generalizability: although the hashtag calls on a specific category of experience, it is sufficiently broad that many interpretations of *acoso* fit the bill, and many users were able to affiliate with the has-

htag without necessarily sharing a personal story of sexual violence. As Papacharissi (2015) notes, the use of hashtags as "open" signifiers allows various publics to affiliate with a movement and "fill in" the open hashtag with their own desired meanings. Women were able to link their own experiences of sexual violence to the individual narratives that had already been shared using the hashtag #MiPrimerAcoso. What, then, of those who did not contribute their original narratives to the library of primer acosos, but instead chose to respond or rebroadcast existing #MiPrimerAcoso content? In responding to a tweet, users may amplify, stifle, or otherwise alter the public life of the digital *acoso*. Although Papacharissi and others have linked the act of retweeting to the expression of solidarity with a movement, this conclusion may prove reductive in the context of #MiPrimerAcoso. However, solidarity does not adequately summarize the act of rebroadcasting another person's *acoso*: it is an expansion of the tweet's intangible audience of ethical witnesses to the tweeted acoso, a "re-telling" of scene of violence. Like any other hashtag, #MiPrimerAcoso needed to meet specific communicative and technical (in the case of Twitter) requirements in order to maximize its "reach" and extend beyond the core audience of (e)stereotipas. Referring to the act of retweeting, Papacharissi argues that refrains reinforce affect (Papacharissi 2015). By posting tweets tagged #MiPrimerAcoso, users spread the affective and contextual implications of the hashtag to their own Twitter audiences: those in digital "earshot" of their tweets. Similarly, the authors of original #MiPrimerAcoso tweets were also invited to act as amplifiers of the larger movement by adding their story to a collaborative, polyvocal narrative of lived violence.

Conclusions

In our study of digital movements, the use of the hashtag is the tip of the iceberg in comparison to the forms of knowledge, feeling, and understanding that emerge from these affective discourses. The results of this research have also suggested that conventional Twitter analysis methods may not adequately assess the affective clout of digital dialogues. For this reason, this analysis has strived to use the concrete metrics of the #MiPrimerAcoso data as guide to direct a "closer" reading of the narrative attributes of the tweets. When examining Twitter data, we must strive to expand the possibilities behind a simple, quantifiable act such as a retweet, and understand the hashtag as a point of contact between the user and digito-phenomenological processes of which we are largely unaware. Of course, there are key characteristics of the hashtag itself that are crucial to our understanding: its connectivity, for example, or its capacity to understand individual content as part of a larger dialogue. The hashtag is a departure point: an entity that gives rise to visible manifestations of trauma, digital acts of vulnerability and moments of personal catharsis, responses of support, condemnation, or indifference.

We should consider the tweet, then, as the execution of a series of digital actions, but also as the manifestation of a confluence of contacts between the ontological and phenomenological worlds of Twitter. To better assess these intangible qualities of Twitter data, we can listen to the testimonies of #MiPrimerAcoso authors, and pay attention to the strategies they employ to construct the *acoso* in relation to their present selves, the ways in which they reflect on the act of tweeting the *acoso* in front of an intangible digital audience. Here, I want to emphasize the diversity of experiences that users bring to the discursive space of Twitter, and the need to pay attention to the varied motivations that drive Twitter users to participate in social campaigns. These experiences do not easily reduce themselves to quantitative metrics, but we can search for their traces in the textual manifestations of our digital activity: the stories we tell, the words we use, the affective investments that we make as observers and participants.

References

Bennett, W. Lance, and Alexandra Segerberg. "The logic of connective action: Digital media and the personalization of contentious politics." *Information, Communication & Society* 15.5 (2012): 739-768.

Gerbaudo, P. (2014) The persistence of collectivity in digital protest,*Information, Communication & Society*, 17:2, 264-268, DOI: 10.1080/1369118X.2013.868504

Lo Que Sigue TV (2016). Tuits de #MiPrimerAcoso disponible en "table_5d787653". Database available on Carto. https://lqs.carto.com/tables/table_5d787653/public

Papacharissi, Z. (2015) *Affective publics: Sentiment, technology, and politics.* Oxford University Press.

Novas faces da arte política: ações coletivas e ativismos em realidade aumentada

Daniela Torres Lima
danielatorreslima@yahoo.com.br
Universidade Federal de Juiz de Fora, Brazil

Da modernidade até aquilo que Lipovetsky e Serroy (2010) nomearam como hipermodernidade, o sistema de construção de imagens e recepção delas passaram por diferentes fases e evoluções, transformando radicalmente as relações individuais e interações sociais. Para Pierre Lévy (2010), a partir da multiplicação de dispositivos móveis e suas funções comunicativas a nível global, as relações sociais deslocaram-se de contextos locais de interação e foram rearranjadas em extensões indefinidas, baseadas

em uma noção de espaço-tempo diferenciados. Essas novas conexões aglutinam indivíduos em afinidades de interesses e conhecimentos, propiciando um processo de cooperação e troca entre eles que independe de proximidades geográficas, mas que evidentemente constitui uma nova forma de organização social (Lévy, 2010:134).

Dessa nova interpretação de lugar, sem fronteiras geográficas e hiperconectado, emergem outras abordagens perceptivas sobre ser cidadão e de atuação sobre esses espaços e assuntos cada vez mais comuns. Percebemos, então, a emergência de formas de engajamento político que ultrapassam a prática do ativismo em partidos, sindicatos e movimentos sociais locais, assumindo também no ciberespaço e na utilização de novas mídias uma postura ativista, tomando tais sistemas como suporte para suas práticas. Na atualidade, notamos que os atos de ativismo e militância tem se apoiado em tecnologias cada vez mais avançadas, não apenas para usufruir dessa fácil e acessível forma de divulgação de informações, mas também explorando o potencial político de pressão e engajamento com a realidade pelo permeio de ambos espaços que mediam interações sociais e culturais atualmente: o virtual e o físico.

Nesse cenário, a tecnologia de realidade aumentada se destaca ao exigir a participação contínua de um interveniente, promovendo a interação entre objetos virtuais tridimensionais e usuários reais, interatuando em tempo-real no espaço. Para Gonçalves (2006), o uso de imagens atraentes e passíveis de interação e manipulação tem a função de mobilizar e despertar o interesse para esse gênero de iniciativas, criando um entusiasmo para o engajamento político, levantando questões e discutindo-as de forma crítica e lúdica ao mesmo tempo. Portanto, a arte tem papel fundamental nessas ações, uma vez que age como um fator de atração e reflexão para questões socioculturais relevantes (Gonçalves, 2006:12).

De acordo com Mark Skwarek (2017), artista multimídia que tem trabalhado na construção e articulação de atos de resistência em redes, a tecnologia de realidade aumentada ganhou notoriedade neste século na mediação de narrativas elaboradas permitindo que artistas usufruam da potência de visualização e comunicação digital para alcançar propósitos reflexivos e experiências estéticas contemporâneas. Segundo o autor, os primeiros ativistas a utilizarem-se desta tecnologia foram inspirados pelo trabalho de *culture jammers* e artistas de grafite dos anos 80, que apoiados em uma semiótica de guerrilha, colocavam em voga técnicas de anti-consumismo e anti-capitalismo a fim de romper ou subverter a cultura *mainstream*. Esses grupos criavam grosseiramente sobreposições e intervenções sem permissão de um estabelecimento, desafiando a noção do espaço público e privado ao serem aplicadas sobre muros, portas de instituições ou no logotipo de corporações (Skwarek, 2014: 17). Na realidade aumentada, entretanto, essas sobreposições ocorrem de forma virtual em ações interativas através da digitalização de um código *Quick Response* (QR) ou de reconhecimento de um objeto pré-codificado via câmera de celular, disparando elementos virtuais tridimensionais (frases, desenhos, vídeos e pichações) que aparecem sobrepostos ao mundo cotidiano através da tela do telefone. Enquanto visualiza elementos virtuais sendo sobrepostos à 'realidade', o usuário tem a possibilidade de registrar sua interação através da câmera fotográfica ou de um *print screen* disponibilizado em algumas aplicações. A etapa seguinte, apesar de não ser o único modo de conferir credibilidade a um movimento visto que a maioria dos aplicativos interativos computam o número de participantes e os identifica por localização, é a divulgação voluntária e em rede dos materiais visuais obtidos durante a interação. Talvez este seja o ponto crucial desses atos de ativismo coletivo uma vez que coloca maior agência e destaque nas mãos de pessoas comuns que se prontificam e se declararam como apoiadoras de um movimento coletivo, expondo suas identidades em uma espécie de fragmentação da cobertura midiática, além de ramificarem um ponto inicial de protesto para diferentes comunidades e pontos do globo de forma instantânea, o que de alguma forma aumentam o reconhecimento de lutas que tem se tornado cada vez mais universais.

Como um exemplo notável dessa nova forma de manifestar-se politicamente encontra-se o protesto *Occupy Wall Street*, organizado através de redes de internet em 2011 nos Estados Unidos, e que contou com a colaboração de artistas e programadores de todo o mundo para seu sucesso. Esse movimento teve estopim na cidade de Nova Iorque em um protesto que reivindicava o fim da desigualdade social e econômica, a corrupção e a grande influência de empresas sobre o governo, particularmente do setor de serviços e o financeiro. Os manifestantes não tiveram permissão para protestar em Wall Street, onde somente parte da calçada estava acessível ao público, sob constante vigilância da polícia. Foi a partir deste impedimento que ativistas de 82 países se organizaram para criar o movimento virtual sobre a hashtag #arOCCUPYWALLSTREET, visando que utilização de aplicativos de realidade aumentada levasse o protesto ao coração do distrito financeiro e desse voz aos manifestantes barrados (Fig.1)

Fig 1. *AR Occupy Wall Street*, em 2011. Imagens retiradas do site do evento.

Já num propósito de discussões de gênero e articulações feministas, *The Whole Story Project* propunha ocupações simbólicas dos espaços a partir da reformulação de imagens femininas, visando discutir a presença das mulheres na sociedade, inclusive na história, atuando num devir entre arte e ativismo. O projeto foi inspirado na campanha *Monumental Women Campaign*, que teve início em 2016 a partir de uma constatação que apenas 7,5% das 5193 estátuas espalhadas pela cidade de Nova York retratavam mulheres. Desde então este projeto passou a ser articulado por artistas com o objetivo de a arrecadar fundos para colocar as primeiras estátuas em homenagem à história das mulheres no Central Park de Nova York, que além de contribuírem para a maior representatividade feminina na sociedade, promoveria a conscientização sobre as contribuições das mulheres para a história compartilhada. Foi no início de 2017, durante as preparações para a maior marcha feminista já realizada no mundo, a *Women's March*, que artistas multimídias se apropriaram do projeto físico para criarem um aplicativo de realidade aumentada que colocasse diferentes mulheres em monumentos públicos, relatando suas histórias de luta e transformações sobre a cidade. O aplicativo *The Whole Story*, embora virtual, pretendia fazer ondas no mundo físico, chamando a atenção para que mulheres comuns se inspirem com a história de outras grandes realizadoras, empoderando-as politicamente em sua comunidade, além de permitir ao público recuperar a narrativa histórica das grandes cidades. (Fig.2)

Diante desses exemplos, e apoiados nos estudos de Ricardo Rosas (2003), podemos dizer que tais movimentos tratam-se de novas formas de ativismo uma vez que se articulam como modos de resistências temporárias e nômades, baseadas em ações coletivas de intervenção em espaços públicos que se fundamentam pelas redes virtuais ou no uso de mídias diversas (Rosas, 2003). De um modo pragmático, esses ativismos em Realidade Aumentada passam a abranger diversas províncias de significado e experimentar universos múltiplos em forma de manifestações, evidenciando que os métodos do passado podem estar se tornando menos efetivos e se faz necessário a reestruturação de um modelo que condiz com uma realidade hipermoderna.

Sobre antigos modelos de manifestação política e suas funcionalidades, o coletivo americano *Critical Art Ensemble* (1996) argumenta sobre a ocupação de espaços públicos em protesto da atualidade. Para o grupo, embora alguns dos monumentos do poder permaneçam fixos, ostensivamente presentes em locais estáveis, o poder já não reside nesses locais, já que a ordem e o controle agora se movem livremente. O que é proposto em seus manifestos é a apropriação de algo que tenha um valor comum na atualidade, tanto para as instituições de poder as quais se combate quanto para a sociedade de forma geral.

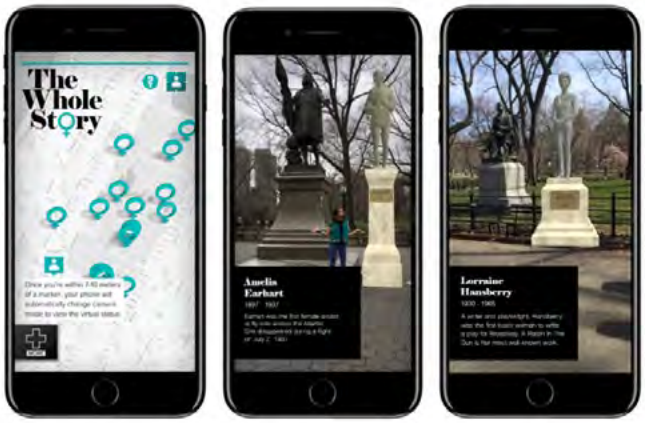


Fig. 2. RA em discursos de empoderamento feminino (2017). Imagem retirada do site do projeto.

Nesse sentido, a ocupação e uso das novas mídias e do espaço político que elas fornecem mostram-se como uma abordagem diferenciada sobre as competências da comunicação, flexibilizando seus usos comuns e trazendo a reflexão sobre as potencialidade de um trabalho colaborativo. Dentro do contexto cultural-midiático existente, isso seria uma forma de combater um sistema através de seus próprios meios, sendo, portanto, uma estratégia válida de enfrentamento, e mostrando um caminho possível para inversões temporárias no fluxo do poder.

Ao tentar compreender as novas formas de exercício político na contemporaneidade, esbarramos com uma grande gama de ações que têm sido realizadas através de tecnologia de Realidade Aumentada. A sobreposição de camadas virtuais e moldáveis às referências visuais do mundo físico tem sido usada como forma de atração e engajamento para atos de ativismos na tentativa de discutir sistemas políticos, econômicos e ambientais da forma mais atual possível. Essas novas formas de envolvimento político coletivo são reflexos de tempos onde as mídias móveis e a hiperconectividade reoganizaram nossas relações com o mundo, motivando indivíduos a ingressarem em novas experiências estéticas e reflexões singulares através da fusão entre arte, tecnologia e política.

References

Critical Art Ensemble (1996), "Electronic civil desobedience and other unpopular ideas". Disponível em: http://critical-art.net/books/ecd/ (acesso em 13 janeiro de 2018).

Gonçalvez, F. (2006). *Resistência nômade: arte, colaboração e novas formas de ativismo na Rede.* Rio de Janeiro: Revista Compós, p. 85-90.

Lemos, A. (2006). *Ciberespaço e tecnologias móveis: processos de territorialização e desterritorialização na cibercultura.*Porto Alegre: Sulina.

Lèvy, P. (2010). *Ciberespaço.*São Paulo: Editora*34*, p.133-134.

Lipovetsky, G., e Serroy, J. (2010). *O ecrã global: cultura midiática e cinema na era hipermoderna.*Lisboa: Edições 70.

Rosas, R. (2003). Que venha a mídia tática. *In: Rizoma.net.* Disponível em: http://www.rizoma.net/interna.php?id=174&secao=intervencao/ (Acesso 17 de dezembro de 2017).

Skwarek, M. (2017). Augmented reality activism. *InAugmented Reality Art.* Berlim: Springer, p. 3-29.

The Whole Story Project (2017). Website oficial. Disponível em: https://thewholestoryproject.com (Acesso em: 10 nov. 2017).

Modeling the Fragmented Archive: A Missing Data Case Study from Provenance Research

Matthew Lincoln
mlincoln@getty.edu
Getty Research Institute, United States of America

Sandra van Ginhoven
svanginhoven@getty.edu
Getty Research Institute, United States of America

Historians grapple with missing information constantly. While there are many statistical tools for gauging the impact of missing source data on quantitative results and conclusions, DH researchers have rarely deployed these tools in their work. This paper presents one implementation of data imputation used in the study of the New York City art dealer M. Knoedler & Co.Demonstrating the significant contribution imputation had on our study and its conclusions, this paper will discuss specific, practical rhetorical strategies, including static and interactive visualization, for explaining this methodology to an audience that does not specialize in quantitative methods.

Missing Data in the Digital Humanities

Miriam Posner has argued that both data structures and rhetorical conventions for computing with missing information, uncertainty, and highly subjective/viewpoint-contingent knowledge remains a key desideratum of DH scholarship. (Posner, 2015) Several attempts have been made by the information science community to express uncertainty in a structured format, ranging from generalized ontologies for reasoning in a networked world (Laskey et al., 2008), as well as more specific projects such as the *Topotime* library for reasoning about temporal uncertainty. (Grossner and Meeks, 2013)

However, many DH projects have sidestepped these approaches. Matthew Jockers, for example, has asserted that the availability of full text is becoming such that literary historians will no longer have to be concerned about drawing a representative sample. (Jockers, 2013: 7–8) More commonly, though, scholars have attempted to carefully constrain their conclusions based on what they know to be missing from their data. Theorizing and documenting the difference between one's data set and one's subject has become a genre of DH work unto itself. Katherine Bode has argued that such documented datasets should be understood as *the* object of DH inquiry. (Bode, 2017)

While statistical literature on the problem of missing-data imputation is quite mature (see Gelman and Hill, 2006 for a valuable review), few DH research projects have openly explored the use of statistical procedures for reckoning with missing data, nor have they grappled with how to theorize and present such imputation in the context of their home disciplines. (An important exception includes Brosens et al., 2016) Bode, for one, has explicitly rejected such approaches, arguing (without specific evidence) that quantitative error assessment cannot be usefully performed in historical analysis. (Bode, 2017: 101)

We argue that such methods should be *central* to data-based digital humanities practice. Simulation and imputation allow us to realize multiple, sometimes conflicting assumptions about the nature of missing data. In doing so, these affordances allow us to evaluate how certain assertions may propagate their assumptions through the transformations we perform on our sources.

Case Study: Modeling M. Knoedler & Co.'s Business from Sparse Stock Books

As part of a research initiative into data-based approaches to the study of the art market, we are investigating the changing strategies of the New York City art dealer M. Knoedler & Co., whose stock books have been encoded by the Getty Research Institute (http://www.getty.edu/research/tools/provenance/search.html). Based on these transaction data, we have built a predictive model that classifies whether a given artwork would result in a profit or a loss, using a host of variables such as how much money the work of art originally cost, the genre and size of the work, their prior relationships with buyers and sellers, and the time the work remained in stock before it was sold, to name but a few. Predictive modeling illuminates complex relationships between these variables and highlights unusual sales for further archival research.

As informative as these stock books are, however, many of their notations are partial: Knoedler's staff may have neglected to record the date of sale; there may be a listed purchase without a description of the type of work (i.e.portrait, landscape, etc.); the identity of the buyer, and whether they were a first-time customer or a well-known shopper, may also have gone unrecorded. Because our random-forest-based model (Liaw and Wiener, 2002) does not allow missing values, we must either discard incomplete records (and thus eliminate nearly half of the records from consideration), or we must find ways to impute values for our predictor variables.

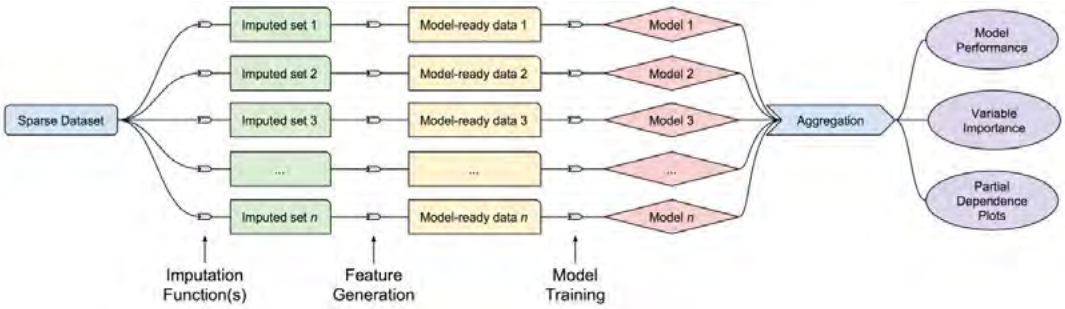


Figure 1 Schematic workflow for imputing missing data, producing derivative features, building models, and then aggregating statistics from the multiple models produced.

While it is impossible to perfectly reconstruct these missing records, it *is* possible to operationalize educated guesses about their possible values. (Figure 1) Purchase and sale dates for artworks, for example, can be predicted with some accuracy based on their location in the roughly-chronological series of stock books. Likewise, unknown genres can also be imputed as a function of the existing distribution of genres across stock books, with, e.g.abstract paintings being far less common in the pre-20th c. books than in the later ones. By defining an informed range of possibilities for these missing data, and then sampling from that range, we can produce ensemble models and results that provide a more nuanced representation.

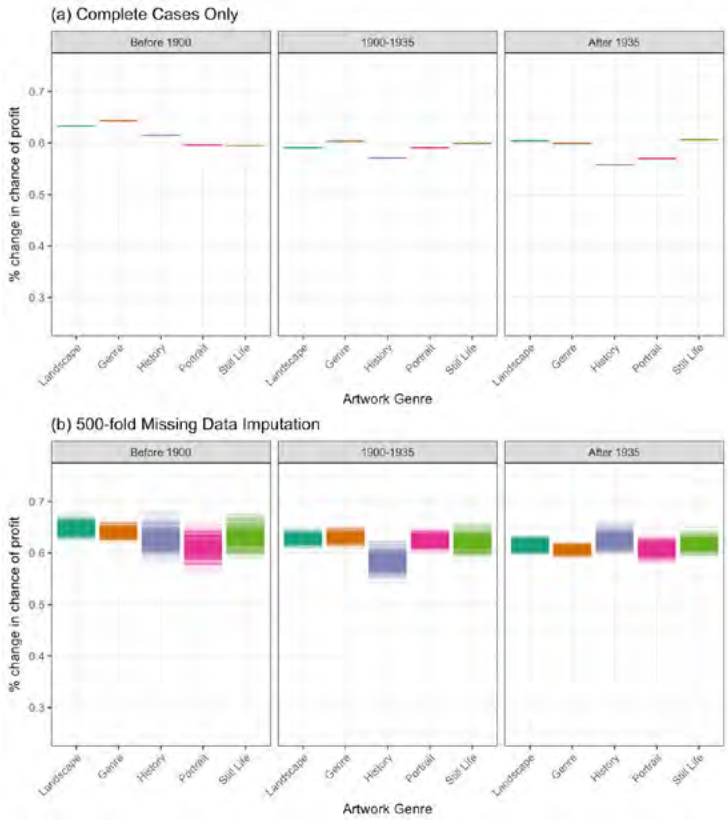


Figure 2 Partial dependence plots illustrating the marginal effect of artwork genre on Knoedler's chance at profitability.

Figure 2(a) shows the marginal effect of artwork genre on Knoedler's chance of turning a profit across three periods of their business, only considering around 20,000 "complete" cases from the Knoedler transaction records (approximately 60% of the known transactions they made.) A first glance suggests that history paintings were markedly less profitable after 1935, while still lifes became comparatively more profitable after 1935.

However, 2(b) shows the results not from 1 model, but from 500 models, each trained on a slightly different set of stochastically-imputed data. By visualizing one bar for each model, this plot drives home the effect of increased uncertainty on these measurements, while visually foregrounding the crucial methodological decision - 500 models instead of 1 - in a way that a box plot or other summary visualization method does not (at least, not in the eyes of a reader unused to reading such idioms.) The apparent advantage of still life in Knoedler's post-1935 business has evaporated, although the notably-lower value of history paintings between 1900-1935 may have withstood this simulation of uncertainty. While this model affirms that genre is largely an anachronistic construct that has little effect on prices, these results complicate a simplistic reading by indicating that, in some cases, there *is* a significant relationship that must be reckoned with.

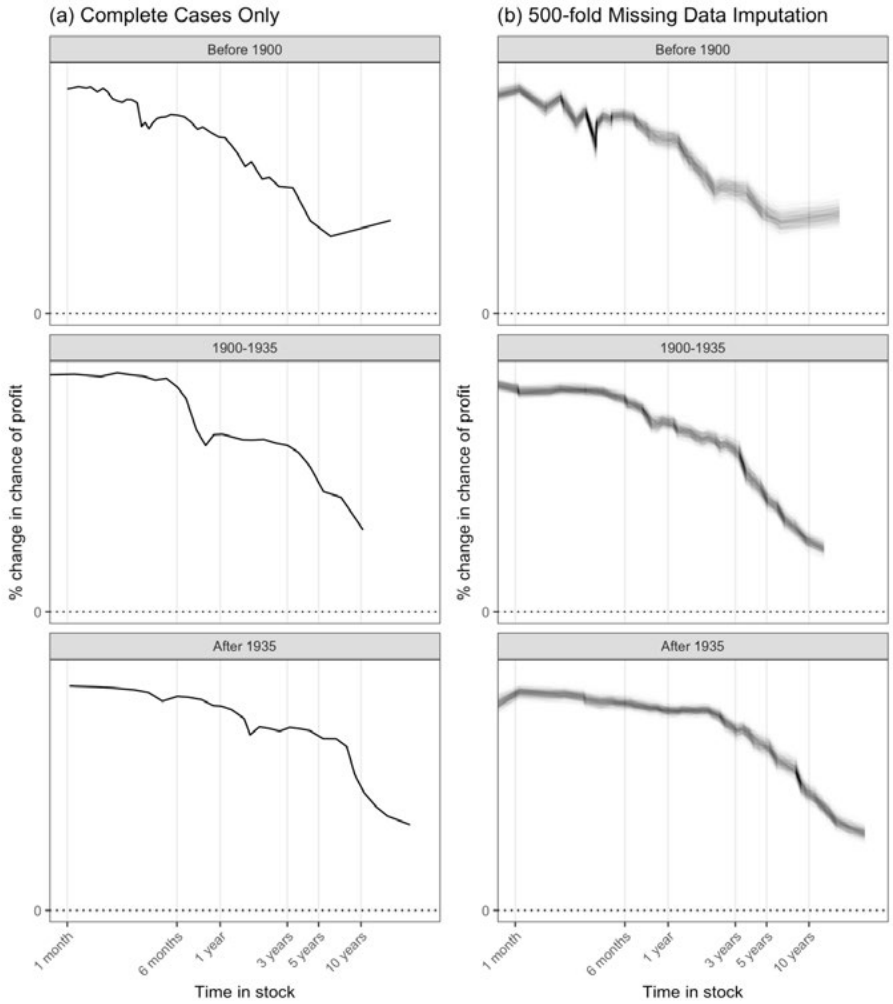


Figure 3 Partial dependence plots illustrating the marginal effect of time in stock on Knoedler's chance at profitability.

Figure 3 shows a similar comparison of complete case vs.imputed data for a continuous variable: the time a painting spent in stock. Both 3(a) and 3(b) support the conclusion that not only did a longer time in stock contribute to lower chances of turning a profit, but that Knoedler's window for making a profitable sale grew throughout the lifetime of the firm, from around 2 years before 1900, to more than 5 years after 1935. The increased uncertainty added by the multiplicity of models in 3(b) discourages the kind of over-interpretation that the seeming-precision of 3(a) allows. However, it also demonstrates that, even in the face of so many missing or imprecise dates in the Knoedler stock books, we can still recover meaningful quantitative conclusions.

Figure 4 Screenshot of an interactive application allowing users to modify imputation assumptions and see the effect on modeling and analysis results.

These static visualizations are easily enhanced through animation that shows the buildup of individual model characteristics into aggregate confidence intervals. (Lincoln, 2015) We have also experimented with interactive applications (Figure 4) that allow the user to specify different imputation assumptions, and then immediately see the downstream results on our predictive models, reinforcing the close relationship between starting assumptions and modeled conclusions. (An early demo of this work: https://mdlincoln.shinyapps.io/missingness/)

Computationally, these imputations are simple, perhaps even simplistic. More complex approaches, such as iteratively modeling every missing variable (Buuren and Groothuis-Oudshoorn, 2011), might lead to more accurate modeling. However, these less parsimonious methods are more opaque to humanities scholars. Operationalizing the historian's habit of educated guessing and thoughtful assumptions, and visualizing those operations straightforwardly, may allow missing data imputation to work its way into the accepted suite of DH methodologies.

References

Bode, K. (2017). The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History. *Modern Language Quarterly*, 78(1): 77–106 doi:10.1215/00267929-3699787.

Brosens, K., Alen, K., Slegten, A. and Truyen, F. (2016). MapTap and Cornelia: Slow Digital Art History and Formal Art Historical Social Network Research. *Zeitschrift Für Kunstgeschichte*, 79: 1–14.

Buuren, S. van and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3) doi:10.18637/jss.v045.i03.

Gelman, A. and Hill, J. (2006). Missing-data Imputation. In, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Oxford: Cambridge University Press, pp. 529–43.

Grossner, K. and Meeks, E. (2013). *Temporal Geometry in Topotime*. Stanford University Libraries http://dh.stanford.edu/topotime/docs/TemporalGeometry.pdf.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Laskey, K. J., Laskey, K. B., Costa, P. C. G., Kokar, M. M., Martin, T. and Lukasiewicz, T. (2008). *Uncertainty Reasoning for the World Wide Web*. W3C Incubator Group Report World Wide Web Consortium (W3C) https://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/ (accessed 26 November 2017).

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3): 18–22 https://www.stat.berkeley.edu/~breiman/RandomForests/.

Lincoln, M. D. (2015). DataGIFs: Animate Your Visualizations for Fun and Clarity *Matthew Lincoln, PhD* https://matthewlincoln.net/2015/12/18/datagifs-animate-your-visualizations-for-fun-and-clarity.html.

Posner, M. (2015). What's Next: The Radical, Unrealized Potential of Digital Humanities *Miriam Posner's Blog* http://miriamposner.com/blog/whats-next-the-radical-unrealized-potential-of-digital-humanities/ (accessed 20 April 2017).

Critical Data Literacy in the Humanities Classroom

Brandon T. Locke
blocke@msu.edu
Michigan State University, United States of America

Humanities data and data in our daily lives

As our world becomes increasingly data-driven, data skills and literacies (including the ability to assess data gaps and coverage, misleading visualizations, and the ethics surrounding data collection, usage, and sharing) are becoming crucial tools to our lives, both inside and outside of higher education. Scholarship across disciples is moving towards more data-intensive work, and scholars are increasingly expected to include open access to the data collected and used. At the same time, devices and software we use, the platforms we use to communicate, and the places we shop are increasingly enabled by the collection of data about our purchasing habits, web history, and contents of our email inboxes. Governments at all levels are increasingly collecting and using data to alter policies and direct day-to-day activities, ranging from transportation infrastructure to policing.

While much (though certainly not all) data-driven scholarship may seem significantly different from third-party data collection and data-driven policing, the former provides an opportunity to prepare students to understand, critique and improve the latter. Learning about the accurate and ethical and collection and usage of data and algorithms is a crucial part of liberal education that can help students better understand the processes around them, and better prepare them to apply those ethics and practices in the workplace and civic realm after graduation.

While many may think of data literacy as being the work of Computer Science departments, or perhaps library workshops targeted at researchers, the author argues that teaching these skills in the humanities classroom is fruitful for both the development of disciplinary knowledge and for developing crucial skills for use outside of the humanities classroom. Humanities data provides an excellent space to think critically about how people, ideas, and culture can and cannot be captured and analyzed through data. Comparing the data structures of colonial record keeping with the structures communities develop to document themselves provides clear lessons in the power of determining who and what gets documented, in the values that each community holds, and in privacy, ethics, and consent. Text mining novels, government records, or newspapers facilitates critical thinking about the value of metadata, the ability (or lack thereof) to derive meaning from large collections of text, and the use of different algorithms and approaches to ask different questions.

At the same time, the ability to think about humanities sources as data, and to properly curate and analyze them as such, provides a productive way to engage more with the way we conceptualize the sources, the way disciplinary knowledge is constructed and practiced, and the affordances provided by digitized and born-digital resources.

Data Challenges in Higher Education

The process of gathering, "cleaning," and organizing data can be incredibly time-consuming and difficult to prepare for. It can be tempting (and in many cases, required), to provide students with pre-prepared data to for analysis. Allotting time, either as in-class instruction or independent, project-based work, can take up weeks of time and can be a grueling disincentive for engagement. However, working critically with data rather than working with pre-packaged, pre-prepared datasets also aides us in the integration of digital humanities methods into the classroom, and better enables us to teach students emerging research methods through the full course of humanities research. Students can get a glimpse of the intellectual labor that goes into data collection, organization, and curation; not just in the final analysis.

There are several data literacy models that have shown success in other contexts. Data curation training often occurs in university-wide workshops or seminars, and are often brief and necessarily divorced from content and community practices (Carlson and Johnston 2015 p. 2-3). The Data Information Literacy (DIL) initiative, led by Jake Carlson and Lisa R. Johnson, is an extension of the ACRL Information Literacy Framework that focuses on

both the creation and consumption of data (ibid.). DIL is designed to be integrated into courses and research labs in the context of subject-specific data and domain-based community practices, but is primarily intended for faculty, staff, and graduate students working on peer-reviewed publication (ibid., p. 2-3). The Library-Led DH Pedagogy: Modeling Paths Toward Information and Data Literacy symposium facilitated productive conversations about the topic of data and information literacy in the digital humanities, but has not produced significant scholarship, models, or frameworks (Padilla et al. 2015).

In addition to making the case for teaching critical data literacy in the digital humanities classroom, the author will discuss both practical and theoretical approaches to data literacy in the undergraduate classroom that speak to the impetus behind teaching data literacy in the humanities: for greater disciplinary knowledge and understanding, to better facilitate digital scholarship and knowledge production, and to prepare students to better grasp, interrogate, and work with data in the public and private sector as citizens, employees, and employers.

References

Carlson, J. and Johnston, L. eds. (2015). *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers. Purdue Information Literacy Handbooks*. West Lafayette, Indiana: Purdue University Press.

Padilla, T., Smiley, B., Miller, S., and Mooney, H. (2015). "Modeling Approaches to Library–Led DH Pedagogy," *DH 2015 Global Digital Humanities Conference Abstract*. http://dh2015.org/abstracts/xml/PADILLA_Thomas_George_Modeling_Approaches_to_Libr/PADILLA_Thomas_George_Modeling_Approaches_to_Library_le.html (accessed 15 August 2017).

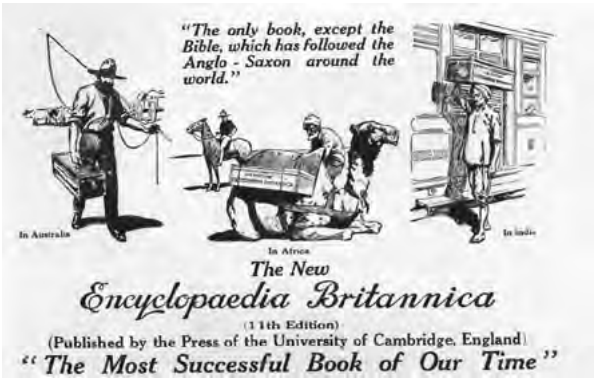
Ontological Challenges in Editing Historic Editions of the Encyclopedia Britannica

Peter M Logan
peter.logan@temple.edu
Temple University, United States of America

First published in 1771, *Encyclopedia Britannica* continues in publication today and is the only encyclopedia in any language to survive that 250-year period. Historical editions of the Encyclopedia offer scholars a unique means of examining the evolution of ideas and beliefs about sensitive cultural topics – such as suicide, race, and hysteria – by studying their treatment in different editions. But what can this curated dataset as a whole can tell us about larger patterns in the social construction of knowledge in the nineteenth-century English-speaking world?

We are creating a data set of all text from these historic editions for use in text mining. The corpus will include over 100,000 entries, all of which need to be tagged with essential metadata fields. How do we identify the different subject areas in this body of knowledge? This article briefly discusses the use of an automatic-metadata-generating algorithm, HIVE, created by the Metadata Research Center at Drexel University. But the central issue it addresses is the theoretical problem encountered in defining a subject vocabulary for this corpus.

The *Encyclopedia* claims to represent the "Sum of Human Knowledge," and while we can dispute this claim, it nonetheless represents the existence of older knowledge taxonomies used in its creation. How do we construct a subject vocabulary without distorting this older organizational scheme for subject categories? Those older vocabularies were clearly biased. For example, the decision to include or exclude entries, as well as the size assigned to entries, were all based on assumptions about what mattered as "legitimate" knowledge. Many of these are assumptions we no longer share; the editors excluded forms of knowledge rooted in folk and tribal cultures, and female authors were wholly absent until 1889. Racism and the perspective of British Imperialism are evident in many entries. These prejudices reflect the social beliefs of the writers and editors, of course, and as such, they illustrate the degree to which knowledge in the nineteenth century was clearly socially constructed. And the invented nature of that taxonomy needs to be captured accurately. The value of the curated content of Britannica to researchers today is that is the most comprehensive representation we have of that older knowledge system in its totality, and so it makes it possible to study that system as a structure and to observe how it changed over time.



The problems in tagging this biased dataset take three forms. First is the danger of historical anachronism. Applying a C21-century ontology, like Library of Congress Subject Areas, to C18 and C19 editions makes it accessible to modern researchers, but it also misrepresents the older system of knowledge. For example, the entries on

"History" from the important 3rd (1797) and 7th (1842) editions present authoritative accounts of human prehistory. While we might tag them under "anthropology," that field of knowledge was not recognized by the Royal Academy of Sciences until the 1880s (as a subset of Biology) and does not appear in the Encyclopedia itself until 1889. In fact, the older references cite the Book of Genesis as their authority, and a tag on applications of scripture to the interpretation of external reality might better represent the entry than an anachronistic "Anthropology, history of" tag could do.

The second difficulty is encountered when trying to reconstruct the older ontology used by the Encyclopedia, because it was a moving target. Subject categories changed over the first 150 years, with new categories added, others (like human prehistory) moving from one field to another, and still others disappearing. While we might construct a stable ontology for one edition, any historically-accurate ontology will have to become a system of multiple ontologies, whose relationships with one another need to be explained at the very least.

Third is the question of how to treat the built-in biases within the corpus. Older ontologies of knowledge are rife with bias, often through omission. Historically-accurate subject terms duplicate that bias. Information on attitudes toward women and national minorities, for example, exists within multiple entries, but there are no subject terms for minorities and no entries for women as such, making that data largely invisible without some form of intervention.

We are in the process of creating this new dataset and by summer of 2018 we will be completing preliminary tests on tagging systems, so the final paper will share preliminary results.

Distinctions between Conceptual Domains in the Bilingual Poetry of Pablo Picasso

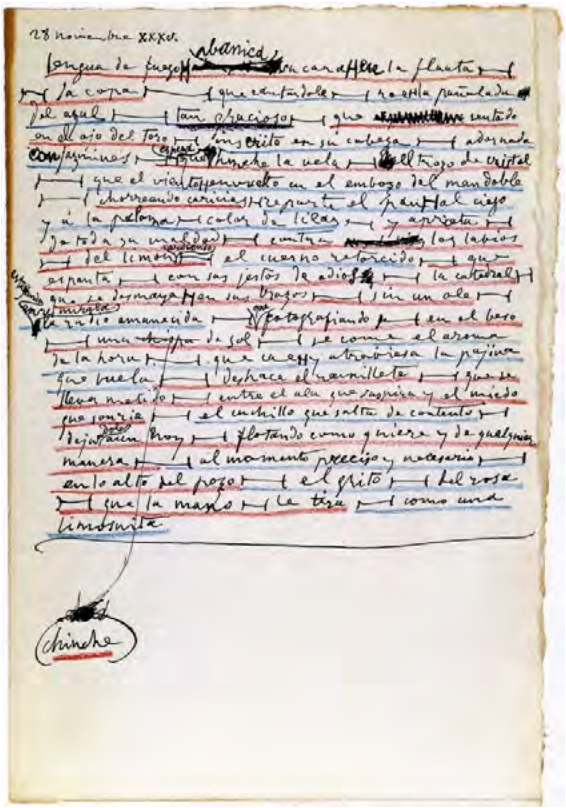
Enrique Mallen
mallen.shsu@gmail.com
Electronic Textual Cultures Lab, University of Victoria, Canada

Luis Meneses
ldmm@uvic.ca
Sam Houston State University, United States of America

Introduction

Picasso started writing poetry in April, 1935 during a period of personal crisis. Many have cited, among other possible causes the political turmoil in Europe in the period between the two wars. These views are predicated on an assumed irreducible conflict between visual composition and verbal expression. However, even before this he had been fascinated by linguistic structure and alternative methods of expression during his cubist period.

His poetry is not only fascinating as a form of communication from someone who is primarily known for his plastic output, it is also puzzling for anyone researching the interconnection between language and writing, i.e. verbal and graphic signs. His poetry is an attempt to expand the expressive power of language, as he adjoins words in unordered strings, following a technique very similar to cubist collage. Figures 1 and 2 show examples of this technique. The relation between words remains open and bidirectional, so that the reader is free to establish multiple semantic relations. And yet, while we see a close correlation between his poems and his artworks (Elizabeth, 2002) (Picasso and Baldassari, 2005), one cannot deny that his texts are primarily verbal; and this is precisely what makes them fascinating, as they provide a window into Picasso's mind that is separate from his own artistic creations –although they share with his artworks a predominance of ambiguity (Rubin et al., 1992) and the presence of unresolved conceptual oppositions.



Example of multiple additions and deletions in Picasso's poetry. P. Picasso, "lengua de fuego abanica ... (7)", Musée Picasso, Paris, 1935.

Example of the visual components in Picasso's poetry. P. Picasso, "si yo fuera afuera ... (2)", Claude Ruiz-Picasso Collection, 1935.

For some time now, our research has taken us through different approaches to analyze Picasso's artistic legacy (Meneses et al., 2008a), his poetry (Meneses et al., 2008b) and its semantic domains (Meneses and Mallen, 2017). In this paper, we propose to investigate how Picasso explored subtle differences between words within specific concepts in French and Spanish –as he composed his poems in two languages. This new perspective allows us to identify how the two languages offered Picasso a wide range of semantic domains to choose from when establishing subtle contrasts.

We have determined that, in Picasso's poems, certain semantic domains are predominant in each of the two languages. For instance, Picasso is more inclined to refer to food items and everyday objects in his Spanish poems. On the other hand, given the influence French Surrealist writers exerted on him, his French poems concentrate on more abstract concepts involving politics, religion and sexuality. Why did he choose to use these languages in the way he did? Daix (Daix and Emmet, 1993) has pointed out that "Picasso did not believe in spontaneous poetry – or painting". Our research will address the question of why did Picasso choose to write in a given language about a specific semantic domain.

Methodology

We have already classified the semantic interconnections between the concepts that Picasso explored (Meneses and Mallen, 2017). For this purpose, we used a taxonomy-based approach to identify the semantic domains in Picasso's poetry. We created a set of database tables that allowed us to specify concepts and then map them to their related terms in Picasso's poetry. It is important to note that these concepts are not bound to a given language per se: we were able to overcome the language barrier by linking concepts using the English translation of relevant terms –a language that Picasso didn't use in his poems.

We observed that some of the existing semantic categories are linked to a higher number of concepts than others. For example, we find a high number of nominal artifacts in his poems. Some are related to art, such as engraving, impression, ornament, paint and palette; others related to war, such as armor, axe, blade, bomb, bow, bullet, camouflage and rapier. These may appear antagonistic, but in Picasso's world there is a close relation between destruction in war and creation in art. Not surprisingly for a painter and writer, nominal communication is another frequent semantic category, with such concepts as advice, agreement, alphabet, fable, language, news, outcry and parable.

Given that we had a refined taxonomy, we decided to approach this problem from a purely computational perspective and expand on our previous efforts based on statistical models and algorithms (Meneses et al., 2016). More specifically, we propose to address our research questions by analyzing Picasso's semantic domains using Latent Dirchlet Allocation (Blei et al., 2003) and Term frequency-Inverse document frequency. We will do this by linking sets of words with their corresponding semantic concepts. Our analysis has shown that these techniques are capable of highlighting patterns and trends in Picasso's poetry that escape other forms of traditional analysis.

Conclusions

Our study is an attempt to further understand the semantic domains that Picasso operated with. Again, we know that Picasso's style, both in his visual and his verbal compositions, was very much inspired by collage. What makes them interesting is that those elements he placed together belonged to a restricted set, so that their interconnection, while not obvious to the viewer/reader, must have been somewhat determined in Picasso's "view" of reality. It is that determined interconnection which Picasso saw that we propose to explore with this study. In other words, we want to get closer to Picasso's "vision" of the world through his poems in order to investigate how that "vision" may differ from what he depicted in his graphic works.

To summarize, in this paper we propose to analyze why Picasso chose to write in a given language about concepts in a specific semantic domain. More so, throu-

gh the use of statistical models we propose to identify and pinpoint representative themes and correlations across different concepts and languages. Picasso once said: "Computers Are Useless. They Can Only Give You Answers". In this case, through our use of computational methods we are attempting to do just that: help us as researchers to get a better understanding of Picasso's poetry and artworks –and consequently, from the artist that created them as well.

References

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**: 993–1022.

Daix, P. and Emmet, O. (1993). *Picasso: Life and Art*. Thames and Hudson.

Elizabeth, C. (2002). Picasso: style and meaning.

Meneses, L., Estill, L. and Furuta, R. (2016). This was my speech, and I will speak it again": Topic Modeling in Shakespeare's Plays.

Meneses, L., Furuta, R. and Mallen, E. (2008a). Exploring the Biography and Artworks of Picasso with Interactive Calendars and Timelines. pp. 160–62.

Meneses, L. and Mallen, E. (2017). Semantic Domains in Picasso's Poetry Paper presented at the Digital Humanities 2017, Montreal, Canada.

Meneses, L., Monroy, C., Mallen, E. and Furuta, R. (2008b). Picasso's Poetry: The Case of a Bilingual Concordance. pp. 157–59.

Baldassari, A. (2005). *The Surrealist Picasso*. Flammarion.

Rubin, W., Varnedoe, K., Reff, T., Cottington, D., Fry, E. F., Poggi, C., Krauss, R. and Bois, Y. A. (1992). *Picasso and Braque: A Symposium*. Museum of Modern Art.

A formação de professores/ pesquisadores de História no contexto da Cibercultura: História Digital, Humanidades Digitais e as novas perspectivas de ensino no Brasil.

Patrícia Marcondes de Barros
patriciamarcondesdebarros@gmail.com
UNESPAR, Brazil

A presente comunicação (resultante de pesquisa em fase inicial) tem como objetivo geral, a análise e a reflexão, sob uma perspectiva metodológica qualitativa, acerca da formação de professores/pesquisadores de História no contexto da cibercultura e na sua esteira, da chamada História Digital (*Digital History*) e Humanidades Digitais (*Digital Humanities*). Os contributos metodológicos e práticos que as diversas tecnologias podem oferecer aos profissionais de História, as competências técnicas necessárias ao usufruto das mesmas e o entendimento das novas subjetividades erigidas, são de suma importância para a análise das mudanças paradigmáticas contemporâneas que abarcam o ensino, a pesquisa e, portanto, a formação docente neste devir.

Desde os anos 60 e 70 do século XX, observam-se mudanças culturais relacionadas aos meios comunicacionais, estudadas por grandes pesquisadores das mais diversas áreas de saber, a exemplo de Marshall McLuhan, filósofo e teórico da comunicação, que postulou a ideia de que a interdependência eletrônica recriaria o mundo numa aldeia global resultando no neotribalismo, erigindo assim, uma nova cultura.

Seus aforismos como "o meio é a mensagem", "os meios como extensões do homem" e "O homem cria a ferramenta. A ferramenta recria o homem", permanecem atuais na análise do mundo contemporâneo com suas múltiplas conexões, dotado da dimensão de universalidade (e assim sendo, "extenso, interconectado e interativo") e, portanto, menos totalizável (LÈVY, 1999, p.120) e de difícil apreensão.

Mcluhan aponta para uma sensibilidade na qual o meio traz consequências sociais e pessoais resultantes do estalão introduzido em nossas vidas por uma nova tecnologia, que é a extensão de nós mesmos. A máquina, por exemplo, independente do tipo de produção que faz, constitui a mensagem e transforma as relações. O autor pretende assim postular que o meio, geralmente pensado como um simples canal de passagem do conteúdo comunicativo, é um elemento determinante da própria comunicação: "o meio é a mensagem" (MCLUHAN, 2006).

A ideia postulada por Mcluhan de "aldeia global", de "ser planetário" relaciona-se aos movimentos de contracultura dos anos 60, que junto à instantaneidade dos meios comunicacionais eletrônicos, construiu uma subjetividade diferenciada (não-linear), denominada por muitos como pós-moderna e que foi a gênese da cibercultura que eclodiu em 1989 (BOLESINA; GERVASONI, 2015, p.08). Surge assim um novo mundo, relacionados à tecnociência e **às** tecnologias de Informação e de Comunicação (TICs).

É no universo educacional, o *locus* de grande visibilidade das mudanças sociais e culturais, tendo em vista a construção das identidades e apreensão da alteridade cultural através dos processos de aprendizagem e socialização. Com a pós-modernidade este universo passa por ressignificações e buscam metodologias que se integrem **às** novas tecnologias da informação, a interdisciplinariedade - entendida como os saberes comuns a uma ou mais matrizes de conhecimento-, e principalmente, a Antropologia, esfera privilegiada que aborda a cultura como dimensão fundadora da sociedade e permite o entendimento da alteridade, importante valor de reconhecimento das diversas culturas que permeiam o ambiente escolar.

A complexidade e a diversidade cultural observada reflete o espaço sem fronteiras, desterritorializado da cultura engendrada no ciberespaço, denominada como cibercultura.

A cibercultura representa um conjunto de técnicas, modos de pensamento e valores que se instituíram no ciberespaço (LÈVY,1999) que especifica não apenas a infraestrutura material da comunicação digital, mas também o universo de informação que ela abriga, assim como os seres humanos que navegam e alimentam esse universo. Pode ser entendido como a união de redes e recursos de comunicação formada pela interconexão global dos computadores pelo qual passou a ser possível o acesso à distância aos recursos de um computador, a exemplo da troca de arquivos digitais de forma simplificada, o envio de mensagens de forma sincrônica ou assincrônica, conferências eletrônicas em tempo real e transmissão de vídeo/som, entre horizonte de outras possibilidades. O conjunto dessas novas práticas, suportadas pelas tecnologias digitais e que foram apropriadas pela sociedade contemporânea transformaram os saberes e as práticas educacionais.

Vale ressaltar que tais transformações culturais se dão não somente com o aparato tecnológico, mas principalmente pelos tipos de signos que circulam nesses novos meios engendrando mensagens e processos de comunicação (SANTAELLA, 2003, p.24).

Com o advento da cultura digital e sua universalização, as interações sociais e a produção de conhecimento são amplamente transformadas através da virtualidade, reverberando na chamada História Digital e Humanidades Digitais.

Dentro do contexto educacional brasileiro esta nova configuração tecnológica não se enquadra ainda a realidade escolar e não apenas devido à questão estrutural de precarização de escolas e universidades públicas, como se observa atualmente. Há também resistência dos profissionais no campo das licenciaturas, especificamente nas das Ciências Humanas, como a História, em relação à inserção e discussão acerca de metodologias do ensino relacionadas ao novo processo comunicacional, o que nos coloca em situação de atraso frente a outros países.

Hansen (2015) assinala, segundo dados da centerNet, que existem 196 centros de pesquisa sobre Humanidades Digitais, sendo 88 na América do Norte, 75 na Europa e os 12 restantes pelo resto do mundo. A História Digital (*Digital History*) e o campo das Humanidades Digitais (*Digital Humanities*) são termos ainda recentes no léxico acadêmico brasileiro e não há consenso sobre seus significados.

A presente pesquisa analisará os processos comunicativos contemporâneos e suas interfaces com a Educação e a História (tanto no ensino, quanto na pesquisa), assim como as tecnologias, interações e convergência (História Digital e Humanidades Digitais) e a produção de linguagens e produção de sentidos no contexto da cibercultura.

Trata-se de forma geral, de repensar sob a **égide** das mudanças paradigmáticas postuladas pela cibercultura, os novos sentidos para a educação e a pesquisa que seja consonante **às** tecnologias, mas também humanista e aberta à complexidade e diversidade que os novos "meios e mensagens" nos trazem na contemporaneidade. Como afirma Hansen (2015), preparar futuros historiadores para o uso de outras mídias, que não as convencionalmente usadas, significa equipá-los com ferramentas que permitam explorar criativamente diferentes formas de apresentação do conhecimento histórico, e também avaliar criticamente produções e recursos disponíveis.

References

HANSEN, P.S. *Digital History* e formação de historiadores: sugestões para um debate.In BUENO,A.;ESTACHESKI, D.; CREMA (organizadores). *Tecendo amanhãs: O ensino de História na atualidade*. Rio de Janeiro/União da Vitória: edição especial *Sobre Ontens*, 2015.

LEVY, P. *O que é o virtual*. São Paulo: Ed. 34, 1996.

_____. *Cibercultura*. São Paulo: Ed. 34, 1999.

MCLUHAN, Marshall. Os meios de comunicação como extensão do homem (understanding media:The Extensions of Man). Editora Cultrix, São Paulo, 2006.

SANTAELLA, L. *Culturas e artes do pós-humano: da cultura das mídias à cibercultura*. São Paulo: Paulos, 2003.

Presentation Of Web Site On The Banking And Financial History Of Spain And Latin America

Carlos Marichal
cmari@colmex.mx
El Colegio de Mexico, Mexico

The purpose of this ten-minute presentation is to present this thematic site which we have constructed in 2016 (and is bilingual) and is of use for professors, students, and general public. The object this academic web page (the first of its kind in this specific field) is to provide a large amount of documents (including links to over 200 working papers) historical statistics (over 600 Excel charts and graphs), bibliographies, guides to archival sources and short historical summaries of the banking histories of many countries in Latin America as well as Spain. Both El Colegio de México and the University of Cantabria (Spain) have collaborated in this project under the direction of Dr. Carlos Marichal. The site will soon be transferred to the electronic resources of the Libraries of both academic institutions.

I argue that this thematic webpage corresponds to an increasing trend in contemporary academics to *combine*

concrete and deep research results in *subdisciplines* with complementary resources of a varied nature, including historical statistical series, reference texts, images (photos and engravings), timelines, and resources for teaching.

Such resources are especially useful for consultation on line by professors and students in local universities, many of which – in Mexico and Latina America- do not have really rich library/digital resources In addition I might remark that there is a demand from schools and universities for advanced online courses in humanities and social sciences that can be especially useful for updating university professors, especially in the provinces, where there is urgent need for support to achieve a substantial improvement in teaching and research in humanities and social sciences in Mexico or other countries in the region.

To accomplish this, a multidisciplinary working group of academics has been set up to gather pertinent information from the various humanities and social sciences in the field of banking and financial history of Latin America and Spain, which explains the international consortium engaged. The interest of the project lies in the pioneering projects in this field in the humanities and social sciences both in academia in Mexico and elsewhere. The site can be consulted at http://codexvirtual.com/hbancaria/

Spatial Disaggregation of Historical Census Data Leveraging Multiple Sources of Ancillary Data

João Miguel Monteiro
joao.miguel.monteiro@tecnico.ulisboa.pt
University of Lisbon, IST and INESC-ID, Portugal

Bruno Emanuel Martins
bruno.g.martins@ist.utl.pt
University of Lisbon, IST and INESC-ID, Portugal

Patricia Murrieta-Flores
p.murrietaflop.a.murrieta-flores@lancaster.ac.uk
University of Lancaster, United Kingdom

João Moura Pires
jmp@fct.unl.pt
Universidade NOVA de Lisboa, FCT / NOVA LINCS, Portugal

Accurate information about the human population distribution is essential for formulating informed hypothesis in the context of several social, economic, and environmental issues. Government instigated national censuses are authoritative sources of population data, subdividing space into discrete areas (e.g., fixed administrative units) and providing multiple snapshots of society at regular intervals, typically every 10 years. Many research institutions or national statistical offices have developed historical Geographical Information Systems (GIS), containing statistical data from previous censuses together with the administrative boundaries (i.e., records of administrative boundary changes) used to publish them over long periods of time. However, using these data can still be quite challenging, particularly when looking at changes over time.

There are multiple reasons why population data aggregated to administrative units is not an ideal form of information about population counts and/or density. First, these representations suffer from the modifiable areal unit problem (Lloyd, 2014), which states that the results of an analysis that is based on data aggregated by administrative units may depend on the shape and arrangement of the units, rather than capturing the theoretically continuous variation in the underlying population. Second, the spatial detail of aggregated data is variable and usually low, particularly in the context of historical data. In a highly aggregated form these data are useful for broad-scale assessments, but using aggregated data has the danger of masking important local hotspots, and overall tends to smooth out spatial variations. Third, there is often a spatial mismatch between census areal units and the user-desired units required for particular types of analysis. Finally, the boundaries of census aggregation units may change over time from one census to another, making the analysis of population change, in the context of longitudinal studies dealing with high spatial resolutions, difficult.

Given the aforementioned limitations, high-resolution population grids (i.e., geographically referenced lattices of square cells, with each cell carrying a population count or the value of population density at its location) are often used as an alternative format to deliver population data. All cells in a population grid have the same size and the cells are stable in time. There is no spatial mismatch problem as any partition of a given study area can be rasterized to be co- registered with a population grid.

Population grids can be built from census data through the application of spatial disaggregation methods (Monteiro et al., 2014), which range in complexity from simple mass- preserving areal weighting, to intelligent dasymetric weighting schemes that leverage regression analysis to combine multiple sources of ancillary data.

Nowadays, there are for instance many well-known gridded datasets that describe the modern population distribution, created using a variety of disaggregation techniques (e.g., the Gridded Population of the World (Doxsey-Whitfield et al., 2015) or the WorldPop databases (Tatem, 2017)). However, despite the rapid progress in terms of disaggregation techniques, population grids have not been widely adopted in the context of historical data. We argue that the availability of high-resolution population grids within historical GIS has the potential to improve the analysis of long-term geographical population changes. Perhaps more importantly, this can also facilitate the combination of population data with other

GIS layers to perform analyses on a wide range of topics, such as the development of the transport network, historical epidemiology, the formation of urban agglomerations, or climate changes.

This work reports on experiments with a hybrid disaggregation technique that combines the ideas of dasymetric mapping and pycnophylactic interpolation (Monteiro et al., 2014), using machine learning methods (e.g., linear regression models, ensembles of decision trees, or deep learning approaches based on convolutional neural networks, which previously have only seldom been used for spatial disaggregation (Robinson et al., 2017)) to combine different types of ancillary data (e.g., historical land-coverage data from the HILDA project (Fuchs et al., 2015), together with modern information that we argue can correlate with historical population), in order to disaggregate historical census data into a 200 meter resolution grid. Apart from few exceptions related to the use of areal interpolation for integrating historical census data, most previous related studies have focused on modern datasets.

We specifically report on experiments related to the disaggregation of historical population counts from three different national censuses which took place around 1900, respectively in Great Britain, Belgium, and the Netherlands. All three statistical datasets, together with the corresponding boundaries for the regions at which the data were collected (i.e., parishes or municipalities), are presently available in digital formats within national historical GIS projects. The obtained results indicate that the proposed method is indeed accurate, outperforming simpler schemes based on mass-preserving areal weighting or pycnophylactic interpolation. Moreover, the obtained results also show that modern data, particularly pre-existing gridded datasets that describe the modern population distribution (i.e., data from the Gridded Population of the World (Doxsey-Whitfield et al., 2015) project), are particularly useful as features for supporting the disaggregation of historical population counts. The best results were obtained with regression models leveraging multiple features (i.e., different models attained the best results in each of the three national territories that were considered), although a simple dasymetric technlque, leveraging the modern population gridded data to define the disaggregation weights, achieved very competitive results.

Acknowledgements

This research was partially supported by the Trans-Atlantic Platform for the Social Sciences and Humanities, through the Digging into Data project with reference HJ-253525. The researchers from INESC-ID also had financial support from Fundação para a Ciência e Tecnologia (FCT), through the project grants with references PTDC/EEI-SCR/1743/2014 (Saturn) and CMUPERI/TIC/0046/2014 (GoLocal), as well as through the INESC-ID multi-annual funding from the PIDDAC program, which has the reference UID/CEC/50021/2013.

References

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O. and Baptista, S. R. (2015). Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1(3).

Fuchs, R., Herold, M., Verburg, P. H., Clevers, J. G. and Eberle, J. (2015). Gross changes in reconstructions of historic land cover/use for Europe between 1900 and 2010. *Global change biology*, 21(1).

Lloyd, C. D. (2014). *The Modifiable Areal Unit Problem. Exploring Spatial Scale in Geography*. John Wiley & Sons.

Monteiro, J., Martins, B. and Pires, J. M. (2017). A hybrid approach for the spatial disaggregation of socio-economic indicators. *International Journal of Data Science and Analytic*s, 5(2-3), pp 189–211.

Robinson, C., Hohman, F., and Dilkina, B. (2017). *A Deep Learning Approach for Population Estimation from Satellite Imagery. Proceedings of the ACM SIGSPATIAL Workshop on Geospatial Humanities*. New York: ACM Press.

Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, 4, 170004.

The Poetry Of The Lancashire Cotton Famine (1861-65): Tracing Poetic Responses To Economic Disaster

Ruth Mather
r.m.mather@exeter.ac.uk
University of Exeter, United Kingdom

Our project will make freely available a searchable webapp, built in eXist-db (http://exist-db.org/exist/apps/homepage/index.html), containing a database of poems responding to Lancashire Cotton Famine of 1861-65, along with audio recitations and musical performances drawing directly on these poems (Rennie, 2017). This poetic response is important in that it often represents labouring-class voices from the mid-nineteenth century, which, in spite of renewed academic interest in such material, remain underappreciated (Goodridge et al, 2012 provides a useful introduction and bibliography). The study of this material and its digital publication will significantly enrich literary scholarship and historical perspectives of this economic crisis, and provides the opportunity to draw public attention to an episode of history that is little known beyond the scholarly sphere. The project seeks to establish a much more detailed understanding of

the nature of Lancashire Cotton Famine poetry: its extent, its intents, and its functions. To date, there is no critical literature specifically addressing the poetry written and published during the Cotton Famine, though the period is touched upon in Brian Hollingworth's anthology *Songs of the People: Lancashire dialect poetry of the industrial revolution* (1972: 98-113).

The project draws predominantly on the local newspaper collections of Lancashire's various civic archives and local studies centres for material. Though some of these newspapers have been digitised in collections such as the British Newspaper Archive and Gale Historical Newspapers, the majority are in hard-copy or microfilm format, so that users attempting to access the poetry encounter practical obstacles relating to both geography and the preservation of materials. Additionally, as the archives and local studies collections in Lancashire are significantly under-resourced, there are long-term concerns about maintenance of the equipment and staffing levels required to ensure that access to these significant historical collections is sustainable. The recovery element of our project therefore aims to ensure long-term, free-of-charge access to a near-complete repository of Cotton Famine Poetry without the requirement to visit multiple archives or local studies centres. As the recovered poems have been transcribed by hand, we are also avoiding replicating the Optimal Character Recognition errors which have been incurred by some of the existing newspaper databases (Joulain-Jay, 2016).

Alongside the vital recovery and collation of this material, the experience of the investigators in the field of labouring-class poetry enables a simultaneous critical analysis of the poetry as it emerges, focussing on local, regional, national, and international fields of interest. We encounter a wide range of poetic styles, written both in Lancashire dialect and standard English, which demonstrate the sophisticated literary engagements of their authors. In terms of subject matter, the poems describe not only the direct, local experience of the Cotton Famine, but also offer more abstract reflections on issues including work, poverty, war, slavery and abolition. We want to determine the extent to which political dissent is present in the poetry, and to what degree opposing discourses relating to slavery and the American Civil War were articulated through literature of this type. We are already beginning to establish that a significant proportion of Cotton Famine poetry represents a labouring-class address to a regional and national middle-class readership, and part of our analysis will involve mapping a transatlantic discourse between the Lancastrian labouring classes and writers on both sides of the American conflict. The popular narrative of the Cotton Famine has Lancashire textile workers staunchly supporting the North in spite of the deprivation caused by the war, because of their strong support for abolition. Early analysis of poetic responses suggests a more complicated engagement with the American conflict, including some elements of support for the south (see also Ellison, 1972). We hope that by making the poetry freely available online, we encourage its further use as an important historical and literary resource for understanding some of these complex themes.

The digitisation of a large and varied body of work such as this presents both challenges and possibilities, and this paper will reflect upon the difference that digital methods make to our interpretation of the material and the ways in which it can be used. The process of marking up text for the database enables us to make our editorial decisions in presenting this body of work transparent, and to group similar poems and themes for the reader's ease of analysis. Nonetheless, in so doing, we impose our own interpretations upon what was a fluid form - often orally transmitted, and published in different versions across different media. A key concern in presenting this material has been the desire to ensure its usefulness for scholars who might have different approaches and questions to our own. In forcing us to grapple with these challenges, the use of digital methods has encouraged us to make explicit our own methodologies and thought processes, and enabled the creation of a resource that could be considered more intellectually 'open' than the traditional analogue anthology.

The design of our webapp therefore reflects our desire for a flexibility which in turn offers better representation of a literature which was originally available in multiple, sometimes changing forms. Some poems were written to be read or sung aloud, while others endeavoured to capture in writing the transient forms of local dialect, and different variations of the same poem appear across the newspapers. The use of XML enables us to continuously add layers of interpretation as they occur in the data for macro analysis, while marking up at word-by-word level enables a careful close reading in which we are forced to be conscious of the decisions we are making about the presentation of material.

An important part of the project is its public-facing element, including the involvement of school groups in finding and transcribing the poetry. We also welcome submissions of potential Cotton Famine poetry from members of the public, local historical societies, and educational projects with an interest in this material. Managing the upload and editing of these submissions, and ensuring that appropriate credit is given, is one of the tasks that the project team has taken on, and it is likely to present its own set of challenges: gathering this data provides an opportunity to involve the public in undertaking research and giving them insights into this process, but we also need to ensure that the results are useful and worthwhile for the project's outputs. At present we are not planning to train people beyond the team in how to encode in TEI, but giving contributors a 'behind the scenes' tour of the database and offering an introduction to how we create and structure our digital materials (and why)

has the potential to enable further discussions and may encourage contributors to get involved with digital humanities activity beyond our immediate project. We feel that this is an important step in ensuring that contributors gain an understanding of what happens to their data once they submit it, and how it is transformed into what they see in the final digital publication. We will discuss these challenges and how we intend to maintain interest and engagement amongst our contributors.

At this stage of the project, in which we are making key decisions about how to manage our own data and that from our external contributors, we would welcome discussion and comments from the wider digital humanities community on how we can ensure that our resource is an effective tool for both research and teaching. We hope, too, that the challenges that we are encountering and some of our proposed solutions might prove helpful for others working with comparable datasets or audiences.

References

Ellison, M. (1972) *Support for Secession: Lancashire and the American Civil War.* Chicago & London: University of Chicago Press.

http://exist-db.org/exist/apps/homepage/index.html

Goodridge, J. et al (2012) 'Introduction', *Labouring-Class Poets Online.* https://lcpoets.wordpress.com/introtobibliography/

Hollingworth, B. (1972) *Songs of the People: Lancashire dialect poetry of the industrial revolution*. Manchester: Manchester University Press.

Joulain-Jay, A. (2016) 'Dealing with Optimal Character Recognition errors in Victorian Newspapers', *British Library Digital Scholarship Blog*, July 20th 2016. http://blogs.bl.uk/digital-scholarship/2016/07/dealing-with-optical-character-recognition-errors-in-victorian-newspapers.html

Rennie, S. (2017) *The Poetry of the Lancashire Cotton Famine (1861-65)* http://cottonfaminepoetry.exeter.ac.uk

READ Workbench – Corpus Collaboration and TextBase Avatars

Ian McCrabb
ian@prakas.org
University Of Sydney, Australia

The Research Environment for Ancient Documents (READ) project commenced in 2013 with development support from a consortium of institutions (University of Washington in Seattle, Ludwig Maximillian University in Munich, University of Lausanne, University of Sydney and Prakaś Foundation) involved in the study and publication of ancient Buddhist documents preserved in Gāndhārī.

READ has been developed as a comprehensive multi-user platform for the transcription, translation and analysis of ancient Sanskrit, Gāndhārī, Pali and other Prakrit texts: manuscripts, inscriptions, coins and other documents. It is based on open source software (Postgres, PHP and JQuery), supports the TEI standard and provides an API for integration with related systems. READ is positioned as a research environment, complementary to existing textual repositories and integrated with existing dictionaries. Existing transcriptions can be imported, elaborated upon, analyzed, and then published as research output in standards-based formats. The defining innovation of READ is atomization to a semantically linked network of objects; a paradigm shift in data structure from strings of marked-up text to sequences of linked objects.

The underlying design and entity model was presented at both the Digital Humanities conference in 2015 and the 2016 Australian Digital Humanities Conference. READ has been publicly released and is supporting a wide range of corpus development projects. Whilst this presentation will follow on to briefly precis the range of research projects currently supported by READ, the focus will be on a related platform. READ Workbench (Workbench) is a server portal hosted at the University of Sydney since 2016 to 'harness' READ. Developed using the same technology stack as READ, it is a comprehensive management framework to support the integration of people and processes in the collaborative development, maintenance and publishing of textual corpora.

The design of Workbench evolved organically as the implementation requirements of READ expanded from a single researcher working on a single text, to the capacity to support multiple projects, each with a team of researchers collaborating on the development of an integrated corpora, with widely divergent research objectives. The fundamental objective was to design a supporting framework with which manage the balance between autonomy and collaboration in large scale projects. The approach adopted was to implement strategies, models and workflow patterns consistent with those applied in the IT consulting industry to digital content design, development and migration projects.

Workbench is a software as a service (SaaS) platform managing multiple READ installations, each with project and language specific configurations, supporting researcher collaborations across multiple institutions. It provides a self-service portal for researchers to develop, maintain, manage and publish texts without requiring technical support or the mediation of a database administrator; critical to the longer-term sustainability of corpora projects. Workbench's three facets (configuration management, database management and corpus workflows) provide a scalable implementation architecture for READ and instantiate a comprehensive corpus development methodology.

Whilst the configuration management services might be bracketed as conventional for a SaaS platform, database management is predicated upon an architectural innovation that enables researchers to build, share, manage, maintain and publish their own texts. The adoption of a single text/single database (TextBase) as the fundamental object of development, collaboration and portability is quite a departure from conventional models where a centralized administrator manages a single monolithic corpus database.

This TextBase architecture underpins a corpus development, analysis and publishing methodology that provides significant flexibility in terms of the iteration and synchronization of three fundamental workflows: text alignment, analysis registration and text aggregation.

The text alignment process integrates image, text and model configuration data to automatically generate a database. This approach allows for the distribution of responsibility to specialists and integration of their research output to align the image and the transcription at their most atomized to generate a 'substrate'. Rather than requiring researchers to command exacting markup schemas, substrate databases can be automatically generated from Word processing and Spreadsheet inputs. Workbench enables each of the specialist roles to work independently and their contributions be separately managed and integrated, ameliorating project risk by minimizing dependencies and bottlenecks.

The analysis registration process synchronizes analysis data with an existing substrate. This approach allows a researcher to work independently and externally to READ in developing their own analysis 'strata' and then register that strata on a substrate. Grammatical analysis, translation, semantic, syntactic and structural analysis can all be independently developed and iteratively registered. Researchers from other disciplines can develop and register their own analysis (archaeological, historical etc.). Each stratum is registered on a particular substrate (an edition) of the text within a TextBase, is separately owned and attributed, and its visibility is controlled by the researcher registering it.

The text aggregation process allows individual researchers to work and innovate in private to the point where they elect to participate in research collaborations or their text is ready for publishing. A TextBase might be aggregated with others to form a 'sequenced' collection, a 'mapped' collection or a 'merged' corpus; a continuum expressing an increasing degree of synthesis and harmonization of analysis ontologies and methodological standards. Researchers may contribute their TextBase to any number of aggregates. This approach allows a researcher to align a TextBase with the analysis standards of an established corpus as a predicate to participation as a constituent of that merged aggregate. In parallel, that same TextBase might be mapped to the analysis ontology of an entirely different collection. The potential exists for the same TextBase substrate to manifest as a constituent of separate aggregates, with alternative configurations of registered analysis strata, supporting widely divergent (aggregate specific) research objectives; the emanation of multiple TextBase avatars.

The strategy adopted with Workbench was to design a solution architecture within which to reframe some of the ubiquitous issues in the conventional corpus development model; ownership, control, confidentiality, innovation, standardization, portability, resourcing and support. The critical innovation in maximizing development flexibility and in balancing autonomy and collaboration across the range of individual, collection and corpora development projects is the TextBase; the target of text alignment, the substrate for registration of analysis and the object aggregated.

Preserving and Visualizing Queer Representation in Video Games

Cody Jay Mejeur
cmejeur@gmail.com
Michigan State University, United States of America

The nascent field of queer game studies has expanded exponentially in recent years thanks to the work of scholars such as Adrienne Shaw, Bonnie Ruberg, and Edmond Chang. This growth in scholarship has paralleled a significant rise in LGBTQ representation in games, including games such as *Gone Home*, *The Vanishing of Ethan Carter*, *Dream Daddy*, and others. Yet, despite growing representation and scholarly attention to queer characters and players, queer game studies continues to face the multivalent marginalizations of queer folks and their experiences in gaming. A prime example of this marginalization is the difficulty of preserving queer gaming cultures: queer representations and gaming communities are recorded largely in ephemeral, unofficial digital forms such as wikis, blogs, and fan-made websites due to a lack of access to mainstream platforms that often minimize and reject queer perspectives and desires. There is some advantage to these forms in that they allow queer gamers to create online communities as "counterpublics," which are communities defined against normative rules and expectations, but this means that queer gaming cultures are also in constant danger of being ignored, becoming outdated, or disappearing suddenly due to lack of resources (Warner, 2002).

A case in point is GayGamer.net, a website dedicated to game news, commentary, and community for LGBTQ gamers that went dark without notice in May 2016. GayGamer.net was a valuable resource for documenting LGBTQ game characters and communities, and while parts of it were captured by the Internet Archive, much of the site is no longer accessible outside of an old Facebook

page (GayGamer.net). While many digital objects face similar issues of compatibility and archiving, queer game artifacts and documentation are especially endangered because of the marginalized status of queer gamers and characters in gaming culture. With fewer individuals (almost all volunteers) and institutional resources to support them, these sources must be actively preserved now before they—and crucial LGBTQ cultural heritage with them—are lost.

The LGBTQ Video Game Archive, founded by Adrienne Shaw at Temple University, was created to address these issues by collecting LGBTQ representations from the 1980s to the present in order to "offer a record of how characters are explicitly coded, what creators have said about these characters, as well as how fans have interpreted these characters" (Shaw). The archive aims to allow easy, comprehensive access to queer gaming sources for queer game scholars, queer gamers, and the interested public, and further to ensure that these sources remain available in the future. To this end, one of the archive's current projects is an ongoing preservation effort in association with the Strong National Museum of Play that seeks to save copies of the many online media artifacts that document queer gaming cultures. The archive's preservation project demonstrates how archiving can be used to further social justice projects in digital spaces by safeguarding the cultural productions (including personal blogs, community forums, wikis, etc.) of marginalized peoples. As part of this presentation, I will share the process I developed for collecting and storing the websites, images, and videos referenced in the archive, and then transferring these materials to the Strong for permanent storage and public access. Using a combination of browser plugins and command line tools such as wkhtmltopdf and youtube-dl, the sources are saved as common file types that are entered into an Omeka database. The database allows the Strong to make the files publicly accessible, and the common file types allow for easier maintenance and curation of the collection. This process could be of use to other scholars and activists working to collect, curate, and sustain digital cultural resources, especially those significant to marginalized communities.

By preserving this cultural heritage, the LGBTQ Video Game Archive allows for new analyses of queer gaming culture and representation that highlight ongoing issues and emerging possibilities in games. For example, Utsch et al. used the archive to create data visualizations of queer representation throughout video game history, and revealed several trends such as a predominance of gay men in LGBTQ representation and an exponential growth in overall number of representations (Utsch et al., 2017: 7). To date, however, an intersectional analysis of the archive that addresses sexuality alongside identity categories of race, class, or disability has not been attempted, and this paper presentation will address these intersections using new interactive data visualizations. Completing these visualizations required revisiting each representation in the archive and recording additional data about the character's identity. The visualizations are interactive in order to make them more fluid and dynamic: in other words, to make them better representations of identity than the static categorizations that intersectionality has sometimes been accused of (Puar, 2005: 125). This intersectional analysis of the archive is only the beginning of the archive's potential, and it has a number of limitations. For example, it only includes games currently in the archive, and only what is observable and documented about each representation. Future work will add more games to the analysis, and provide more granular analysis of particular genres, developers, and intersectional identities in games. Together, preservation and critical analysis are essential tools for developing archival practices that support social justice in digital humanities, and both are much needed forms of public, academic, community-oriented activism.

In sharing the LGBTQ Video Game Archive's ongoing efforts to preserve and visualize queer representation in games, this paper presentation calls for increased attention in digital humanities to the needs of marginalized groups such as queer gaming communities. Concepts and design practices such as imagining a QueerOS can help guide the field's attempts at better inclusion, but we as digital humanities scholars can and must do more (Barnett et al., 2016). As we build and make with our digital tools, we must constantly confront the question of who we are building and making for. I argue that digital humanities should be the digital theories and practices of social justice, and it should do the crucial work of engaging with communities and supporting their efforts to make and shape themselves. Representation in queer games and queer gaming communities provides some practical methods for doing so, and contributes to ongoing discourse of what digital humanities can be.

References

Barnett, F., Blas, Z., Cárdenas, M., Gaboury, J., Johnson, J. M. and Rhee, M. (2016). QueerOS: A User's Manual. In Gold, M. K. and Klein, L. F. (eds), *Debates in the Digital Humanities: 2016*. University of Minnesota Press.

Chang, E. (2017). Queergaming. In Shaw, A. and Ruberg, B. (eds), *Queer Game Studies*. University of Minnesota Press, pp. 15–24.

Condis, M. (2015). No homosexuals in Star Wars? BioWare, 'gamer' identity, and the politics of privilege in a convergence culture. *Convergence*, **21**(2): 198–212 doi:10.1177/1354856514527205.

GayGamer.net. Social Media. *Facebook*. https://www.facebook.com/gaygamernet/?ref=br_rs. (accessed 28 Nov. 2017).

Gold, M. K. and Klein, L. F. (eds). (2016). *Debates in the Digital Humanities: 2016*. Minneapolis London: University of Minnesota Press.

Malkowski, J. and Russworm, T. M. (eds). (2017). *Gaming Representation: Race, Gender, and Sexuality in Video Games*. (Digital Game Studies). Bloomington: Indiana University Press.
Puar, J. K. (2005). Queer Times, Queer Assemblages. *Social Text*, **23**(3-4-85): 121–39 doi:10.1215/01642472-23-3-4_84-85-121.
Shaw, A. (2014). *Gaming at the Edge: Sexuality and Gender at the Margins of Gamer Culture*. Minneapolis: University of Minnesota Press.
Shaw, A. *LGBTQ Video Game Archive*. https://lgbtqgamearchive.com. (accessed 28 November 2017.
Shaw, A. and Friesem, E. (2016). Where is the queerness in games?: Types of lesbian, gay, bisexual, transgender, and queer content in digital games. *International Journal of Communication*, **10**: 13.
Utsch, S., Braganca, L. C., Ramos, P., Caldeira, P. and Tenorio, J. Queer Identities in Video Games: Data Visualization for a Quantitative Analysis of Representation.
Warner, M. (2002). *Publics and Counterpublics*. New York: Zone Books.
Warner, M. and Social Text Collective (eds). (1993). *Fear of a Queer Planet: Queer Politics and Social Theory*. (Cultural Politics v. 6). Minneapolis: University of Minnesota Press.

Segmentación, modelado y visualización de fuentes históricas para el estudio del perdón en el Nuevo Reino de Granada del siglo XVIII

Jairo Antonio Melo Flórez
jairom@colmich.edu.mx
El Colegio de Michoacán, Mexico

Introducción

Una de las características de la cultura jurídica del antiguo régimen era la evidente polisemia de sus conceptos (Hespanha, 2002). Términos que actualmente pertenecen al plano teológico, moral, histórico y literario; tenían un valor normativo dentro del lenguaje jurídico que afectaba la práctica del gobierno y la justicia. Como explica Alejandro Agüero, la perspectiva crítica develó una lógica del orden natural del mundo y del poder político en la que el derecho escrito si bien no es irrelevante tampoco cumple el papel protagónico en la organización de las ciudades y los reinos (Agüero Nazar, 2012: 84).

El giro metodológico, de la exégesis a la auscultación del pensamiento jurídico a través de los conceptos centrales del discurso normativo desde la edad media europea, ha implicado cruzar las fronteras de la historia del derecho como campo especializado del estudio del derecho para entrar a discutir con la disciplina histórica. La hermenéutica jurídica trasciende por lo tanto el ejercicio de la interpretación del texto para remitirse al "sentido" (*sinn*) del derecho en un enfoque cercano a la filosofía hermenéutica (Costa, 1972: 46), cuya derivación más conocida por los historiadores la representan Gadamer y Koselleck (1997).

Con este trabajo pretendemos explorar las posibilidades que brindan los métodos de análisis computacional para la historia de los conceptos y en general del lenguaje jurídico-político anterior al siglo XIX, en el entendido que el perdón permite analizar un elemento fundamental del poder político de la Edad Moderna (Foucault, 1975: 56–57), así como el proceso de secularización y de pretendida tecnificación del indulto en el marco del proyecto de modernización legislativa decimonónico (Prodi, 2000).

La construcción del corpus: modelado básico de la información.

El corpus textual objeto de esta muestra está compuesto de documentos seleccionados de fondos de justicia y gobierno de archivos españoles y colombianos. Al no existir una serie documental consistente para el problema del perdón, fue necesario realizar una exploración y recolección de documentación en diversos repositorios que permitiera representar el universo de la clemencia en el ámbito del virreinato del Nuevo Reino de Granada.

Con el propósito de estructurar la información, se aprovechó el entorno Omeka para facilitar tanto la transcripción como la asignación de metadatos al contenido y a los elementos (Melo Flórez, 2016). Se identificaron distintos tipos documentales que se agruparon en cédulas, peticiones y concesiones de indulto, legislación, doctrina, prisiones, perdones particulares, expedientes judiciales y biografías. Para estos dos últimos elementos se construyó un tipo de elemento (ítem-type) con lo cual se puede recuperar información específica como suplicaciones, vistas fiscales, sentencias, alegatos de defensores, testimonios.

La transcripción de los documentos se realizó de manera tradicional intentando conservar el valor fonético o literal de las fuentes, por lo cual el texto digitalizado no fue modernizado en su ortografía ni en la acentuación. Un problema consistió en el desarrollo de abreviaturas, las cuales por lo general se indican haciendo uso de corchetes, por ejemplo, N^{vo} R^{no} de Granada se desarrolla como N[ue]vo R[ei]no de Granada. Por el momento se ha optado por imitar la etiqueta <expan> del modelo TEI del modo [expan = Nuevo Reino de Granada] con lo cual se adelanta la identificación de algunos elementos semánticos y por otra parte solventa la lectura automática del texto. La misma operación se realiza con etiquetas como <abbr>,

<gloss>, <note>, <corr>, <sic>, <placeName>, <geo>, <textLang mainLang> y <name type>.

Finalmente, la información se recuperó mediante la función *metadata* de Omeka que permite seleccionar entre diferentes tipos de metadatos para luego exportarlos en HTML y convertirlos a texto plano (Turkel and Crymble, 2012). Con el propósito de visualizar el cambio de significado el corpus se segmentó en seis grupos temporales: 1739-1775, 1776-1789, 1790-1807, 1808-1818, 1819-1829, 1830-1842.

Segmentación

Antes del análisis textual, el texto requiere ser *tokenizado*, es decir, segmentado en unidades lingüísticas con la intención de conocer las métricas de las fuentes (Mikheev, 2005). Este proceso tiene el propósito de agrupar caracteres alfanuméricos en palabras, diferenciar tipos (número de palabras diferentes en un corpus), la frecuencia de cada palabra representada como *tokens*, y aplicar el proceso de *stemming* (reducir las palabras a su raíz) y la lematización (formas flexionadas de una palabra). Por lo tanto, en esta etapa, el análisis se reduce a la estructura básica del texto, su construcción y la medición del peso de sus elementos (Jockers, 2013: 4). El resultado se presenta en la tabla 1, aunque el primer segmento (1700-1775) revela la disparidad temporal respecto a las demás divisiones, por lo cual se comprende deberá corregir esta discrepancia en un futuro ejercicio. Los periodos con mayor cantidad de tokens están representados por aquellas etapas más convulsas del periodo: la rebelión de los comunes de 1781 y el proceso de revolución e independencia desde 1808 hasta 1830.

Corpus	Tokens	Types	Lemmas
1700-1775	112136	15514	15514
1808-1818	93301	12198	12538
1776-1790	80015	11438	11714
1819-1830	51548	8510	8605
1830-1842	32968	6578	6736
1790-1808	25885	5932	6043

Tabla 1. Resultados del proceso de segmentación del corpus por segmentos

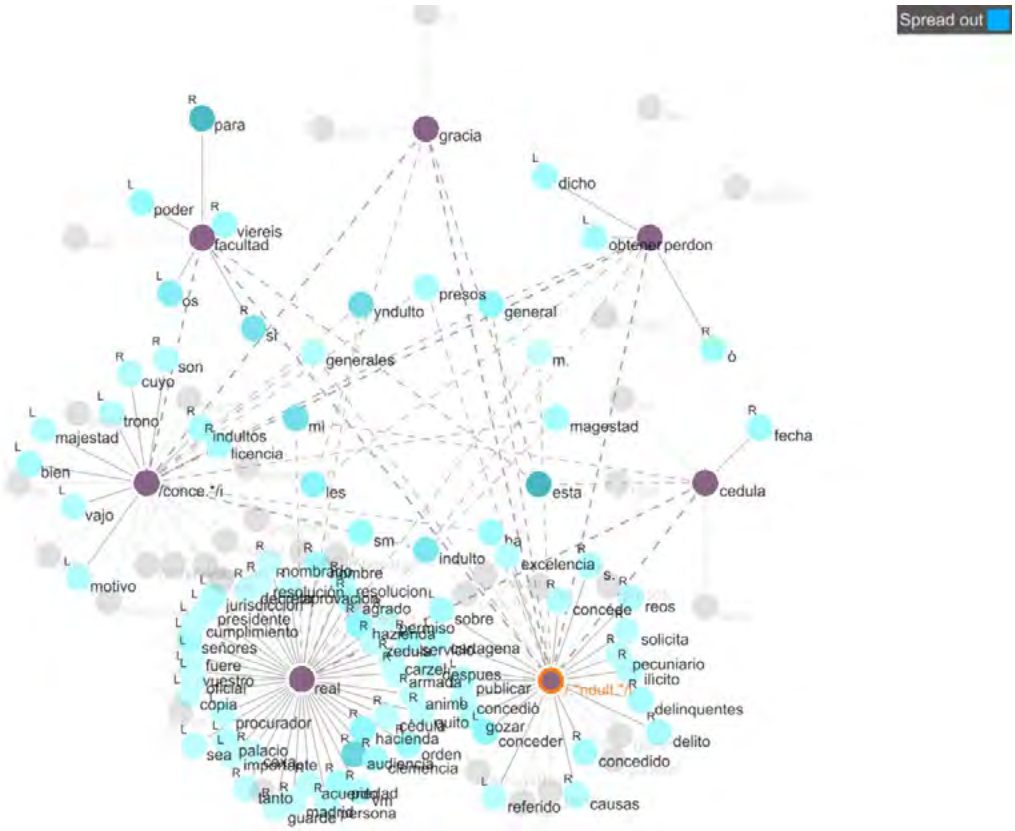
La abundancia de tipos y lemas se deriva de la cantidad de variaciones que el software no tiene la capacidad de deducir formas de una misma palabra, por ejemplo, indulto e yndulto es leído como dos vocablos separadas. La manera más simple de solucionar este inconveniente consiste en modernizar la ortografía de las expresiones arcaicas, sin embargo, esto disociaría el corpus de las fuentes doctrinales y legales impresas, cuya información se recupera por técnicas de OCR. En este caso nos remitimos nuevamente a la representación de grafemas, tarea propia de la paleografía, y su uso por parte de escribanos en la Edad Moderna, así como las posibilidades de semi-automatización y estandarización de esta tarea.

Análisis y visualización

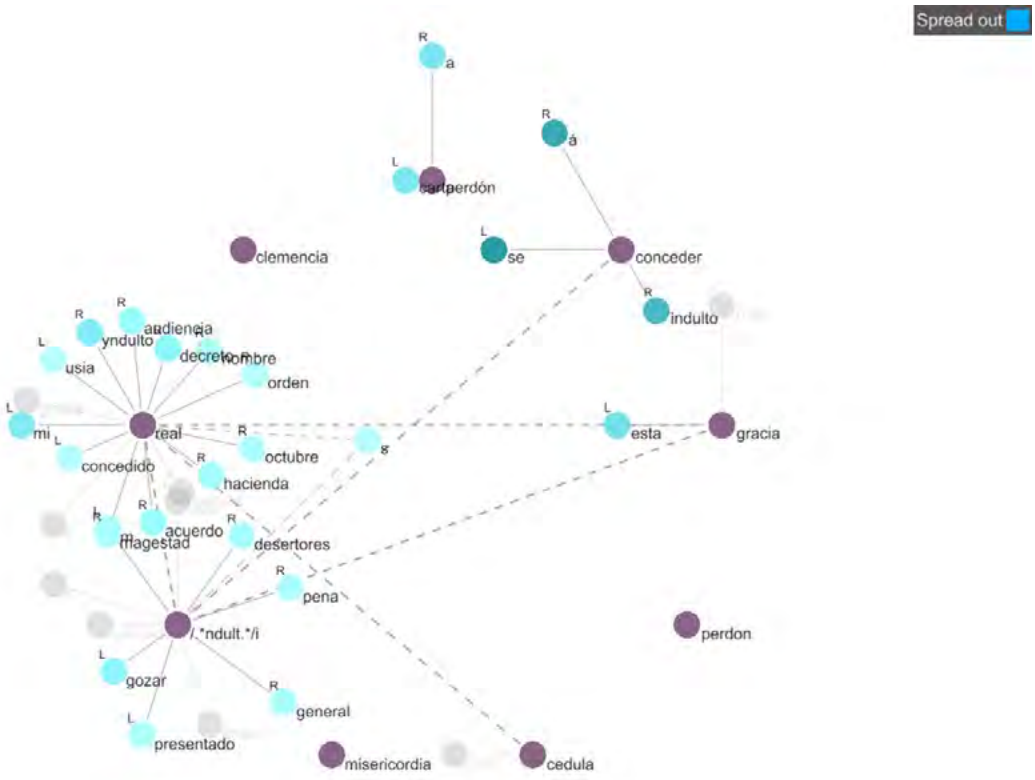
La colocación es un fenómeno léxico que da cuenta de las unidades fraseológicas más allá de las locuciones con significado propio. Su interpretación se fundamenta en la frecuencia estadística con la cual ciertos vocablos se relacionan entre sí y cuál es la relevancia de dicha asociación (Alonso Ramos, 1995: 9–28). En este sentido, consideramos esta es una de las estrategias de la lingüística que mejor podemos aprovechar para percibir un posible cambio semántico (Pazos-Bretaña, 2016). Para realizar el análisis de colocación nos servimos de la herramienta informática *LancsBox* (Brezina et al., 2015), así como de las propuestas metodológicas del lingüista Paul Baker (2016: 142–48). Cada segmento del corpus fue interpretado con la opción "GraphColl" del mencionado programa, la cual se configuró con una estrategia estadística MI (*mutual information statistic*) que favorece las relaciones léxicas entre palabras evitando al mismo tiempo artículos de uso frecuente como "el", "la" o "de". Se utilizó la extensión de análisis estándar de cinco palabras hacia la izquierda y la derecha del término.

El resultado de cada segmento se asemeja al presentado en la gráfica 1, en el cual se despliegan los valores más significativos de la colocación. En este ejemplo, el término indulto (representado con caracteres comodín para solventar los problemas de *semmatization* y lematización) se despliega en una red que comprende en un primer nivel los términos real, facultad, gracia, perdón, cédula, delito, reos, presos, concesión, entre otros. Todos estos son términos que coinciden con el discurso tradicional del perdón real que dominó durante la Edad Moderna castellana (Rodríguez Flores, 1971; Sandoval Parra, 2014).

Si se compara con la gráfica 2, el vocablo perdón es reemplazado por el nombre gracia, separándose completamente los términos perdón, clemencia y misericordia del nombre indulto, aunque siguen formando parte de las impetraciones y de las cartas de perdón de parte. Esto parece indicar que en el lenguaje de la práctica jurídico-política hubo un fenómeno de metonimia en el cual la gracia reemplazó al perdón, en este caso, gracia era equivalente a perdón pero no a indulto (por ello se añadiría la conjunción copulativa "indulto y gracia").



Gráfica 1. Red semántica centrada en el concepto indulto (1739-1775). R5-L5, MI(5)



Gráfica 2. Red semántica centrada en el concepto indulto (1790-1807). R5-L5, MI(5)

Este ejercicio abarca varios procesos relevantes para el análisis de corpus de información histórica no estructurada. Con el avance de la digitalización de fuentes documentales en archivos y bibliotecas en ambos lados del Atlántico la tarea de recuperación y macroanálisis de la información se hace más compleja, por lo cual es necesario ya no sólo introducir metodologías computacionales utilizadas en contextos anglosajones, sino construir estrategias propias que permitan lidiar con una tradición paleográfica y archivística particular.

Las tareas inmediatas que se plantean para este proyecto incluyen el resolver la transcripción de documentos para lo cual se está explorando la plataforma Omeka S, así como la posible exportación de elementos y modelarlos en XML-TEI. Del mismo modo, se pretende mejorar el modelo de segmentación construyendo una estrategia para resolver las disparidades grafológicas. Se espera que estos ejercicios en un futuro pueden ser relevantes para los proyectos de digitalización actuales en Latinoamérica.

Referencias

Agüero Nazar, A. (2012). Historia política e Historia crítica del derecho: convergencias y divergencias. *PolHis*, 5(10): 81–88.

Alonso Ramos, M. (1995). Hacia una definición del concepto de colocación: de J. R. Firth a I. A. Mel'ĉuk. *Revista de Lexicografía*, 1: 9–28.

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21(2): 139–64 doi:10.1075/ijcl.21.2.01bak.

Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2): 139–73 doi:10.1075/ijcl.20.2.01bre.

Costa, P. (1972). Semantica e storia del pensiero giuridico. *Quaderni fiorentini per la storia del pensiero giuridico moderno*, 1(1): 45–87.

Foucault, M. (1975). *Surveiller et punir: naissance de la prison*. Paris: Gallimard.

Gadamer, H.-G. and Koselleck, R. (1997). *Historia y hermenéutica*. (Ed.) Villacañas, J. L. & Oncina Coves, F. Barcelona: Paidós.

Hespanha, A. M. (2002). *Cultura jurídica europea: síntesis de un milenio*. (Trans.) Soler, I. and C. Valera Madrid: Tecnos.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. (Topics in the Digital Humanities). Urbana: University of Illinois Press.

Melo Flórez, J. A. (2016). Metadatos *Cibercliografía* http://ciberclíografia.org/manuales/crear-un-fichero-de-investigacion-con-omeka/metadatos/ (accessed 28 April 2018).

Mikheev, A. (2005). Text Segmentation. In Mitkov, R. (ed), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199276349.001.0001/oxfordhb-9780199276349-e-10 (accessed 25 November 2017).

Pazos-Bretaña, J.-M. (2016). El efecto de la historia sobre el cambio semántico en el español peninsular. *Itinerarios: revista de estudios lingüisticos, literarios, históricos y antropológicos*(23): 123–39.

Prodi, P. (2000). *Una Storia Della Giustizia: Dal Pluralismo Dei Fori Al Moderno Dualismo Tra Coscienza e Diritto*. (Collezione Di Testi e Di Studi). Bologna: Il mulino.

Rodríguez Flores, M. I. (1971). *El perdón real en Castilla (siglos XIII-XVIII)*. Salamanca: Universidad de Salamanca.

Sandoval Parra, V. (2014). *Manera de Galardón: Merced Pecuniaria y Extranjería En El Siglo XVII*. (Sección de Obras de Historia). Madrid: Fondo de Cultura Económica : Red Columnaria.

Turkel, W. J. and Crymble, A. (2012). From HTML to List of Words (part 2). *Programming Historian* https://programminghistorian.org/lessons/from-html-to-list-of-words-2 (accessed 28 April 2018).

Part Deux: Exploring the Signs of Abandonment of Online Digital Humanities Projects

Luis Meneses
ldmm@uvic.ca
Electronic Textual Cultures Laboratory - University of Victoria, Canada

Jonathan Martin
jonathan.d.martin@kcl.ac.uk
King's College London, United Kingdom

Richard Furuta
furuta@cse.tamu.edu
Center for the Study of Digital Libraries, Texas A&M University, United States of America

Ray Siemens
siemens@uvic.ca
Electronic Textual Cultures Laboratory - University of Victoria, Canada

Introduction

Building online research components for projects in the digital humanities is a common practice. However, not many researchers have a plan for these online components once the project halts or comes to an end. Consequently, many of these projects become abandoned and

slowly degrade over time –some more gracefully than others. Additionally, there is a certain inherent fragility associated with software and our online research tools. In turn, this fragility threatens the completeness and the sustainability of our work over time.

Previous studies have attempted to harness and manage the fragility of online resources. Studies have been carried out to address their potential reconstruction (Klein et al., 2011), the overall decay of websites (Bar-Yossef et al., 2004) and the decomposition of their shared resources (SalahEldeen and Nelson, 2012). Recently, our research has been focusing on analyzing the perceptions of change in distributed collections (Meneses et al., 2016)NY, USA","page":"273–278","source":"ACM Digital Library","event-place":"New York, NY, USA","abstract":"It is not unusual for documents on the Web to degrade and suffer from problems associated with unexpected change. In an analysis of the Association for Computing Machinery conference list, we found that categorizing the degree of change affecting digital documents over time is a difficult task. More specifically, we found that categorizing this degree of change is not a binary problem where documents are either unchanged or they have changed so dramatically that they do not fit within the scope of the collection. It is in part, a characterization of the intent of the change. In this paper, we present a case study that compares change detection methods based on machine learning algorithms against the assessment made by human subjects in a user study. Consequently, this paper will focus on two research questions. First, how can we categorize the various degrees of change that documents endure? And second, how did our automatic detection methods fare against the human assessment of change in the ACM conference list?","URL":"http://doi.acm.org/10.1145/2914586.2914628","DOI":"10.1145/2914586.2914628","ISBN":"978-1-4503-4247-6","author":[{"family":"Meneses","given":"Luis"},{"family":"Jayarathna","given":"Sampath"},{"family":"Furuta","given":"Richard"},{"family":"Shipman","given":"Frank"}],"issued":{"date-parts":[["2016"]]},"accessed":{"date-parts":[["2017",4,12]]}}], "schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} . However, we believe that the inherent characteristics of online digital humanities projects present an interesting (and unique) area for inquiry for two reasons. First, the research aspect of digital humanities projects hinders previous approaches –as the methods for identifying change in the Web do not fully apply. And second, digital humanities projects have a limited useful life –which is accompanied by research from primary investigator, which may or may not be indicated by updates in the project's content and tools.

We presented a paper in Digital Humanities 2017 that explored the abandonment and the average lifespan of online projects in the digital humanities (Meneses and Furuta, 2017). However, we believe that managing and characterizing the online degradation of digital humanities projects is a complex problem that demands further analysis. In this abstract, we propose to explore further the distinctive signs of abandonment of online digital humanities projects. For this second instalment of our study we took a different direction: we departed from strictly using retrieved HTTP response codes and incorporated additional metrics such as number of redirects, DNS metadata and a detailed analysis of content features.

This study aims to answer four questions. First, can we identify abandoned projects using computational methods? Second, can the degree of abandonment be quantified? Third, what features are more relevant than others when identifying instances of abandonment? Our final question is philosophical: can an abandoned project still be considered a digital humanities project?

Methodology

A complete listing of research projects in the Digital Humanities does not exist. However, the Alliance of Digital Humanities Organizations publishes a Book of Abstracts after each Digital Humanities conference as a PDF. Each one of these volumes can be treated as a compendium of the research that is carried out in the field. To create a dataset, we downloaded the Books of Abstracts corresponding from 2006 to 2016 –except for 2015 which was not available for download. We must thank and acknowledge Dr. Jason Ensor from Western Sidney University for providing us the abstracts for the 2015 Digital Humanities conference –which completes our dataset of conference abstracts. We obtained these abstracts after we had carried out our preliminary analysis. Therefore, we will present our findings using the complete dataset of abstracts in the presentation of our paper.

Then we proceeded to extract the text from these documents using Apache Tika and parse the 5845 unique URLs that we found using regular expressions. Then we used Python's Requests Library to retrieve the HTTP response codes and headers corresponding to the URLs, which we used to classify the websites into two groups depending on their correctness: valid (correct) and decayed (showing signs of degradation). Figure 1 shows the distribution of decay for each year. Based on our preliminary findings we approximate the average lifespan of a research project to 5 years, which aligns with reports from previous work (Goh and Ng, 2007). The average time in years since the last modification of the websites in the study is shown in figure 2.

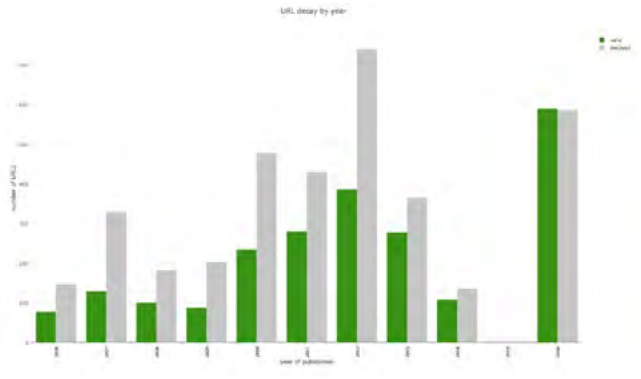


Figure 1: URL decay by year.

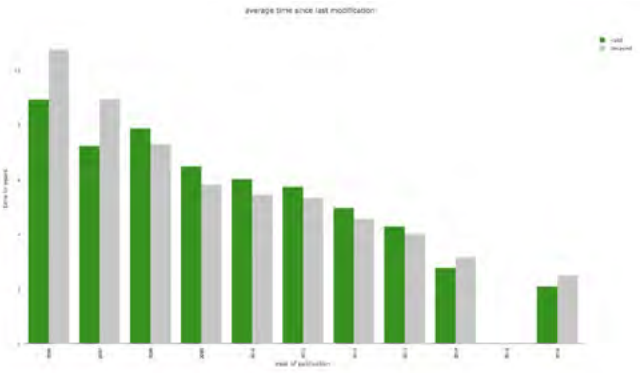


Figure 2: Average time in years since last modification.

Developing classifiers

To develop classifiers for the degradation identified in the previous section, we considered features computed based on DNS metadata, the initial HTTP request, number of redirects, and the contents and links returned by traversing the base node. The features we included are divided into topology, content-type, anchor-text and child node features. These features stem from concepts we used in our previous work (Meneses et al., 2016)NY, USA","page":"273–278","source":"ACM Digital Library","event-place":"New York, NY, USA","abstract":"It is not unusual for documents on the Web to degrade and suffer from problems associated with unexpected change. In an analysis of the Association for Computing Machinery conference list, we found that categorizing the degree of change affecting digital documents over time is a difficult task. More specifically, we found that categorizing this degree of change is not a binary problem where documents are either unchanged or they have changed so dramatically that they do not fit within the scope of the collection. It is in part, a characterization of the intent of the change. In this paper, we present a case study that compares change detection methods based on machine learning algorithms against the assessment made by human subjects in a user study. Consequently, this paper will focus on two research questions. First, how can we categorize the various degrees of change that documents endure? And second, how did our automatic detection methods fare against the human assessment of change in the ACM conference list?","URL":"http://doi.acm.org/10.1145/2914586.2914628","DOI":"10.1145/2914586.2914628","ISBN":"978-1-4503-4247-6","author":[{“family":"Meneses","given":"Luis"},{“family":"Jayarathna","given":"Sampath"},{“family":"Furuta","given":"Richard"},{“family":"Shipman","given":"Frank"}],"issued":{“date-parts":[[“2016"]]},"accessed":{“date-parts":[[“2017",4,12]]}}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} .

The text associated with resources is the most obvious feature for determining the topics. Given that we are dealing with a very specialized domain, we developed a domain-oriented expectation model. In particular, we generated topic and term frequency models to examine the similarity among the documents in a given project (the contents of the base node and the metadata and the contents of the child nodes). We used Latent Dirichlet Allocation to model the content of the text (Blei et al., 2003) and a simple Tf-Idf ranking function to measure and compare them. This ranking function is based on adding the Tf-Idf values for the documents, which were calculated using the terms from the topic modelling as a vocabulary. We will present a detailed version of our results on the longer version of our paper.

Discussion

This study is an attempt to categorize change in a very specific domain. More so, this study constitutes one step towards addressing potential strategies for the archival and the long-term preservation of abandoned digital projects. It is important to highlight that not all projects are equal and thus require different approaches towards long-term preservation. In the case of dynamically generated projects, a common approach nowadays is to produce a static set of HTML files which are easier to store. However, this approach assumes the backwards compatibility of Web browsers over time –something that has not always been the case.

To summarize, in this study we aim to computationally identify the indicators of the abandonment of digital humanities projects –as well as quantify their degrees of neglect. To address our philosophical question, we believe that an abandoned project can still be considered a valid digital humanities project depending on its audience. However, this has several nuances that should be considered. Digital online projects in the humanities have unique characteristics that make them impervious to the metrics that used in the Web as a whole –which make them worthy of study. In the end, we intend this study to be a step forward towards better preservation strategies and for the planned obsolesce of digital humanities projects.

References

Bar-Yossef, Z., Broder, A. Z., Kumar, R. and Tomkins, A. (2004). Sic transit gloria telae: towards an understanding of the web's decay. *ACM* doi:10.1145/988672.988716.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3: 993–1022.

Goh, D. H. and Ng, P. K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, 58(1): 15–24.

Klein, M., Ware, J. and Nelson, M. L. (2011). Rediscovering missing web pages using link neighborhood lexical signatures. ACM doi:10.1145/1998076.1998101.

Meneses, L. and Furuta, R. (2017). Shelf life: Identifying the Abandonment of Online Digital Humanities Projects Paper presented at the *Digital Humanities 2017*, Montreal, Canada.

Meneses, L., Jayarathna, S., Furuta, R. and Shipman, F. (2016). Analyzing the Perceptions of Change in a Distributed Collection of Web Documents. *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. (HT '16). New York, NY, USA: ACM, pp. 273–278 doi:10.1145/2914586.2914628. http://doi.acm.org/10.1145/2914586.2914628 (accessed 12 April 2017).

SalahEldeen, H. M. and Nelson, M. L. (2012). *Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?*

A People's History? Developing Digital Humanities Projects with the Public

Susan Michelle Merriam
merriam@bard.edu
Bard College, United States of America

In this short paper I will explore some of the problems—particularly those having to do with power and access—inherent in collaboratively produced community digital history projects. I will focus on two projects currently in development in which I (working from within an academic institution and digital media lab) have partnered with people from marginalized populations located in geographic areas that have been given relatively little attention. One of my goals in initiating these projects has been to explore how to use institutional resources, including grants and IT support, to work outside of institutional structures. In each instance, my community partners and I have created projects centered on individual, personal narratives as they relate to place. Our objective has been to develop a kind of "people's history," giving voice to those who have traditionally been excluded from historical research and writing. In the course of conceptualizing and beginning to make these projects, however, we've encountered a number of thorny questions about notions of community, access, and narrative form and content.

Conceptually, these projects have been fundamentally enabled by digital technologies that allow new makers to produce historical narratives. Indeed, digital media has fed an emerging industry in small-scale, creative historical projects, many of which academic historians would term "micro histories." Explored perhaps most famously by Carlo Ginzburg, micro histories can be viewed as correctives to "great man" theories of history or macro narratives that are easily undermined when challenged by specific circumstances. Focusing on seemingly "small" events—a day in an individual's life, for example—micro histories often make transparent the point of view of the researcher, thus destabilizing hegemonic forms of historical writing. Micro histories can also bring attention to, or use, lacunae in the historical record, as well as offer narrative forms for "regular" people to engage in the construction of history.

Working with digital tools and a loose concept of micro history, last spring I founded Bard College's "Mobile History Van," which operates under the umbrella of Bard's Digital History Lab (http://eh.bard.edu/dhl/). Both are funded by a major grant from the Mellon Foundation. The Mobile History Van uses digital technology to record and publish local history, and has worked closely with a local library and museum on digitizing archival materials and recording community history. While these projects have excavated important aspects of the historical record, they were executed with institutional partners—a college and historical society—and are thus still inscribed in easily recognized power structures.

In pursuit of developing projects outside of institutional structures, I approached students from Bard's Clemente program (https://www.clementecourse.org), a college credit granting year-long course for people who earn below a certain income level. Many of these students are struggling with substance abuse issues, criminal records, and post-traumatic stress disorders, but they wish to find ways to engage in the world. My intention: to work together with them and develop digital projects from the ground up, including working with other people who might not consider themselves part of the community described by the local historical society.

The first part of this talk will briefly introduce the audience to the genesis of the projects, and to their current state. By the time of the conference, both projects will be near completion, one as a series of personal narratives mapped onto the city of Kingston NY, the other in the form of a podcast about storytelling.

The second half of this talk will examine a series of thorny questions we have encountered in the process of our work: Who controls the projects? What role does the institution play in supporting the projects, and how does

this institutional affiliation shape the outcome in each case? How do we develop the projects for maximum input from collaborative partners? As we are working outside of an institutional structure, how do we define "community"? When should we, or is it necessary to, protect the storyteller? And finally, what does this type of project reveal about access to digital media?

Peer Learning and Collaborative Networks: On the Use of Loop Pedals by Women Vocal Artists in Mexico

Aurelio Meza
meza.aurelio@gmail.com
Concordia University, Canada

PoéticaSonora is a research group interested in the study of sound, listening, and legibility at the intersection of art and literary studies. The group has worked together for two years, organizing several events and projects that operate under two main axes, activation and preservation. The most important project on the preservation axis is dedicated to the design, creation, and development of a digital audio repository (DAR) for sound art and sound poetry in Latin America. During the first phase of the project to gather audio files for the repository, we have conducted fieldwork and archival research in different public centers and private collections, mostly in Mexico City. The DAR prototype was designed and developed by graduate students at Concordia University, in Montreal, with an initial sample of 317 audio tracks, performed or composed by around 180 artists, most of them from Mexico but also Argentina, Brazil, and the US. These tracks have previously been classified as sound art, sound poetry, radioart, experimental music, spoken word, poetry slam, performance, hip hop, indigenous language poetry, among many other terms.

The interest of PoéticaSonora members has focused on studying audio recordings of poetry readings, as well as organizing and curating sound art and experimental music events. This presentation, however, studies how musical instruments and other sound-generating devices accompany, even modify the human voice, and how the DAR contributes to understand the understanding of these works and the context where they are developed. Texts by scholars who have taught, studied, and/or conducted fieldwork in Montreal, such as Mark J. Butler, Jeremy Wade Morris, Tara Rodgers, and Jonathan Sterne will serve as a theoretical framework to discuss some artistic practices by Mexican women vocal artists who participate in collaborative creative networks (sometimes called "bands," "collectives," "jams," among other labels) and use sound-generating devices as a fundamental element in their performance. As a case study, I will focus on the path of two such artists, Edmée García and Leika Mochán, who combine spoken word and singing with the use of loop pedals. For several years they collaborated on the LP *Frágil*, along with jazz songwriter Iraida Noriega, and have ever since worked in different creative projects, both solo and with other artists. Some of the pieces where they use loop pedals are also analyzed here, such as García's *Chilanga habla*, described by herself as a "piece for poet and Line6," and Mochán's "Kaleidocycle," consisting of an amplifier and a Line6 DL4 attached to a customized bicycle.

The work of García and Mochán contributes to the discussion about what is intuitive and what is not in the use and adaptation of digital devices to produce sound with artistic or aesthetic purposes. These artists generate their own learning networks, transmitting to each other the empirical knowledge they acquire from free experimentation with a device. García calls Mochán "the loop pedal guru," and learned from her and Noriega how to use it during the creation of *Frágil*. This experience completely shaped the way García would perform her next poetry book, *Chilanga habla*, up to the point of deciding not to publish the text-based version, as it did not portray the project's whole scope and shape. As for the "Kaleidocycle," it allows Mochán to interact with the audience in a direct way, and posits questions about the false distinction between liveness and recording, particularly at the moment of performance. The different paths followed by García and Mochán after *Frágil* are a good example of how knowledge is not always prefigurative (from an elder to a youngling), but also configurative (among peers) and sometimes even postfigurative (from a youngling to an elder). This presentation sheds light on how Mexican artists face a device's intended use and how their actual uses diverge and become mainstreamed within certain collaborative networks.

Frágil and some works by García have already been integrated to PoéticaSonora's DAR. The presentation will start with a brief showcase of how their collaborative networks are illustrated in the prototype, as well as the roles and instruments each participant plays in a particular composition. It will then discuss how to integrate new works by García and Mochán, how to possibly solve some of the prototype's limitations, and reflect upon the next steps in the project, considering the implications it may have on the prototype's data schema. As it stands, does the DAR help us visualize these collaborative networks? Is it necessary to have an entity for groups and collectives, or can it be inferred from other categories in the database? It will finally discuss a brief genealogy of loop pedals to understand how such a marginal guitar effects unit (like the Boss RC 20 or 30, Line6 DL4) evolved into a device for singers (Boss RC 500, but more specifically the TC Helicon series), and in so doing re-purposed this device. With these case studies I will explain how the functions delegated to the loop pedal allow these artists to overco-

me the fact of not being a "musician," even though both have a strong musical background, and to perform "solo" despite holding a creative relation with the loop pedal.

Next Generation Digital Humanities: A Response To The Need For Empowering Undergraduate Researchers

Taylor Elyse Mills
mills@hope.edu
Michigan State University, United States of America

Introduction

Integrating the digital humanities (DH) into undergraduate level higher education programs has often been a difficult and ambiguous process. Faculty sometimes struggle to create syllabi that incorporate technologies but that do not require constant redesign as technologies evolve. Institutions may lack systems to connect students with faculty and staff who are interested in collaborative research, and collaboration beyond one's own institution can be complicated or inaccessible for students. These are real challenges; as institutions increasingly develop DH courses and degrees, the impact on undergraduate students is diverse, ranging in minimal involvement, to career-altering. So, what should the role of the undergraduate in DH be, and how can we address these challenges? For the past three years I have explored these questions. This exploration has led to helping redesign and teach the foundational seminar for Hope College's Mellon Scholars DH Program, as well as co-founding and chairing the Undergraduate Network for Research in the Humanities (UNRH), an undergraduate-led organization with the mission of reimagining the undergraduate role in DH through the establishment of a network of digital humanists who present research, collaborate, and share ideas. On the basis of these experiences as an alumna of Hope's DH Program and UNRH Chair, I have been considering the ways in which faculty, staff, and institutions might support undergraduate DH researchers. My work has culminated in a series of models, programs, and initiatives that address the need for fostering the next generation of digital humanists in the classroom, at the institution, and beyond.

Method

Classroom

The first challenge I consistently identified in DH courses was an incohesive structure that treated the digital and the humanities as separate units rather than an interconnected academic space. Secondly, seminar themes grounded in particular technologies had to be redesigned frequently as these technologies evolved or became outdated. This was the case for the year-long introductory seminar for Hope's DH Program. Each year students felt that the seminar was two unrelated courses, one focusing on a particular area in the humanities, the other, teaching technologies like GitHub and data analysis. The course was a noble attempt but ultimately inconsistent, incohesive, and not a truly interdisciplinary approach to DH. I set about designing a seminar model that was adaptable to new technologies yet still focused on an intersectional theme. I consulted with educators at conferences and researched seminar formats at other institutions, but unsurprisingly there was a wide range of approaches that seldom emphasized independent research quite like Hope's program. Thus, I grounded the seminar model in that very aspect: a chronological approach to independent research in the humanities. Over course of four units students engage with the evolution of humanities-based research and with the research process from beginning to end. During the first unit, students work in the archives, practice cataloging primary sources with tools like Zotero, develop strong but focused research questions, and discuss literature to answer the ever-present question "What is DH?" The second unit follows the progression in humanities-based research, moving from sources like libraries and datasets into the first examples of DH: text analysis. Students curate their own text-based datasets, analyze and visualize them, present them with Omeka, and discuss research project methodologies of source compilation and argumentation. The third unit it titled: CCP-Collaboration, Communication, & Presentation. It involves group research collaboration and finalizing research projects through effective communication and presentation. Students complete writing workshops in which they must adapt a piece of writing for different audiences and styles, from conference abstracts to blogs and tweets; they also practice oral and web presentation skills. The final unit addresses advanced topics and tools which require students to focus on race, gender, sexuality, politics, and socioeconomic status. Students learn that equity and accessibility are paramount when creating public scholarship, digital or otherwise, and they are exposed to a survey of technologies in efforts to broaden their concept of what form research can take. The outcome of this course should be a comprehensive and diverse approach to humanities-based research projects through the chronological progression that research in the humanities has followed.

Institution

For collaborative research, students and faculty alike find it challenging to make necessary connections with one another in the four short years that students have on campus.

My solution is Bin(d)r: the Baccalaureate Interdisciplinary Network for (Digital) Research. Stemming from an initial idea of a physical binder with pages featuring the profiles of faculty, staff, and students interested in collaborative research, Bin(d)r: is ideally implemented as a searchable database of anyone on campus with research interests and skills. It is like Tinder for academics. All faculty and staff interested in collaborating simply create a profile on a site with tools like WordPress's "Ultimate Member" Plugin. Students are invited to create profiles if they are interested in research. By including specific research interests and skills, faculty and students can get "matched" in a timely manner. Bin(d)r: has parentheses around "digital" because this tool does not have to be exclusively for digital projects, but it would provide an extra level of support for digital projects, connecting computer science students with humanities faculty, for example. Bin(d)r: is capable of being entirely free, low maintenance, highly interdisciplinary, and ultimately a tool for encouraging undergraduate research. Furthermore, if the digital Bin(d)r: takes off at numerous institutions, searching others' databases would foster cross-institutional collaboration.

While considering the institutional level, I would also argue that institutions must make space to hear the voices of their students. I propose that institutions establish a quarterly forum for undergraduates, faculty, and administrators to gather and discuss how the institution can better support students. Academic institutions are designed first and foremost to educate their students, so I assert that students have the right to tell institutions how they can improve, and institutions have the responsibility to listen. Simply creating space for dialogue is empowering.

Beyond

I also argue that empowering undergraduate researchers means providing agency, accreditation, and opportunities to join a community. Because DH is emerging at different rates across the globe, many students never meet other students engaging in their work. Furthermore, exposure to different methodologies, technologies, and project ideas has a profound impact. Faculty and staff gain this exposure at academic conferences and within their departments. UNRH aims to give this space and community to students, too.

Our method of creating UNRH relied heavily upon initial organization, forming a Steering Committee, review system, and website. The format of our conference was meticulously designed. We created a "speed-dating" session for rapid introductions and elevator pitch practice, a formal project presentation session, informal poster-style presentation sessions, a keynote address, and workshop sessions. These workshops include technology tutorials, panel discussions about different students' roles and experiences at their institutions, and design-thinking sessions to address the needs and concerns of students striving to develop DH projects.

Beyond the conference we have been developing an online network space in which students create profiles and can share project updates, articles, conference opportunities, and requests for peer review. In essence, each of our decisions was an effort to create space and flexibility for students to answer for themselves the question of what the undergraduate role in DH can be.

Results

Classroom

The feedback from my students who experienced my seminar model have been positive. The survey results indicate that the seminar has largely met the learning outcome goals, and students indicated increases in confidence and preparedness in conducting independent research (approximately 30% average increase) and using new technologies (approximately 37% average increase) according to a seven-point scale. Those who indicated having less prior experience (1-4) had an average increase of about 33% in independent research and about 39% in technology use. I plan to track program retention rates in the coming years to hopefully see improvements as the sophomore students navigate from the structured seminar into the independent research spaces of their junior and senior years.

Institution

Bin(d)r: has not yet been implemented but is in development for implementation at Hope College in the coming year.

Beyond

The results of our efforts exceeded expectations. Since our first conference in 2015, we have accepted over 50 projects, involving over 80 undergraduates from 31 institutions all across the United States, Canada, Nigeria, and Pakistan. According to in-person comments and our post-conference evaluations, students have felt empowered, encouraged, and independent in their research. Moreover, students were amazed at what they learned and accomplished by interacting with undergraduates from other institutions.

Through our initial design and modifications over the years, we feel confident in the model for an organization and conference that grants agency to undergraduates, and space to understand their own roles. Now in my third year as Project Manager/Chair, when I consider again the undergraduate role in DH, I think of students as connected learners and independent researchers pursuing their own interests while learning from peers and mentors ali-

ke. Within and beyond this space, each student must determine her role for herself.

Instructors, institutions, and organizations, invest in these students, for they are the next generation of digital humanists.

La creación del Repositorio Digital del Patrimonio Cultural de México

Ernesto Miranda
mirandatrigueros@gmail.com
Secretaría de Cultura, Mexico

Vania Ramírez
vaniasa.ramirez@gmail.com
Secretaría de Cultura, Mexico

Introducción

La creación de la Secretaría de Cultura Federal del Gobierno de México, trajo consigo la creación de la Dirección General de Tecnologías de la Información y Comunicaciones, y con esta, el mandato de construir la Agenda Digital de Cultura. Dentro de las atribuciones que tiene la Dirección se encuentra la interoperabilidad de las colecciones digitales albergadas y administradas por la Secretaría de Cultura. Para poder responder a este mandato se ha puesto en marcha el desarrollo del "Repositorio Digital de Patrimonio Cultural de México (RDPCM)", primer esfuerzo de largo alcance y con visión integral desde el gobierno de México para integrar los acervos digitales de museos, bibliotecas, televisoras, radiodifusoras y diferentes instituciones culturales que son coordinadas por la Secretaría de Cultura.

En el presente artículo, se describirán los diferentes módulos de trabajo que se han planteado para sentar las bases de este Repositorio y el grado de avance que cuenta al día de hoy. Asimismo se plantearán los retos técnicos, económicos y de gestión que ha implicado e implicará un proyecto de esta envergadura.

Contexto y antecedentes

Uno de los retos prioritarios para la Agenda Digital de Cultura, es la integración de los acervos culturales y ofrecerlos a los mexicanos para su divulgación, difusión y a través de una herramienta digital, convertirlos en una poderosa herramienta educativa de apoyo para la formación de la población.

El desafio es enorme, ya que actualmente los acervos en su gran mayoría no se encuentran normalizados bajo ningún tipos de esquema de datos, los contenidos descriptivos carecen de información y los objetos digitales son precarios o inexistentes, lo cual implica no únicamente una difusión carente de estructura, libre y abierta al público, sino que también, no se cuenta con ningún programa de preservación a largo plazo.

Objetivos

Algunos de los objetivos que el desarrollo del RDPCM tiene contempledos son:

- Generar una base sólida tecnológica, interoperable, libre y sustentable para la institución.
- Estandarizar los acervos bajo un mismo modelo de datos, para ser utilizado no sólo en la SC sino que sea extensivo en todo el país.
- Preservar el enorme y valioso patrimonio cultural de México de forma digital.
- Generar una plataforma web que permita a las audiencias acceder al vasto patrimonio cultural mexicano de manera enriquecida, sencilla y atractiva.
- Definición de derechos de los objetos digitales para la difusión y divulgación. Cabe destacar que actualmente en México no se cuentan con buenas prácticas en este tema.

Módulos de trabajo

Para el desarrollo del RDPCM, se plantearon los siguientes módulos que generaron productos específicos para la preservación y esquematización del patrimonio cultural digital:

1. Modelo de Datos México: creación de un modelo de datos único que permita normalizar e interoperar los metadatos de las instituciones, para coadyuvar a la mejor gestión de información producida desde la administración pública.
2. Normalización de registros: heterogeneidad de los registros y vocabularios, así como enriquecimiento editorial.
3. Tesauro Regional Mexicano:creación de un tesauro que incluirá terminologíamexicana especializada disponible a través del modelo LOD (Linked Open Data) e incluirá autores, obras, términos y relaciones de distintas instituciones.
4. Desarrollo: creación de sistema que cumplan con la visión de la Web Semántica, que permita exponer en formatos estándar todo la información. Contará con un meta-modelo ontológico, una herramienta de extracción y normalización, un cosechador-Indexador, buscador, CMS, API EndPoint y un módulo de preservación.
5. Sistemas gestores de colecciones de museos mexicanos: definidos en sistemas transparentes para el intercambio de datos con el RDPCM, pero que cuentan con total independencia al Repositorio.
6. Declaratorias de derechos: análisis del caso mexicano vs. el panorama internacional para la definición de derechos, según las leyes mexicanas.

Figura 1. Primera maquetación del RDPCM

Retos y responsabilidades

En la primera etapa de conceptualización y desarrollo del RDPCM, contendrá los acervos de 14 instituciones de la Secretaría de Cultura, que ascienden a más de 600,000 objetos digitales que represetan los acervos arqueológicos, históricos, artísticos, videográficos y sonoros de México. Uno de los retos más importantes en esta etapa es seguir incrementando proveedores de datos institucionales y continuar aumentandolos registros y objetos digitales de los que ya se encuentran en el Repositorio, además de seguir desarrollando contenidos de alto nivel.

También ser desarrollarán contenidoscurados por expertos, que permitan a los usuarios finales entender mejor las colecciones digitales y conectarse virtualmente con el patrimonio.

Prospecciones

- Sustentabilidad técnica y financiera: crear un sistema sólido y escalable en módulos, que permita adaptarse a las necesidades futuras y permitir reducir los costos de desarrollo, para poder ser aplicados en recursos humanos que administre las colecciones digitales.
- Aumento de audiencias: creación de una estretegia en medios y vinculación ciudadana a través de Hackatones y activaciones en redes sociales.
- Investigación y creación de contenidos: reconocer el valor de los investigadores y creadores de contenidos en las intituciones mexicanas, vinculado sus conocimientos para hacerlos participes en la creación y edición de contenidos.
- Proveedores de datos y agregadores: incrementar y exponer el mayor material disponible libre de derechos de autor en el RDPCM y generar salidasinnovadoras para las las nuevas audiencias.
- Creación del programa de digitalizción permanente: que permita incrementar el acervo digital para su posterior integración al repositorio de difusión y preservación.
- Profesionalización de los recursos humanos: generar redes para compartir conocimiento y solventar procesos a través de la creación de estrategias para el mejoramiento del sector cultura mexicano en el ambito de catalogación y preservación digital.

Conclusiones finales

México es un país con un enorme y valioso patrimonio cultural, la creación del RDPCM es una medida prioritaria y necesaria para el estado mexicano, que permitirá coadyuvar al acceso universal al patrimonio cultural mexicano para beneficiar a más audiencias educando, compartiendo conocimiento y transformando el mundo a través de la cultura.

Esta es una oportunidad para enriquecer la web con acciones positivas que refirmen la cultura en el mundo digital.La estrategía del RDPCM a 10 años, es proyectar el mayor número de objetos con una excelente calidad y a través de colaboraciones creativas e innovadoras, ofrecer un gran número de colecciones curadas en línea de acceso público para la investigación, el aprendizaje y la sociedad.

Referencias

Organización de las Naciones Unidas (2003). "Carta sobre la preservación del patrimonio digital". París.

Scholz, Henning, Devarenne,Cécile, Freire,Nuno, Kyrou,Panagiotis,Pekel, Joris (2017). "Europeana Content Strategy. Getting the right content to the right user at the right time".https://pro.europeana.eu/post/europeana-content-strategy

Digital Public Library of America (2015). Strategic Plan 2015-2017. Boston: DLPA.https://dp.la/info/wp-content/uploads/2015/01/DPLA-StrategicPlan_2015-2017-Jan7.pdf

Towards Linked Data of Bible Quotations in Jewish Texts

Oren Mishali
oren.mishali@gmail.com
Technion, Israel Institute of Technology, Israel

Benny Kimelfeld
bennyk@cs.technion.ac.il
Technion, Israel Institute of Technology, Israel

Introduction

The Hebrew Bible (the Tanakh) is the most ancient and sacred collection of Jewish texts. Throughout the history, additional religious Jewish texts have been written such as the Mishna, the Babylonian Talmud, and many more. These additional texts are often related to (or inspired by)

the Bible. As such, many of them quote verses[1] from the Bible (as in Figure 1). Depending mostly on their frequency and location within the text, the quotations may indicate a weak or strong semantic relation between a given text and a specific portion of the Bible. Knowing these semantic relations may be beneficial for those interested in studying or investigating the Bible.

Nowadays, a variety of Jewish texts are publicly available over the Internet, yet the identification of Bible quotations within them is often sparse and sometimes entirely absent. Moreover, the existing identification lacks a rigorous representation, which makes it difficult to automatically infer semantic correspondence and to develop supporting software applications.

We report an ongoing project that aims to establish the machinery for the automatic detection and rigorous representation of quotations of Bible verses within Jewish texts. The project consists of three interleaving components. In the first component, an algorithm for identifying Bible quotations in text is developed. In the second, the results of executing the algorithm on a large and open text corpus are represented as a Linked Data graph (RDF dataset). In the third component, we develop a web frontend for making the dataset accessible to end users. Exposing the data to end users may also engage their participation in data-driven crowdsourcing (Ched et al, 2015), and hence, will serve to collectively help in improving the dataset quality. In what follows, we elaborate on each of the project components.

Algorithm

Quotation detection is gaining popularity in fields such as copyright enforcing and political analysis, and within ancient texts (Ernst-Gerlach andCrane, 2008; Gesche et al, 2016). The algorithms in use share common characteristics, yet each domain brings its own specificities and challenges. Given an input text, our algorithm first matches maximal n-grams[2] in the text to candidate Bible verses. For example, the green bigram (2-gram) in the first line of Figure 1 will have one matching verse, since its text ('לך לך') appears in exactly one Bible verse. This matching is maximal, since the words that appear before and after the bigram are not part of the quoted verse.

```
PREFIX jbo: <http://jbs.technion.ac.il/ontology/>
PREFIX jbr: <http://jbs.technion.ac.il/resource/>

SELECT ?uri ?text FROM <http://jbs.technion.ac.il> WHERE {
    ?uri a jbo:Text; jbo:text ?text.
    ?uri jbo:quotes jbr:text-tanach-1-1-1.
}
```

A portion of ancient Jewish Text (from Midrash Raba), that quotes two Bibles verses. Quotations to the same verse are marked in a similar color. Note that each quotation refers only to a part of the verse (1-4 words of it).

A first challenge that we face is related to variations found between the quoting text and the original Bible text, mostly related to the omission (or inclusion) of Hebrew vowel letters. As an example, consider the red quotation in the second line of the figure, that contains the word 'ומה הירו', where in the original Bible source the 'ו' (vav) vowel is omitted. We have implemented two alternative solutions, one is based on *fuzzy search* (Levenshtein distance), and the other on *exact search* performed simultaneously on two versions of the Bible, with and without vowels.

Not all verse candidates are valid quotations of Bible verses in the text. For instance, the phrase 'ויבא תיב' in the third line of the figure (underlined) is a common phrase that appears in eleven different Bible verses. Nevertheless, the phrase is mentioned in a different context, which is not related to any of them. False candidates occur mostly in bigrams and trigrams (3-grams), and the algorithm makes an effort to filter them out. One approach is to keep a candidate if a matching candidate appears in a larger n-gram in the same text. For instance, the green bigrams and trigram shown in the figure are reported as valid quotations since there is a 4-gram that quotes the same verse in the text ('וארא רשא קראה לא', line 3). We are considering additional filtering approaches related to statistical data inference and machine learning. We are also creating collections of labeled data for a systematic evaluation of the algorithm.

Linked Data

The detected quotations are represented as RDF Linked Data, making them accessible to machines for standard consumption and integration. We use a lightweight ontology that we have defined, augmented with standard properties taken from known ontologies such as RDF, RDFS, and Dublin Core (DC). We are working on the integration of additional ontologies such as CIDOC-CRM, FRBR, and SPAR. Key ontology classes are *Book*, *Section*, *Text*, and *Quotation*. A *Book* is composed of *Sections*, that may be composed of other *Sections*, and eventually of *Text* elements. Each Bible verse is a node of type *Text* in the RDF graph. To date, our graph contains a total of 23,206 *Text* nodes of Bibles verses. Additional 355,181 *Text* nodes represent text elements within other Jewish books (where quotations are searched for). An edge from a *Text* node of the latter kind to one of the former kind indicates a 'quotes' relationship. Nodes of class *Quotation* hold additional details such as the exact location wherein a quotation appears in the text.

A SPARQL query that retrieves all text elements quoting the first verse of the Bible.

1 The Bible is divided into basic text elements called *verses*.
2 An *n-gram* is a contiguous sequence of n words from a text.

A Linked Data graph may be accessed by expert users using the SPARQL query language. An example SPARQL query is shown in Figure 2. To make our data widely accessible, we have implemented a graphical web frontend that acts like a search engine for Bible verses. A user selects a set of verses from the Bible, and then being presented with all text elements that quote one or more verses from the set. (The elements are retrieved from the RDF graph.) The results are sorted by significance, and may be filtered using predefined categories. We plan to enhance the web interface with data-driven crowdsourcing support, where the crowd will help in improving the accuracy of the algorithm by marking false negatives (places in the text that the algorithm has missed), as well as false positives (incorrect detections). The web tool, as well as the detection algorithm and related artifacts, are accessible via our main GitHub repository.

References

Ched, L. and Lee, D. and Milo, T. (2015). *Data-driven Crowdsourcing: Management, Mining, and Applications*. International Conference on Data Engineering (ICDE), Tutorial.

Ernst-Gerlach, A. and Crane, G. (2008). *Identifying Quotations in Reference Works and Primary Materials* . Research and Advanced Technology for Digital Libraries, 78-87.

Gesche, S. and Egyed-Zsigmond, E. and Calabretto , S. (2016). *Was it better before? Automated Quotation Detection in Ancient Texts*. CORIA-CIFED, 167-182.

Towards a Metric for Paraphrastic Modification

Maria Moritz
mmoritz@gcdh.de
University of Goettingen, Germany

Johannes Hellrich
johannes.hellrich@uni-jena.de
Graduate School "The Romantic Model", Friedrich-Schiller-Universität Jena, Germany

Sven Buechel
sven.buechel@uni-jena.de
JULIE Lab, Friedrich-Schiller-Universität Jena, Germany

Introduction

Clarifying the genesis of a passed down text is of outmost importance for many scholarly disciplines within the humanities such as history, literary studies, and Bible studies. Often, historical text sources have been copied over and over for hundreds or even thousands of years, thus being subjected to paraphrasing and other kinds of modifications, repeatedly. Despite the significance of source criticism for the humanities as a whole, algorithmic support in this matter is still limited. While current approaches are able to tell **if** and **how frequent** a text has been modified—to the best of our knowledge—there has been no work on determining the **degree** of paraphrastic modification. To a human reader, the introduction of, say, spelling variations is indubitably a minor modification compared to substituting entire words. Yet, how can the different "degrees" of alterations, which are intuitively clear to scholars, be captured in an algorithmic way?

To this end, we present a first approach for designing a metric for paraphrastic modification in text (henceforth paraphrasticality). Based on an English Bible corpus (three literal Hebrew and Greek translations and three standard translations) we measure the frequency of different classes of textual variations between each pair of Bibles. We then use the probability of these variations in a machine learning experiment to derive weights for these classes of modifications. Ultimately, this allows us to define a metric for paraphrasticality which we validated with promising results.

Related work

Measuring the **similarity** or **distance** between two spans of text is relevant to many areas in and related to natural language processing (NLP). One of the earliest approaches is Levenshtein's (Jurafsky and Martin, 2009) edit distance which is based on character-level removal, insertion, and replacement operations. BLEU (Papineni, 2002) is the most common evaluation metric in machine translation, capturing the difference between gold and automatic translations based on (word-level) n-gram overlap. In **stylometry**, different kinds of delta metrics are used to compute the difference between the writing style of authors or texts (Jannidis et al., 2015). These are typically based on the frequency distribution of the most frequent words. These first three approaches have in common that they rely on surface features (token and character-level) alone and do not incorporate semantic proximity. In contrast to that, computing the **semantic similarity** between two sentences is a popular task within NLP (Xu et al., 2015). However, approaches in this field are typically not indented for manual inspection and are thus less suited for applications in the humanities. Lastly, Moritz et al. (2016) quantify modification operations on a parallel Bible corpus yet do not present a way to aggregate these counts into a distance metric. In contrast to these related contribution, here, we aim to develop a metric which is both semantically informed as well as human interpretable.

Data

We use a parallel corpus of the Old Testaments of six English Bible translations[3] from the 19th century, half of them being literal translations that closely follow the primary source texts' language and syntax while the other half are standard translations (see Table 1).

name	abbr.	publication	source	translation
The Webster Bible	WBT	1833	bst	standard
Brenton's English Septuagint	LXXE	1851	mys	literal
Young's Literal Translation	YLT	1862	bst	literal
Smith's Literal Translation	SLT	1876	mys	literal
English Revised Version	ERV	1881-1894	mys	standard
Darby Bible	DBY	1890	ptp	standard

Table 1: Bible editions used for investigation. Sources: bst: https://www.biblestudytools.com/; mys: https://www.mysword.info; ptp: Parallel Text Project (Mayer and Cysouw, 2014)

Literal translations: Robert Young, the translator of **YLT,** created a highly literal translation of the original Hebrew and Greek texts. Thus, Young tried to be as consistent as possible in representing Greek tenses with English ones, e.g., he used present tense where other translations used past tense (see Young, *1898a; Young, 1898b*) as in: 'In the beginning of God's preparing the heavens and the earth —' (Genesis 1:1). **SLT**: Upon publication, Julia Smith's Bible translation was considered the only one directly translating the historical source texts to contemporary English. She aimed at complete literalness and tried to translate each original word with the same English word, consistently, and tended to translate the Hebrew imperfect to English future tense (Malone, 2010). **LXXE** by Sir Lancelot Charles Lee Brenton is an English translation from the Codex Vaticanus version of the Greek Old Testament, which itself is a translation of the Hebrew Old Testament (Roger, 1958).

Standard translations: WBT by Noah Webster is a revision of the King James Bible mainly eliminating archaic words and simplifying Grammar (Marlowe, *2005*). **ERV** is today's only officially authorized revised version of the King James Bible in Britain (no author, 1989). The most recent edition in our study is **DBY**, Darby's translation of the Bible. The Old Testament was published by his students in 1890 and is based on Darby's German and French versions (Marlowe, 2017).

Methods

Preprocessing and alignment: We use MorphAdorner (Burns, 2013), a specialized toolkit for early modern and modern English, to tokenize and lemmatize the Bibles. After removing punctuation and verse identifiers, we pair up our six Bibles in every possible combination (15 in total). Since the different Bible versions are inherently aligned on the verse-level (by their verse identifier), our next step builds up a statistical alignment at the token level for each pair of bibles using the Berkeley Word Aligner (De Nero and Klein, 2007), a tool originally designed for machine translation.

Counting modification operations: Building on these word-aligned pairs of Bibles, we can describe the divergence between a pair of verses in terms of the **modification operations**—such as replacing a word by its synonym—which would be necessary to convert one version into another. We automatically apply and count the modification classes introduced by Moritz et al. (2016) for each verse and Bible pair (see Table 2). Synonyms, hypernyms, hyponyms and co-hyponyms, are identified based on BabelNet (Navigli and Ponzetto, 2012).

abbr.	operation	estimated coefficient θ_{relative}
lower	case-folding matches	0.060
lem	lemmatizing matches	0.195
low_editdist	writing variant	0.068
syn	synonyms match	0.190
hyper	source word is hypernym of target word	0.117
hypo	source word is hyponym of target word	0.170
co-hypo	co-hyponyms match	0.122
fallback	other	0.078

Table 2: Operations used as features together with normalized estimated weights (coefficients) of the fitted model

Weight identification: By counting modification operations, we gain a fine-grained description of the *exact* differences between two spans of text. However, to construct a metric, we had to find a way to condense these modification frequencies down to a single number. For that we exploit the fact that we deal with two classes of Bible translations, literal and standard ones. Thus, to estimate a human judgment of deviation, we assume that standard translations are more homogenous to each other than literal translations (since the latter demand for more creative language use; see Section 3). Hence, we can train a classifier to distinguish whether a pair of Bible verses is from the same class (both Bibles being standard or literal translations, respectively) or from different classes. For this task, we train a maximum entropy classifier[4] where we use the relative frequencies of the modification operations as features. Now, the key part of our contribution is that we can exploit the coefficients of our fitted model as the first ever presented empirical estimate of the relative importance of these modification operations for paraphrasticality.

3 Note that our approach is not limited to applications on historical text and that our choice of textual material is based on technical reasons only. In fact, any paraphrastic, parallel corpus would work equally well for our proposed method.

4 Using the scikit-learn.org implementation. Training for this binary classification task was done using 10-fold cross-validation achieving an accuracy of .68.

Results

Feature weights: Table 2 lists the final, normalized (summing up to 1) feature weights of our fitted model. Lemmatization, hyponym and synonym relations turn out to be especially important for the classification task.

Metric: Based on these coefficients, we define the paraphrasticality metric *par* between two word-aligned text spans *a* and *b* as

$$par(a,b) = \sum_{i=0}^{n} \theta_i x_i^{a,b}$$

where n is the total number of features (or classes of operations), θ_i is the absolute weight for feature i determined via the classifcation experiment and $X_i^{a,b}$ is the relative frequency of the respective operation. In order to gain face validity for this newly defined metric, we compute the paraphrasticality score for each one of the 15 Bible pairs in our corpus (as average of their verse paraphrasticality). The results are presented in Table 3.

	DBY	ERV	WBT	LXXE	YLT	SLT
DBY	-	0.13	0.13	0.29	0.31	0.29
ERV	-	-	0.09	0.3	0.32	0.31
WBT	-	-	-	0.28	0.33	0.29
LXXE	-	-	-	-	0.42	0.37
YLT	-	-	-	-	-	0.31
SLT	-	-	-	-	-	-

Table 3: Deviation between each pair of Bibles in terms of our newly developed paraphrasticality metric; higher values indicate higher distance

Bible pair	operation 1	operation 2	operation 3	classes
DBY-ERV	lem (1.6%)	syn (1.1%)	cohypo (.9%)	standard
DBY-WBT	lem (1.6%)	syn (1.1%)	cohypo (.9%)	standard
ERV-WBT	lem (1.6%)	syn (.7%)	cohypo (.6%)	standard
DBY-LXXE	lem (3.1%)	syn (2%)	cohypo (1.9%)	standard/literal
DBY-YLT	lem (6.6%)	low (4.7%)	syn (2.6%)	standard/literal
DBY-SLT	lem (5.9%)	syn (2.6%)	cohypo (2.2%)	standard/literal
ERV-LXXE	lem (3.5%)	low (2.1%)	syn (1.9%)	standard/literal
ERV-YLT	lem (6.6%)	low (4.7%)	syn (2.5%)	standard/literal
ERV-SLT	lem (5.9%)	syn (2.6%)	cohypo (2.2%)	standard/literal
WBT-LXXE	lem (3.4%)	low (2.2%)	syn (1.9%)	standard/literal
WBT-YLT	lem (6.8%)	low (4.8%)	syn (2.7%)	standard/literal
WBT-SLT	lem (5.8%)	syn (2.6%)	cohypo (2.2%)	standard/literal
LXXE-YLT	lem (7.%)	low (4.4%)	syn (2.6%)	literal
LXXE-SLT	lem (5.8%)	cohypo (2.6%)	syn (2.6%)	literal
YLT-SLT	lem (5.4%)	low (4.8%)	syn (2.5%)	literal

Table 4: Top 3 most frequent operations (without fallback) per Bible pair

Qualitative validation: We can identify three regions in the plot. The upper left triangle shows that our standard translations do not differ much from each other (as expected), especially since WBT and ERV are revisions of the same Bible. The 3x3 rectangle in the upper right corner represents pairs of one literal and one standard translation, respectively. We can see that the distance between those is about 0.3 thus displaying increasing paraphrasticality compared to pairs of *only* standard translations. The highest deviation however is between the literal translations by Smith (SLT) and Young (YLT) compared to the English Septuagint (LXXE). This can be explained by the choice of vocabulary by each translator and by the purpose they follow in their translations. For example, SLT and LXXE use "firmament" when YLT uses "expanse", SLT and YLT use "rule" when LXXE uses "regulating". We thus conclude that our metric yields valid and—perhaps even more important for applications in the humanities—interpretable results.

Our approach also enables to judge distance on a fine-grained level based on pure operation counts. In Table 4 we show the top 3 operations for each Bible pair. As we can see, most of the top 3 operations per Bible pair relate to semantic relations between the aligned word pairs (matching lemma, synonymy, or co-hyponomy) underscoring the advantage that our metric has as opposed to more surface feature-dependent approaches (to textual similarity) such as Levenshteindistance or delta measures.

Conclusion

We presented the first study on designing a metric for paraphrasticality. Different from existing approaches on measuring distance or similarity between texts, we describe paraphrasticality as frequency of specific modification operations for which we tried to find empirically adequate weights via a machine learning experiment. As demonstrated, our approach is specifically useful for applications in the humanities as operation frequencies, and feature weights, as well as paraphrasticality scores are open to manual inspection. A more comprehensive comparison against existing similarity metrics and a human judgment is left for future work.

References

Burns, P. R. (2013). Morphadorner v2: A java library for the morphological adornment of English language texts. *Northwestern University, Evanston,* IL, no page numbers.

De Nero, J. and Klein, D. (2007). Tailoring word alignments to syntactic machine translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, pp. 17–24.

Jannidis, F., Pielström, S., Schöch, C. Vitt, T. (2015). Improving Burrows' Delta—An empirical evaluation of text distance measures. *Digital Humanities Conference 2015*. Sydney, no page numbers.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics.* Englewood Cliffs: Prentice-Hall.
Malone, D. (2010). Julia Smith bible translation (1876), https://recollections.wheaton.edu/2010/12/julia-smith-bible-translation-1876/ *(accessed November 2017).*
Marlowe, M. *(2005).* Webster's Revision of the KJV (1833), http://www.bible-researcher.com/webster.html*(accessed November 2017).*
Marlowe, M. (2017). John Nelson Darby's Version, http://www.bible-researcher.com/darby.html*(accessed November 2017).*
Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* Reykjavik, pp. 3158–61.
Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y. and Büchler, M. (2016). Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. *Empirical Methods in Natural Language Processing.* Austin, pp. 1849–59.
Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(2012): 217–50.
Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, pp. 311–18.
Roger, N. (1958, 1959). New Testament Use of the Old Testament. In Henry, C. F.H. (ed.), *Revelation and the Bible. Contemporary Evangelical Thought.* Grand Rapids: Baker, 1958. London: The Tyndale Press, 1959, pp. 137–51.
Xu, W., Callison-Burch, C. and Dolan, B. (2015). SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT). *SemEval@ NAACL-HLT.* Denver, pp. 1–11.
Young, R. *(1898a). Young's Translation: Publisher's Note and Preface,* http://www.ccel.org/bible/ylt/ylt.htm*(accessed November 2017).*
Young, R. *(1898b). Young's Literal Translation*, http://www.bible-researcher.com/young.html (accessed November 2017).
No Author. (1989). *The Holy Bible. Revised Version.* London: Cambridge University Press. Synopsis.

Temporal Entity Random Indexing

Annalina Caputo
annalina.caputo@adaptcentre.ie
Adapt Centre, Trinity College Dublin, Ireland

Gary Munnelly
gary.munnelly@adaptcentre.ie
Adapt Centre, Trinity College Dublin, Ireland

Seamus Lawless
seamus.lawless@adaptcentre.ie
Adapt Centre, Trinity College Dublin, Ireland

Introduction

In this exploratory research, we sought to investigate how we might identify and quantify the contextual shift surrounding significant entities in news based corpora. For example, might we be able to see changing public opinion such as that experienced by George W. Bush Jr. after the events of 9/11 and thus note how a population can rally behind their leader in the face of cultural trauma?

Our method of identifying these changes has its roots in the field of distributional semantics and the measurement of semantic shift. A typical approach to solving this problem involves building multiple word models across subsets of the sample corpus which are organized by date. By comparing the outputs of the different models we can see how the definitions of words have evolved. We adopt Temporal Random Indexing (TRI) (Basile et al., 2014) as our method of measuring semantic shift over time as it allows for a direct comparison between word representations on the basis of simple cosine similarities.

Method

In order to apply our method of measuring contextual shift in relation to entities we require a consistent representation of each entity that will span the entire collection e.g. the algorithm will need to know that "President Bush", "G.W." and "Dubyah" all refer to the same individual. In order to achieve this, an Entity Disambiguation process is applied to the source text prior to building the semantic space. This step substitutes mentions of each entity with a URI obtained from DBpedia, allowing the algorithm to track an individual through the collection irrespective of how they are referenced. We use CogComp NER[5] (Ratinov and Roth, 2009) for entity recognition and AGDISTIS[6] (Usbeck et al., 2014) for disambiguation.

Given the output from the disambiguation tools, a different semantic space for each year in the collection's timespan is built using the TRI implementation by Basile[7]

5 https://github.com/CogComp/cogcomp-nlp/tree/master/ner
6 https://github.com/dice-group/AGDISTIS
7 https://github.com/pippokill/tri

(Basile et al., 2014). Each space provides a semantic representation of words and Named Entites (NE) in terms of their proximity in space, which reflects their semantic relatedness. A time series for each NE is extracted by computing the cosine similarity between two consecutive semantic spaces (e.g. 2001 and 2002). Finally, candidate dates for the shift in meaning are extracted using the Change Point Detection algorithm as implemented by Kulkarni[8] (Kulkarni et al., 2015).

Evaluation

For test data we utilized the New York Times collection curated by LDC[9] (Sandhaus, 2008) which spans 20 years of American news from 1987 to 2007. While methods which measure semantic shift in word sense typically require collections which span hundreds of years, because circumstances evolve more quickly than language, we believe that a 20 year span is more than enough to produce interesting results when the same methods are applied to the examination of entities.

The collection was preprocessed and analysed using the method described in Section 2. This yielded a series of 20 language models which provided semantic representations for each entity identified and linked by CogComp NER and AGDISTIS. We computed the temporal shift for all the entities in the corpus and ranked them by the magnitude of this shift (p-value from the Change Point Detection algorithm). We selected the top 100 entities from this ranking (i.e. those with the greatest semantic shift) and selected the largest group of entities which underwent a semantic shift in the same year from within that group.

Results

The evaluation methodology described in Section 3 yielded a shortlist of 12 entities which undergo a sizeable semantic shift in 2001: Federal_Bureau_of_Investigation, Pentagon, White_House, New_York, Congress, Department_of_Justice, George_H._W._Bush, Texas, West, Saddam_Hussein, Republican_Party_(United_States), and American_Motors. Almost all of them are related to politics and have strong connections with the happenings of 9/11. Notably, while a member of the Bush family is connected with these events and does indeed undergo a shift in semantic representation, it is the wrong individual - the father rather than the son. This assignment of a semantic shift to George_H._W._Bush in 2001 is certainly due to the disambiguation process.

While we believe the inclusion of the entity disambiguation step is an interesting contribution of this work, we observed a number of problems with the process.

The contents of the knowledge base, which informs the disambiguation software, has a dramatic impact on the quality of the results obtained. So too does the nature of the entities being disambiguated. One notable example of this was our results with regards to mentions of "the Internet". Our method showed a dramatic increase in discourse surrounding the Internet from the mid 90s up into the second millennium. However, while the representation was consistent, the referent chosen by the disambiguation software was an American band known as "The Internet", rather than the network of computers we use today.

While the error with the Internet is obvious, more challenging was distinguishing between mentions of George W. Bush Jr. and George H. W. Bush Sr. The former's role in the events post 9/11 (reports of which were included in our corpus) made him an important entity for the disambiguation software to correctly annotate. However, in many cases this proved to be extremely difficult. This is understandable given the similarity in context surrounding both Bush Jr. and Bush Sr., We can work with an incorrect annotation provided it is consistently incorrect. However the unpredictability surrounding the name "Bush" presents a difficult problem when this information is used as part of the Random Indexing process.

Conclusion

We have presented a preliminary case study, which although not robust enough to infer any conclusions, highlights the potential of this type of analysis. We conducted our preliminary investigation guided by a major cultural trauma that occurred between 1987 and 2007, and which caused a sudden reaction and change in the public discourse. It is clear that a weakness in the method is the disambiguation process. Future work will focus on improving the quality of disambiguation as well as investigating the possibility of building time series models over shorter spans of time e.g. months or weeks.

References

Basile, P., Caputo, A. and Semeraro, A. (2014). Analysing word meaning over time by exploiting temporal Random Indexing. *Proceedings of the First Italian Conference on Computational Linguistics CLiCit 2014 and of the Fourth International Workshop EVALITA 2014 911 December 2014 Pisa* doi:10.12871/CLICIT201418. http://www.pisauniversitypress.it (accessed 25 April 2018).

Kulkarni, V., Al-Rfou, R., Perozzi, B. and Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. ACM Press, pp. 625–35 doi:10.1145/2736277.2741627. http://dl.acm.org/citation.cfm?doid=2736277.2741627 (accessed 25 April 2018).

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. Association for Computational Linguistics, p. 147 doi:10.3115/1596374.1596399. http://portal.acm.

8 https://github.com/viveksck/langchangetrack
9 https://catalog.ldc.upenn.edu/ldc2008t19

org/citation.cfm?doid=1596374.1596399 (accessed 25 April 2018).
Sandhaus, E. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12): e26752.
Usbeck, R., Ngomo, A.-C. N., Röder, M., Gerber, D., Coelho, S. A., Auer, S. and Both, A. (2014). AGDISTIS-graph-based disambiguation of named entities using linked data. *International Semantic Web Conference*. Springer, pp. 457–471 http://link.springer.com/chapter/10.1007/978-3-319-11964-9_29 (accessed 12 February 2017).

IncipitSearch - Interlinking Musicological Repositories

Anna Neovesky
anna.neovesky@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Germany

Frederic von Vlahovits
frederic.vonvlahovits@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Germany

Open research data is facilitating broader ways of using, reusing, enriching, and linking research results. Many services use metadata to bring the information of different repositories together. Europeana, for example, links material from various thematic focal points with diverse origins and makes a wide range of collections, archives and source objects searchable. Other platforms interlink and aggregate material for one distinct discipline or thematic interest.

To connect musicological collections and repositories, we created a metasearch that builds up on annotated music. IncipitSearch is a tool and a service specifically tailored for research on music incipits, the initial sequences of notes that characterise a work. It is simultaneously a centralised data endpoint, where multiple aggregated catalogues can be accessed and searched by their music incipits, as well as a decentralised software and data cluster.

Open Data and Meta Search Engines: Perspectives for Digital Musicology?

Open research data is facilitating broader ways of using, reusing, enriching, and linking research results. Many services use metadata to bring the information of different repositories together. Europeana (https://europeana.eu), for example, links material from various thematic focal points with diverse origins and makes a wide range of collections, archives and source objects searchable. Other platforms interlink and aggregate material for one distinct thematic interest such as Ariadne (http://ariadne-infrastructure.eu), which makes manifold archaeological contents accessible, or correspSearch (http://correspsearch.net), which enables to search through collections of editions of letters.

Meanwhile, musicological projects do not only often have digital components, too. Ambitious global catalogue projects like the Répertoire International des Sources Musicales (RISM, https://opac.rism.info) or national library services such as the catalogues of the Italy's Servizio Bibliotecario Nazionale (SBN, http://opac.sbn.it) or the Deutsche Nationalbibliothek, (DNB, https://portal.dnb.de) substantially rely more and more on the digital representation of their data. In addition, overall demand of digital research platforms has lead to born digital editorial projects, e.g. Freischütz Digital, a genuinely digital edition of Carl Maria von Weber's Freischütz (http://freischuetz-digital.de) exploring the possibilities of multimedial digital musicological work editions, or the digital thematic work catalogue of the complete edition of Gluck's works (http://gluck-gesamtausgabe.de). The researcher's stronger trust and belief in the benefits of open and accessible research data has led to a stronger emergence of open data policies in musicological projects. In order to interlink existing data repositories and encourage new proposals, a digital data hub is needed. But how can musicological data collections be connected and linked together? In our approach, we concentrated on musical incipits, the initial sequences of notes, that function as identifier for works, and created IncipitSearch, a metasearch for musical incipits.

Encoding Music Incipits

One of the main goals of musicological catalogues is making musical works findable and researchable. The main problem that often occurs, especially for music composed before 1800, is that it originally was composed for a singular religious or secular cultural event, e.g. at an aristocratic court to be performed only once or just a few times. Music was additionally bound to formalised genre standards and therefore unambiguous titles were not required. But how to search for a Sonata in D of a composer who has composed 20 sonatas in D? As early as the 1960s, Music librarians introduced the idea to generate a human and machine readable standardised format to identify music by its melodic beginning. For that purpose, Barrey S. Brook and Murray Gold developed the Plaine & Easie Code that allows the transcription of the beginning notes of a musical piece into a combination of numbers and letters. What Brook and Gould pointed out in 1964 was already a distinct definition of and guide to the Plaine & Easie code system. They introduced it as "an accurate shorthand for musical notation, especially useful for incipits and excerpts." With some foresight they also stated that "it must be so devised to be readily transferable to electronic data-processing equipment for key transposi-

tion, fact-finding, tabulaturing and other research purposes." (Brook and Gold, 1964)

Plaine & Easie Code is a simple to parse plaintext format and therefore suitable to deliver important metadata for manifold musicological interests. IncipitSearch adopts this standard and at the moment is purely concentrated on Plaine & Easie. However, the future goal is to be capable of reading incipits notated in other formats as well, e.g. MEI (http://music-encoding.org) or abc notation (http://abcnotation.com).

Searching Music Incipits

Musc information retrieval systems either build up on audio or symbolic music notation. In digital musicology, that deals with notation and critical digital edition of works, the search in notated music is widely used (Typke et. al. 2005).

RISM is undoubtedly the most established repository for musical data. It contains over one million records of historic music materials and over 1,7 million musical incipits (for manuscripts only), which can be accessed using an incipit search (RISM search). Further incipit search engines build up on the RISM datasets. For example, Utrecht University has developed an extended and experimental search approach offering extended functionalities for user input as well as using sophisticated matching and ranking methods (Van Nuss et. al. 2017).

But other musical incipits exist which cannot be accessed via RISM because they either have not been implemented as data yet or because they are not a type of resource the RISM collection is focusing on and will not be added to the catalogue, such as work catalogues.

IncipitSearch

Scope and Functionalities

The efforts to implement incipits in the digital work catalogue of the complete edition of Gluck's works and to make them searchable have led to the idea of connecting this research data with other repositories and creating even easier ways to instantiate new machine readable incipit repositories. Both digital and analogue catalogues, editions and collections which provide their data in a standardised format can be interlinked with IncipitSearch.

IncipitSearch addresses music that can be displayed in common western music notation. Its main focus lies on music composed prior to 1800. Nevertheless, through its openness it can be furthermore used as a platform to explore challenges in researching culturally and historically different forms of musical notation.

IncipitSearch is a tool and a service specifically tailored for research on music incipits. It is simultaneously a centralised data endpoint where multiple aggregated catalogues of incipits can be accessed as well as a decentralised open source software that can be integrated as stand-alone search in other platforms. A microservice based software architecture allows high flexibility in usage and extension of individual components (Haft et. al. 2015).

IncipitSearch enables users to enter search queries in the search field by playing them on a virtual piano keyboard while Plaine & Easie Code can also be directly entered into the search field. Search with transposition or with exact matching can be selected (https://incipitsearch.adwmainz.net). Next to the found concordant incipts, the result list displays backlinks to the entry in the respective catalogue.

Screenshot of the search interface of IncipitSearch.

Metadata Schema

To enable a standard suitable for the different types of musicological repositories such as digital and analogue catalogues, editions and collections and to provide an output of the collected data, we have developed an easy to understand RDF schema using the schema.org vocabulary. Besides being recommended by the W3C, cross-linking possibilities for data and the possibility to rely on various vocabularies for specific topics, the interoperability and the multiple serialisation formats for RDF are advantageous.

Schema.org provides a vocabulary for the description of web pages. The initiative of several major search engine companies aims to develop a simple vocabulary to add semantic information to webpages. These vocabularies were designed in collaboration with domain experts. For the markup of music information, the data type MusicComposition (http://schema.org/MusicComposition) supplies most elements to describe a work and its parts. To add the possibility of describing music incipits, we have expanded the vocabulary with further elements. The format can be used directly for data interchange - a feature request for the extension of schema.org with incipit declaration is planned.

The metadata format functions as an acquisition format for the repositories. It can be used to add information to the catalogue by adding music incipits to existing re-

source as well as a schema for the annotation and digital publication of analogue catalogues. Moreover, it will provide the aggregated data in a standardised format to enable further usage.

Conclusion

At the moment, IncipitSearch aggregates the incipit data of the catalogue of Gluck's works, the SBN OPAC, the RISM OPAC and includes a sample data set of the thematic Breitkopf Catalogo delle Sinfonie 1762.

IncipitSearch builds on the potential of open musical data and provides the possibility to interlink musicological repositories of various types. This is accomplished by concentrating on musical incipits and using a standardised data interface, a straightforward metadata schema and encapsulated software components.

Through consistent usage of authority control and metadata standards, IncipitSearch is an open source tool and service warranting sustainability, transparency, and accessibility of research data.

External Links

- Europeana: https://europeana.eu
- correspSearch: http://correspsearch.net
- Deutsche Nationalbibliothek (DNB): https://portal.dnb.de
- Freischütz Digital: http://freischuetz-digital.de
- IncipitSearch: https://incipitsearch.adwmainz.net
- Répertoire International des Sources Musicales: https://opac.rism.info
- schema.org: http://schema.org
- Servizio Bibliotecario Nazionale (SBN): http://opac.sbn.it
- Work catalogue of the complete edition of Gluck's works (GluckWV): http://gluck-gesamtausgabe.de

References

Brook, B.S., Gold, M. (1964). Notating Music with Ordinary Typewriter Characters (A Plaine and Easie Code System for Musicke). *Fontes Artis Musicae,* vol. 11, no. 3, 1964, pp. 142–159. www.jstor.org/stable/23504533.

Haft, M., Neovesky, A. and Reimers, G (2016). Digitale Nachhaltigkeit von Forschungsdaten durch Microservices. FORGE 2016. Forschungsdaten in den Geisteswissenschaften: Conference Abstract, pp. 23 -24. https://www.fdm.uni-hamburg.de/ueber-uns/a-nachrichten/aktivitaeten/forge16/praesentationen/programmheft.pdf#page=23.

Typke, R., Wiering, F. and Veltkamp, R.C. (2005). A survey of music information retrieval systems. Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR 2005, pp. 153 - 160. http://ismir2005.ismir.net/proceedings/1020.pdf.

Van Nuss, J., Giezeman, G.-J., Wiering, F. (2017). Melody retrieval and composer attribution using sequence alignment on RISM incipits. Proceedings of the International Conference on Technologies for Music Notation and Representation. Universidade da Coruña, pp. 79–90. http://doi.org/10.5281/zenodo.924135

OCR'ing and classifying Jean Desmet's business archive: methodological implications and new directions for media historical research

Christian Gosvig Olesen
c.g.olesen@uva.nl
University of Amsterdam, The Netherlands

Ivan Kisjes
I.kisjes@uva.nl
University of Amsterdam, The Netherlands

This paper discusses the endeavours of the research project *MIMEHIST: Annotating EYE's Jean Desmet Collection* (2017-2018) - funded by the Netherlands Scientific Research Organisation - to do optical character recognition (OCR) and apply various computer vision techniques on the business archive of film distributor and exhibitor Jean Desmet (1875-1956).

The Desmet collection consists of approximately 950 films produced between 1907 and 1916, a business archive containing around 127.000 documents, some 1050 posters and around 1500 photos. The Collection is unique because of its large amount of rare films from the transitional years of silent cinema, and because of the richness of its business archive which holds extensive documentation of early film exhibition and distribution practices in the 1910s. These features contribute to its immense historical value which was one of the main reasons why it was inscribed on UNESCO's Memory of the World Register in 2011.

By OCRing and classifying Jean Desmet's business archive, MIMEHIST will allow scholars to browse and annotate its documents - all scanned in high resolution - in the new 'Media Suite' of the Dutch national research infrastructure (CLARIAH). The results will be integrated in a search interface enabling media historians to identify word frequencies and topics as a basis for research on early film distribution and exhibition and, the paper argues, open for media historical research which productively builds on and expands the collection's use in previous scholarship.

Throughout the past decades, Desmet's business documents have offered a rich source for socio-economic

cinema history. Media historians such as Karel Dibbets and Rixt Jonkman have studied parts of the collection's (related) data by manually transcribing and organising it into databases (Jonkman, 2007; Dibbets, 2010). This work produced an empirical, quantitative foundation for network analysis of Dutch film distribution and exhibition in cinema's earliest years. However, this research also made evident that the archive is too large and diverse to organise and transcribe manually. A particular challenge is that collection contains many different kinds of documents: personal letters, business letters, records of film rentals, postcards, newspaper clippings, telegrams, scraps of paper with notes, photographs etc. Furthermore, some documents are typewritten, others handwritten.

To allow scholars to research and annotate larger amounts of the archival documents' data in CLARIAH's Media Suite, automated information extraction from the documents seemed challenging yet promising. MIMEHIST took up this challenge by trying OCR, document classification, topic modelling, named entity recognition and other visual and linguistic tools on the set of scans in order to extract as much data and metadata from the individual documents as possible. Different document types required different treatment. For instance, we quickly determined that it did not make much sense to do OCR on a tiny handwritten note, while handwriting detection on the other hand would be possible and could yield productive results on such an item.

Experiments were conducted in visual document classification, visual document analysis and distant reading. Visual document classification was performed by clustering a combination of color and texture histograms derived from the scans. This step was taken mostly because the existing index of the archive is incomplete: it has information on the folders in the archive, which contain the documents, but not the documents themselves. The Media Suite works with individual documents, not folders, so it became necessary to, for instance, discern sub-folder covers from the documents inside.

A second reason to do classification was that each type of document needs a different kind of processing - typed letters can be OCR-ed, but not photos, while handwritten letters could be classified by comparing handwriting styles. By separating different document types it became possible to employ the most effective information extraction tools on them. This procedure also allowed for finding visually similar documents, making it possible for researchers to look for similarities in for instance texture or color.

The typewritten documents were OCR-ed, then classified on the basis of the recognized text in order to differentiate e.g. personal letters from business correspondence. Named entity recognition on the texts provided us with a network of people and places, with links to the letters. Attempts at handwriting recognition on the basis of 'image texture' histogram comparisons provided mixed results, - for the instances where larger samples of a single person's handwriting were available it worked reasonably well, but for handwriting types occurring only a few times the confidence of the classifier was too low and such documents were classified as one of the more frequently occurring types. The results of these steps, in combination with the existing index's metadata, provided a rich enough metadata structure for the use of individual documents in the tool.

In addition to a discussion of these steps, our paper reflects on the results' epistemological implications for future research, by discussing them in relation to previous quantitative approaches to the Desmet Collection. From this vantage point, our paper argues that while previous quantitative studies of Desmet's business documents were premised in the coding and transcription procedures of Cliometrics and *Annales* historiography, MIMEHIST's results nurture exploratory and qualitative research coupled with serendipitous search and annotation procedures focusing also on visual features. Consequently, the paper argues, researchers may to a greater degree than hitherto highlight data contingencies and multiplicity of viewpoints in the Desmet business archive.

The 91st Volume — How the Digitised Index for the Collected Works of Leo Tolstoy Adds A New Angle for Research

Boris V. Orekhov
nevmenandr@gmail.com
National Research University Higher School of Economics, Russian Federation

Frank Fischer
ffischer@hse.ru
National Research University Higher School of Economics, Russian Federation

Introduction

The collected works of Leo Tolstoy were printed and published in 90 volumes of some 46,000 pages between 1928 and 1958. The visibility and usability of these volumes were increased by the project "Tolstoy Digital", a TEI-encoded version of this vast resource (Skorinkin & Mozhaev 2016).

This talk, however, is not about the 90 volumes themselves, but about the 91st volume of this edition, a supplement volume containing indexes of works and proper names, from both the fictional works and the many volumes containing Tolstoy's letters.

"The 91st Volume" is a web application based on the digitised index of proper names for the 90-volume collection of Tolstoy's collected works (http://index.tolstoy.ru/). The digitised data features additional properties,

which can be explored by the enthusiast as well as the specialist.

This talk tries not just to present a new tool for literary scholars, but tries to generalise how this kind of resources can be used to gain new insights into larger text collections.

Level 1: Enhanced Searches

First and foremost, the index retains its original functionality, which is to map names to volumes and pages. Collected works of a canonical writer are not primarily meant to be read one by one, line by line. A 90-volume collection of books does not only contain entertaining narratives, but it can also be viewed as a set of facts, dates, names, mentions, etc. An index is the key to this data, and it was the only means to gain some orientation in the pre-digital age.

In the web app version of the "91st Volume", the index is even more convenient to use than in the paper version, as it allows "fuzzy" searches. By entering "ava" it will list among the results terms like "Poltava", "Bavariâ", or "Abdulla-al'-Mamun Zuravardi". The higher the frequency of a name within the whole collection, the higher up it will be displayed in the results. These types of searches are already an enhancement over the traditional index search.

If we cannot define in advance what we are looking for, we still have the lists of all names in the index (which sum up to more than 16,000 entries). Once we've found what we were looking for, we don't need to remove any book from its shelf and open the right page, but can jump directly to the corresponding page.

A graphical word-cloud representation is also featured and conveys a first idea about the most frequent words in the corpus.

Level 2: Studying Life and Works of Leo Tolstoy by Means of Network Analysis

Turning an index of names into a network is a new approach to facilitate the study of contexts. The co-occurrence of names in the same environment (on the same page, in the same chapter, etc.) reveals similarities and relations between different entities, which on the scale of 90 volumes, helps us to understand larger contexts.

"The 91st Volume" unfolds a rather unconventional social network of Leo Tolstoy. It shows not only Tolstoy's connections with other people (e.g., his pen pals), but also the connections of people from the point of view of Tolstoy.

The co-occurrence of proper names on the same page within the 90 volumes establishes an edge of the emerging network as it creates a link between two entities. For example, the Hindu scripture "Bhagavat-gita" can be found five times on the pages of the Complete Works, and it shares these five pages with a total of 43 other names mentioned. The proximity of these mentionings is not accidental, of course, in our example they form some kind of "Indian cluster" containing works like "Gitopadeša", "Dhammapada", "Vamana Purana", or names like Ramakrišna Šri Paramagamza.

For Tolstoy, the mentioned texts are part of a set of carriers of philosophical knowledge, and are associated with names like Xenophon, Montaigne, Montesquieu, Pascal, Skovoroda, Socrates. These networks provide great opportunities for understanding the whole range of Tolstoy's interests and ideas. It presents a panoramic picture revealing general trends and larger thematic clusters. For each individual name there is also a small graph showing the most significant names associated with it.

Another new kind of access to the 90 volumes is a heat map that shows the density of proper names used in each of them (the more names mentioned, the warmer the colouring).

In the first volume of the collection containing youth experiments, a red splash suddenly appears in the middle of a rather calm blue background on page 269. You can view this page and will find that it contains a list of European cities: Rome, Naples, Dresden, Berlin.

Level 3: Editorial Evolution of the "Complete Works"

The index also allows scholars to study the coming into life of the "Complete Works of Leo Tolstoy", i.e., the difficulties that had to be overcome when working on this edition (as they are laid out in Osterman 2002). The "91st Volume" allows us to understand how editorial principles have changed over time, especially as regards the depth of commenting.

For example, the 13th volume, with draft editions of "War and Peace", has a weak commentary, and the 47th volume (diaries and notebooks) features such detailed comments that it is the most detailed in the entire 90-volume edition. Quantifications like this allow us to draw conclusions to the process of editing the Complete Works over three decades.

Like mentioned above, the web app retains all the capabilities of the traditional index, and at the same time extends its potential through computer-based information management, a multi-purpose search engine and different kinds of visualisations. The app is to be understood as a suggestion to apply the newly developed methods to the Collected Works of other authors.

References

Osterman L. (2002): *The Battle for Tolstoy: History of the Publication of Tolstoy's Complete Works*. [Srazhenie za Tolstogo. Istorija izdanija Polnogo sobranija sochinenij Tolstogo.

Skorinkin D., Mozhaev E. (2016). TEI markup for the 90-volume edition of Leo Tolstoy's complete works. In: *TEI Conference and Members' Meeting 2016. Book of Abstracts*. Vienna: Austrian Centre for Digital Humanities, pp.107–109.

Adjusting LERA For The Comparison Of Arabic Manuscripts Of _Kalīla wa-Dimna_

Beatrice Gründler
beatrice.gruendler@fu-berlin.de
Freie Universität Berlin, Germany

Marcus Pöckelmann
marcus.poeckelmann@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Introduction

In this paper, we present the collaboration between the pilot project *Kalīla wa-Dimna – Wisdom encoded*[1] and LERA.[2] In the project's first phase, devoted to experimenting with existing tools and identifying necessary adjustments, we adopted and generalized LERA for the classical Arabic language. This modification worked well and thus will become a cornerstone for future research within the ERC-Advanced Grant Project "AnonymClassic" (Gruendler 2017).[3] The benefit is already apparent, yielding first observations of the text's development, and the tool was successfully applied in an undergraduate academic course at the Seminar for Semitic and Arabic Studies of FU Berlin.

Kalīla wa-Dimna – Wisdom encoded

Using Sanscrit sources, *Kalīla wa-Dimna* was composed in Middle Persian (version lost) in the 6th century CE and ultimately translated into in forty languages worldwide. Its key phase is the Arabic translation from the 8th century, from which all later translations derive. But this version has proliferated in many variants, which prevents a conventional critical edition by stemma (Gruendler 2013). This project seeks to assess via a (partial) synoptic critical edition the range of variation between selected Arabic manuscripts of this work. These derive from the 13th to the 19th century CE.

LERA – Locate, Explore, Retrace and Apprehend complex text variants

LERA is an interactive, digital tool for analyzing variations between multiple versions of a text in a synoptic manner with several differences to other well known collation tools (Schütz and Pöckelmann 2016). It was first developed for printed texts of the French Enlightenment (Bremer et al. 2015) within the SaDA-project[4] at Martin Luther University Halle-Wittenberg and since then adopted to other texts and languages.

The tool follows three major steps for generating the synopsis. The first is a segmentation of the given manuscripts. In the case of *Kalīla wa-Dimna,* the text-units are narrative steps. Second, an automatic alignment of these segments is done by the built-in algorithm. The researcher can intervene afterwards by moving, cutting or merging the segments if necessary. The final step is the detailed comparison of the aligned segments' words by the system. The identified variants are highlighted with colors depending on the kind of difference. Besides, a variety of filters are available, e.g., hiding all changes that purely concern punctuation. On this basis, a comprehensive and easily readable apparatus is generated. The proto-edition thus produced can be downloaded in several formats.

Moreover, LERA provides further assistance for the analysis of the variants. Most prominent is CATview (Pöckelmann et al. 2015), a graphical representation of the alignment that facilitates overviewing and navigating within the synopsis.[5] It is also associated with the word clouds and search functions of LERA.

Modification of LERA for Kalīla wa-Dimna

In this project, LERA made its debut in classical Arabic, which has required some language-specific adoptions. Processing the Arabic alphabet was rather uncomplicated, because the system already uses Unicode. Regarding the backend, the processes for tokenizing, indexing (for search) and language recognition were extended. On the other hand, modifications for the frontend included adding a font for the alphabet, displaying the correct writing direction, and revising the download function.

More important, some specific needs for the *Kalīla wa-Dimna* project have already been implemented. LERA now allows the manual alignment by experts using unique segment identifiers, which are encoded within the manuscripts' XML/TEI files. On this basis, we also added identifiers for the segments that can be edited and displayed in the synoptic view. On major improvement is the visualization of transposed segments. They occur if the order of the segments within one manuscript was changed or when similar segments appear somewhat distant to each other, but aligning them is blocked due to other aligned segments. We included an option to display copies of them in the synopsis that will be shown grayed and are linked to their actual position. They will also appear in CATview.

1 E-Learning/E-Research project, located at and funded by Freie Universität Berlin, homepage http://www.geschkult.fu-berlin.de/e/kalila-wa-dimna

2 Information and a demonstrator can be found at https://lera.uzi.uni-halle.de

3 No. 742635, "The Arabic Anonymous in a World Classic," PI Beatrice Gruendler, Freie Universität Berlin, see http://www.geschkult.fu-berlin.de/forschung/erc/anonym_classic/index.html

4 See the project's homepage: https://sada.uzi.uni-halle.de

5 Further information on CATview is also available at https://catview.uzi.uni-halle.de

In respect to the project's goal to investigate the interrelation of the manuscripts of *Kalīla wa-Dimna*, we developed two new modes for coloring the variants. The first one only highlights passages that are unique to one manuscript, which points to some independence of the copyist-redactor regarding the other manuscripts. The second mode only highlights those passages that appear in exactly two manuscripts. Finding such pairs is important evidence that suggests that both manuscripts are related to each other.

Another helpful extension is the so called 80%-filter, which leads to treating words as identical if they share at least 80% of their letters. This approximating similarity measure is grounded in the property of the Arabic language that words with identical roots tend belong to one semantic field.

Benefits

LERA could be adjusted for the first phase of this interdisciplinary collaboration. Based on this, we already made interesting observations: against our assumption, the first analysis shows that there are no distinct groups of manuscripts. Instead, variations fluctuate, forming continua in which some manuscripts cumulatively assemble reformulations that appear scattered among others.

Furthermore, in the summer semester of 2017 the project was integrated into an undergraduate academic course on *Kalīla wa-Dimna* at Freie Universität Berlin. The students used the synopsis to explore the variants of five aligned manuscripts in class and wrote papers applying this method individually.

Conclusion

With the work presented here, we established a foundation for a comprehensive analysis of *Kalīla wa-Dimna*. Owed to the text's complex history and manifold variants, this ambitious project is planned for a timespan of ten years. With the ongoing research, more features of analysis will be needed. This includes an advanced utilization of language specific information for the comparison, e.g., a root extraction for Arabic words. Moreover, the comparison and visualization of manuscripts of *Kalīla wa-Dimna* in other language is being considered. Finally, the functionality to comment on the identified variants is crucial for their scientific investigation.

References

Bremer, T., Molitor, P., Pöckelmann, M., Ritter, J. and Schütz, S. (2015). "Zum Einsatz digitaler Methoden bei der Erstellung und Nutzung genetischer Editionen gedruckter Texte mit verschiedenen Fassungen - Das Fallbeispiel der *Histoire philosphique des deux Indes* von Guillaume Thomas Raynal." In v. Nutt-Kofoth, R., Plachta, B. and Woesler, W. (eds), *Editio,* 29(1), pp. 29–51.

Gruendler, B. (2013). "Les versions de *Kalīla wa-Dimna*: une transmission et une circulation mouvantes." In Ortola, M. (eds), *Énoncés sapientiels et littérature exemplaire: une intertextualité complexe*. Nancy, pp. 385-416.

Gruendler, B. (2017). "The Arabic Anonymous in a World Classic (Acronym: AnonymClassic). Presentation of a Research Project." *Geschichte der Germanistik,* 51/52, pp. 156-57.

Pöckelmann, M., Medek (⁎Gießler), A., Molitor, P. and Ritter, J. (2015) "CATview - Supporting the Investigation of Text Genesis of Large Manuscripts by an Overall Interactive Visualization Tool Digital Humanities." *Digital Humanities 2015: Conference Abstracts.* Sydney: UWS. http://dh2015.org/abstracts/xml/POCKELMANN_Marcus__CATview____Supporting_The_Inve [accessed 11/27/2017]

Schütz, S. and Pöckelmann, M. (2016) "LERA - Explorative Analyse komplexer Textvarianten in Editionsphilologie und Diskursanalyse." In: *Book of abstracts of the third annual conference of Digital Humanities for German-speaking regions (DHd 2016)*, pp. 249-253.

Afterlives of Digitization

Lily Cho
lilycho@yorku.ca
York University, Canada

Julienne Pascoe
julienne.pascoe@gmail.com
Library and Archives Canada; Canadiana.org, Canada

This paper is based on our commitment to the possibilities of re-thinking the processes of digitization such that digitization does not end with the uploading the scanned object and archivally-mandated metadata. Rather, that point is merely the beginning of the life of any particular digital collection. The ways that any collection is used by academic researchers, community groups, and members of the public should contribute to the processes of digitization. These collections live when they are used and these uses should be reflected in the collection so that other researchers can see and build on this work. Every collection that has been digitized has an afterlife. But how can we use new technologies – in particular, the International Image Interoperability Framework (IIIF) and linked data – in order to make these afterlives visible and usable? How can we develop infrastructure and protocols so that the metadata *lives*?

Our project focuses on building a platform for annotations based around a specific collection of images: the Chinese Immigration records, initially captured by Library and Archives Canada (LAC), and subsequently digitized, preserved and made accessible online by Canadiana.

There are approximately 41,000 images in this particular set of archival images that we work with. Canadiana has recently completed the digitization of this collection. Because it is comprised of a nearly complete set of immigration certificates for individual Chinese migrants collected between 1910 and 1953, the collection is particularly rich for researchers working in the area of race, immigration history, and citizenship.

In working with these materials, Lily Cho's research team has identified several layers of annotations that would be pertinent to this material. For example, the research team has transcribed names of each immigrant on the record. Each image contains two names: the anglicized Chinese name written down by an immigration agent, and a name in Chinese script written by the migrants themselves. In our transcriptions of the first several hundred images, there is no correspondence between the name written by the immigration agent and the name in Chinese script. Because these records were used to identify individual immigrants for the purposes of allowing them to exit and enter Canada (and thus functioning much like a passport for Chinese immigrants who were, during this period, denied the rights of citizenship), this finding radically changes our understanding of how Chinese immigrants navigated racist immigration controls during this historical period. However, there is currently no way for her research team to contribute to the metadata already attached to this collection.

Such contributions to the metadata already in place function as annotations in this project. Working in partnership, Cho and Julienne Pascoe, who has been the Lead Metadata Architect for Canadiana and is now serving as a Digital Archivist at LAC, are developing a platform for supporting annotations for this archive using the Web Annotations standard, the IIIF, and linked data. Canadiana is currently in the process of implementing IIIF as well as the initial stages of developing a data model that would provide the foundation for such a partnership. In short, this project uses IIIF as a framework for enabling open standards for annotations that can then be reused as linked data - all three areas coming together to support the linking, sharing and re-use of metadata.

This paper reports on the progress we have made in developing this platform, and will also briefly outline the possibilities for the use of this platform beyond this specific collection of images. Although museums and archives are under enormous pressure to digitize their collections, and are rapidly in the process of doing so, these digitization initiatives are rarely undertaken in conversation with some of the primary users of these digitized texts and objects: academic researchers. For example, metadata that meets archiving standards is not necessarily useful for researchers, and is often based on hegemonic archival practices that reinforce colonial structures and narratives. At the same time, academic researchers often have resources to contribute to, and enrich, the digitization that has been accomplished as well as facilitate postcolonial interpretations of the archive. This project brings academics and digital archivists together in order to develop protocols so that digitized collections can be dynamically connected to the communities using them. Once digitized, a collection does not need to remain static. It can respond to, and include, the findings of researchers in the community; and these findings could and should be made available to other users of the collection. However, protocols for curating, organizing, and disseminating this information must be developed. This project will use one specific collection, an archive of approximately 40, 000 head tax certificates held by LAC and digitized by Canadiana, as a test case for developing precisely the kinds of protocols that would allow a digitized collection of materials to leverage the findings emerging from people using these materials.

Rapid Bricolage Implementing Digital Humanities

William Dudley Pascoe
bill.pascoe@newcastle.edu.au
University of Newcastle, Australia

This paper presents a practical approach to building digital humanities (DH) at a university, across disciplines with diverse requirements, starting without institutional support, with scarce staff on a low budget. Examples are provided from the Centre For 21st Century Humanities (C21CH) at University of Newcastle, Australia (UON).

Digital humanities (DH) requires expertise that crosses many fields from specific humanities disciplines to software development and production management. DH has a broad range in scale – from a scholar learning basic programming to hack a Python script, to multi-institutional collaborations on neural network learning. Few people are experts in all these fields meaning DH is often a collaboration. The requirements for any individual DH project can differ greatly also requiring IT skill sets that may not be easy to find in any one individual. This makes it difficult for university humanities departments with no spare cash, and often reluctant to invest heavily in IT, to successfully support DH, yet DH projects present problems beyond standard service offerings and provisioning and different to STEM. The Digital Lab of C21CH at UON has evolved an approach, here called 'rapid bricolage', that has successfully delivered a range of sustained internationally recognised DH projects influencing national debate. Some comparison will be made with other approaches, and while not necessarily suiting all circumstances, 'rapid bricolage' has proved an effective approach catering to characteristic issues in DH research, drawing from but differing to established IT practice.

This 'rapid bricolage' approach draws on 'rapid' software development and 'bricolage', both common practice in software development and humanities, but modifies them to meet the unique needs of Digital Humanities. These modifications are epistemic, structural, methodological and a matter of degree. It has also crucially involved consultative processes to identify and Pareto prioritise inter-disciplinary interests and achievable, feasible, high impact projects. The success of these feeds back to build interest and support for DH towards funding and growth, and results in project driven infrastructure, bridging the gap between projects without infrastructure and infrastructure without projects by beginning with demonstrable utility and developing with shared human and technical resources.

C21CH projects include (http://c21ch.newcastle.edu.au):

- Colonial Frontier Massacres (v1.2) – map of massacres in Australia.
- EMWRN archive (v2) – innovative archive of material cultures of early modern women's writing.
- Intelligent Archive (v3beta) – stylometry software.
- ELDTA site (v1alpha) – linguistics web player for media with tiered glosses, translations etc.
- Text To Map (prototype) – online automatic recognition and mapping of places in texts, linking to and from the text and the map.
- Scriptopict (v1) – annotations for images eg: *Battle of Kurukshetra Mural* and *Mixtec Glyph.*

Rapid

Rapid application development is an established methodology for software development focusing on getting a prototype working as early as possible followed by regular review with clients and incremental feature additions and bug fixes. For humanities departments this approach ensures that at least some software exists as an outcome of initial spending when the budget is tenuous and provides an encouraging proof of concept. For the cost of a meeting with several professors or executives a working prototype can be developed, making it worth simply trying it out rather than lengthy discussions about the value of proceeding. A rapid approach also helps greatly when the client is unclear of what is needed or has little understanding of IT. An early prototype establishes confidence and commitment early. Gaps in desired functionality immediately become clear through interaction. In particular because research is heuristic and highly changeable it allows for speculative requirements to change as the project progresses. Because of this, an even more rapid than usual approach is suited to humanities research because, as research, not all requirements cannot be known in advance. The process necessarily involves taking some action with ongoing revision, addition and enhancement. Not all aspects of humanities research activity, such as thorough rumination on a nuanced argument on a complicated problem, fit this 'rapid' model, but:

- the speculative, heuristic activities necessary to research are enhanced;
- some slower methodical activities essential for rigour and completeness can be sped up, sometimes making research possible that otherwise would not have been, or improved through the need for clear structures and definitions;
- the 'slow' process of rumination, of considering complex problems and developing arguments, while irreplaceable, can be augmented.

Bricolage

Bricolage is a well-established approach in software development. Software is typically put together from pre-existing libraries, frameworks and cut and pasted code that is modified and added to, to produce something that works in ways that conventional intellectual property and copyright are not practically applicable to. This approach is in sympathy with developments in critical theory in the late 20th century and after, with 'bricolage' and the problematization of authorship being major themes in describing postmodernity and in contemporary humanities methodology. Just as a very rapid approach suits humanities research so too is bricolage especially suitable for DH.

As research, DH typically requires constant and regular modification and adjustment, rather than delivery of a working system according to contracted specification. Much software is developed for a STEM or commercial purpose, or has a STEM like approach to problem solving. STEM and the commercial sector have larger budgets and devote larger budgets to software. This means that humanists are often in a pragmatic situation of re-using software from different disciplines despite having divergent requirements. Humanities often focus on complexity, exceptions, structural change and highly contingent historical (not repeated) events, while STEM and commerce focus on systemisation, normalisation, ceteris paribus and repeatability, for example. If humanities researchers are to avoid fitting research to the software limitations this means constantly adapting systems to their own different epistemic, ontological and methodological paradigm, ie: bricolage.

The DH research need for these two approaches, rapid application development and bricolage, combined in extremis presents challenges to established IT practice. These challenges can be met with appropriate staffing, strategy and a 'rapid bricolage' approach to build DH at a University despite diverse demands and resourcing adversity.

References

Craig, H., Pascoe, W. (2018). *Intelligent Archive v3.0* Newcastle: Centre For 21 Century Humanities

Ryan, L., Debenham, J., Brown, M., Pascoe, W. (2017). *Colonial Frontier Massacres* Newcastle: Centre For 21 Century Humanities http://hdl.handle.net/1959.13/1340762

Smith, R., Pender, P., Pascoe, W. (2017). *Early Modern Women's Research Network Digital Archive* Newcastle: Centre For 21 Century Humanities http://hdl.handle.net/1959.13/1326860

The Time-Us project. Creating gold data to understand the gender gap in the French textile trades (17th–20th century)

Eric de La Clergerie
eric.de_la_clergerie@inria.fr
ALMAnaCH, Inria, France

Manuela Martini
manuela.martini@univ-lyon2.fr
LARHRA, Université Lyon 2, France

Marie Puren
marie.puren@inria.fr
ALMAnaCH, Inria, France

Charles Riondet
charles.riondet@inria.fr
ALMAnaCH, Inria, France

Alix Chagué
alix.chague@enc-sorbonne.fr
ALMAnaCH, Inria, France

The role of women in industrial development is now largely recognized in sociological and economic studies on developing countries during the first industrial revolution in Europe. Yet data on their remuneration, schedules and domestic work, and that of men working in the same sectors, remains deficient for many regions, especially for France. The Time-Us project aims to reconstruct the remuneration and time budgets of women and men working in the textile trades in four French industrial regions (Lille, Paris, Lyon, Marseille) in a long-term perspective, by bringing together a multidisciplinary team of historians, natural language processing (NLP) experts and sociologists. It will create comparable series on the remuneration and time allocation of employed men and women (i)through classical sources and company and trade association archives, and (ii)by piecing together a series of qualitative sources identifying words and actions associated with work in both domestic and non-domestic activities. The project will provide keys to understanding the gender gap by analyzing changes in work and time uses during the first industrialization process.

The Time-Us team works on a heterogenous corpus of French handwritten and printed sources spanning from the seventeenth to the twentieth century. These documents are mainly preserved in French local archives, from the four industrial regions that have been mentioned above (for instance, Archives municipales de Lyon, Bibliothèque municipale de Lyon, or Bibliothèque nationale de France in Paris, etc.). The analyzed corpus brings together numerous historical sources, and includes court decisions, petitions, police reports and files, and sociological surveys on living conditions of the working class (especially *Les monographies de famille de l'École de Le Play* or Le Play's families' budgets (Hincker, 2011)). Many of these documents are manuscripts, written by various hands over long periods of time (more than a hundred years for the "Registre de contraventions aux règlements des métiers" that begins in 1670 and ends in 1781 (Lyon, Archives municipales).

This unpublished set of documents constitutes an important corpus of historical sources that is well-suited for applying computational analysis. In this paper, we will present the approach adopted by the Time-Us team to analyze this corpus. We will also discuss the prospects opened up by this project for historical research in terms of digital research workflow.

Our goal consists in applying NLP methods to heterogeneous historical documents, in order to identify and analyze the relevant semantic or syntactic patterns that describe work, remuneration and time budgets. The application of such methods, mainly parsing, will facilitate the analysis of the corpus by creating series of comparable quantitative and qualitative data:

Quantitative data on remunerations, household budgets and time spent for domestic (or unpaid) and non-domestic (or paid) work by women and men.

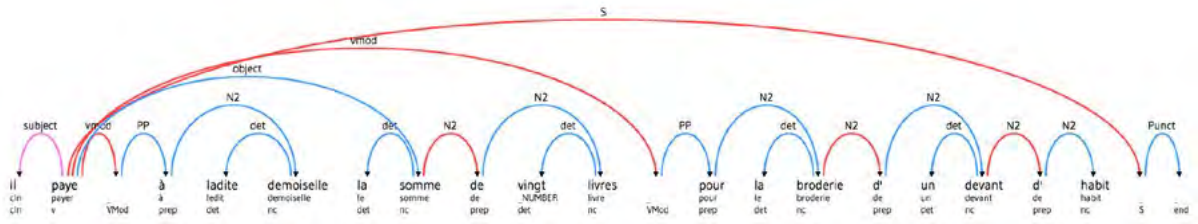
Qualitative data on paid and unpaid tasks realized by women at home and at work, namely information on the type of the task, its description, its duration and its results. Computational methods will also be used to extract statements describing the women performing these tasks (occupation, social status, age, marital status, family composition), and the relationships between the actors involved in these tasks, especially between men and women (family relationships such as husband and wife, brothers and sisters mothers and sons, or working relationships such as employers and employees).

The analyzed sources can take a number of varied forms. Thus, we chose to work closely with economic and labour historians in the data modelling process. As the corpus gathers together diverse historical sources, the definition of a light and flexible annotation schema, bringing together the history and language processing experts, is a major step to create "gold data" to train parsing models. This gold data take the form of annotated texts encoded in TEI (*Text Encoding Initiative*). TEI can be seen as

a bridge representation for historians and NLP experts: in this approach, historians annotate a first set of documents in TEI, in order to create training data that can be easily processed and analyzed by NLP experts. Besides, the choice of the TEI allows for the creation of sustainable data, that can be reused in the long term by other projects and researchers. Our aim consists also in creating a flexible TEI data model that will be relevant to modelize differents types of data, and that will enable NLP experts to extract comparable information such as quantitative data (amounts of money, period of time…). In this way, this model could be reused by other research projects especially, but not only, projects of economic and labor history.

A first step is the transcription of the manuscripts into a simple TEI representation, covering the text and a set of metadata. This task is nothing but trivial, due to the diversity of sources mentioned above, but it is not the scope of this paper. Then, the representation is enriched by annotation layers. The first annotation layer is the recognition of tasks and occupations, linked to their associated amounts of money, and the actors of the transaction. The extraction of Named Entities such as person and place names is also necessary in order to properly analyse how gender and localization influence remuneration.

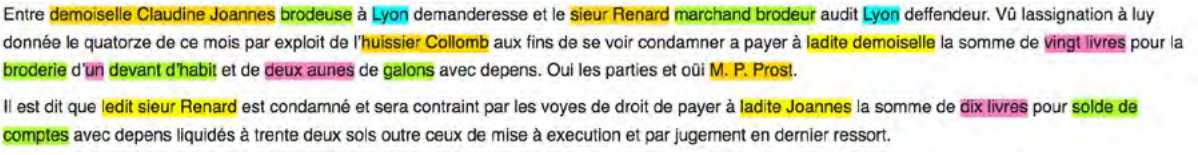
The annotation process will start as a collaborative effort, in order to get a first dataset that could possibly be used to train/configure NLP tools, but also to help designing a precise annotation guide between the NLP people and historians. At a later stage, we will progressively deploy more automatic NLP tools to create these annotations. In this regard, we plan to identify the elements of vocabulary (tasks, products) and the interesting phrases (e.g."*someone was paid (this amount) for (this product) for a (given amount of time)*"), using knowledge acquisition techniques based on the distributional hypothesis and syntactic analysis of the corpus. The knowledge of the domain will allow us to define syntactic extraction pattern to be applied on the corpus to detect and annotate specific instances of tasks, products, money, people and relationships between these pieces of information. Some human validation will still be needed to filter the vocabulary, refine the patterns, and propose missing elements (vocabulary and patterns). Language processing will be conducted with the French processing chain developed by the INRIA Almanach team, and in particular with the FRMG parser (Morardo and de La Clergerie 2014). Parsing produces dependencies between words, allowing us to identify who does what, when, how for some event. The processing chain has already by used several times for knowledge acquisition over specific domains (legal, medical). In our case, specific issues may arise because of the quality of the transcriptions and the peculiarities of the language used, which contains archaic constructions, whereas our parser was designed for contemporary French.



Example of a parse for one sentence of the corpus

The annotation task is therefore mainly collaborative, so the need for a shared framework has emerged. Several digital projects have already taken into account the specific needs of historians in terms of image visualization, transcription and collaboration. For instance, the *Transkribus* interface enables Humanities scholars to transcribe handwritten and printed historical sources, and offers a very powerful Handwritten Text Recognition engine. The project *Transcribe Bentham* takes account the collaborative dimension in transcribing historical documents. The *Old Bailey* transcription project uses a combination of hand encoding an automatic recognition and extraction systems. Nevertheless, they do not address all the requirements of Humanities scholars working on primary sources, and the need of comprehensive Digital Humanities-based publishing systems is emerging. We have chosen to setup a specific digital workflow enabling historians and NLP experts to work together. We will present the solution that has been put in place, and especially a customized wiki with:

the Transcribe Bentham transcription desk, adapted to our needs,
and a TEI toolbar, specifically customized for tagging named entities and measures.


Entre demoiselle Claudine Joannes brodeuse à Lyon demanderesse et le sieur Renard marchand brodeur audit Lyon deffendeur. Vû lassignation à luy donnée le quatorze de ce mois par exploit de l'huissier Collomb aux fins de se voir condamner a payer à ladite demoiselle la somme de vingt livres pour la broderie d'un devant d'habit et de deux aunes de galons avec depens. Oui les parties et oüi M. P. Prost.

Il est dit que ledit sieur Renard est condamné et sera contraint par les voyes de droit de payer à ladite Joannes la somme de dix livres pour solde de comptes avec depens liquidés à trente deux sols outre ceux de mise à execution et par jugement en dernier ressort.

Customization of the TEI toolbar

References

Clergerie, É. D. L., Sagot, B., Stern, R., Denis, P., Recourcé, G. and Mignot, V.
(2009). Extracting and Visualizing Quotations from News Wires. vol. 6562. Springer, pp. 522–32 doi:10.1007/978-3-642-20095-3_48. https://hal.inria.fr/inria-00607463/document (accessed 24 April 2018).

Hincker, L. (2001). Les monographies de famille de l'École de Le Play. Les Études sociales, n 131-132, 1er et 2e semestres 2000. *Revue d'histoire du XIXe siècle. Société d'histoire de la révolution de 1848 et des révolutions du XIXe siècle*(23): 274–76.

Morardo, M. and Clergerie, É. V. de L. (2014). Towards an environment for the production and the validation of lexical semantic resources. https://hal.inria.fr/hal-01005464/document (accessed 24 April 2018).

Seaward, L. and Kallio, M. (2017). Transkribus: Handwritten Text Recognition technology for historical documents. Montréal https://dh2017.adho.org/abstracts/649/649.pdf (accessed 24 April 2018).

Thomasset, F. and Clergerie, É. D. L. (2005). Comment obtenir plus des Méta-Grammaires. *Proceedings of TALN'05*. Dourdan, France.

University College London UCL Transcribe Bentham http://blogs.ucl.ac.uk/transcribe-bentham/ (accessed 24 April 2018).

Old Bailey Online - The Proceedings of the Old Bailey, 1674-1913 - Central Criminal Court https://www.oldbaileyonline.org/ (accessed 26 April 2018).

Modeling Linked Cultural Events: Design and Application

Kaspar Beelen
k.beelen@uva.nl
University of Amsterdam, The Netherlands

Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, Netherlands, The

Julia Noordegraaf
j.j.noordegraaf@uva.nl
University of Amsterdam, The Netherlands

Harm Nijboer
harm.nijboer@huygens.knaw.nl
Huygens Institute for the History of the Netherlands, The Netherlands

Thunnis van Oort
t.vanoort@uva.nl
University of Amsterdam, The Netherlands

Claartje Rasterhoff
c.rasterhoff@uva.nl
University of Amsterdam, The Netherlands0

Introduction

This paper discusses the promises and pitfalls of linking historical data on cultural events. Quite a few datasets on historical European music, theatre and film are now publicly available online (Baptist 2017). The ones that contain programming information are, at least to some extent, already event-based. However, they are highly heterogeneous in scale and scope, and they generally do not use the same definitions for, for example, venues, events, or companies. Conceptualizing and embedding cultural events such as concerts or theatrical performances in a linked data framework helps to overcome such issues without forcing an overarching ontology, and it enables researchers to acknowledge the performative and interactive nature of cultural expressions within their (local) societal context (Nijboer and Rasterhoff 2018).

By linking event data internally as well as to external knowledge bases such as DBpedia and Wikidata by means of shared vocabularies, researchers are invited to systematically analyse cultural life cross-sectorally (i.e. theatre, music), internationally (European comparisons and connections), and contextually (in relation to local social, economic, political and cultural features) (cf. EPAD: European Performing Arts Dataverse). In this paper we discuss the conceptual and practical requirements for such a linked-data approach on the basis of a series of research projects on historical cinema, musical, and theatrical events in the research program Creative Amsterdam: An E-Humanities Perspective (CREATE).

Cultural events

Events play a key role in historical scholarship, and have gained even more urgency with the increasing importance of digital resources in humanities research. Many projects on historical events, however, employ them as devices to structure data collections and do not explicitly aim to develop analytical frameworks in relation to event data collection and data modeling (De Boer et al. 2015; Van Hage et al. 2011; Shaw 2013. An exception can be found in a statistical method known as event history analysis, which treats events as dependent variables, seeking to statistically describe, explain, or predict their occurrence (Allison 2004). Most research on (urban) arts and culture, however, does not try to statistically identify variables that predict or explain an event, for example the staging of the opera Norma or the screening of the movie Casablanca. Rather, historians may seek to identify (series of) events that have contributed to, for example, the canonical status of specific expressions or genres, to the shaping of local and international cultural taste cultures, or to the emergence of some places as particularly creative and cultural.

We therefore emphasize that (networks and series of) events should also be considered as independent variables that can help us identify and disentangle processes

of cultural change and continuity. Central in this view is the assumption that 1) events can be seen as units of analysis with structural properties (notably, a time and place) with links to, for example, actors, institutions, other events, and local properties, and 2) that these interlinkages are key to analysing their role in shaping, for instance, local cultural or social life (Tilly 2002). Turning individual event datasets into linked data versions would provide instantaneous insight into how much performing arts datasets overlaps, ontologically, with any of the others. This provides a roadmap for integrating these still scattered data and studying them in conjunction. A systematic analysis of cultural events therefore requires a data structure which allows for querying connections.

Linking cultural event data

A first analysis of performing arts datasets demonstrated that normalizing even the most basic data across datasets is tricky and that trying to completely harmonize and link all the relevant datasets is futile (Baptist 2017). Fortunately, the structure of linked data provides a way to transparently query heterogeneous data, without enforcing an overarching ontology. Breaking events down into variables such as 'people', 'venues', 'place', and 'time', for instance, circumvents the issue of formally defining a 'performance'. Linked data also allows researchers to test various different link-ups of two data sets so they can evaluate the results when they adjust their queries. In the case of cinemas, for example, one of the problems is that the typology of cinemas differs across countries and periods. In the Netherlands cinemas are divided into types 'A' and 'B' according to frequency of screenings; in Flanders the cinemas are classified according to how soon they tend to new films after their premiere. If the data was put into a relational database it would be necessary to 'reconstruct' either of the classifications for the other dataset. But linked data, because of its model of loose connections, allows querying both datasets, defining a classification only during the query.

For the datasets on cultural events such as historical musical and theatrical performances we build on a rigorous relational data model by Karel Dibbets et al. for the Cinema Context database (Van Vliet et al. 2009). All movies (often circulating under various titles), persons and companies in in this dataset have been identified and aligned to a master record, and where possible linked to the well known and well maintained Internet Movie Database (IMDb). We develop this approach for other datasets and by linking data on cultural events to other datasets and to other knowledge bases using shared vocabularies such as schema.org and Vocabulary of a Friend (VOAF). In this paper we illustrate research potential, but also practical issues by discussing a recent project on the establishment of movie theatres in the city of Amsterdam in the early twentieth century. By linking data on the history of cinema and movie-going to local contextual data (e.g. census data, municipal election data), we assess how linked data might be used to analyse how specific local historical characteristics shaped form and function of urban cultural life.

References

Allison, P. (2004). Event History Analysis. In Hardy, M. and Bryman, A. (eds.), *Handbook of Data Analysis*. Sage Research Methods, pp. 369-385

Baptist, V. (2017). Mapping European Performing Arts Databases. Presentation at the symposium *European Performing Arts Dataverse*, 9 November 2017, Amsterdam. http://www.create.humanities.uva.nl/epad

Cinema Context. http://www.cinemacontext.nl

De Boer., V., Oomen, J., Inel, O., Aroyo, L., Van Staveren, E., Helmich, W., De Beurs, D. (2015). DIVE into the event-based browsing of linked historical media. *Journal of Web Semantics*, 35(3), 152-158

European Performing Arts Dataverse (EPAD). http://www.create.humanities.uva.nl/epad

Nijboer, H. and Rasterhoff, C. (2018). Linked cultural events: Digitizing past events and its implications for analyzing and theorizing the creative city. In Münster, S., Friedrichs, K., Niebling, F. and Seidel-Grzesińska, A. (eds.), *Digital Research and Education in Architectural Heritage. 5th Conference DECH 2017 and First workshop UHDL 2017*, Dresden, Germany, 30-31 March 2017, Springer CCIS series, pp. 22-33

Tilly, C. (2002). Event Catalogs as theories. *Sociological Theory* 20(2), 248-254

Shaw, R. (2013). A Semantic Tool for Historical Events. In *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. Atlanta, Georgia, 14 June 2013, pp. 38–46

Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). *Web Semantics*, 9(2), 128-136

Van Vliet, H., Dibbets., K., Gras, H. (2009). Culture in Context: Contextualization of Cultural Events. In Ross, M., Grauer, M., Freisleben, B. (eds.), *Digital Tools in Media Studies: analysis and research*.Transcript Verlag: Bielefeld, pp. 27-42

Bridging Divides for Conservation in the Amazon: Digital Technologies & The Calha Norte Portal

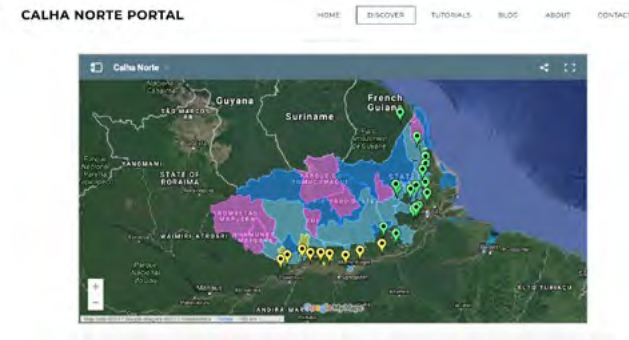
Hannah Mabel Reardon
hannahmreardon@gmail.com
McGill University, Canada

Calha Norte is the northernmost region of the Brazilian Amazon, and constitutes the largest mosaic of protected

areas in the world, encompassing nearly 14 million hectares of land. It stretches from the Amazon river in the south to the frontier regions between Brazil and the Guianese and Surinamese borders in the north, encompassing half of Pará and the state of Amapá in its entirety. Due to the vastness of the area, enforcement of protected areas can be poor, and far too little is done to involve local communities in the decision-making processes that inform conservation policy. Increasingly, digital technologies are helping to overcome some of the challenges to conservation.

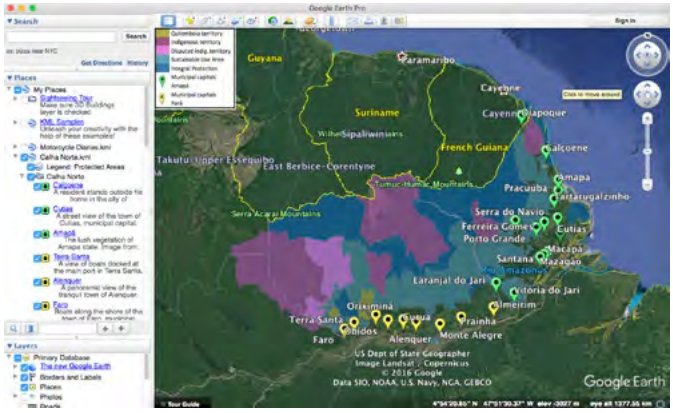
In the past ten years the Amazonian Institute for Man and the Environment (Imazon), an environmental NGO based in Belém, has made a name for itself by developing a "logging feasibility map" (Souza et al., 2010). The feasibility project uses GIS mapping software to chart data on transportation networks (including existing roads and navigable rivers), topography, biodiversity figures, deforestation areas, conservation zones and timber processing facilities. The combination of this data allows for reliable predictions of at-risk areas for deforestation, which has led to better zoning of different areas for community use, conservation and even tourism. By charting the data in a format easily interpreted by stakeholders, the feasibility maps produced by Imazon have become invaluable tools for raising public awareness and building consensus between enforcement officials, regulatory agencies, loggers, conservation groups, and local communities.

Given the immensity of the Calha Norte region and the isolation of many of its inhabitants, communication and monitoring are pressing challenges for authorities charged with the enforcement and protection of conservation zones. Initiatives like Imazon's feasibility map offer innovative solutions at a relatively low cost. Digital tools such as GIS mapping, data analytics software and tele-communications technologies are helping to bridge gaps in the knowledge and understanding of the ecological, political, and social realities of the region, thereby improving predictors of at-risk zones and response times to threats. This short paper presentation will discuss some of the ways in which digital technologies are helping environmentalists overcome the challenges of conservation

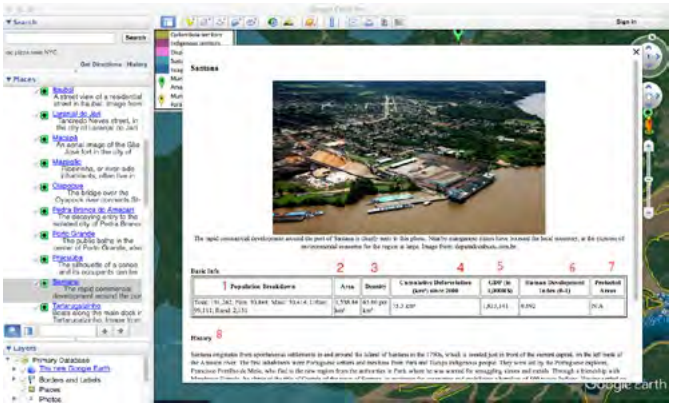


Ex. 1 & 2: Screenshots from CalhaNortePortal.com. Top: The online version of the map application, located on the "Discover" page. Bottom: The CalhaNortePortal.com homepage

The data used for Calha Norte Portal was gathered during my work with the Social Policy department at Imazon. In accordance with the department's focus on communities in the Calha Norte region, I compiled data from various sources about the region's history, cultural diversity, transportation networks, governing bodies, development indices, demographics, economic activities, protected area implementation, and accessibility. This data was then plotted in the creation of the Calha Norte Portal.



Ex. 3: Calha Norte in the Google Earth map. This application allows the user to navigate through the region by clicking on protected areas, indigenous territories, maroon communities and municipal capitals to access historical information, demographic statistics, economic and political data, photos, deforestation figures and an implementation index for protected areas. Users can also look back in time at satellite images from 1960 to the present and visualize patterns of deforestation, and urban sprawl over time.



Ex. 4: Example of the pop-up window for the municipality of Santana, Amapá.

The Calha Norte map focuses mostly on political, economic, historical, cultural and social data for populations in protected areas and municipalities. As an anthropologist, I am particularly interested in dispelling the myth of Amazonia as an uninhabited biological entity. Advocating for conservation policy which involves the participation of local communities has been the focal point of recent critical reports from the International Union for the Conservation of Nature (IUCN) (Cisneros and Orellana, 2017; Orellana, 2017). These reports reflect a general shift in attitudes on the management of protected areas away from traditional approaches which sought to isolate designated conservation zones and often neglected the history of interactions between local populations and the land in question.

My motivation in developing the Calha Norte map was to bridge some of the divides between disciplines in conservation studies. Inspired by Susanna Hecht's political ecology approach to the study of the Amazon (Hecht and Cockburn, 2010; Hecht, 2013), I wanted to create an interdisciplinary platform which would incorporate elements from the fields of anthropology, political science, history, sociology and geography. By breaking down the data I collected on the Calha Norte region in a visual, interactive format I hoped to facilitate an engagement with the region's political, social and cultural characteristics, and emphasize the importance of tailoring environmental policy to the realities of the region's inhabitants. The stand-alone map in Google Earth is meant to be played with, manipulated and explored, in ways that dismantle the linear narrative format of most textual information on the area and incorporate elements of critical cartography studies. Crampton and Krygier state that critical cartography demonstrates its political nature by "linking geographic knowledge with power" (2005: 1), suggesting that a democratization of mapping tools through digital technologies can also result in new avenues for democratizing political power. My project on the Calha Norte region aims to engage with this idea by reformulating knowledge of the region with a focus on mapping its social, cultural and historical characteristics. By advancing a more holistic vision of human interactions with the environment, I hope to make an argument for conservation policy which advocates for greater local-level management of natural resources by the inhabitants of designated protected areas.

Beyond this engagement with the critical cartography literature, my short paper presentation will also raise questions on open source access to information and digital technologies related to environmental issues in the Amazon. Particularly, my paper focuses its commentary on the complexities of conservation in the region, and the importance of greater transparency in the creation and management of protected areas, indigenous territories and traditional community lands. In relation to the over-arching theme of the conference, I hope that my presentation will demonstrate the immense potential for digital technologies to bridge divides of communication and understanding between institutional bodies, environmentalists, policymakers and the communities they serve, as well as bridging gaps in knowledge for scholars and researchers of the Amazon region.

References

Coronel Cisneros, M. y. Solórzano Orellana. J. (2017). *Comunidades locales y pueblos indígenas: Su rol en la conservación, mantenimiento y creación de áreas protegidas,* Quito: Iniciativa Visión Amazónica, REDPARQUES, WWF, FAO, UICN, ONU Medio Ambiente.

Crampton, J. and Krygier J. (2005). An Introduction to Critical Cartography. *ACME: An International Journal for Critical Geographies,* 4(1), pp. 11-33.

Hecht, S. (2013). *The Scramble for the Amazon and the "Lost Paradise" of Euclides da Cunha.* Chicago: University of Chicago Press.

Hecht, S. and Cockburn, A. (2010). *The Fate of the Forest: Developers, Destroyers and Defenders of the Amazon, Updated Edition.* Chicago: University of Chicago Press.

Reardon, H. (2018). *Calha Norte Portal.* [Online] Available at: calhanorteportal.com

Souza, C. J., Brandão, A. J. and Lentini, M. (2010). The feasibility of logging in the Pará Calha Norte region of the Brazilian Amazon. In: *Mapping Forestry.* Redlands(California): ESRI Press, pp. 1-5.

Solórzano Orellana, J. (2017). *El aprovechamiento de los bienes comunes en los bosques amazónicos: Impactos económicos, sociales y culturales de la creación y funcionamiento de áreas protegidas en dos paisajes amazónicos fronterizos,* Quito: Iniciativa Visión Amazónica, REDPARQUES, WWF, FAO, UICN, ONU Medio Ambiente.

Measured Unrest In The Poetry Of The Black Arts Movement

Ethan Reed
ecr6nd@virginia.edu
University of Virginia, United States of America

Introduction

"Anger is loaded with information and energy," says Audre Lorde ina 1981 speechon its political uses—but the nature of this affective information, sparked by a given political present, becomes highly vexed when articulated through literary objects (Lorde, 1997: 280). On the one hand, the cool detachment of aesthetic mediation keeps the politics of experimental works from being seen as mere propaganda, but runs the risk of appearing elitist or self-indulgent. On the other hand, the red-hot political outrage of a protest poem by Amiri Baraka or Sonia Sanchez grounds itself in the present, but may be attacked for subordinating aesthetic sophistication to political agendas.

Building on recent scholarship (like the work of Lauren Berlant and Sianne Ngai) suggesting that feeling gives structure to cultural formations, my research investigates the provocation and articulation of emotions like frustration, anger, and discontentment within recent US literary history as they relate to systemic injustice. An agitprop play that ends with shouts for workers to unite in class revolution; a poetic broadside that vents frustrations against white supremacy in America; a novel that indulges in a revenge fantasy against America's colonial history. Unlike plays, poems, or novels that seem to obscure, submerge, or confound their own political dimensions, these works wear their hearts on their sleeves: they are frustrated, fed up with how things are, and unafraid to speak truth to power in a direct, seemingly "un-literary" way.

"Measured Unrest in the Poetry of the Black Arts Movement" offers a proof-of-concept for performing sentiment analysis on some of the most politically and affectively charged poetry of the 20th century in America, that of the Black Arts Movement of the 1960s and 1970s. The BAM first took shape at the height of the Black Power Movement with the foundation of the Revolutionary Theatre by Amiri Baraka in 1965. As Larry Neal—one of BAM's principal theorists—says in a 1969 manifesto, the "Black Arts movement seeks to link, in a highly conscious manner, art and politics" toward "the liberation of Black people" (Neal, 1969: 54). Moreover, what Neal calls the movement's "black esthetic" is famous for its affective dimensions, often exploring the limits and political uses of anger, frustration, and militant poetic rage. But while BAM writers sought to link art and politics through explicitly racial terms, many—though by no means all—were marked by a failure to attend to the intersections of gender with racial injustice.

In this project I ask two questions in particular: first, how are the feelings associated with injustice in this corpus coded in terms of race and gender? And second, what can natural language processing techniques like sentiment analysis show us about the relations between different dimensions of poetry—like affect and gender—given that poetry is highly figurative and notoriously difficult to quantify in terms of sentiment or opinion?

Method

In addressing both these questions, this project uses a small corpus of poetry—currently 26 books—from prominent BAM authors. I employ both close reading as well as machine reading techniques, combining the powerful scale of sentiment analysis with the granularity of traditional literary analysis in an effort to explore the intersections of feeling, gender, race, and injustice in the radical poetry of this period. My goal in this project is not to develop a sentiment classifier that works on experimental poetry in English. Rather, it is to see what existing classifiers can show us about a specific corpus of poetry.

In this sense, I use pre-existing sentiment classifiers like VADER and Pattern (via TextBlob) to perform a kind of exploratory computational analysis on my corpus (Hutto and Gilbert, 2014; De Smedt, and Daelemans, 2012). Rather than use these tools to make general claims about this incredibly diverse body of poetry, I test, experiment, and make targeted use of sentiment analysis techniques to pursue research questions already present in existing scholarly conversations—for example, how poets might tie heightened affects to an explicitly political quest for racial justice in America. The insights I draw from my computational analyses, then, go hand in hand with more traditional literary practices. Moreover, my methodology aims to acknowledge the fact this poetry was written in the shadow of government surveillance programs, active FBI counterintelligence operations, and a larger culture fearful of radical thought. Because of this, my project explores the fraught methodological implications of using distanced, potentially decontextualizing computational text analysis techniques to think through BAM poetry, and how these methods might best be used to pursue questions, problems, and lines of inquiry centered around black thought and experience.

The already vibrant conversations on sentiment analysis and natural language processing more generally have been illuminating in forming these thoughts and questions. The discussion between Matthew Jockers and Annie Swafford on the*Syuzhet*package and "archetypal plot shapes" has helped me not only to consider the currentpossibilities and limitationsof sentiment analysis as applied to literary corpora, but also to think through the kinds of results we expect from digital projects and how we verify those results as an academic community (Swafford, 2015; Jockers, 2015). With regards to poetry and NLP more specifically, Lisa Rhody'stopic modeling of highly figurative ekphrastic poetryis a great model for

how unexpected failures in textual analysis can also be productive, prompting us towards new questions as well as new understandings of familiar methods like close reading (Rhody, 2012).

Results

I have implemented NLP techniques with NLTK and TextBlob, a text-processing Python library, on my collection of 26 books of poetry. I have also used two sentiment classifiers—Pattern (via TextBlob) and VADER—to evaluate my corpus for sentiment and interpret my results. While this work is ongoing, so far my work comprises explorations and experiments in the smaller-scale uses of sentiment analysis in the study of poetry and affect.

For example, Pattern considers Quincy Troupe's "Come Sing a Song"—from his 1972 collection *Embryo Poems, 1967-1971*—to be the most negative poem in my entire corpus. In a corpus of poetry containing direct attacks, extreme invective, and explicit takedowns of individuals, groups, and institutions, I did not find this poem to contain an exceptional amount of negative sentiment. On the contrary, I found "Come Sing a Song" to be positive and celebratory with regards to black life and black artistic expression. VADER, meanwhile, considers Nikki Giovanni's "The True Import of the Present Dialogue, Black vs. Negro"—from her 1968 *Black Feeling, Black Talk*—to have the most negative sentiment in the corpus. These results are very much in keeping with other human readers of this poem: critics consider it to be one of the most significant and famous examples of a certain type of angry, militant, even aggressive poem. Where Pattern and I disagree strongly over the feel of Troupe's "Come Sing a Song," critics and VADER seem to agree that Giovanni's "The True Import" has, on the surface, an exceptional amount of negative sentiment compared with its contemporaries.

Among other things, my project analyzes discrepancies and correspondences such as those described above. Already, my findings have revealed an interpretive disjoint between the denotative affective impact of words—what might be called their surface sentiment—and their more nuanced affective import as shaped by poetic, literary, social, and political contexts. A sentiment classifier like VADER, for example, highlights the intensity of negative sentiment in a poem according to the words and phrases it contains without the literary and historical context of their use. This kind of surface reading, attuned specifically to words' immediate affective impact, anticipates the space between a surface anger that can spark feelings regardless of context and a poetic form that, in the case of Giovanni's "The True Import," leverages negative sentiment to address meaningful social issues in a productive, ultimately positive way. By investigating these poems through conventional literary methods (i.e., historical contextualization, close reading, consideration of relevant scholarship) and computational methods (in this case Pattern and VADER), while also investigating the histories, intended use contexts, and potential biases of the chosen computational methods, this project provides an opportunity to examine what it is, exactly, that provides a book, poem, or poetic line with its emotional charge.

References

De Smedt, T. and Daelemans, W. (2012). "Pattern for Python." *Journal of Machine Learning Research* 13: 2063–67.

Hutto, C. J. and Gilbert, E. (2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Eighth International Conference on Weblogs and Social Media*. Ann Arbor, MI, June 2014.

Jockers, M. (2015). "Revealing Sentiment and Plot Arcs with the Syuzhet Package," February 2, 2015. http://www.matthewjockers.net/2015/02/02/syuzhet/ (accessed 27 February 2018).

Lorde, A. (1997). "The Uses of Anger." *Women's Studies Quarterly* 25, no. 1/2: 278–85.

Neal, L. (1969). "Any Day Now: Black Art and Black Liberation." *Ebony* 24, no. 10: 54–62.

Rhody, L. (2012). "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2, no. 1. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ (accessed 27 February 2018).

Swafford, A. (2015). "Why Syuzhet Doesn't Work and How We Know," March 30, 2015. https://annieswafford.wordpress.com/2015/03/30/why-syuzhet-doesnt-work-and-how-we-know/ (accessed 27 February 2018).

Does "Late Style" Exist? New Stylometric Approaches to Variation in Single-Author Corpora

Jonathan Pearce Reeve
jon.reeve@gmail.com
Columbia University, United States of America

Does "late style" exist? That is, do novelists exhibit a well-defined and distinctive stylistic shift as they reach old age, artistic maturity, or both? Edward Said's *On Late Style: Music and Literature Against the Grain* argues not only that such a style does exist, but that it has well-defined characteristics. Said describes late style as, somewhat paradoxically, involving "a nonharmonious, nonserene tension, and above all, a sort of deliberately unproductive productiveness going *against*" (Said 22). The term "late style," derived from Thodor Adorno's concept of Beethoven's *Spästil*, is one which Adorno conceives of as "catastrophic" (Adorno 567). As Adorno puts it, "the maturity

of the late works of significant artists does not resemble the kind one finds in fruit. They are, for the most part, not round, but furrowed, even ravaged. Devoid of sweetness, bitter and spiny, they do not surrender themselves to mere delectation" (564). To determine whether this claim is more than just anecdotally true, it deserves to be experimentally tested. Using new techniques of computational stylometric analysis, I test whether a writer's late works are statistically dissimilar to the rest of their corpus. I find that late style is not a statistically quantifiable phenomenon. Instead, the opposite is true: the novelists tested exhibit very distinctive early styles.

Twelve single-author corpora were prepared for this study. These include three novelists Said cites at length: Marcel Proust, Thomas Mann, and Jean Genet, as well as nine novelists from the 19th and 20th centuries, chosen for their prolificacy and electronic availability: Charles Dickens, Joseph Conrad, Ernest Hemingway, Henry James, Walter Scott, George Meredith, Willa Cather, Arnold Bennett, and Mary Augusta Ward. Two samples were taken from each novel in these corpora, so that the internal stylistic similarity of the samples serve as a metric check for the validity of the method. These samples were randomly chosen, to ensure that no text is longer than the shortest text in each corpus, and that that the analysis will compare equal amounts of text.

Each of these samples was then vectorized to 500-dimensional vectors, according to their top 500 word frequencies. These samples were then reduced to five dimensions using principal component analysis (PCA). Five dimensions were used here, instead of the usual two, since a cross-validated grid search in a previous study determined this value to be the most effective at clustering documents according to voice and style. This study also introduces two new metrics for stylistic difference. First, the "distinctiveness score" of a novel sample is calculated by determining the distance of the vector from the mean in five-dimensional space, using the Pythogorean theorem. A late novel that shows a high distinctiveness score, therefore, could correctly be called an instance of "late style."

Second, I introduce a metric representing the "periodicity" of the writer's style. This is calculated by first inferring prior category labels of early, middle, and late using publication years. Then, the novel's reduced vectors are clustered using a Baysian Gaussian mixture model, which probabilistically infers three or fewer clusters. These assignments are finally compared using a mutual information score, which calculates the similarity of these clusters with the prior inferences, regardless of label. A high periodicity score indicates that a novelist exhibits distinct stylistic periods, whereas a low score indicates that a novelist has a relatively unchanging or unpredictable stylistic progression.

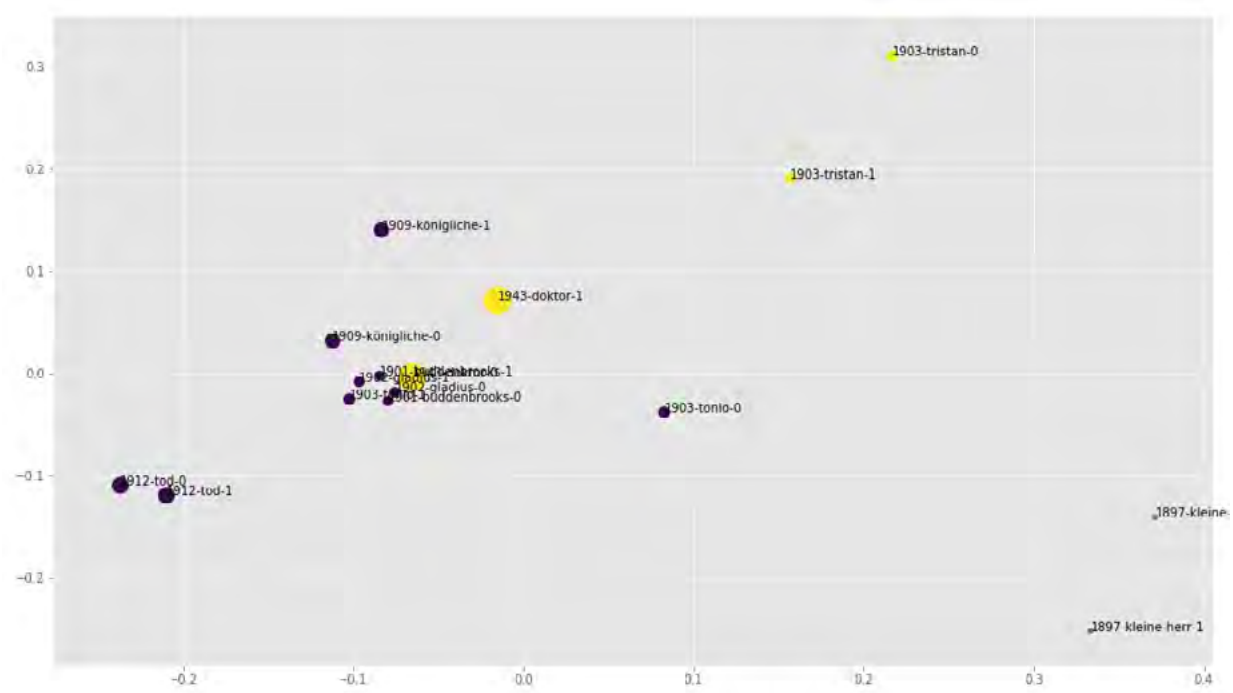


Figure 1: Thomas Mann

Figure 1 shows a projection of the first two dimensions of the vectors generated from Thomas Mann novels. The sizes of the points represent their relative publication years: small circles are early works, and large circles are late works. The colors represent the clusters predicted from the Bayesian Gaussian mixture model. The samples with the highest distinctiveness scores are from his first work *Der Kleine Herr*

Friedemann and his early work *Tristan*. The samples showing the least distinctiveness, are from *Doktor Faustus*, the very work Said cites as an example of a distinctive late style.

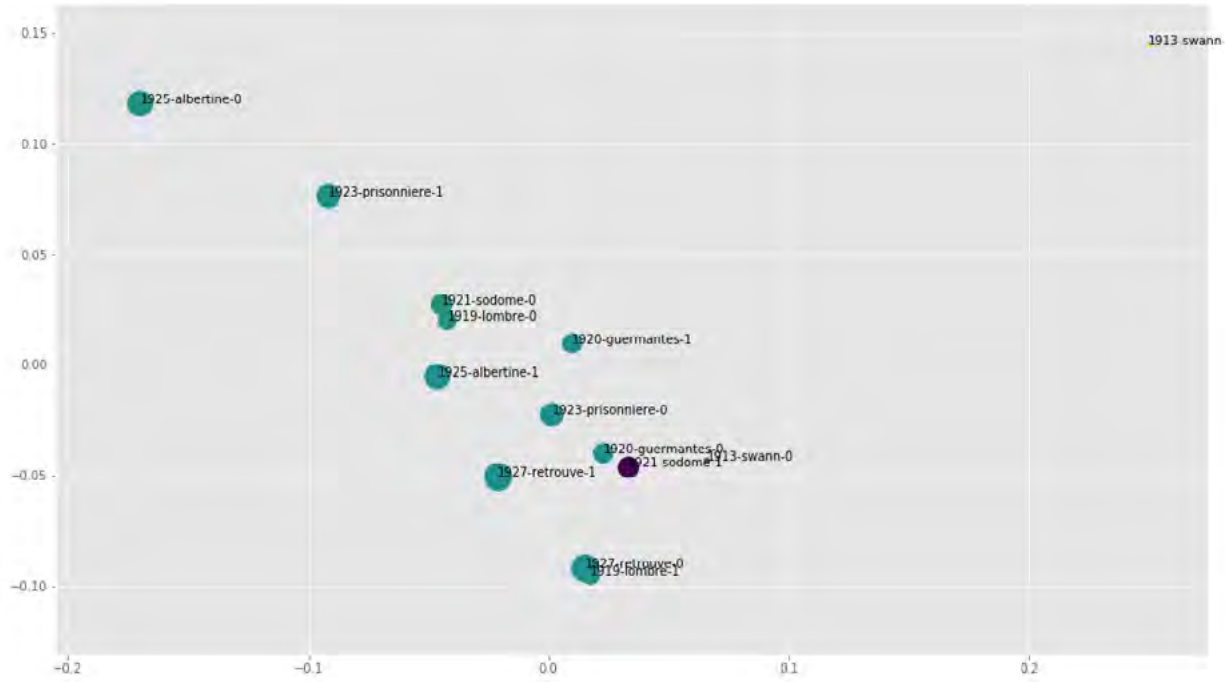


Figure 2: Marcel Proust

Figure 2 shows the same projection for samples from the works of Marcel Proust. Proust's first work, *Du côté du chez Swann*, is the most distinctive. Proust's last published work, *Le temps retrouvé*, which Said cites as an example of late style, is in fact very non-distinctive. Proust's middle works, however, *La prisonnière* and *Albertine disparue*, are only intermediary with respect to publication dates, since they were the final novels he wrote. Here, Said is somewhat correct that Proust has a late style, but misidentifies the works that exemplify it. Again, however, Proust's early style shows a stronger signal than his late.

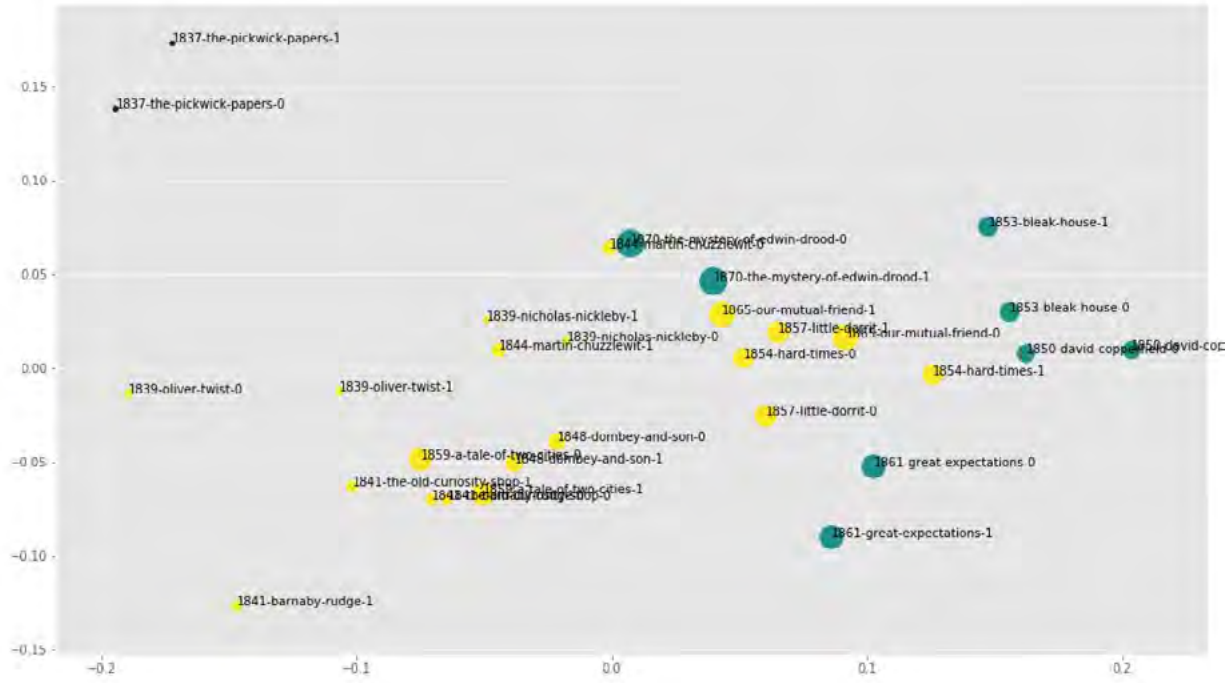


Figure 3: Charles Dickens

Figure 3 shows vectors generated from Charles Dickens novels. Here again, the early work *The Pickwick Papers* has the highest distinctiveness score, followed by *David Copperfield*. Late works like *Our Mutual Friend* are among the least distinctive. As the alignment of the point colors and sizes here suggests, Dickens shows a strong periodicity. At 0.469, his is the second-highest periodicity score.

Author	Periodicity Score
Proust	0.023
Meredith	0.028
Ward	0.166
Cather	0.177
Conrad	0.177
Bennett	0.220
Hemingway	0.326
Scott	0.360
Mann	0.367
Genet	0.457
Dickens	0.469
James	0.472

Table 1

Table 1 shows the periodicity scores of all the novelists studied here. Those novelists with well-known early and late styles, such as James and Dickens, have high periodicity scores. Writers like Proust, on the other hand, whose novels all form part of the series *À la recherche du temps perdu*, and were all published within about a decade, show the lowest periodicity scores.

This study, beyond simply testing and ultimately disproving the claims of Said and Adorno, provides a framework for stylometric analysis of textual difference, one which could be used to enhance authorship detection techniques and the techniques of forensic text analysis more generally. More experiments are needed, of course, to test the validity of these techniques beyond the domain of literature.

References

Adorno, T. (2002). Late Style in Beethoven. In: *Essays on Music*. Berkeley: University of California Press, 2002. pp. 564–568.

Said, E. (2006). *On Late Style: Music and Literature Against the Grain*. New York: Pantheon Books.

Keeping 3D data alive: Developments in the MayaCityBuilder Project

Heather Richards-Rissetto
richards-rissetto@unl.edu
University of Nebraska-Lincoln, United States of America

Rachel Optiz
rachel.opitz@glasgow.ac.uk
University of Glasgow, United Kingdom

Fabrizio Galeazzi
fabrizio.galeazzi@york.ac.uk
University of York, United Kingdom

Digital data preservation is complex and multi-layered. The digital humanities brings unique challenges and opportunities to „keeping data alive" that are leading to innovative cross- disciplinary solutions. Data preservation involves standards, guidelines, open-source vs. proprietary software, accessibility, and much more. While establishing best practices, cultivating a community of experts, and developing infrastructure for 3D data used in cultural heritage has been the focus of several coordinated efforts in Europe over the past decade (Campana and Remondino 2014; Fresa and Prandoni 2015; Vecchio et al. 2015), efforts have been less systematic in the United States. Recently, however, digital humanities practitioners have spearheaded 3D data preservation and sharing in the United States.

While scholars working with 3D data must deal with management and sustainability issues (Galeazzi 2016; Richards-Rissetto and von Schwerin 2017), endeavors are typically tailored to individual projects. To broaden and coordinate efforts, the Community Standards for 3D Data Preservation (CS3DP) project is bringing together librarians, curators, technical specialists, and scholars to begin the process of developing standards for preservation and sharing of digital 3D data. While long-term archival of these data, for example, in a dark archive, is integral to our research (Koller et al. 2010), the MayaCityBuilder project is contributing to "keeping data alive" by developing workflows to supporting reuse and repurposing of procedurally-generated 3D data in the humanities.

While many types of 3D models are being used in humanities scholarship, the case study focuses on 3D models of ancient Maya architecture generated from multiple data sources including architectural drawings, excavation reports, Geographic Information Systems (GIS), and airborne LiDAR. To contribute to 3D data preservation efforts, while maintaining realistic goals, the MayaCityBuilder Project focuses on procedural modeling—rapid prototyping of 3D models from a set of rules. Procedural modeling is ideally suited for the development of 3D modeling standards that promote data interoperability, dissemination, and reuse because they bring with them the underlying metadata, paradata (information about modeling choices) (Bentkowska-Kafel et al. 2016), and descriptive data (e.g., data sources, textures, building type).

Within these circumstances, the two objectives of the "keeping data alive" component of the MayaCityBuilder Project, supported by a Tier I Research and Development Grant from the Division of Preservation and Access of the National Endowment for the Humanities (NEH), are to de-

velop **workflows**: (1) to generate, store, and make accessible 3D models of ancient architecture in open-source and proprietary software to foster data (re)use and (2) to host, deliver, and visualize these 3D models, linked to metadata, paradata, and descriptive data, in 3D visualization environments. These objectives are part of a larger goal to contribute to *innovative methods of materials analysis and new modes of discourse using interactive 3D web visualizations.* To achieve this goal requires not only data accessibility but also data compatibility—scholars must also be able to combine and recombine data for reuse and repurposing.

Building on previous research and development and lessons learned from the MayaArch3D Project (von Schwerin et al. 2013), Gabii Goes Digital (Opitz et al. 2016), and the Archaeological Data Service (ADS) (Galeazzi et al. 2016), we present technical workflows to dynamically host, deliver, and visualize 3D models that are linked to metadata, paradata, and descriptive data in two 3D environments: (1) an open source 3D web-based environment based on 3DHOP (3D Heritage Online Presenter—an open-source software package for hosting interactive, high-resolution 3D models on the web that uses HTML, JavaScript, and WebGL (Web Graphics Library) (2) Unity—a proprietary and widely-used gaming engine that offers free access to many of its powerful tools. We present an overview of the workflows we have developed explaining how the steps serve our objective of data reuse and more broadly access and preservation of 3D data. Additionally, we discuss how these workflows relate to the next phase of the project, i.e., prototype development. The prototype will take advantage of recent developments in web technology, namely the adoption of WebGL that renders interactive 2D and 3D computer graphics in browsers without plugins.

The ability to efficiently generate, store, deliver, and visualize models in an interactive 3D web- based environment will help keep data alive by fostering collaborative and comparative humanities research. We focus on procedural models because they can be quickly generated and are directly linked to metadata and paradata. 3D models allow scholars to test architectural reconstructions and situate them within landscapes to investigate spatial relationships at multiple scales while providing a sense of embodiment (Barcelo et al. 2000; Dylla et al. 2010; Frischer and Dakouri-Hild 2008; Richards-Rissetto and Plessing 2015; Saldana 2015). However, the diversity of 3D data types, tools, and technologies in combination with a lack of standards requires workflows to promote reuse and repurposing of 3D data to contribute to long-term access and preservation of 3D data.

References

3D Heritage Online Presenter (3DHOP). http://vcg.isti.cnr.it/3dhop/index.php; last accessed on 04/24/18

Barcelo, J., M. Forte, and D. Sanders. (2000). Virtual Reality in Archaeology. *BAR Int. Series* 843.

Bentkowska-Kafel, A., Denard, H., Baker, D. (Eds.), (2016). *Paradata and Transparency in Virtual Heritage – Digital Research in the Arts and Humanities Series*. Routledge Taylor & Francis, London.

Campana, S., & Remondino, F. (2014). 3D modelling in archaeology and cultural heritage: theory and best practice. *BAR Int. Series* 2598.

Dylla, K., B. Frischer, P. Mueller, A. Ulmer, and S. Haegler. (2009). Rome Reborn 2.0: A Case Study of Virtual City Reconstruction Using Procedural Modeling Techniques. In *Making History Interactive*, pp. 62-66. Oxford: Archaeopress.

Fresa, A., Justrell, B., & Prandoni, C. (2015). Digital curation and quality standards for memory institutions: PREFORMA research project. *Archival Science*, 15(2), 191-216.

Frischer, B. and A. Dakouri-Hild (eds). (2008). *Beyond illustration: 2d and 3d digital technologies as tools for discover in archaeology*. Oxford: Archaeopress.

Galeazzi, F, M. Callieri, M. Dellepiane, M. Charno, J. Richards, R. Scopigno. (2016). Web-based visualization for 3D data in archaeology: The ADS 3D viewer. *Journal of Archaeological Science*: Reports 9: 1-11.

Galeazzi, F. (2016). Towards the definition of best 3D practices in archaeology: Assessing 3D documentation techniques for intra-site data recording. *Journal of Cultural heritage* 17: 159-169.

Koller, D., Frischer, B. and G. Humphreys. (2010). Research challenges for digital archives of 3D cultural heritage models. *Journal on Computing and Cultural Heritage* 2(3):7:1-7:17.

Opitz, R., Marcello Mogetta, and Nicola Terrenato. (2016). *A Mid-Republican House from Gabii.* Ann Arbor: University of Michigan Press.

Richards-Rissetto, H. and R. Plessing. (2015). "Procedural Modeling for Ancient Maya Cityscapes: Initial Methodological Challenges and Solutions." *2015 Digital Heritage International Congress*, Volume 2: 85-88. IEEE Conference Publications.

Richards-Rissetto, H. and J. von Schwerin. (2017). A Catch 22 of 3D Data Sustainability: Lessons in 3D Archaeological Data Management & Accessibility. *Journal of Digital Applications in Archaeology and Cultural Heritage*. 6: 38-48.

Saldana, M. (2015). An Integrated Approach to the Procedural Modeling of Ancient Cities and Buildings. *Digital Scholarship in the Humanities*, Volume 30, Issue suppl_1, 1 December 2015, Pages i148–i163,

Vecchio, P., Mele, F., De Paolis, L. T., Epicoco, I., Mancini, M., & Aloisio, G. (2015). Cloud Computing and Augmented Reality for Cultural Heritage. In *Augmented and Virtual Reality* (pp. 51-60). Springer International Publishing.

von Schwerin, J. , H. Richards-Rissetto, F. Remondino, and G. Agugiaro. (2013). "The MayaArch3D Project: A 3D WebGIS for Analyzing Ancient Maya Architecture and Landscapes at Copan, Honduras." *Literary and Linguistic Computing* 28(4):736-753.

Finding Data in a Literary Corpus: A Curatorial Approach

Brad Rittenhouse
bcrittenhouse@gatech.edu
Georgia Institute of Technology, United States of America

Sudeep Agarwal
hello@sudeep.co
Georgia Institute of Technology, United States of America

PI and Presenter: Brad Rittenhouse
Others involved: Taha Merghani, Sudeep Agarwal, Vidya Iyer, Madison McRoy, Sidharth Potdar, Nate Knauf, and Kevin Kusuma

In this short paper, I will discuss an ongoing text analysis project, which applies NLP, topic modeling, mapping, and other methodologies to the Wright American Fiction corpus. From a theoretical standpoint, the project is an extension of my qualitative work, which tracks a notable historical shift in literary data management strategies through the works of two canonical American writers: Herman Melville and Walt Whitman. Both wrote in New York as it grew from a small market town of around 60,000 residents to a global metropolis of nearly 1,000,000 and had to imagine strategies of data management to integrate newly urban, consumerist surroundings into their writings in an effective, efficient manner. Translating increasingly crowded material realities—populated by people, products, and print—into literary data, these writers illustrate an important ontological shift from the positivist data strategies of the Enlightenment to digital logics of aggregation, organization, and metonymic indexing that increasingly address the impossible scale of modern infospheres.

As relatively privileged subjects, however, these writers' very ability to integrate and innovate with this information was largely based upon a free access to information (and indeed information overload) that many contemporaries did not enjoy. In short, critics have historically apportioned literary status upon hegemonic standards of information, with prestigious genres like "encyclopedic writing" preferring masculinist topics and knowledge bases such as ballistics (Pynchon), cetology (Melville), violence (Bolaño) over spheres of knowledge historically more accessible and immediate to women and people of color.

My quantitative work looks to sidestep these biases, using an assortment of natural language processing techniques to recover works from the archive that may be performing similarly impressive literary acts of aggregation, but which critics may have overlooked because the works exist in alternative thematic and affective registers. By measuring the accretion of material information across the corpus, and identifying areas of relative density, my process points to writing which humans readers have overlooked but which machines are able to see as substantially similar to canonical encyclopedic works.

We intentionally made a very broad measurement of the text to identify a broader range of artistic expression. The process itself involves chunking all the texts into 500-word segments, performing a parts-of-speech tag with OpenNLP, then rendering these tags in "baby binary": a "0" for all non-nouns, a "1" for all nouns. We then summed the segments and divided by the total length of each to obtain a noun density measurement, which generally indicates an aggregation of material information. Though it is possible to use more specific grammatical measures (subjects, objects, etc.), we used nouns at-large so as to capture a fuller spectrum of thought, sentiment, and other immaterial objects that accompany the human masses of urbanization.

We also assembled a fair amount of demographic metadata for the corpus, which has allowed us retrieve relatively forgotten works from the archive. After identifying the densest chunks of text, we attempted to identify author gender with the use of the machine learning platform SexMachine. We cross-referenced these results with those derived from the noun-density analysis to pinpoint female authors of interest. To conduct this analysis, we first performed exploratory data analysis to understand the underlying distribution of noun ratios across the corpus, which appeared to be normally distributed, although with a slight right skew. Then we compared this distribution with that of the noun ratios identified for authors of each gender. The distributions seem to be largely similar. This naturally led to an outlier analysis within each gender, which identifies outliers as works with a noun ratio 1.5 interquartile ranges either above or below the median, yielding 71 outliers for male authors and 47 outliers for female authors (43 and 26 on the high-end, respectively). We then performed additional analyses on these outliers to get a better understanding of what differentiated them from the rest of the corpus.

One case study I will present from among these outliers is that of Emma Wellmont, a nineteenth-century temperance writer who the academy has largely ignored, I suspect because of the emotional, sensationalist overtones of her chosen genre. Nonetheless, her work is quantitatively similar to Walt Whitman's, with many extracts in the highest quadrant of noun density across the corpus and packed with what the latter evocatively refers to as "stuff." Unlike Whitman, however, her densest passages are often emotional, pathetic scenes of death and suffering. Critics, if they read Wellmont's work (and most do not), would likely label it sensationalistic or melodramatic, and therefore, unserious, writing. My methodology, on the other hand, makes an argument for her as an important encyclopedist, albeit of canonically unlikely subject matter. I will present the case study through a prototype interactive visualization that allows users to explore the corpus at-large, all the way down to significant passages within individual works (Figure 1).

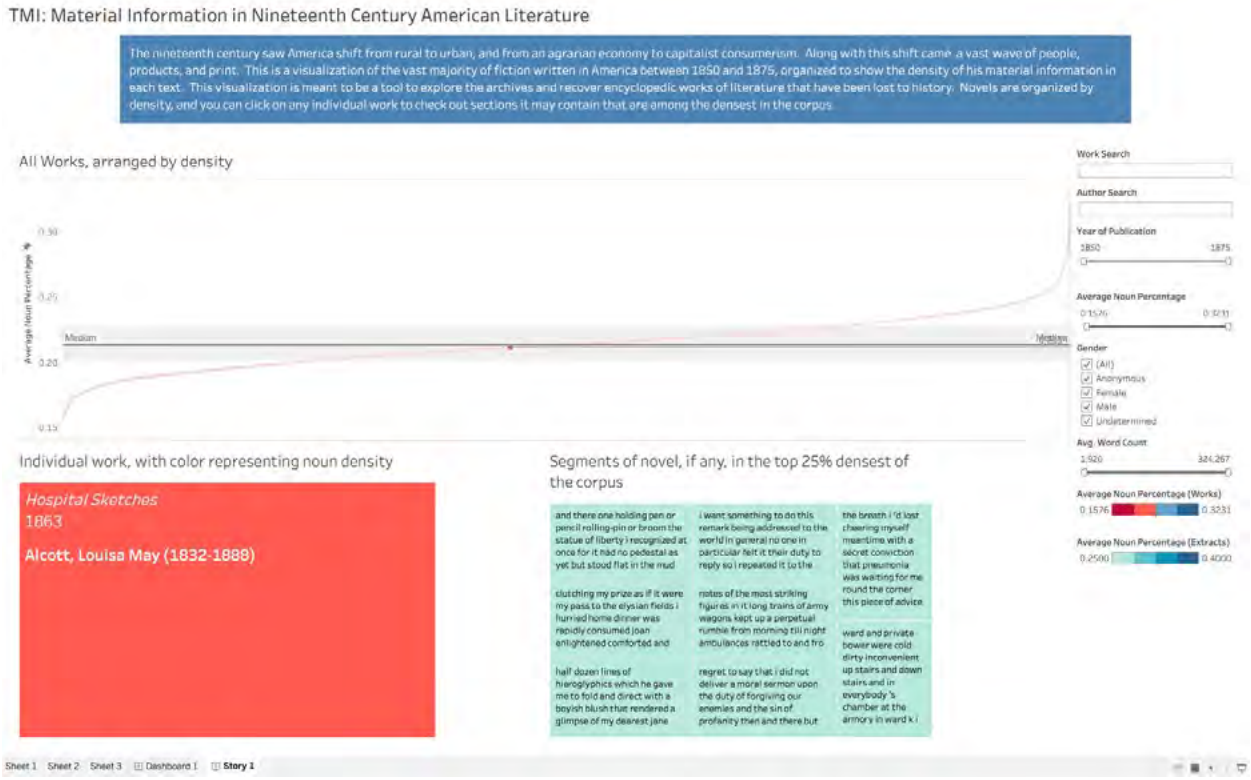


Figure 1

This curatorial process builds upon the methods described by Long and So in their recent article "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," using high-powered computing and statistical analysis on a corpus scale to identify information-dense passages for later close reading and analysis. Reading literature as information, the methodology is flexible in not only illuminating macro-scale trends, but also identifying human-readable works and passages for literary critics who also value critical reading practices. The project also runs in parallel to Dennis Yi Tenen's recent work in its "articulation of 'effect spaces' via material density," though it pulls from a broader range of quantitative, grammatical measures in its attempt to broaden the generic construct of encyclopedic writing.

References

Long, H. and So, R. (2016). Literary pattern recognition: modernism between close reading and machine learning. *Critical Inquiry*, 42(2): 235-267.

Yi Tenen, D. (2018). Toward a computational archaeology of fictional space. *New Literary History*, 49(1): 119-147.

Mapping And Making Community: Collaborative DH Approaches, Experiential Learning, And Citizens' Media In Cali, Colombia

Katey Roden
rodenk@gonzaga.edu
Gonzaga University, United States of America

Pavel Shlossberg
shlossberg@gonzaga.edu
Gonzaga University, United States of America

Engaging the "bridges/puentes" theme central to the conference, this paper presents first-hand knowledge and practical insights garnered from a collaborative digital mapping project between North/South academics, students, and community activists engaged in community-based social justice activism in Cali, Colombia. A foundational goal of this Digital Humanities project is thus to create intercultural and communicative bridges between not only the academic communities of Gonzaga University and Pontifica Universidad Javeriana, but also to provide a platform by which Colombian community organizers shape their presence in local as well as digital communities.

The paper discusses our goals and methods, and also the roadblocks we encountered, in establishing collabo-

rative pathways to embed Digital Humanities mapping tools as central elements within a field-based Communication and Community Development course. The Digital Humanities project at the heart of this course aligns with pedagogy as well as practical fieldwork in the area of Development Communication, which holds that communication processes and projects that support or foster the growth of grassroots civil society are essential elements of community development and empowerment. In this vein, Digital Humanities perspectives and methodologies that privilege the bottom-up democratization of access and information inform course content assigned to students from the Global North, who come to Cali, Colombia as part of an intensive immersion.

As such, the course invites students from Cali and the United States to engage, accompany, and shadow community-based organization that work in areas such as citizens' radio; street theatre and community-based performances; and grassroots documentary production. The work undertaken by the community-based organizations seeks to displace hegemonic media and dominant culture imaginaries, which routinely render these resource-deprived communities as being inherently abject, dangerous, chaotic, and pathological. The community organizations with whom we partner engage community problems by creating and claiming spaces for public expression that amplify popular voices within their own communities and beyond. The Digital Humanities mapping project developed for this course responds to these community initiatives, in that it serves as a community-academy collaborative space. The digital map produced collaboratively, provides a platform that presents, promotes, captures, and renders visible popular or grassroots media, communication activities, and products through which "citizens can learn to manipulate their own languages, codes, signs, and symbols, empowering them to name the world in their own terms" (Rodriguez 2011). The paper documents a central element of our work, the mutual efforts engaged in creating active and equitable roles for each party involved in the digital product's production, whether those groups come from the academy, the community, the Global North or the Global South.

Beyond discussing the compatibility and fit of Digital Humanities tools with the articulation of community-based approaches to citizens' media and popular communication, this paper also discusses the significant ways in which Digital Humanities mapping tools can be mobilized to foster or promote community-based experiential learning experiences in intercultural contexts bridging the global North and South. Experiential learning emphasizes the acquisition of knowledge through interactive processes of action and reflection; where students can take an active part in the creation of knowledge (Hale 1999). We contend that producing Digital Humanities projects in the contexts of an international immersion and hand-in-hand with local partners whose voices, perspectives, and needs drive project conceptualization and the mapping process, presents a unique opportunity for experiential learning that extends well beyond the classroom and into the lived realities of all the parties involved. In this vein, the experiential learning opportunities developed in such an environment embrace the broader humanistic agenda of Digital Humanities as a field, where people come together through and with technology to "produce a collaborative, connected, and relational knowledge production, of making and learning and learning through making" (Goldberg 2015). Accordingly, our project seeks to facilitate an experiential learning opportunity for our students, but in doing so we also seek to diminish the sometimes too rigid boundaries that privilege academic institutions as the sole purveyors and producers of knowledge. By collaboratively creating a digital map with and for local community members while in their communities, our project aims to decentralize knowledge production and encourage our students to become conscious of diverse forms of knowledge and authority.

Furthermore, our experience also suggests that with effective planning and development, community-based Digital Humanities mapping projects can productively alleviate issues and problems that commonly arise in the context of experiential- or service-learning courses taking place in intercultural contexts across the North/South boundary. It is well known that "service-learning can reinforce stereotypes and paternalism among students. Some scholars argue that many applications of service learning do little to question the role of students as providers of resources…" (Chupp & Joseph 2010). Additionally, service- or experiential-learning is "often implemented with a sole focus on the potential beneficial impact on the student, with little or no emphasis on the possible longer-term beneficial impact on those served by the activity and their broader community" (Chupp & Joseph 2010). The collaborative mapping project we have developed engages Digital Humanities approaches within an embedded community context, with the explicit intention of addressing potential problems linked to the implementation of experiential service learning project in partnership between the North and South.

In sum, the Digital Humanities mapping project nested within this Communication and Community Development course remains an experimental and open collaboration. Well-established and emergent issues and challenges continue to exist. With that caveat in mind, experience and evidence also suggests that digital technology mapping tools provide a set of ready enhancements to experiential learning, study abroad, and Communication and Community Development courses. These features begin to realize the promise and purpose of Digital Humanities by creating bridges that foster global collaboration, create open access platforms, and generate academy-community, North/South collaborations that equalize access to the generation and circulation of knowledge locally and globally.

KEYWORDS: Global South/North, Experiential Learning, Mapping, Community Development, Citizens' Media

References

Chupp, Mark G., and Mark L. Joseph. "Getting the most out of service learning: Maximizing student, university and community impact." *Journal of Community Practice* 18.2-3 (2010): 190-212.

Goldberg, David Theo. "Deprovincializing Digital Humanities." In *Between Humanities and the Digital*. Eds. Patrik Svensson and David Theo Goldberg. MIT Press, 2015. 163-71.

Hale, Aileen. "Service-learning and Spanish: A missing link." *Construyendo Puentes (Building Bridges): Concepts and Models for Service-Learning in Spanish*. Ed. Josef Hellebrandt and Lucía T. Varona. Washington, DC: American Association for Higher Education (1999): 9-31.

Rodríguez, Clemencia. *Citizens' media against armed conflict: Disrupting violence in Colombia*. U of Minnesota Press, 2011.

The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings

Pablo Ruiz Fabo
pablo.ruiz@linhd.uned.es
Universidad Nacional de Educación a Distancia, Spain

Helena Bermúdez Sabel
helena.bermudez@linhd.uned.es
Universidad Nacional de Educación a Distancia, Spain

Clara Martínez Cantón
cimartinez@flog.uned.es
Universidad Nacional de Educación a Distancia, Spain

Elena González-Blanco
elenagbg@gmail.com
Universidad Nacional de Educación a Distancia, Spain

Borja Navarro Colorado
borja@dlsi.ua.es
Universidad de Alicante, Spain

Introduction

Digital resources in poetry in Spanish are scarce, particularly for certain periods. This poses difficulties for Digital Humanities studies in Spanish.

Some digital editions of medieval poetry exist, e.g. BiDTEA (Gago Jover et al, 2015), ADMYTE (Marcos Marin and Faulhaber, 1992), PoeMetCa (Escribano et al, 2016), besides resources containing partial editions like ReMetCa (González-Blanco and Rodríguez, 2014). For the Golden Age, Navarro-Colorado et al. (2015) presented the *Corpus of Spanish Golden-Age Sonnets*. For later periods, we are not aware of poetry collections, although other genres are covered in Textbox (Schöch et al, 2017), BETTE (Santa María Fernández et al, 2017), Aracne (Álvarez and Martín, 2015) or Revistas Culturales 2.0 (Ehrlicher and Rißler-Pipka, 2015).

This paper describes the DISCO corpus and how it complements available digital materials for poetry in Spanish in several respects: First, the author and period range. Second, metadata concerning the authors and their works expressed in TEI-RDFa, given the importance of interoperability between literary datasets and the advantages of Linked Open Data as a paradigm. Finally, example findings that can be obtained with our corpus are provided, regarding metrical patterns diachronically.

The corpus is available on GitHub[1] and Zenodo.[2]

Corpus description

The corpus contains 4087 sonnets in Spanish by 1204 authors (15th to 19th century),[3] extracted from HTML sources at Biblioteca Virtual Cervantes (García, 2005, 2006a, 2006b) and Wikisource. Sonnets were chosen given the form's importance in European poetry, where it is even considered as its own genre. The form's clear restrictions make it easily amenable to computational treatment, facilitating meaningful comparison across poems. Several computational linguistics studies on the sonnet exist (Navarro-Colorado et al., 2015, 2016, 2017a, 2017b; Agirrezabal, 2017). A new sonnet corpus complements earlier work on both traditional and computational poetry analyses.

We focused on canonical and non-canonical authors, from different Spanish-speaking countries (Figure 1).

Period	Nbr of Sonnets	Nbr of Authors			Sources
Golden Age (15th-17th)	1088	477	Female	31	García (2006a)
			Male	446	
			America	12	
			Europe	458 (+7)	
18th century	323	42	Female	1	García (2005) Wikisource
			Male	41	
			America	6	
			Europe	36	
19th century	2676	685	Female	46	García (2006b)
			Male	637	
			America	334	
			Europe	348 (+3)	

Sonnet and author distribution per period, including the number of female and male authors, and the continent where they developed

1 https://github.com/postdataproject/disco/
2 https://doi.org/10.5281/zenodo.1012567
3 About 125 sonnets by approx. 20 authors whose production took place in the early 20th century (with date of death prior to 1936) are also included in the corpus; the documentation on the GitHub repository (footnote 1 above) provides more details.

their literary activity. Numbers in parentheses indicate authors which were probably active in Europe.

Encoding Paradigms: TEI and Linked Open Data

The poems are encoded in XML-TEI P5. A plain-text version is also provided. Together with the TEI-semantics, this corpus provides a layer of Linked Open Data (LOD) expressed in RDFa (Herman et al., 2015). To our knowledge, no out-of-the-box tools exist for publishing literary TEI corpora as LOD.[4] In this context, the enrichment of TEI with RDFa attributes is a solid approach to translate TEI semantics to the web (see precedents like Jewell, 2010) and benefit from the wide range of possibilities of the Semantic Web: First, we enrich our dataset by linking to third-party ones (as DBpedia), providing additional resources to complement the corpus. Second, we publish our data openly using standard schemas, thus supplying semantic interoperability that allows third-party applications to automatically use our data.

Author metadata

Author metadata were extracted or inferred from unstructured source content, and specified in the teiHeader: Year, place of birth and death, and gender. Two versions of the texts are available: one collecting every sonnet per author, the other with a single sonnet per file.

For the current corpus release we augmented the TEI annotation with URIs and class/property information, expressing them in RDFa. The most straightforward information concerns authors and their works, and the DCMI Metadata Terms (DCMI Usage Board, 2012) provides an appropriate scheme. Most features regarding authors' biographical data were formalised with the FOAF vocabulary (Brickley and Miller, 2014). Links to other resources were supplied. For instance, authors were assigned *Virtual International Authority File* (VIAF) identifiers, by querying VIAF's API supplemented with manual validation. Since the corpus includes non-canonical authors, LOD is an important asset to share their work thanks to the enhanced display of this type of data implemented by search engines.

Our documentation[1] provides further details.

Metrical encoding and enjambment

Using the **met** attribute, each line was annotated for scansion (strong and weak syllables) with the ADSO tool[5] (Navarro-Colorado, 2017), which specializes in Spanish fixed-meter forms, attaining a performance of 0.95 F1. A heuristic was used to automatically annotate the quatrains' rhyme-scheme, i.e. enclosed (ABBA) or alternate (ABAB).

Using an **enjamb** attribute, lines were annotated for enjambment[6] with the ANJA tool[7] (Ruiz-Fabo et al., 2017). The tool's performance at detecting enjambment is above 0.8 F1, and its efficacy at classifying enjambment types varies across periods and types. A *cert* attribute specifies the expected certitude for each enjambment type annotated.

The corpus documentation[1] provides more details.

How's this corpus different?

The metadata mentioned in 2.2. were unavailable in structured, machine-readable format in the corpus sources, or in other sonnet collections, like Sonnet-Archiv (Elf Edition). Regarding coverage, the corpus complements Navarro-Colorado et al's (2015) Golden Age Sonnet corpus, by including minor Golden Age authors. For later periods, we cover more poems and authors than existing digital corpora, up to the 19th century. Our corpus integrates RDFa annotations, which in a second version will be fully compliant with the POSTDATA model.[8] This is a pioneering model that will provide means to publish European poetry materials as Linked Open Data. Finally, combining the annotation of metrical patterns, stanza types and enjambment is not offered by prior corpora.

Some metrical findings

Corpus data on stress patterns (Figure 2) agree with existing descriptions[9] of the Spanish hendecasyllable based on small-sample analyses: A maiori patterns (with 6th-syllable stress) predominate, and a minori patterns (with 4th-syllable stress) follow. However, our data show an increase of a minori patterns in the 19th century, which might suggest an interest in metrical variety in that period.

Regarding diachronic data on the number of stressed positions (Figure 2), patterns with three stresses are

4 Whereas the publication of literary corpora in Linked Open Data formats is not widespread, inspiration could be drawn from the linguistics community, which has been especially successful in building the means to convert resources with linguistic annotations to the Resource Description Framework model (see McCrae et al., 2011; Chiarcos and Ejavec, 2011). In addition, more general projects, not limited to linguistic analysis, are being developed as well: see work on building a TEI ontology in Ciotti et al (2016).

5 https://github.com/bncolorado/adsoScansionSystem

6 The tool detects different types of enjambment (i.e. a mismatch between syntactic and metrical structure) as characterized by Quilis (1964). The tool also detect's Spang's (1983) concept of *enlace*, which takes place when a subject or direct object occur in a line adjacent to their governing verb's line, and which triggers a less noticeable effect than the enjambment types defined by Quilis

7 See https://sites.google.com/site/spanishenjambment/ for details

8 See Bermudez-Sabel et al. (2017). Version 0.2 of the POSTDATA model is available at https://doi.org/10.5281/zenodo.832906

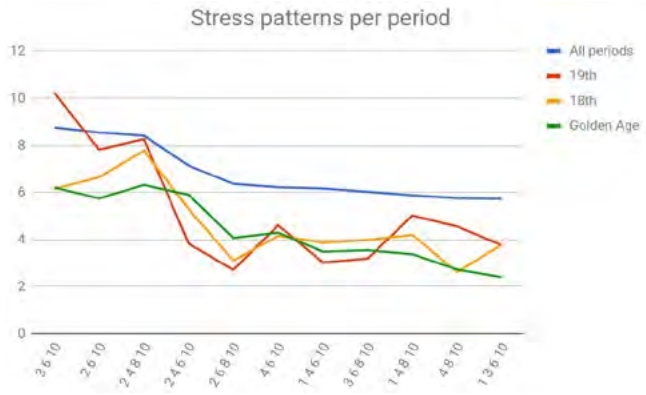
9 See Domínguez Caparrós (2014: 143) or Henríquez Ureña (1919: 132) for details on *a maiori* and *a minori* patterns. The main *a maiori* variants as described in previous literature are 2 6 10 and 3 6 10; this is confirmed in our data. Patterns are formalized as a series of numbers indicating stressed syllables, e.g. 2 6 10 for the second, sixth and tenth syllables. Note that 10th-syllable stress is mandatory in all patterns.

highly used across periods. However, most a maiori patterns with four stresses decrease in the 19th century. This might indicate a 19th-century preference for "lighter" patterns, with stresses further apart from each other.

Whereas the predominant meter for sonnets is naturally the hendecasyllable, alexandrines[10] are attested, mostly in the 19th century, preferentially used by American authors. The alexandrine sonnet uses an alternate rhyme scheme (ABAB) more often than the usual enclosed scheme (ABBA). See Figure 4.

Pattern	Pattern Class	Stress Count	Percentage of lines			
			All periods	19th	18th	Golden Age
3 6 10	mai	3	8.76	10.23	6.16	5.21
2 6 10	mai	3	8.55	7.82	6.65	5.75
2 4 8 10	*min*	*4*	*8.41*	*8.26*	*7.77*	*6.32*
2 4 6 10	mai	4	7.14	3.83	5.30	5.90
2 6 8 10	mai	4	6.37	2.71	3.10	4.07
4 6 10	mai	3	6.23	4.61	4.16	4.30
1 4 6 10	mai	4	6.17	3.03	3.87	3.49
3 6 8 10	mai	4	6.03	3.19	3.98	3.55
1 4 8 10	*min*	*4*	*5.88*	*5.02*	*4.2*	*3.38*
4 8 10	*min*	*3*	*5.76*	*4.56*	*2.62*	*2.73*
1 3 6 10	mai	4	5.73	3.76	3.79	2.40

Distribution of stress patterns per period (percentage of lines for each pattern) for the 10 most frequent patterns in the corpus, sorted by decreasing percentage of occurrence in the complete corpus. Pattern classes are also provided (*mai*: *a maiori*, i.e. stress on 6th syllable, *min*: *a minori*, i.e. stress on 4th and 8th syllable). Rows for *a minori* patterns are in italics. *Stress count* refers to the number of stresses in the pattern. Patterns with three stresses are widely used in any period. Most *a maiori* patterns with 4 stresses decrease in the 19th century, whereas *a minori* patterns increase in that century.



Distribution of stress patterns per period (percentage of lines for each pattern) for the 11 most frequent patterns in the corpus.

10 In Spanish, the alexandrine has 14 metrical syllables. In sonnets, the hendecasyllable predominates almost exclusively. However, particularly since the 19th century, alexandrine sonnets have been written.

Meter Length	Quatrain Rhyme	Sonnet Count		
		total	American	European
hendecasyllable	Enclosed	2269	1218	1051
	Alternate	122	96	26
	Total	**2391**	**1314**	**1077**
Alexandrine	Enclosed	122	98	24
	Alternate	145	121	24
	Total	**267**	**219**	**48**

Count of hendecasyllable vs. alexandrine sonnets according to the authors' continent of production, in the **19th century** (alexandrine sonnets are very rare before). The type of rhyme scheme in the quatrains (enclosed or alternate) is also specified. The alexandrine sonnet is preferentially used by American authors, and there's a preference for alternate rhyme for this meter length.

Acknowledgements

Supported by the project 'Poetry Standardization and Linked Open Data: POSTDATA' (ERC-2015-STG-679528), funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, and led as a Principal Investigator by Dr. Elena González-Blanco, LINHD-UNED (http://postdata.linhd.es/).

References

Agenjo, X. (2015). Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos. Ínsula: revista de letras y ciencias humanas 822: pp. 12–15.

Agirrezabal, M. (2017). *Automatic Scansion of Poetry. PhD Thesis*. University of the Basque Country.

Álvarez Mellado, E. and Martín-Fuertes, L. (2015). *Aracne Project,* http://www.fundeu.es/aracne/ (Accessed 22 Sep. 2017).

Bermúdez-Sabel, H., Curado Malta, M. and González-Blanco, E. (2017). Towards Interoperability in the European Poetry Community: The Standardization of Philological Concepts, in Jorge Gracia et al. (ed.) *Proceedings of Language, Data, and Knowledge: First International Conference (LDK 2017)*: pp. 156–65. Springer International Publishing doi:10.1007/978-3-319-59888-8_14.

Biblioteca Virtual Miguel de Cervantes (1999). *Biblioteca Virtual Miguel de Cervantes,* http://www.cervantesvirtual.com/ (Accessed 22 Sep. 2017).

Biblioteca Virtual Miguel de Cervantes (2007). *Biblioteca del Soneto [Sonnet Library]*, http://www.cervantes-

virtual.com/bib/portal/bibliotecasoneto/ (Accessed 22 Sep. 2017).
Brickley, D. and Miller, L. (2014). FOAF Vocabulary Specification, http://xmlns.com/foaf/spec/ (Accessed 22 Nov. 2017).
Chiarcos, C. and Erjavec, T. (2011). Owl/dl formalization of the multext-east morphosyntactic specifications. *Proceedings of the 5th Linguistic Annotation Workshop, LAW V* '11, pp. 11–20 Stroudsburg: PA, USA,
Ciotti, F., Peroni, S., Tomasi, F., and Vitali, F. (2016). An OWL 2 Formal Ontology for the Text Encoding Initiative. *Digital Humanities 2016: Conference Abstracts,* pp. 151–153
DCMI Usage Board (2012). DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms (Accessed 22 Nov. 2017).
Domínguez Caparrós, J. (2009). *El moderno endecasílabo dactílico, anapéstico o de gaita gallega*. Sevilla: Padilla Libros.
Domínguez Caparrós, J. (2014). *Métrica española*. Madrid: UNED.
Ehrlicher, H., and Rißler-Pipka, N. (2015). *Revistas Culturales 2.0*, https://www.revistas-culturales.de/es. (Accessed 22 Sep. 2017).
Elf Edition: *Sonett-Archiv,* http://sonett-archiv.com. (Accessed 22 Sep. 2017).
Escribano, J., González-Blanco, E. and Río Riande, G. del (2016). *PoeMetCa—Recursos digitales para el estudio de la Poesía Medieval Castellana,* http://poemteca.linhd.es (Accessed 22 Sep. 2017).
Gago Jover, F. (2015). La biblioteca digital de textos del español antiguo (BiDTEA). *Scriptum Digital* 4: pp. 5–36.
García González, R. (2005). *Sonetos del siglo XVIII*. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xviii--0/html/. (Accessed 26 Nov. 2017).
García González, Ramón (2006a). *Sonetos del siglo XV al XVII*. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xv-al-xvii--0/html/ (Accessed 26 Nov. 2017).
García González, R. (2006b). *Sonetos del siglo XIX*. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xix--0/html/ (Accessed 26 Nov. 2017).
González-Blanco, E. and Rodríguez, J. L. (2014). ReMetCa: A Proposal for Integrating RDBMS and TEI-Verse. *Journal of the Text Encoding Initiative*, 8 https://jtei.revues.org/1274 (Accessed 22 Sep. 2017), doi:10.4000/jtei.1274.
Henríquez Ureña, P. (1919). El endecasílabo castellano. *Revista de Filología Española*, 6: pp. 132–157.
Herman, Ivan, Asida, B., McCarron, S., Birbeck, M. (2015). RDFa Core 1.1 - Third Edition, https://www.w3.org/TR/rdfa-core (Accessed 22 Nov. 2017).
Jewell, M. O. (2010). Semantic screenplays: Preparing TEI for Linked Data. In *Digital Humanities 2010*. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-878.html (Accessed 22 Nov. 2017).
Marcos Marín, F. and Faulhaber, C. B. (coord.) (1992). *ADMYTE. Archivo Digital de Manuscritos y Textos Españoles*, http://www.admyte.com/admyteonline/contenido.htm (Accessed 22 Sep. 2017).
McCrae, J., Spohr, D. and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications*, V (Part I): pp. 245–259, Berlin: Springer-Verlag.
Navarro-Colorado, B. (2015). A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. *ACL Workshop on Computational Linguistics for Literature*.
Navarro-Colorado, B. (2017). *ADSO project – Análisis distante del soneto castellano de los Siglos de Oro [Distant analysis of the Spanish Golden Age sonnet]*, http://adso.gplsi.es/index.php/es/proyecto-adso (Accessed 22 Sep. 2017).
Navarro-Colorado, B., Ribes Lafoz, M. and Sánchez, N. (2015). *Corpus of Spanish Golden-Age Sonnets*. Alicante: University of Alicante, https://github.com/bncolorado/CorpusSonetosSigloDeOro (Accessed 22 Sep. 2017).
Navarro-Colorado, B., Ribes Lafoz, M. and Sánchez, N. (2016). Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. *Proceedings of the Language Resources and Evaluation Conference* http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf (Accessed 22 Sep. 2017)
Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities.* https://doi.org/10.1093/llc/fqx009 (Accessed 22 Sep. 2017)
Quilis, A. (1964). *Estructura del encabalgamiento en la métrica española.* Consejo Superior de Investigaciones Científicas, Patronato Menendez y Pelayo, Instituto Miguel de Cervantes.
Ruiz Fabo, P., Martínez Cantón, C., Poibeau, T. and González-Blanco, E. (2017). Enjambment detection in a large diachronic corpus of Spanish sonnets. *LaTeCH-CLFL 2017, Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.* Vancouver, Canada.
Santa María Fernández, M. T., Jiménez Fernández, C. M. (2017). *Biblioteca Electrónica Textual Del Teatro Español, 1868-1936*. Universidad Internacional de la Rioja, Spain.
Schöch, C. Henny, U., Calvo Tello, J. Popp, S. (2015). *The CLiGS Textbox*, https://github.com/cligs/textbox (Accessed 22 Sep. 2017)
Wikisource: *Categoría: Sonetos.,* https://es.wikisource.org/w/index.php?title=Categor%C3%ADa:Sonetos (Accessed 26 Nov. 2017)

Polysystem Theory and Macroanalysis. A Case Study of Sienkiewicz in Italian

Jan Rybicki
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland, Poland

Katarzyna Biernacka-Licznar
katarzyna.biernacka-licznar@uwr.edu.pl
Uniwersytet Wrocławski, Wrocław, Poland

Monika Woźniak
moniwozniak@gmail.com
Università degli Studi di Roma „La Sapienza", Italy

Introduction

Even-Zohar's polysystem theory is a well-established approach to understanding how entire translated literatures interact (or not) with the body of the receiving native literary culture. Even-Zohar identifies a number of possible interactions depending on the relative "strength" and "age" of the two (or more) literatures, and translated literatures may assume "peripheral" or "central" positions within the target literary polysystem. According to this scholar, translations are usually peripheral to native literature; but he also cites examples where a given literary polysystem places some imported subsystems in a central position, while other "foreign imports" remain in the periphery (Even-Zohar 1990).

Even-Zohar thus deals with literary creation en masse rather than, as is often the case in academic approaches to literary translation, on single books original and translated. The obvious parallel to Distant Reading has already been drawn (Helgesson and Vermeulen 2015, 25-26); but it might also be tempting to do the same for a related approach, macroanalysis, if we are to follow the distinction made by the exponent of the latter term (Jockers 2013, 48). Both bring together investigations into masses of literary material unattainable by traditional close reading; yet macroanalysis looks inside many books at once using quantitative methods applied to their lexical layers that have been called "stylometry" well before both Moretti and Jockers.

Material

From our personal mixed Polish-Italian perspective, few cases could serve as a better pretext to try to negotiate this marriage between polysystem theory and computational stylistics than that of *Quo vadis* (1896), the historical romance by Henryk Sienkiewicz, Poland's first literary Nobel Prize winner of 1905. Its international success – long gone with the wind but unparalleled by any other Polish novel to this day – resulted in a veritable explosion in terms of numbers of translations into various languages. In many countries, several different translations simultaneously vied for the public's attention. Yet "several" does not even begin to describe the situation in Italy, where at least three hundred different editions can be still found today (Woźniak 2016). In the first two years of the existence of *Quo vadis* on the Italian market (1899-1900), as many as eight different translations were already available to the readers (Berti and Gagetti 2016).

No wonder: not only was the novel set in the Italian capital and not only did it deal with a subject already very present in Italian culture old and new; the book's (and its author's) brand of conservative Catholicism must have appealed to some of the most influential circles of the country. Yet the novel was also praised by some of Italy's progressive critics, who saw, in Sienkiewicz's persecuted Christians, the struggle of their contemporary revolutionary movements, and who liked to read his depiction of Imperial Rome's decadence as a diatribe against the existing power structure (Marinelli 1984).

This profusion of Italian renderings is also the reason why building their representative selection was no easy task. Only a single translation was available online; a search in Polish and Italian libraries provided almost seventy candidate texts: signed or unsigned by a translator, published by a variety of publishers, often in several somewhat different editions. In the end, twenty-four translations produced until mid-20th c. have been identified as more or less independent of each other, although some of these still share over 50% of material, as evidenced by comparison of texts for identical 5- or longer word clusters with *WCopyFind* (Bloomfield 2011-2016). When applied to genuinely different translations, the similarity ratio is of the order of 5-7%.

The natively Italian literary polysystem was represented by close to 1300 different literary texts, mostly selected and adapted from *Progetto Manuzio*, one of the most comprehensive Internet collections of electronic texts in Italian. To include as many texts as possible, this set of Italian writing included dramas, epic poems and opera libretti as well as novels and novellas from the 15th to the 21st century. Several translations of other novels by Sienkiewicz were also added to the collection, and another big body of translations of a single author, Shakespeare, was included as well.

Methods

The stylometric method applied has been described by Eder (2017) and applied to other literary corpora by Rybicki (2014, 2016). Basing on Burrows's Delta procedure (2002), a list of most-frequent words (MFWs) is produced for the entire corpus. These words are then counted in the individual texts, and their frequencies are compared in text pairs to produce a matrix of distance measures; in this study, the distances were established by means of the modified Cosine Delta (Smith and Aldridge 2011), which is now seen as the most reliable version (Evert et al. 2017). The distance matrix then undergoes Cluster Analysis (Ward's hierarchical clustering), resulting in grouping the texts into "clus-

ters" of greatest similarity; this is repeated for reiterations from 100 to 2000 MFWs at 100-word increments, and a consensus between the individual iterations is produced to show each text's most consistent nearest neighbors, next-to-nearest neighbors and next-to-next-to nearest neighbors. The procedure is performed by means of *stylo* (Eder et al. 2016), a stylometric package for *R* (R Core Team 2016). The results are visualized by means of network analysis, applying the "Force Atlas 2" gravitational algorithm (Jacomy et al. 2008) in *Gephi* (Bastian et al. 2009) to the above-mentioned scores. Instead of applying a "human-made" classification of the resulting network of nodes and edges (i.e. identifying authors, genres and literary periods based on external and traditional literary history), the task of dividing the network into groups of greatest internal similarity was entrusted to *Gephi*'s modularity function, which finds communities within a weighted network (Blondel et al. 2008). The main experiment was conducted by successively increasing the number of communities shown until the expected separate cluster of translations of *Quo vadis* became a separate entity in the network, and the degree of its discreteness could thus be assessed.

Results

Dividing the network into just two modularity groups failed to isolate Sienkiewicz from the main Italian community. Instead, the main division was that between 19th/20th-century novels, translated or originally Italian, and everything else – the one notable exception to this rule was the prose of Pirandello, classified with the earlier texts. At three modularity groups, Italian drama detached itself from early prose. At four, the first writer became a separate community, but this was the native Deledda rather than the alien Sienkiewicz. At five, 19th- and 20th/21st-century novels became two distinct groups; at six, another native Italian, Salgari, received his own class; at seven, pre-19th-century works detached themselves from later prose. It is only at ten communities that a translated rather than an Italian author became a separate subsystem (to use Even-Zohar's term) – in fact, not one but two: Sienkiewicz (not just his *Quo vadis*) and Shakespeare (Fig. 1).

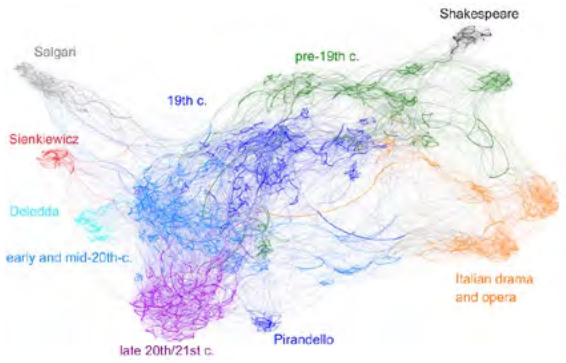


Figure 1. Network analysis of distances between most-frequent-word usage. Thick and short lines (edges) denote small distance (or high similarity). For simplicity, only the final 10-community modularity is shown.

Discussion

It seems too much of a coincidence that two major subsystems (translations of Shakespeare and translations of Sienkiewicz) become separated from the main body of literature in Italian at the same time, and that this happens only after two native authors receive their own subsystems. If such a mechanism were to be observed in even more extensive collection of texts (when they finally become available), Even-Zohar's hypothesis of the usually peripheral position of translated literature could find its stylometric illustration. At the same time, this experiment confirms not only that original novels are more similar to translated ones than the former to original drama; but also that certain original authors are more different from other original authors than those translated from another language.

Obviously, this hypothesis must be tested in the future in other literary polysystems to claim that the affinity between polysystem theory and macroanalysis is anything more than metaphorical. Even-Zohar speaks of reception of literary works within a broader national culture; macroanalysis counts context-free words. Still, in its attempts to bring distant and close reading together, stylometry has been clutching at even weaker straws. Stylometrists continue to make similar leaps (of faith?) between their graphs and trees and networks on the one hand, and traditional literary history on the other. They usually believe that frequencies of very frequent words provide insights into more abstract characteristics of texts than their mere lexical or even grammatical difference: and these abstracts so far include authorship, genre, chronology, or gender. This study might just have added a new one. At the very least, it is an invitation to apply Even-Zohar's concepts in various "distant" approaches to literature.

Acknowledgements

This research was made as part of the project: "Miejsce *Quo vadis*? w kulturze włoskiej. Przekłady, adaptacje, kultura popularna" (0136/ NPRH4/H2b/83/2016), funded by Poland's National Program for Advances in the Humanities (NPRH).

References

Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: an open source software for exploring and manipulating networks." *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, Ca.

Berti, G. de, and Gagetti, E. (2016). "La fortuna di 'Quo vadis' in Italia nel primo quarto del Novecento." In Woźniak, M., Biernacka-Licznar K., eds, *Quo Vadis. Da caso letterario a fenomeno di massa. Ispirazioni - adattamenti - contesti.* Roma: Ponte Sisto, 50-59.
Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment* 10: 1000.
Bloomfield, L. (2011-2016). *WCopyFind. The Plagiarism Resource Site*, http://plagiarism.bloomfieldmedia.com. Accessed 24. Nov. 2017.
Burrows, J.F. (2002). "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17: 267-287.
Eder, M. (2017). "Visualization in stylometry: Cluster analysis using networks." *Digital Scholarship in the Humanities* 32(1): 50-64.
Eder, M., Kestemont, M., and Rybicki, J. (2016). "Stylometry with R: A package for computational text analysis." *The R Journal* 8(1): 107–121.
Even-Zohar, I. (1990). "The Position of Translated Literature within the Literary Polysystem." In *Polysystem Studies [= Poetics Today]* 11(1): 45-51.
Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. (2017). "Understanding and explaining Delta measures for authorship attribution." *Digital Scholarship in the Humanities* 32 (sup. 2): 4-16.
Helgesson, S. and Vermeulen, P. (2015). „Introduction. World Literature in the Making." In Helgesson, S. and Vermeulen, P. eds, *Institutions of World Literature, Writing, Translation, Markets*. London: Routledge, 1-22.
Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2008). "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." *PLoS ONE* 9(6): e98679. DOI=10.1371/journal.pone.0098679.
Jockers, M. (2013). *Macroanalysis. Digital Methods and Literary History*, Champaign: University of Illinois Press 2013.
Marinelli, L. (1984). "'Quo vadis.' Traducibilità e tradimento," *Europa Orientalis* 3: 131-146.
R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Rybicki, J. (2014). "Pierwszy rzut oka na stylometryczną mapę literatury polskiej," *Teksty drugie* 2: 106-128.
Rybicki, J. (2016). "Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies," *Digital Scholarship in the Humanities* 31(4): 746-761.
Smith, P. and Aldridge, W. (2011). "Improving authorship attribution: Optimizing Burrows' Delta method." *Journal of Quantitative Linguistics*, 18(1): 63–88.
Woźniak, M. (2016). "Quo vadis: da caso letterario a fenomeno di massa. Dove ci ha portato Sienkiewicz?" In Woźniak, M., Biernacka-Licznar K., eds, *Quo Vadis. Da caso letterario a fenomeno di massa. Ispirazioni - adattamenti - contesti*. Ponte Sisto, Roma, 6-15

Interrogating the Roots of American Settler Colonialism: Experiments in Network Analysis and Text Mining

Ashley Sanders Garcia
asanders@cmc.edu
The Claremont Colleges, United States of America

Even as the United States fought for independence in the American Revolution, it was already in the process of becoming a settler colonial power in its own right. This short paper interrogates the origins of American settler colonialism through text mining three corpora of personal and official documents. In order to understand and address present structural inequity in the United States, scholars, policy-makers, educators, and the public need to examine the country's long history as a settler colonial society.

Through topic modeling and text mining methods, my research highlights the underlying goals and desires that prompted land acquisition, settlement, and cycles of violence between Euro-American settlers and Native Americans in the trans-Appalachian west between 1776 and 1820. This project explores three collections, or corpora, of documents, separated by the positions of the historical authors and document type: settler correspondence and records; official government documents; and writings of political elites in the eastern United States. The first corpus for this study consists of correspondence, journals, and memorials from settlers, colonial officials and military leaders in the territories (colonies) between 1776 and 1820. This is the smallest corpus of the three, at two million words. Few documents from representative settlers have been transcribed and published, so the corpus over-represents leaders in the settler communities, however the petitions from the settlers to Congress give voice to the most pressing challenges, needs, and hopes of the settlers themselves. The documents included in each corpus were transcribed and published in bound volumes during the nineteenth century and are now in the public domain. A second corpus, of approximately four million words, consists of official government records, including treaties with Native American communities, military records, documents related to public lands and governance of the territories, as well as pension and other petitions submitted to Congress in the late eighteenth and early nineteenth centuries. The third corpus is, by far, the largest of the three, at approximately 39 million words, and consists of the papers of the foremost political leaders in the eastern United States. The letters of the members of the Continental Congress are included, as are the writings of George Washington, James Madison, Thomas Jeffer-

son, Benjamin Franklin, and John Adams. Not surprisingly, these statesmen wrote far more than settlers, who were primarily concerned with agricultural cultivation, hunting, and defending their families on the frontier.

The aforementioned sources form the corpora for text mining and analysis experiments. My study extracts and compares American settler, administrator, and political leaders' perspectives on significant topics in the study of settler colonialism, such as land value; property acquisition and sales; as well as the presence, actions, and views of Native Americans. Early experiments using the LDA algorithm in MALLET to topic model the corpora and Lexos to visualize the topic clouds have already revealed significant patterns (Blei, 2012).

While recognizing that topic models are more effective with large corpora, my research began with a small experiment. Using MALLET, I created a topic model of ten topics of the twenty-five published petitions from settlers to Congress (1787-1798) from the *Territorial Papers of the United States.* This model suggests that one of the primary motivations for Euro-American emigrants to move to the western territories was to achieve what they described as a competency, or the means to rear their children "in a comfortable manner" and "raise a subsistence by their [own] industry" (Petition from the Inhabitants of Vincennes to Congress, 1787). The topic related to land reveals the dominant concerns that settlers expressed. They implored Congress to recognize their existing land claims, ensure reasonable land prices, provide military protection from Native American raids, and ensure justice through the provision of judges. These measures, they believed, would foster access to land, enable trade, establish legitimacy, and provide settlers with the means to achieve their modest goals.

Even though their objectives differed from those of the settlers, government officials both in the east and on the ground, in the western territories, were equally motivated to acquire land beyond the Appalachian Mountains. In the aftermath of the American Revolution, the government was in dire financial straits. Political leaders urged agents to obtain western lands from Native communities so that the territory could be sold to pay off the burdensome war debts. Consequently, backcountry government officials decried settler violence against neighboring Indigenous communities, even as they took advantage of the unruly settlers' actions to compel land cessions that the United States government desperately needed.

Figure 1: Topics related to land in the Continental Congress members' correspondence

There was a high price to be paid for white American independence though, as is demonstrated in the topics generated from the Continental Congress members' correspondence records (Figure 1). The words "transmitted, negotiations, ceding, extinguishment, extinguishing," and, ominously, "funeral" stand out among the more benign "northwest, lands, and western." Most of these words are more or less neutral when considered out of context, but, given their use in relation to the settler colonial endeavor, they evidence the brutal effects of American land acquisition and expropriation from Native communities.

These topics and the related documents both direct attention to specific sources for close reading, but also yield new terms of interest to explore at a distance and in a broad comparative framework. In addition to the results of topic modeling the aforementioned corpora, this presentation will also share experiments using part-of-speech tagging and collocations to explore concepts, such as land, family, independence, competency, and war to understand the ways in which settlers, and political and military leaders conceived each of these topics.

This talk offers an initial glimpse into the early stages of a much larger project that seeks to create an interactive interface for documents from the first four decades of the United States' formation as a nation and nascent empire based on topic models and text mining approaches, such named entity recognition, and collocates. The interface will eventually allow users to drill down into documents that contain specific sought-after features, such as individuals' names, gender identity, topics of interest, etc.

This interface, it is hoped, will enable historians, students, genealogists, and interested members of the public to explore some of the most important documents related to the complicated, conflicting, and, occasionally, complementary objectives of American settlers and other political actors. The policies these agents developed between 1776 and 1820 not only shaped American settler colonialism in the eighteenth and nineteenth centuries, but they continue to reverberate more than two centuries later.

References

Blei, D. M. (2012) "Probabilistic Topic Models." *Communications of the ACM* 55.4 (April 2012): 77-84.

Blei, D. M. (2012) "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2.1 (Winter 2012). Web. http://journalofdigitalhumanities.org/2-1.

The Inhabitants of Vincennes to Congress, July 26, 1787, in *Territorial Papers of the United States,* Volume 2 (Washington, D.C.: United States Government Publications Office, 1934): 58-60.

¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata?

Teresa Santa María
teresa.santamaria@unir.net
Universidad Internacional de la Rioja, Spain

Elena Martínez Carro
elena.martinez@unir.net
Universidad Internacional de la Rioja, Spain

Concepción Jiménez
concepcionm.jimenez@unir.net
Universidad Internacional de la Rioja, Spain

José Calvo Tello
jose.calvo@uni-wuerzburg.de
University of Würzburg, Germany

Resumen

¿Los nodos centrales de una red social de personajes son los protagonistas de una obra de teatro? Para responder a esta pregunta utilizamos diferentes medidas de centralidad junto con otros valores cuantitativos textuales en un corpus anotado de obras dramáticas de teatro español correspondientes a la Edad de Plata (1868-1936). Los resultados señalan que la centralidad está en correlación moderada con la importancia, siendo mayor la correlación con valores cuantitativos textuales.

Introducción

La representación de personajes literarios mediante grafos y redes sociales (Marcus 1973, Moretti 2011) aporta nuevas herramientas al estudio literario. La interpretación del concepto de centralidad en grafos (Jannidis et al., 2017) ha sido investigada en su aplicación a las obras literarias (Moretti 2011; Rochat 2014; Trilcke et al. 2015 y 2016; Jannidis et al., 2016, Rodríguez 2016; Algee-Hewitt 2017). En la tradición hispánica, se han utilizado enfoques cuantitativos para analizar la densidad versal en obras del Siglo de Oro (Hermenegildo 1994 y Espejo 2002), estudiar tanto contenido simbólico y sociopolítico de los personajes de Galdós (Menéndez 1983), así como el origen social o caracterización de los personajes de Lope de Vega (Oleza 1984 y Oleza 2013).

En este trabajo queremos evaluar cuatro preguntas:

1. ¿Qué tipo de correlación hay entre las medidas de centralidad y la importancia del personaje?
2. ¿Aparecen los personajes más importantes al comienzo del *dramatis personae*?
3. ¿Hay correlación entre importancia y valores textuales (cantidades de unidades textuales del personaje)?
4. ¿Qué valores podríamos utilizar para distinguir a los protagonistas del resto?

Textos y metadatos

A diferencia de otras lenguas europeas, el español no cuenta con un gran corpus teatral anotado en XML-TEI. El proyecto *Biblioteca Electrónica Textual del Teatro en Español de la Edad de Plata (1868-1936)* (BETTE) ha publicado veinticinco obras en XML-TEI de Lorca, Valle, Galdós, Clarín o Muñoz Seca, como repositorio GitHub (María Jiménez et al., 2017). En la versión 2.0 cada personaje ha sido anotado con diferentes metadatos:

- Sexo
- Papel en la obra (protagonista, amante, antagonista u otro)
- Naturaleza (persona, animal, no humano...)
- Importancia (personaje primordial, secundario o terciario)
- Persona individual frente a grupo

Además, se añadieron una serie de valores textuales cuantitativos de manera automática:

- Posición en el dramatis personae (castList)
- Cantidad de texto que pronuncia
- Cantidad de intervenciones
- Cantidad de referencias a su nombre
- Cantidad de escenas en las que aparece

Aquí un ejemplo de esa información en XML-TEI:

```
<person n="1" role="protagonist" sex="M" xml:id="max">
   <persName>Max Estrella</persName>
   <affiliation type="nature">person</affiliation>
   <affiliation type="importance">primary</affiliation>
   <ab>
      <measure unit="characters">15813</measure>
      <measure unit="sps">278</measure>
      <measure unit="rss">96</measure>
      <measure unit="scenes">11</measure>
   </ab>
</person>
```

Fig. 1. Metadatos de personaje en XML-TEI

El valor de importancia fue asignado según los siguientes criterios:

- Minor: si el personaje no aparece en el resumen (contenido también en el archivo TEI)
- Secondary: si aparece en el resumen
- Primary: si pertenece al grupo de entre dos y cuatro personajes esenciales

De esta manera por cada personaje (con un total de 516) tenemos:

1. Un valor de su importancia dentro de la obra (que puede ser utilizado como *ground truth*)
2. Diferentes valores cuantitativos textuales
3. Posición en dramatis personae
4. Diferentes valores según medidas de centralidad

Metodología

La implementación para extraer, analizar, evaluar y visualizar los datos se realizó en Python mediante librerías como *lxml* y *networkx*. Para la creación de las redes sociales se definió la arista no direccional como la coaparición en escenas (la definición más frecuente en trabajos de este tipo):

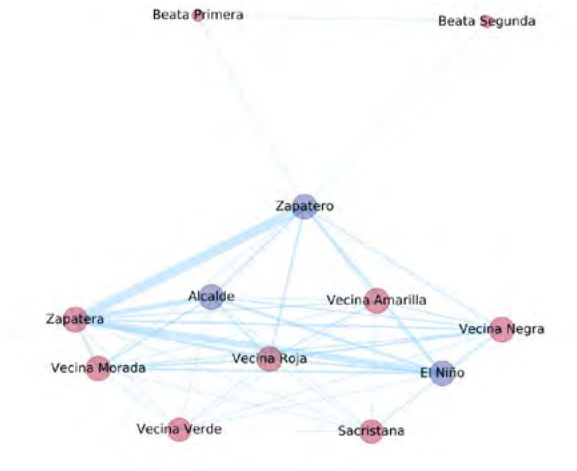


Fig. 2. Red social de personajes en La zapatera prodigiosa de Lorca

A partir de estas redes sociales, calculamos diferentes medidas de centralidad e información sobre los nodos:

- Degree
- Betweenness centrality
- Eccentricity
- Closeness centrality
- Load centrality
- Current flow betweenness centrality
- Eigenvector centrality
- Approximate current flow betweenness centrality
- Communicability centrality exp

Resultados

Analizamos la dependencia entre la importancia y el resto de valores, calculado su correlación (Spearman)

Ninguna de las medidas de centralidad tiene una correlación fuerte (> 0.6 o < -0.6 según Evans 1996). El valor máximo (0.51 en correlación negativa) es de *current flow betweenness centrality*, también conocida como *information centrality* (Brandes and Fleischer 2005; Stephenson and Zelen 1989), medida que no está entre el repertorio usual de las HD.

En cuanto a la posición en el *dramatis personae*, la correlación es solo de 0.42, con una fuerte dispersión, aunque los primeros y terceros cuartiles de personajes primarios y terciarios se posicionan en rangos totalmente diferentes. Es decir, la posición en el dramatis personae sí parece aportar cierta información sobre la importancia, aunque no podemos utilizarlo de manera exclusiva (p.ej. Muñoz Seca los ordena por sexo).

En tercer lugar, las medidas de cuantitativas textuales tienen todas correlaciones notablemente más altas, llegando hasta 0.67 en la cantidad de intervenciones.

Ante estos resultados, nos hemos preguntado si las medidas cuantitativas textuales tienen el mismo tipo de correlación con las medidas de centralidad, en concreto si la information centrality tiene una correlación más fuerte que el resto (calculando Spearman o Pearson, dependiendo si las variables son continuas u ordinales):

Como se observa *current flow betweenness (o information) centrality*, de nuevo, es la medida de centralidad con la correlación más fuerte con la cantidad de intervenciones.

Finalmente hemos observado si la distribución de centralidad o valores textuales son diferentes para los personajes protagonistas de los del resto:

La mayor diferenciación de ambos *boxplots* entre las medidas de centralidad se consigue mediante *current flow betweenness* (o *information*) *centrality*. El solapamiento menor se consigue mediante la cantidad de texto pronunciado (pers_mes_characteres). La posición relativa en el *dramatis personae* en este caso consigue diferenciar de manera bastante clara los protagonistas del resto de personajes.

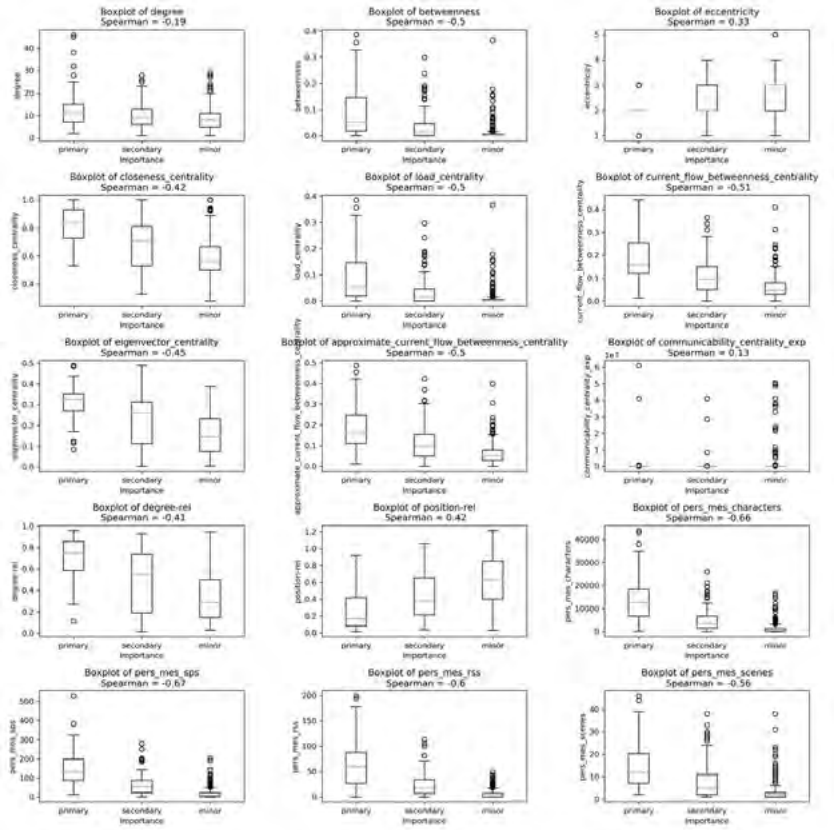


Fig. 3. Boxplots y correlaciones con importancia de todas las obras de BETTE

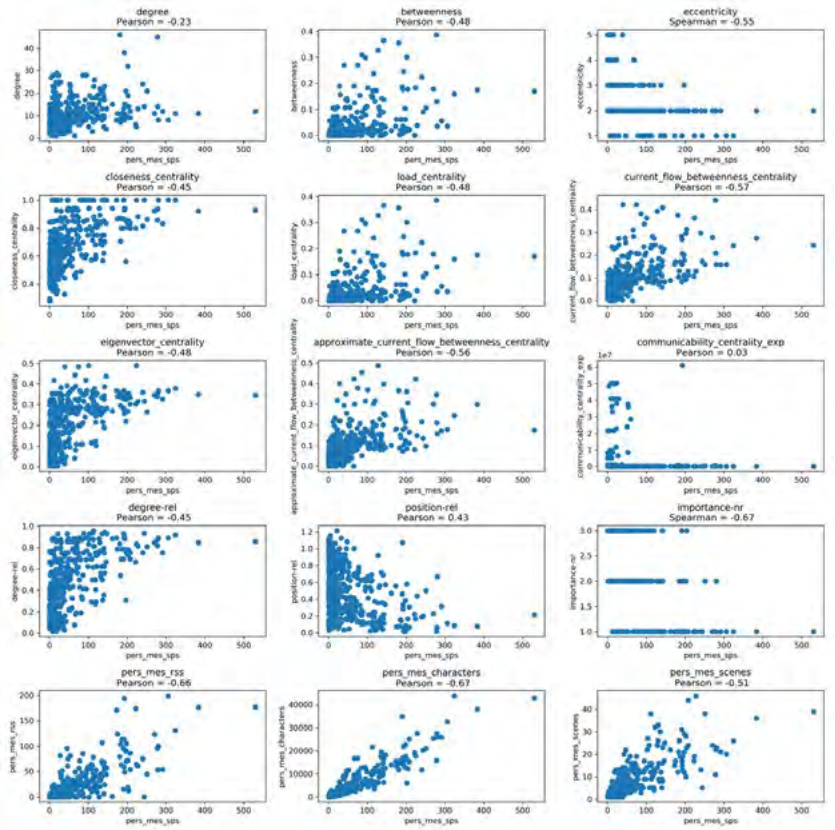


Fig. 4. Scatterplots mostrando correlación entre las veces que un personaje habla (<sp>s) y otros valores

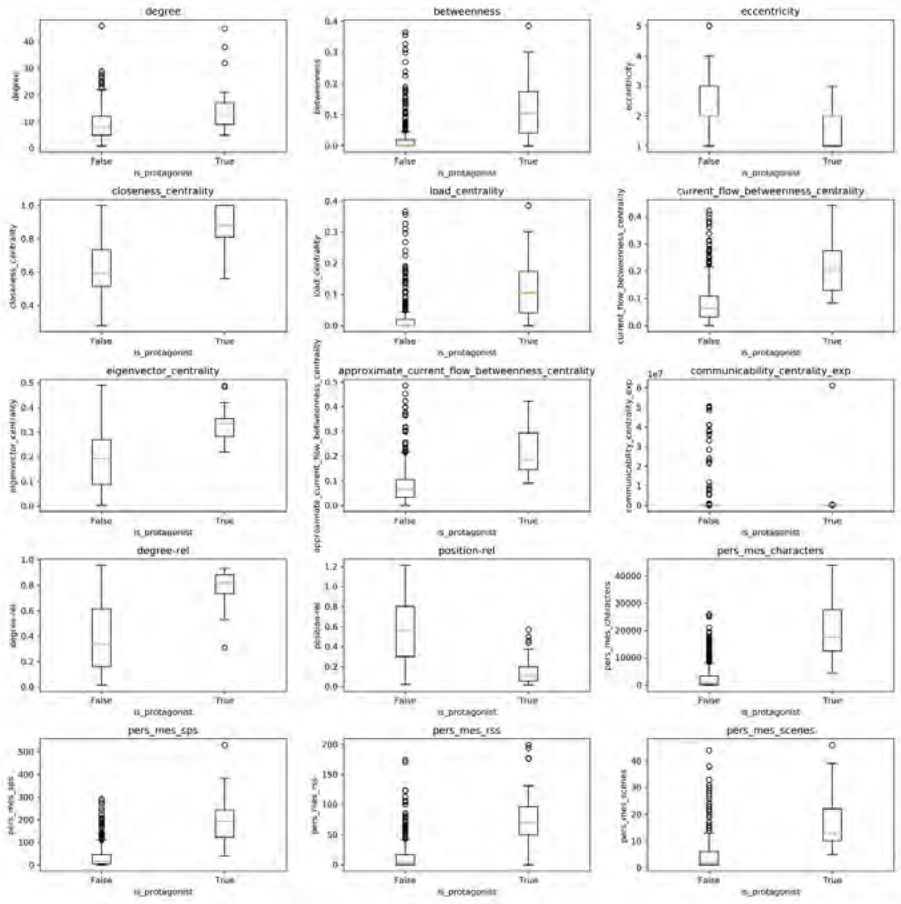


Fig. 5. Boxplots de protagonistas frente al resto de personajes

Conclusiones y futuros pasos

La anotación en detalle de información sobre los protagonistas nos permite evaluar métodos digitales. En concreto seguimos la propuesta de Moretti (2013) de abandonar la división binaria de personajes, incluyendo en nuestro caso los valores de personajes secundarios.

Nuestros resultados muestran que, para el caso del corpus BETTE y con las formalizaciones arriba explicadas:

1. La importancia tiene una correlación solamente entre débil y moderada con cualquier formalización de centralidad, teniendo la correlación más fuerte la *information centrality*
2. La posición en el *dramatis personae* puede ser un indicador sobre el protagonismo de personajes o la diferenciación entre primarios y terciarios, pero no para diferenciar a estos de los secundarios
3. Los valores cuantitativos textuales tienen correlaciones más fuertes. Este tipo de unidades son también las que mejor clasificarían personajes entre protagonistas y no protagonistas
4. Es sorprendente que unidades textuales más sencillas que la centralidad en redes aporten más información tanto sobre la importancia de los personajes, así como su papel de protagonistas.

Como otros trabajos en redes sociales (cf. Moretti 2011 y 2013; Rochat 2014) hemos trabajado con una cantidad reducida de textos. Nos gustaría comprobar estas hipótesis en mayores corpus literarios. También nos gustaría analizar los efectos que subgéneros literarios, períodos y autores ejercen sobre estos valores.

References

Algee-Hewitt, Ma. (2017). *Distributed Character: Quantitative Models of the English Stage, 1500-1920*. Montréal: McGill University & Université de Montréal, pp. 119–21.

Brandes, U. and Fleischer, D. (2005). Centrality Measures Based on Current Flow. *Theoretical Aspects of Computer Science (STACS '05)*. Springer-Verlag, pp. 533–44 http://www.inf.uni-konstanz.de/algo/publications/bf-cmbcf-05.pdf.

Espejo, J. (2002). Algunos aspectos sobre la construcción del personaje en el teatro conservado de Hernán López de Yanguas (1487-¿?). *Scriptura,* 17, pp. 113-132.

Evans, J. D. (1996). *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove: Brooks/Cole Pub. Co.

Gómez, S., Calvo Tello, J., González, J. M. and Vilches, R. (2015). Hacia una biblioteca electrónica textual del teatro en español de 1868-1936 (BETTE). Texto Digital, 11(2), pp. 171–84.
Hermenegildo, A. (1995). Personaje y teatralidad: la experiencia de Juan del Encina en la Égloga de Cristino y Feba. In Pedraza, F.B. y González, R. (ed.*). Los albores de teatro español: actas de las XVII Jornadas de teatro clásico Almagro, julio de 1994*. Almagro: Universidad de Castilla-La Mancha, pp. 90-113.
Jannidis, F., Reger, I., Krug, M., Weimer, L., Macharowsky, L. and Puppe, F. (2016). Comparison of Methods for the Identification of Main Characters in German Novels. *DH2016*. Krakow: ADHO, pp. 578–82 http://webcache.googleusercontent.com/search?q=cache:LjYz88cQhboJ:dh2016.adho.org/abstracts/297+&cd=1&hl=es&ct=clnk&gl=de&client=ubuntu.
Jannidis, F., Kohle, H. and Rehbein, M. (eds). (2017). *Digital Humanities: eine Einführung*. Stuttgart: J.B. Metzler Verlag.
Jiménez, C., Martínez Carro, E., Santa María, M. T., Calvo Tello, J., Simón Parra, M., Martínez Nieto, R. B. and García Sánchez, M. (2017). BETTE: Biblioteca Electrónica Textual del Teatro en Español de la Edad de Plata. *Sociedad, Políticas, Saberes*. Málaga: HDH, pp. 88–91 http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf.
Marcus, S. (1973). *Mathematische Poetik*. (Trans.) Mândroiu, E. Bucureşti; Frankfurt/Main: Editura Academiei ; Athenäum Verlag.
Menéndez, C. (1983). *Introducción al teatro de Benito Pérez Galdós*. Madrid: CSIC.
Moretti, F. (2011). Network Theory, Plot Analysis. *The New Left Review (*68), pp. 80–102.
Moretti, F. (2013). "Operationalizing": or, the function of measurement in modern literary theory. *The New Left Review (*84), pp. 103-119.
Oleza Simó, J. (2013). *Biblioteca Digital Arte Lope*. Valencia: Universitat de València. artelope.uv.es/biblioteca.
Rochat, Y. (2014). *Character Networks and Centrality*. N.p. Web.
Rodríguez, D.I. (2016) *Análisis de grafos en paralelo mediante Graphx*. Trabajo de titulación. Universidad Católica de Loja. Ecuador.
Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, 11(1): 1–37 doi:10.1016/0378-8733(89)90016-6.
Trilcke, P., Fischer, F., Göbel, M. and Kampkaspar, D. (2016). Dramen als small worlds? Netzwerkdaten zur Geschichte und Typologie deutschsprachiger Dramen 1730-1930. In Burr, E. (ed), *DHd 2016 Modellierung, Vernetzung, Visualisierung*. Leipzig: DHd/nisaba, pp. 254–57.
Trilcke, P., Fischer, F. and Kampkaspar, D. (2015). Digitale Netzwerkanalyse dramatischer Texte. *DHd-Tagung. Graz* http://gams.uni-graz.at/o:dhd2015.v.040.

Cultural Awareness & Mapping Pedagogical Tool: A Digital Representation of Gloria Anzaldúa's Frontier Theory

Rosita Scerbo
scerbo@asu.edu
Arizona State University, United States of America

This project looks at the work of American-Chicana poet and fiction writer Gloria E. Anzaldúa, author of *This Bridge we Call Home* (2002). My research proposes a digital representation of Gloria Anzaldúa's Frontier theory as part of my scholarly investigation. This study will include the creation of a mapping tool that will reflect the rhizomatic spaces analyzed by the author, raising awareness about the multiple cultural identities found in the United States. Through personal narratives, theoretical essays, poetry, letters, works of art and fiction, *This Bridge we Call Home* examines issues such as classism, homophobia, racism, political identity, native sovereignty, lesbian pregnancy and motherhood, transgender issues, Arab-American stereotypes, Jewish identities and spiritual activism. These stories are written by women and men, both of color and white, and motivated by a desire for social justice. *This Bridge We Call Home* invites feminists of all colors and genres to develop new forms of transcultural dialogues, practices, and alliances. The anthology, object of study is the last work produced by the author before she passed away and undertakes a more inclusive essence compared with her earlier writings. The book includes women and men of different classes, nationalities, races, ages and sexual orientations, reflecting the desire of inclusivity and dialogue promoted by the author and editor. This project also attempts to bring together multiethnic voices and promotes a interdisciplinary resource that interest not only the literature and culture discipline, but also other humanities fields, such as history, anthropology, sociology and gender studies.

The result of this project will be a powerful new online education and research tool for undergraduate and graduate students as well as the world community at all levels of expertise. To create this public resource I will use the mapping tool "Google Lit Trips", a site affiliated with Google. Normally this tool is used to recreate and mark the journeys of fictional characters from famous literature works. In my case I will use the various sections of Gloria Anzaldúa's anthology that reflect real life experiences of the writers. I will then provide geospatial representations of the true stories narrated by the authors that live some kind of political, racial, sexual or class struggle in the United States. In the book 87 writers are given a space to celebrate their diversity.

In the mapping tool, at each location along the journey there will be placemarks with pop-up windows con-

taining a variety of resources including relevant media, thought provoking discussion starters, and links to supplementary information about 'real world' references made in that particular portion of the text.The author voice herself emerges beyond the limits of either American or Mexican culture and provides a voice to the people of the borderlands. Her work is based on multiple experiences to create a universal history that transcends the social barriers that connect us collectively with each other. While the politics of identities requires subjecting ourselves to specific categories of identity, spiritual activism requires that we get rid of all these barriers.

This project has the objective to put the reader inside the stories, provoking reflections and awareness about contemporary social, political, sexual and racial issue that affect our modern society. The reader will travel alongside with the protagonists of the autobiographical stories through the recreation of 3D geographic tours of the narratives that have been described. At the same time the mapping tool creates an engaging and relevant literary experiences for students. At each location I will be able to include web links, videos, audios, images, annotations and critical activities related with the different sections of the anthology. The experience of the pop up windows provide a range of supplementary information, such as links that give additional information about the 87 authors or cultural traditions that have been mentioned by the characters. The students find themselves seeing the settings almost how they were there. The pop up windows provide engaging content, such as audios, videos or activities related to the story line. These activities are designed to help readers discover connections between their culture and the different cultures that have been described in the story.

One of the primary goals of this project is to emphasize the relevance of cultural diversity in the University environment in the context of the Hispanic world. My objective is to initiate contemporary debates over themes such as immigration, globalization, discrimination, acceptance and inclusion. The mapping tool will explore ways of bringing its unique materials to a wider audience inside and outside the United States. The contribution of this project is not only to continue expressing a dialogue within and between women, women of color, and among people that live in the borderlands, but also to expand visions and theoretical spaces in general. The different stories told in the anthology explore the different shades of the mixed-race identity of women and men that are often perceived as outsiders within their own country.

The digital representation of the anthology and its multiple resources proposes a new attitude towards the learning process of college students and the public sensitivity outside the academia. One of the primary intentions is to dismantle traditional forms of identity, and destroy social boundaries, by embracing difference and otherness as a unique component of every single individual part of our society. The focus on themes such as the effects of migration and globalization are evident in the transnational, transcultural and transgender identities represented though the voices of the 87 writers. The external links provided as resources bring the readers beyond the stories. The students become travelers discovering the similarities and qualities of the characters from cultures beyond their own. This could be an effective way to make students feel part of the stories and hopefully inspire them to fight against the different levels of discrimination that the writers are describing. The final goal will be to include this online platform as an integrative portion of a culture and literature class at the university level.

Corpus Linguistics for Multidisciplinary Research: Coptic Scriptorium as Case Study

Caroline T. Schroeder
carrie@carrieschroeder.com
University of the Pacific, United States of America

The Coptic language is the last phase of the Egyptian language family, descending ultimately from the ancient hieroglyphs. Coptic Scriptorium has developed a multidisciplinary research platform using core Corpus Linguistics tools and methods in collaboration with other disciplinary methods. This paper will argue that this collaborative, interdisciplinary approach allows for the creation of research resources that enrich even *disciplinary* work.

Coptic Scriptorium has created the first open source natural language processing tools for any phase of the Egyptian language family, including a tokenizer, normalizer, part of speech tagger, language of origin tagger (for loan words from Greek, Latin, and other languages), and lemmatizer. We have also contributed annotated data to the universal dependency Treebank project. A fully searchable corpus annotated with these tools is available online at copticscriptorium.org, and all tools and corpora can be downloaded from our GitHub repositories.

This paper will argue that multidisciplinary collaboration improves even disciplinary research. Three examples are provided here; these and others will be demonstrated live in the short paper.

Collaboration with Egyptologists creating a TEI Coptic lexicon file enabled the creation of an online Coptic Dictionary, in which words in our searchable database are hyperlinked to the dictionary entries. The dictionary entries likewise show frequency statistics for the terms in our database. This collaboration benefits Egyptology, by providing an open source corpus for teaching and research linked to a dictionary, and it benefits corpus linguistics, by providing clear frequency data and lexical resources for linguists.

Collaboration with Religious Studies scholars has enabled including in our corpora transcriptions of Coptic manuscripts that have never before been published in print. Scholars in Religious Studies have provided transcriptions of texts to the project, enabling scholars in other disciplines, such as Linguistics, to conduct computational corpus research on important, previously inaccessible texts. Likewise Religious Studies scholars can use the database to conduct philological and historical research on religious texts.

Coptic Scriptorium also annotates manuscript information of interest to archivists, philologists, and codicologists within a multilayer annotation model. This enables codicologists, philologists, and archivists to use the query syntax of our corpus linguistics database (ANNIS) to investigate research questions about scribal practices, spelling and morphology, and other manuscript-related issues over multiple manuscripts, including utilizing metadata such as repository information, dates and locations of the original manuscripts, etc.

We presented the very beginnings of the Coptic Scriptorium project at DH 2014 in Switzerland. This short paper will demonstrate the extensive progress made as a result of collaboration and interdisciplinary partnerships.

Extracting and Aligning Artist Names in Digitized Art Historical Archives

Benoit Seguin
benoit.seguin@epfl.ch
EPFL, Switzerland

Lia Costiner
lia.costiner@epfl.ch
EPFL, Switzerland

Isabella di Lenardo
isabella.dilenardo@epfl.ch
EPFL, Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
EPFL, Switzerland

The largest collections of art historical images are not found online but are safeguarded by museums and other cultural institutions in photographic libraries. These collections can encompass millions of reproductions of paintings, drawings, engravings and sculptures. The 14 largest institutions hold together an estimated 31 million images (Pharos). Manual digitization and extraction of image metadata undertaken over the years has succeeded in placing less than 100,000 of these items for search online. Given the sheer size of the corpus, it is pressing to devise new ways for the automatic digitization of these art historical archives and the extraction of their descriptive information (metadata which can contain artist names, image titles, and holding collection). This paper focuses on the crucial pre-processing steps that permit the extraction of information directly from scans of a digitized photo collection. Taking the photographic library of the Giorgio Cini Foundation in Venice as a case study, this paper presents a technical pipeline which can be employed in the automatic digitization and information extraction of large collections of art historical images. In particular, it details the automatic extraction and alignment of artist names to known databases, which opens a window into a collection whose contents are unknown. Numbering nearing one million images, the art history library of the Cini Foundation was established in the mid-twentieth century to collect and record the history of Venetian art. The current study examines the corpus of the 330'000+ digitized images.

Image Processing Pipeline

Photo/Cardboard Extraction

The records in the Cini Foundation consist of a photographic reproduction mounted on a cardboard card onto which metadata information is recorded. The initial scan of these records is a 300 dpi picture produced on a scanning table, and includes the digitized cardboard and color balance markers. The first task consists in separating the cardboard backing and the photographic reproduction from the raw scanned image.

Despite the apparent simplicity of such a task, it proved challenging on account of the multiple layouts of the metadata information on the cardboard cards, and the variations in the sizes and positions of the attached images. In the end, what proved most effective in the extraction of the image was a Convolutionnal Neural Network (CNN) architecture designed for semantic segmentation (Ronneberger, O. et al 2015). For this, an accurate model was trained on scans which had been annotated in the course of 2 hours. The details oft he approach are part of another study (Ares Oliveira, S. and Seguin, B. 2018).

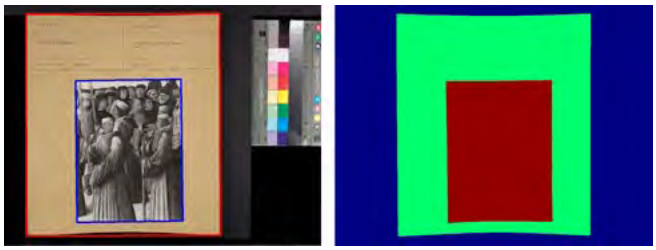


Figure 1 Left: original scan with the extracted areas highlighted with red and blue rectangles.
Right: the prediction mask generated by the neural network.

Text Extraction

The second part of the pipeline consists of extracting and reading the metadata. For this task, the open-source Tesseract toolkit and the commercial Google Vision API were tested, with the latter having better performance.

The OCR system provided a list of words and their positions, which were then clustered into blocks of text representing the different metadata fields (authorship, title of painting, location etc.). A layout model was used to represent the expected positions of these different fields. This allowed the assignment of each block of text to its corresponding metadata field.

A precise analysis of the performance of this step is presented in another publication (Seguin, B. 2018).

Figure 2 Illustration of the OCR process. The extracted words (top-left) are clustered into blocks of metadata (top-right) and then assigned to their corresponding label (bottom).

Automatic Alignment of Artist Names

In order to leverage the extracted metadata to get insights into a collection, it is important to link them to a knowledge database. This can allow, for example, city names to be placed geographically on a map. Here, we focus on aligning artist names with a knowledge database: the Union List of Artist Names (ULAN), managed by the Getty. This opens up a wealth of new information for the contextual understanding of the artwork's creation.

The alignment process is depicted on Figure3, it is a complex two-pass process that integrates automatic matching with collection specific knowledge in an efficient manner. The first pass tries to perform an exact match with a large name dictionary. For the second pass, a list of candidates are generated from the correctly matched elements of the first pass, and approximate matching is used to correct small OCR errors.

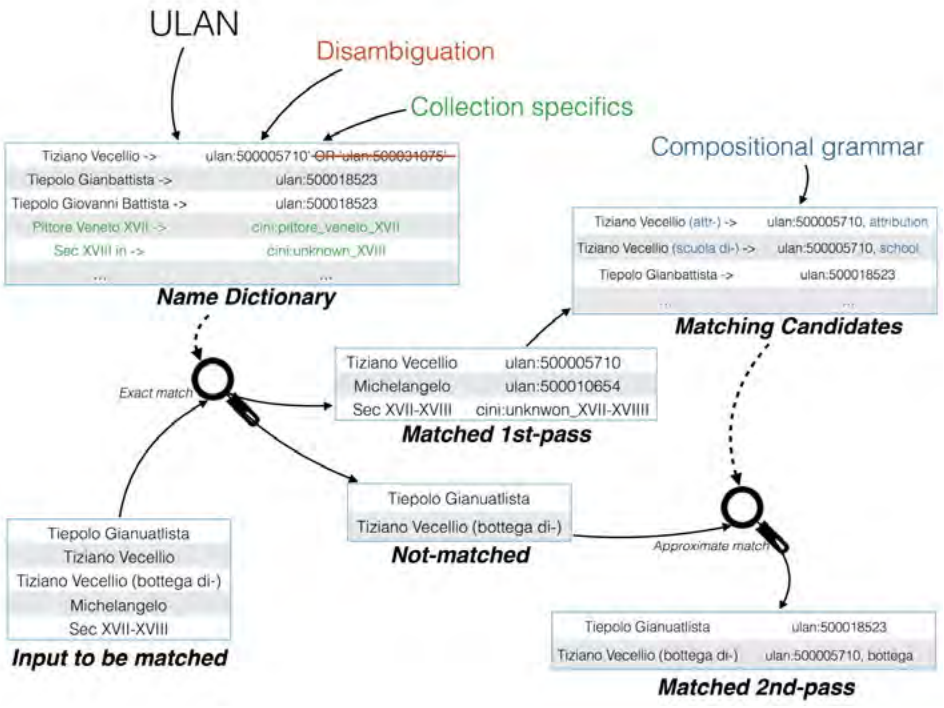


Figure 3 Alignment process. The parts in color correspond to collection-specific knowledge.

There are three challenges that needed to be tackled during this alignment process :

- *Names variation* : one major issue that arises is that a given artist may be called by different names, depending on regional variations and pseudonyms. Many variations are recorded in ULAN (i.e "*Tiepolo Gianbattista*" and "*Tiepolo Giovanni Battista*" both corresponding to the same artist), although some have to be added to the name dictionary. Furthermore, the naming conventions for elements whose dating or provenance is known but not authorship, which may be specific to a collection, can be added to the dictionary.
- *Implicit knowledge* : one related challenge is linked with the pragmatics of the annotation process. Understanding that if one archivist writes "*Leonardo*" on a file, he or she is referring to *Leonardo da Vinci* implies modeling a series of implicit assumptions which are changing depending on the evolution of local cataloging practices and that of the art historical field itself. In our case, we tackle this by disambiguating unclear names. For instance "*Tiziano Vecellio*" could technically refer to the well-known "*Tiziano*", or his relative "*Tizianello*", but the first is much more prominent than the second.
- *Compositional structure* : the last challenge is linked with the practice of archivists to describe particular unknown authors using specific syntactic process like ("*Tiziano (bottega di-)*", "*Tintoretto (Maestro di)*" or "*Michelangelo (copia da-)*"), referring to workshop productions or copies. Understanding and modeling this "grammar" permits to generate, in a compositional manner, potential matching strings to be considered when looking for possible alignments. Such strings do not only give a link to an artist but also qualify relationships (how strongly an artist was involved in the creation process of a painting, whether the piece is an original or a copy, etc.).

Results

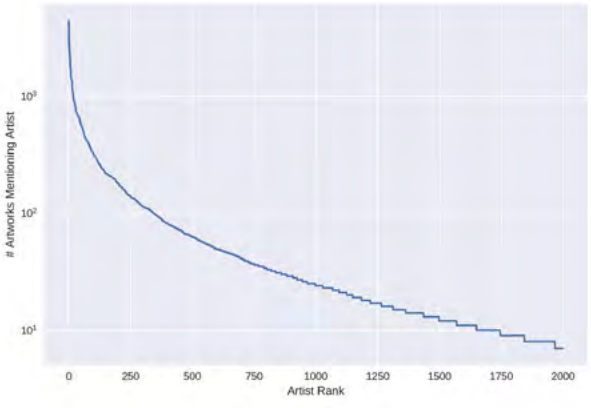
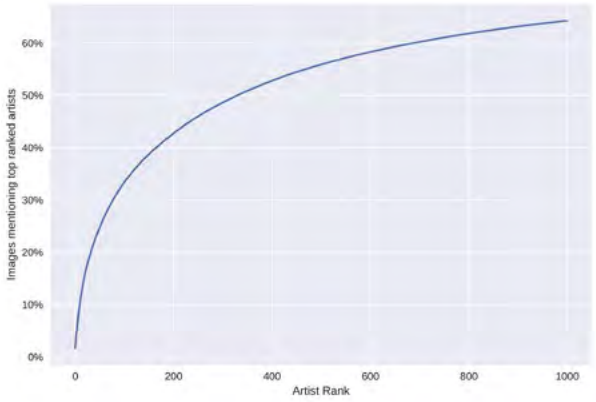


Figure 4 : Distribution of number of artworks assigned for each artist.



Proportion of images assigned with respect to the most common artists. The 200 most represented artists represent 43% of the collection.

Of the 330,078 scans composing the corpus of study, 14.6% had an empty author field, mostly because the photographs represented architecture or aerial city views. Out of the remaining 85.4% with an authorship field, 73.8% were automatically matched to an author (61.6% after the first pass), with an additional 1.4% representing ambiguous situations which could be resolved. This accounts for 208'510 elements automatically matched. At the end of pre-processing, the potential author names can be divided into three categories :

- (A) Author names which have been matched with a reference record of another database
- (B) Author names which may have been matched if the algorithm were to be improved (e.g. in terms of author name variation or possible compositional structure)
- (C) Authors undocumented in standard databases of artists.

Figure 5 shows the global matching results for category A. The geographical composition of aligned authors is dominated by Venetian artists (Tiepolo, Tintoretto, Palladio, Tiziano, Veronese, etc.) showing the rationale behind the creation of the collection. In terms of chronology, the collection is focused on the sixteenth century, as shown by the distribution of year of death of the aligned artists. This is in line with the period referred to as the "Venetian Golden Age". Figure 4 shows the very uneven representation of artists, with only 346 having more than 100 images, representing more than 50% of the whole collection.

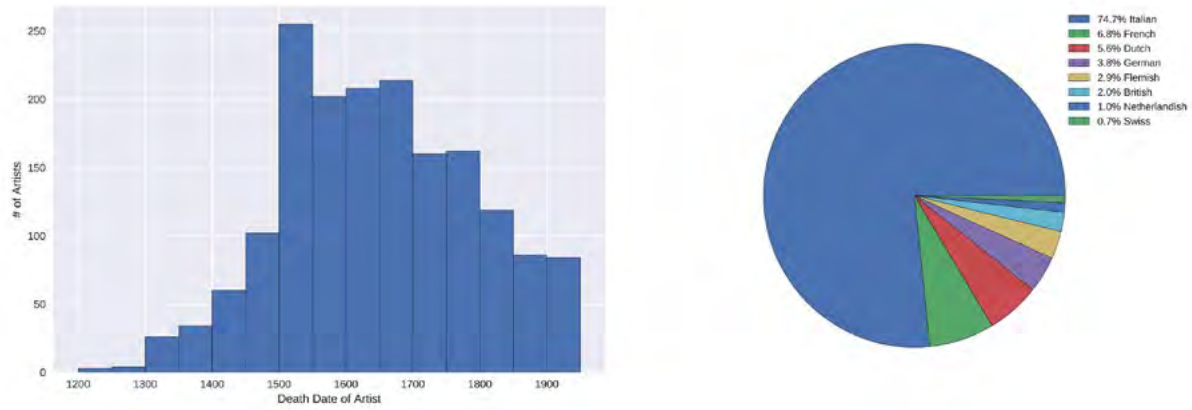


Figure 5 Spatial (right) and temporal (left) distribution of the 1'746 artists with at least 10 images assigned.

Category B is predominant in the elements that were not matched. Apart from OCR errors, the most typical unmatched string corresponds to collective works in which several authors are named. For instance, the string "*Bassano Jacopo e Francesco*" (his son) corresponds to 134 records. Adding additional parsing capabilities to the system could enable the resolution of such cases in the future.

Names in category C, which were not matched with ULAN, are in fact not a product of misalignment but represent new discoveries in the collection. In the present study, a number of artists who do not feature in ULAN were uncovered in the Cini archive. These include, Augusto Caratti, a minor artist from nineteenth-century Padua, who is represented by 65 images in the Cini collection, and Natale Melchiori an early eighteenth-century painter from Castelfranco, Veneto, represented by 39 images. Another artist who does not feature in the ULAN database but nevertheless has a significant presence in the Cini archive with 106 drawing, is Antonio Contestabile, an eighteenth-century draftsman from Piacenza.

Conclusion

These early results show the potential of the systematic processing of a large number of art historical records, leading to the mapping of unknown collections, and to new discoveries. It also highlights for the first time the challenges inherent in the process. Such challenges, it is important to note, are not purely technical but rather linked with the complexity of modeling local archiving traditions and the historical practices of art history.

References

Pharos. *PHAROS: The International Consortium of Photo Archives.* http://pharosartresearch.org/

Ronneberger, O. and Fischer, P. and Brox, T. (2015) *U-Net: Convolutional Networks for Biomedical Image Segmentation.*

D. A. Brown, D. A. and Ferino-Pagden, S. and Anderson, J. and Berrie, B. H (2006) *Bellini, Giorgione, Titian, and the Renaissance of Venetian painting*

Ares Oliveira, S.* and Seguin, B.* and Kaplan, F. (2018) *dhSegment: A generic deep-learning approach for document segmentation.*

Seguin, B. (2018) *New Techniques for the Digitization of Art Historical Photographic Archives—the Case of the Cini Foundation in Venice,* Proceedings of Archiving 2018.

A Design Process Model for Inquiry-driven, Collaboration-first Scholarly Communications

Sara B. Sikes
sara.sikes@uconn.edu
University of Connecticut, United States of America

Even as the scholarly communications field pursues the opportunities presented by digital technology, its routine operations remain anchored in print-centric regimens. For those working to evolve scholarly communications in the Internet age, particularly as it bears upon long-form scholarship, there is compelling need to productively disrupt and reconfigure the workflows and work cultures that have naturalized around the production of printed products. It is precisely this complex, systemic issue that Greenhouse Studios | Scholarly Communications Design at the University of Connecticut (UConn) addresses with its design-based, collaboration-first model of scholarly production.

With funding from the Andrew W. Mellon Foundation, the Digital Media & Design Department at UConn, the University Library and UConn Humanities Institute launched Greenhouse Studios in 2017. As a transdisciplinary collective, Greenhouse Studios employs design-thinking methodology to long-form digital scholarship. With its first two cohorts of collaborative projects, the Studios implemented an inquiry-driven approach that addresses the

divided workflows and counter-productive labor arrangements that have complicated scholarly communications in the digital age.

While the introduction of digital tools across the "information chain" model of scholarly communications has altered activities from research and writing through to preservation and reading, it has not reconfigured the larger workflow in which the various actors remain interlinked but largely independent save for key transactional, or "handoff," moments (CNI, 2016). Simply put, the "information chain" of scholarship begins with a knowledge creator, passes through to a publisher and culminates with accessibility secured by libraries and use by readers (Owen, 2002: 275-88). This transactional model has contributed to the persistence of an increasingly detrimental division of activities into those of the knowledge creation, or "domain," side and those of the production, or "build," side (Sosin, 2016).

By disrupting and reconfiguring divided workflows that have naturalized around the production of printed products,Greenhouse Studios brings together project teams on the "domain" side versus the "build" side. Each year, a new theme or problematic frames the work of the project teams, and diverse groups of collaborators are brought together, including designers, developers, editors, faculty and librarians. Starting with a problematic or issue rather than a faculty interest flattens counterproductive hierarchies and bringing in partners early in the process lends itself to the collaboration-first approach of the creation and expression of knowledge. Digital formats for the projects are not presupposed, as the format—digital or analog—that best represents the long-form scholarly work is taken under consideration. The first cohort of Greenhouse Studios teams developed projects in diverse formats including a documentary film, a virtual reality environment and an electronic decision-making novel.

Guiding the work of the teams, the Greenhouse Studios design process model provides a workflow for each project through five major sprints or phases. The design process modelwas developed through a series of exercises to elicit individual mental models of the scholarly design process from the perspective of a project manager, scholar, designer, repository manager, digital scholarship librarian, developmental editor and MFA student/research assistant. Comparisons of the mental models highlighted similar project phases for each participant, although the points of intersection were often differently identified. In looking at these points of overlap, neutral descriptors for shared activities were adopted, both for mutual intelligibility and to eliminate the kinds of value judgments that domain-specific terms may inscribe.

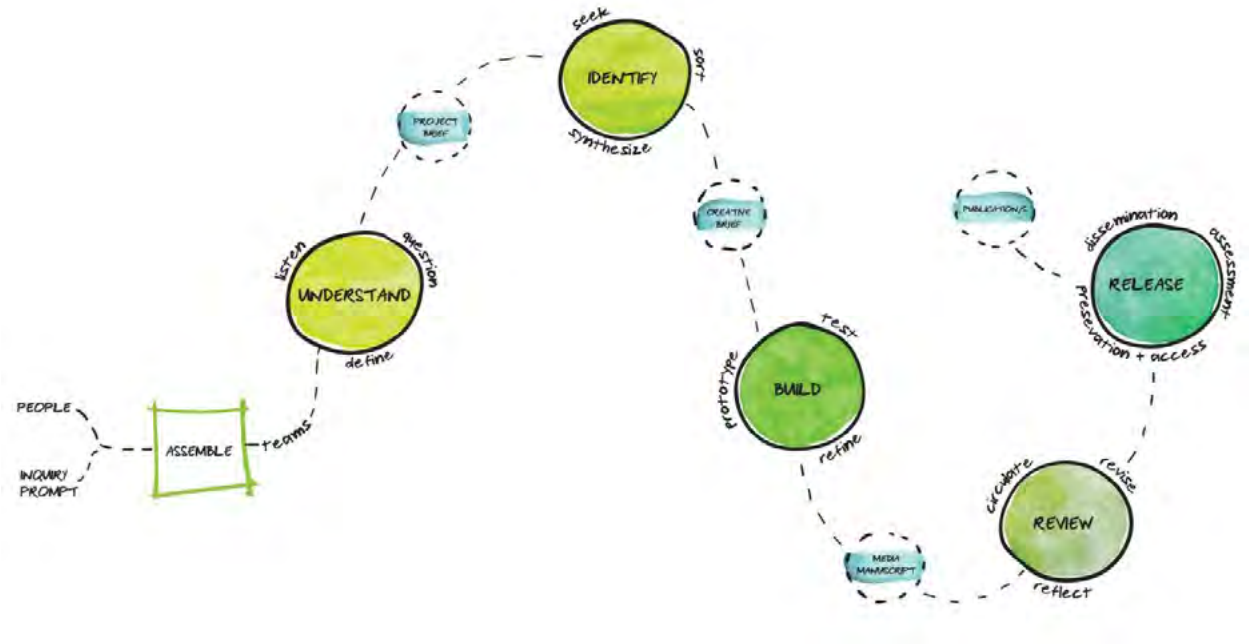


Figure 1. Greenhouse Studios Design Process Model

Being mindful of the Greenhouse Studios goals for workflow and work culture, the design process model adopts elements from the long tradition of design thinking as applicable across diverse fields. Design thinking as taught, practiced and disseminated by its most well-known and long-standing academic and corporate proponents, Stanford University's Hasso Plattner Institute of Design (aka the d.school) and the design firm IDEO, traces its roots to the 1960s' merger of a Stanford program that joined arts and mechanical engineering (Miller, 2015). Today, it has extended to endeavors as far afield as finance, films, museum exhibition, journalistic communications, education, and critical making in the digital humanities. Across various incarnations, design thinking processes

typically involve a series of iterative discovery and development cycles, each characterized by a subset of activities designed to facilitate that cycle's goal.

Work through the Greenhouse Studios design process model begins with an inquiry or prompt and brings together team members in response to a central problematic. During the catalyst phase of **Assemble,** team members gather, meet fellow participants and review the guidelines for project teams. The relevant human talents and other resources are defined during the first full sprint, or the **Understand** phase, which produces a *project brief* framing the project's aims and audiences. During the subsequent **Identify** phase, relevant sources of knowledge and inspiration are researched and synthesized. The resulting *creative brief* outlines the media formats of the project, as well as the formal peer review and assessment plans for the work. Iterative prototyping and refining of a project takes place during the following **Build** phase, producing a *media manuscript*, which could be a website, book manuscript, documentary film, exhibition or other format. During the **Review** phase, the project is revised, edited and submitted for peer review. The final phase is the **Release** or launch of the project, as well as the longer-term work of dissemination, assessment and preservation. Adjacent to this phase, there may also be *other publications* produced by individual project team members.

This design process model guides each of the Greenhouse Studios inquiry-driven, collaboration-first projects. The implementation of the process began before the launch of the first cohort of projects,and the model has undergone subsequent iterations across several development cycles. The team participants and an inquiry prompt act as catalysts for the workflow, which places collaboration at the center of the process, rather than an individual scholar's research goals. The emphasis on the "collaboration-first" nature of the process allows participants to collectively imagine scholarly projects from the outset and serves as a corrective to divided workflows, even digital-centric ones, where collaborators are only brought on board for the final implementation of projects.

References

Coalition for Networked Information. (2016). Supporting the Digital Humanities: Report of a CNI Executive Roundtable, 3. https://www.cni.org/wp-content/uploads/2016/05/CNI-SupportDH-exec-rndtbl.report.F14.pdf.

Miller, P. N. (2015). Is "Design Thinking" the New Liberal Arts? *The Chronicle of Higher Education*, *The Chronicle Review*, 61 (29). http://chronicle.com/article/Is-Design-Thinking-the-New/228779/.

Owen, J. M. (2002). The New Dissemination of Knowledge: Digital Libraries and Institutional Roles in Scholarly Publishing. *Journal of Economic Methodology*, 9 (3): 275-88.

Sosin, J. (June 29, 2016). Associate Professor, Department of Classical Studies and Director of the Duke Collaboratory for Classics Computing, Duke University. Interview by Tom Scheinfeldt, Clarissa Ceglio, and Sara Sikes.

Métodos digitales para el estudio de la fotografía compartida. Una aproximación distante a tres ciudades iberoamericanas en Instagram

Gabriela Elisa Sued
gabriela.sued@gmail.com
Tecnológico de Monterrey Ciudad de México, Mexico

Introducción

Debido a la generalización del uso de cámaras fotográficas en teléfonos móviles y a la posibilidad de su inmediata publicación en redes sociales, la fotografía compartida ha devenido una parte fundamental de la comunicación on-line. Sin embargo, ha sido menos estudiada que los objetos textuales publicados en algunas plataformas sociales, por ejemplo en Twitter. (Highfield y Leaver, 2016). Recientemente la investigación académica comienza a analizar el contenido visual generado por los usuarios. Estas temáticas y nuevos modos de producción de información son crecientemente abordados en análisis científicos y críticos con metodologías innovadoras, como la analítica cultural (Manovich, 2009) los métodos digitales (Rogers, 2009) y la visualización de información (Niederer y Taudin Chabot, 2015).

Nos proponemos investigar empíricamente en el modo en que las ciudades iberoamericanas son representadas en Instagram. A tal fin hemos recolectado un conjunto de fotografías etiquetadas como #buenosaires, #cdmx (México) y #madrid, publicadas en la mencionada plataforma durante la primera semana de octubre de 2016. El estudio profundiza en las formas en que las tres ciudades son representadas desde el punto de vista de los usuarios de la plataforma, describe las especificidades de cada una, e identifica el uso social de las etiquetas o hashtags de gran porte, donde se publican miles de fotos diariamente.

Empleamos una aproximación metodológica distante (Moretti, 2007, 2015) que considera tanto la dimensión digital de esas interacciones como una interpretación crítica que pueda identificar el papel que los objetos digitales juegan en la producción de la cultura contemporánea. Emplea una metodología de investigación mixta que combina el análisis cuantitativo, el empleo de software de procesamiento de datos textuales y numéricos y la interpretación de resultados desde una perspectiva sociocultural.

La aproximación distante como método digital

Interrogamos el corpus a través de un conjunto de técnicas que denominamos genéricamente aproximación distante. En el campo de los estudios literarios cuantitativos Moretti (2007, 2015) distingue dos tipos de lecturas: la distante y la cercana. La primera es de tipo exploratoria, opera con la masa, la generalidad y los hechos comunes. El crítico identifica en esa masa patrones de regularidad y frecuencia, grandes agrupamientos o clústers, ciclos temporales, y estructuras reticulares (Manovich, 2009). Por otro lado, la aproximación cercana usa técnicas procedentes de diferentes corrientes interpretativas a fines de atribuir un sentido a la producción analizada, y establecer relaciones con la cultura en la que estas manifestaciones tienen lugar. Las teorías interpretativas otorgan a las producciones simbólicas y de registro un lugar fundamental para la comprensión de las culturas.

La aproximación distante propone interrogar los datos y metadatos desde diferentes técnicas basadas en software. Aplicamos el análisis de contenido (Rose, 2016) y la analítica visual (Thomas y Cook, 2005 Manovich, 2011b) al corpus fotográfico con el fin de identificar temáticas recurrentes y patrones estéticos. Empleamos la analítica textual (Moreno y Redondo, 2016) para identificar las palabras frecuentes en las descripciones o Captions. El análisis de redes (Venturini, Jacomy y Carvalho, 2015) nos fue útil para establecer conexiones y clústers o agrupamientos entre etiquetas co-ocurrentes. Finalmente estudiamos las reacciones en relación al consumo activo de las fotografías una vez publicadas en la plataforma, también denominado *engagement* (Turner, 2014 y Rogers, 2016).

Hallazgos empíricos y discusión metodológica

En el corpus estudiado se evidencia la recurrencia de elementos temáticos, estéticos y textuales. Las palabras frecuentes evidencian una práctica homogénea, cuyo significado se fija en pocas redes semánticas asociadas a la fotografía, el viaje, la arquitectura, el consumo. Los patrones textuales demuestran diferentes maneras de representar y experimentar las ciudades. En *#madrid* las fotografías de personas en el ámbito urbano cobran mayor importancia que en otras etiquetas, en consecuencia merecen ser estudiadas en profundidad. El uso publicitario y la promoción del consumo también son recurrentes. En #buenosaires lo son la experiencia de práctica fotográfica y los estilos de vida, así como en #cdmx el patrón dominante es el de la estilización del entorno urbano. Las recurrencias pueden interpretarse en varias direcciones: en relación a las características propias de los objetos representados, debido a la homogeneidad de imaginarios sociales, o también como emergencia de nuevos géneros narrativos asociados a la fotografía compartida. La co-ocurrencia de etiquetas esboza redes de intereses, temáticas y comunidades acordes a la definición de la fotografía compartida como elemento de exhibición y mensaje comunicativo de intercambio efímero El análisis de de reacciones evidencia las diferencias culturales y comunicativas entre el acto de fotografiar y el de mirar una fotografía. En #cdmx se destaca la alta cantidad de fotografías de amaneceres y atardeceres, pero es la temática urbana y arquitectónica la que recibe mayores reacciones.

Las tres ciudades se distinguen por el uso publicitario de la imagen generado por los propios usuarios. Los que reciben mayores reacciones por otro lado también hacen un uso económico del hashtag pues su fin es el de la autopromoción. Además desarrollan estrategias para lograr la visibilidad de sus fotos publicando en múltiples hashtags y deslocalizando los territorios representados a partir del etiquetado. Al menos una parte de ellos concibe su práctica como parte de una comunidad, evidenciada por la aparición recurrentes de hashtags asociados a la publicación en Instagram: "instagrammers", "mextagram" y otras similares. Podemos suponer entonces que Instagram instaura una suerte de "economía de la visibilidad", donde las reacciones son la moneda con la que se paga la creatividad vernácula.

La alta homogeneidad y recurrencia sugieren la emergencia de codificaciones semióticas propias de la plataforma y demuestra la existencia de una gramática de acción (Agre, 1994), donde las acciones aisladas adquieren sentido cuando se las analiza colectivamente. Estos elementos pueden indicar el surgimiento de la fotografía compartida no sólo como práctica cultural sino como género discursivo con sus propias temáticas, estéticas y prácticas. Las producciones recurrentes de los usuarios pueden considerarse codificaciones que se siguen para ser parte de diversas comunidades de práctica materializadas en el uso de hashtags.

A partir del estudio realizado podemos observar que la aproximación distante resulta efectiva para abrir la caja negra de los medios sociales y mapear los principales temas y patrones estéticos de la fotografía compartida, aunque este abordaje debe en un futuro someterse a mayor investigación empírica y profundización epistemológica sobre varios de sus componentes. Entre los puntos que requieren mayor investigación se encuentran las técnicas de investigación de datos, la comprensión del modo en que funciona el software que se emplea en el procesamiento de los datos y metadatos, y la determinación de la importancia del volumen de datos que se producen en las redes para la investigación social.

El estudio de las mediaciones en Latinoamérica siempre ha relacionado la práctica cultural con las estructuras sociales, identificando en los consumos culturales o bien prácticas de subordinación, o bien de resistencia (Martín Barbero, 2001). En este trabajo la fotografía compartida sobre ciudades emerge como una práctica donde la búsqueda de visibilidad y el uso de autopromoción resultan evidentes. Será entonces necesario plantear su función social en el contexto de una economía global de intercambios simbólicos, línea que deberá ser profundizada tanto teórica como empíricamente.

References

Highfield, T. y Leaver, T. (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*. 2: 47-62.

Manovich, L. (2009). Cultural Analytics: Visualizing Patterns in the era of more media. Recuperado el 14 de mayo de 2017, a partir de http://www.manovich.net

Manovich, L. (2011b). What is visualization? *Visual Studies*, 26(1): 36–49.

Martín Barbero, Jesús. (2001). *De los medios a las mediaciones: comunicación, cultura y hegemonía*. Naucalpan, Mexico: G. Gili.

Moreno, A., Redondo, T. (2016). Text Analytics: the convergence of Big Data and Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, 3 (Special Issue on Big Data and AI): 57–64.

Moretti, F. (2007). *La literatura vista desde lejos*. Barcelona: Marbot.

Moretti, F. (2015). *Distant reading*. London: Verso.

Niederer, S., y Taudin Chabot, R. (2015). Deconstructing the cloud: Responses to Big Data phenomena from social sciences, humanities and the arts. *Big Data and Society*, 2(2).

Rogers, R. (2009). The End of Virtual. Digital Methods. Recuperado a partir de http://www.govcom.org/rogers_oratie.pdf

Rose, G. (2016). *Visual methodologies: an introduction to researching with visual materials*. London: Sage

Thomas K., y Cook, K., eds. (2005). Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, recuperado de http://vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf

Turner, P. (2014). The figure and ground of engagement. *AI and Society*, 29(1): 33–43.

Venturini, T., Jacomy, M., y Carvalho P. D. (2015). Visual Network Analysis. Recuperado a partir de http://www.tommasoventurini.it/wp/wp-content/uploads/2014/08/Venturini-Jacomy_Visual-Network-Analysis_WorkingPaper.pdf

Revitalizing Wikipedia/DBpedia Open Data by Gamification -SPARQL and API Experiment for Edutainment in Digital Humanities

Go Sugimoto
go.sugimoto@oeaw.ac.at
Austrian Academy of Sciences, Austria

Introduction

The Linked Open Data (LOD) community is growing In Digital Humanities (DH). Important datasets are being published in RDF. SPARQL endpoints have been progressively created in many cultural heritage organizations (Edelstein et al., 2013). However, the use of those datasets in real research is still not prevalent. Although there are several DH projects (Boer, V. de et al., 2016), SPARQL query exploitation is often limited within small technology-savvy communities (Lincoln, 2017). The situation is better for less-complicated Application Programming Interfaces (APIs) (XML and JSON). However, Sugimoto (2017b) suggests the needs of API standardization and ease of data reuse for ordinary users. In a broader context, the underuse of data, tools, and infrastructures seems to be a common phenomenon in DH. For example, the use of the Virtual Language Observatory in CLARIN is rather low (Sugimoto, 2017a). In case of the limited use of SPARQL endpoints, there could be different reasons for this:

- Lack of awareness of existence
- Lack of skills to use SPARQL
- Opened data is too narrow in scope
- Lack of computing performance to be usable
- Interdisciplinary research is not widely exercised

It is a pity that the benefit of Open Data is only partially spread, although data is available. To this end, the author has experimented with Wikipedia/DBpedia to explore the potential use of and/or the revitalization of Open Data in and outside research community.

Revitalization of Wikipedia/DBpedia by gamification

The choice of Wikipedia/DBpedia is rationalized by taking into account the above-mentioned issues. The broad scope of their datasets would solve the problem of datasets in DH being too specific to be used by third party researchers (or the researchers do not know how to use data and/or what to do with them (Edmond and Garnett, 2014; Orgel et al., 2015). In addition, interdisciplinary research could be more easily adopted, using a more comprehensive yet relatively detailed level of knowledge.

The keyword of the approach of this project is **gamification**. In order to showcase a social benefit of Open Data and DH, gamification would be a catalyst to connect the scholars and the increasingly greedy public consumers. Kelly and Bowan (2014) stated that limited attention has been paid to digital games until recently, although this is changing rapidly (see Hacker, 2015). Although there are a few projects such as Cross Cult which uses elaborate semantic technologies (Daif et al., 2017), this article contributes to this discourse from a web innovation perspective in a simplified DIY project environment.

The game developed for the project is quite simple. It is a quiz that requires users to guess the age of a randomly selected person by looking at a portrait of the person (born between 1700 and 2002) (Figure 1). Apparently, the age of a person in a particular image is provided neither

by Wikipedia, nor by DBpedia. It is, in fact, calculated programmatically by comparing the birthdate and the date of image. The random selection of data is sometimes costly for data processing, but it is the key to developing a game application. The application is intended for fun, thus, includes all types of contemporary persons such as politicians, sport athletes, musicians, actors, and businesspersons. In addition, the inclusion of historical figures is very important in DH in that the user would learn history.

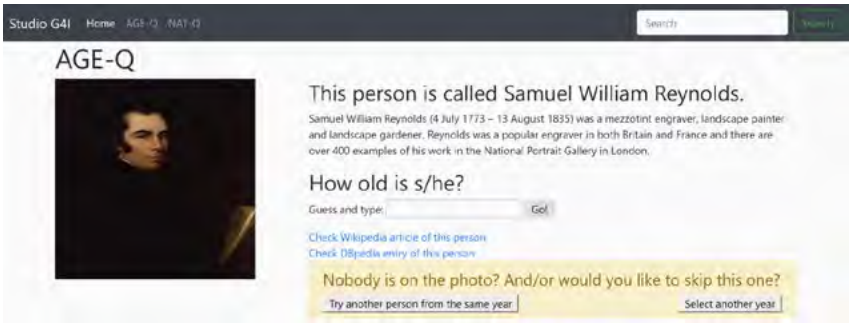


Figure 1 Quiz to guess the age of a person found in a Wikipedia article

When the user cannot guess the age, there is a help function. A hint section is equipped with a face detection API of IBM Watson, suggesting the estimate age and gender of the person in the image by machine learning. Finally, this game is extended into another quiz to guess the nationality of a person. Indeed, any interesting data of Wikipedia/DBpedia can be used for gamification, and the method is easily adoptable.

Potential for Citizen Science

As a reflection of critics of Linked Data quality, Daif et al. (2017) reckon that human supervision is needed to manage the data. In our case, the application is sometimes not able to calculate the age of a person, due to several reasons of metadata quality. For instance, data may be not numeric ("16th century") (Figure 2), malformed (not ISO compliant: "05/11/88"), confusing (the creation date of digital image is used instead of that of analogue image), inaccurate, wrong, or missing, resulting in an error message. This is normally regarded as an optimization problem of the code. However, it is possible to take advantage of this error. When it occurs, it is a sign of data quality problem. Therefore, users are persuaded to follow the provided links to Wikipedia/DBpedia and able to double-check the original data (Figure 3). This scenario creates a dual possibility. In other words, the application can be used as:

- A curation tool of Wikipedia/DBpedia for existing active editors of Wikipedia.
- A tool to transform normal users into new curators of Wikipedia

Although this scenario has not happened due to the project setting, if the users are able to correct data, the impact for data curation could be considerable. Not only is it to the benefit of correcting and/or adding data in Wikipedia, but DBpedia will also be improved, leading to the higher quality of datasets of this LOD magnet, affecting hundreds of applications worldwide. In this way, this application opens up the potential to **crowdsource the curation** of Wikipedia/DBpedia. The success of the crowd data curation has been proven in DH (see Brinkerink, (2010) and NYPL Labs).

Figure 2 Wikimedia metadata displaying "16th century"

Figure 3 The game persuades users to improve Wikipedia

Conclusion

In conclusion, this article demonstrates an experimental case study of mixing gamification (entertainment) with data-driven research (education) and the possibility for data curation (crowdsourcing), showcasing cutting-edge technologies such as SPARQL and Deep Learning API, with the help of Open Data in the framework of DH. It also displays a potential for a new digital research ecosystem among humanities research and digital technologies, connecting various stakeholders including humanities researchers and the public.

References

Boer, V. de, Penuela, A. M. and Ockeloen, C. J. (2016). Linked Data for Digital History: Lessons Learned from Three Case Studies. *Anejos de La Revista de Historiografía*(4): pp139–62.

Brinkerink, M. (2010). Waisda? Video Labeling Game: Evaluation Report. *Images for the Future – Research Blog* http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/index.html (accessed 12 April 2018).

Daif, A., Dahroug, A., López-Nores, M., Gil-Solla, A., Ramos-Cabrer, M., Pazos-Arias, J. J. and Blanco-Fernández, Y. (2017). Developing Quiz Games Linked to Networks of Semantic Connections Among Cultural Venues. *Metadata and Semantic Research*. (Communications in Computer and Information Science). Springer, Cham, pp. 239–46 doi:10.1007/978-3-319-70863-8_23.

Edelstein, J., Galla, L., Li-Madeo, C., Marden, J., Rhonemus, A. and Whysel, N. (2013). Linked Open Data for Cultural Heritage: Evolution of an Information Technology. http://www.whysel.com/papers/LIS670-Linked-Open-Data-for-Cultural-Heritage.pdf (accessed 24 April 2018).

Edmond, J. and Garnett, V. (2014). Building an API is not enough! Investigating Reuse of Cultural Heritage Data. *LSE Impact Blog* http://blogs.lse.ac.uk/impactofsocialsciences/2014/09/08/investigating-reuse-of-cultural-heritage-data-europeana/ (accessed 27 February 2018).

Hacker, P. (2015). The Games Art Historians Play: Online Game-based Learning in Art History and Museum Contexts. *The Chronicle of Higher Education Blogs: ProfHacker* https://www.chronicle.com/blogs/profhacker/the-games-art-historians-play-online-game-based-learning-in-art-history-and-museum-contexts/61263 (accessed 12 April 2018).

Kelly, L. and Bowan, A. (2014). Gamifying the museum: Educational games for learning | MWA2014: Museums and the Web Asia 2014 https://mwa2014.museumsandtheweb.com/paper/gamifying-the-museum-educational-games-for-learning/ (accessed 12 April 2018).

Lincoln, M. (2017). Using SPARQL to access Linked Open Data. *Programming Historian* https://programminghistorian.org/lessons/graph-databases-and-SPARQL (accessed 12 April 2018).

NYPL Labs Whats on the menu? http://menus.nypl.org/about (accessed 12 April 2018).

Orgel, T., Höffernig, M., Bailer, W. and Russegger, S. (2015). A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries*, 15(2–4): pp189–207 doi:10.1007/s00799-015-0138-2.

Sugimoto, G. (2017a). Number game -Experience of a European research infrastructure (CLARIN) for the analysis of web traffic. *CLARIN Annual Conference 2016*. Aix-en-Provence, France: CLARIN ERIC and Laboratoire Parole et Langage and Laboratoire des Sciences de l'Information et des Systèmes (LSIS) and Aix-Marseille Université and Centre National de la Recherche Scientifique (CNRS) https://hal.archives-ouvertes.fr/hal-01539048 (accessed 17 November 2017).

Sugimoto, G. (2017b). Battle Without FAIR and Easy Data in Digital Humanities. *Metadata and Seman-*

tic Research. (Communications in Computer and Information Science). Springer, Cham, pp. 315–26 doi:10.1007/978-3-319-70863-8_30.

The Purpose of Education: A Large-Scale Text Analysis of University Mission Statements

Danica Savonick
danicasavonick@gmail.com
The Graduate Center of the City University of New York, United States of America

Lisa Tagliaferri
ltagliaferri@gradcenter.cuny.edu
Fordham University, DigitalOcean

What is the purpose of higher education? In the United States, this question dates back to at least the nineteenth century with the passage of the Morrill Acts of 1862 and 1890, and has taken on new urgency in an era of manufactured austerity and neoliberal crisis. In particular, scholars of critical university studies such as Christopher Newfield, Fred Moten and Stefano Harney, Sara Ahmed, Craig Steven Wilder and Roderick Ferguson critique the ways higher education often reproduces the very conditions of inequality it claims to challenge. Often, these compelling analyses are based on the investigation of conditions at a few representative universities, but through leveraging digital methodologies we can gain a wider perspective that enables a more comprehensive analysis of what universities put forward as their purpose.

Our research advances these conversations through large-scale textual analyses of two data sets: university mission statements included in the U.S. Department of Education's Database of Accredited Postsecondary Institutions and Programs and recent demand statements put forth by activist students. Mission statements offer a public-facing proclamation that bridge universities to larger communities and educational contexts. Often, they present idealized claims that reflect the university's marketed brand. We use "university" broadly; our data set includes community colleges, public universities, private universities, research institutions, teaching-focused institutions, for-profit and nonprofit schools, and our analysis highlights the variation in their commitments to education. The second data set is a collection of student demands compiled by WeTheProtestors and the Black Liberation Collective, two social justice groups that are working to address institutional inequality across U.S. universities. In many cases, these demands are written to address the institutions of the official university mission statements we are working with, and range from private institutions like Yale University and Ithaca College, to public universities such as Iowa State and UCLA. These demands challenge existing institutional language and require analysis in their own right.

With this research, we seek to answer two questions: 1) What do contemporary U.S. universities claim as their mission and vision? 2) How do these stated aims of education intersect or diverge with the demands of activist students calling for pedagogical, institutional, and social change? In analyzing this data, we draw from the insights of critical race, gender, and sexuality studies, which have long been sites of institutional critique. Coupling digital tools with a theoretical lens informed by activist pedagogy enables us to better apprehend the power structures and social dynamics at play in public-facing institutional documents and how those interface with the communities they are tasked with serving. By better understanding the professed commitments of academic institutions, we aim to contribute to the project of making education more just, equitable, and inclusive.

This work is carried out through the web scraping of data, topic modeling, and statistical analysis in Python. Once analyzed, the raw data and findings are also rendered as interactive web-based data visualizations in JavaScript to make the research more accessible to the public and available for refactoring. Initial statistical textual analysis and data visualization that we have conducted has revealed interesting trends among public universities in contrast to demands put forth by students. Mission statements from state universities emphasize a commitment to the objectives of research, knowledge, and professionalism, and the endeavors of providing and serving, and learning and teaching. However, student demand statements have a more expansive understanding of education that stresses inclusivity and community, while also voicing concerns about race, gender, workers, and resources. By comparing and contrasting across data sets, we examine what each type of institution and group is seeking to achieve, and work to determine whether universities are serving the needs of student populations. When universities are more concerned with vocational skills training rather than challenging power hierarchies, structural inequalities, and the distribution of resources along embodied axes of race and gender, there is a clear disconnect between what institutions are offering students and what students in turn demand. If universities aren't serving their students and communities, who are they serving?

Our data set and programming files will be made publicly available in a code repository so that others who investigate higher education can perform their own research. While our focus is grounded in the specific histories of higher education in the United States, we hope that sharing this research at an international conference will encourage others to perform similar analysis of institutional and popular discourse in their countries, thus allowing for a more vibrant understanding of how higher education functions in different contexts. By inviting

others to add data to our public repository from international institutions, we can begin to consider how globalization impacts learning institutions.

In an effort to advance intercultural scholarly exchange, a Spanish translation of this research will be available online.

Digital Humanities Integration and Management Challenges in Advanced Imaging Across Institutions and Technologies
Nondestructive Imaging of Egyptian Mummy Papyrus Cartonnage

Michael B. Toth
mbt.rbtoth@gmail.com
University College London, United Kingdom; R.B. Toth Associates, United States of America

Melissa Terras
m.terras@ed.ac.uk
University College London, United Kingdom; University of Edinborough, United Kingdom

Adam Gibson
adam.gibson@ucl.ac.uk
University College London, United Kingdom

Cerys Jones
cerys.jones.15@ucl.ac.uk
University College London, United Kingdom

This rapid development and testing project brought together international partners, scholars and collections in an exploratory, pilot effort from November 2015 to March 2017. The international, multidisciplinary team demonstrated that some nondestructive digital imaging techniques and technologies (Fig. 1) have potential to make texts visible in Egyptian Ptolemaic papyrus mummy mask cartonnages. A major challenge in working across the different technologies, disciplines and institutions was integrating data from diverse technical imaging systems and work processes, requiring new and proven digital humanities data management capabilities.

Before this project, other scholars destroyed the masks to access the papyri, denying future researcher access to the primary historical artefacts (Mazza, 2014). This project capitalized on digital humanities skills and data management techniques in assessing the integration of non-destructive digital imaging technologies to make texts visible in layers of papyrus in mummy cartonnages for open research and analysis. Intermediate goals, such as detecting the presence of text, also proved valuable in highlighting the destructive techniques used to study mummy masks and offering scientifically valid approaches for documenting the initial state of objects and their production for future research.

Figure 1. Multispectral Imaging of Mummy Mask at UCL Centre for Digital Humanities, one of the advanced imaging techniques researched during this project.

A global team pulled together expertise from science and the humanities, including: digital humanities, Egyptology and papyrology, medicine, dentistry, particle physics, imaging science, data and project management, and systems engineering. Team members rapidly implemented a phased and agile approach at multiple institutions to develop and apply increasingly complex imaging, processing and data integration techniques to penetrate the paint and papyrus layers in mummy cartonnage and host all data online (Fig. 2).

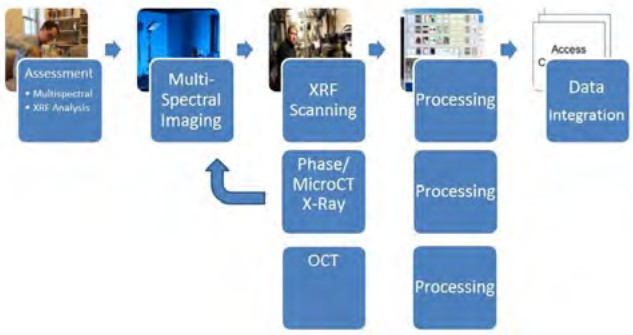


Figure 2. Mummy cartonnage advanced imaging process flow

Data Integration

Project data integration was dependent on common data and metadata standards for ease of image correlation and integration, as well as effective data and project management across disciplines, technologies and institutions. All

the different imaging modalities (Multispectral imaging, X-ray fluorescence, Optical Coherence Tomography, X-ray microCT, Terahertz and others) yielded very different data sets from each technology and institution. Integration of images from multiple imaging sources offered potential to apply the strengths of multiple imaging techniques for ease of visualization by scholars and curators. Integrating data from a variety of equipment required significant planning and collaboration across institutions and disciplines (Fig. 3). This required streamlined standardization processes and/or more time and resources to devote to this part of a program.

Imaging Technology	Imaging Institutions	Contributing Objects	Principal Investigators
Multispectral Imaging	UCL, Manchester, Duke, UC Berkeley, RB Toth Associates	UCL Petrie, UC Berkeley, Duke, UCL*	Melissa Terras, Adam Gibson, Bill Christens-Barry, Michael Toth
Spectral Domain Optical Coherence Tomography	Duke University	Duke, UCL*	Sina Farsiu, Adam Wax, Cynthia Toth
X-ray Fluorescence Scanning	SLAC SSRL	Berkeley†, UCL*	Uwe Bergmann
X-ray Micro-Computer Tomography	*University of California at Berkeley*	Berkeley†, UCL*	Dula Parkinson
X-ray Micro-Computer Tomography	*Queen Mary University of London*	Berkeley†, UCL*	David Mills, Graham Davis
Terahertz Imaging	*University of Western Australia*	UCL*	Vincent Wallace, Shuting Fan, Anthony Fitzgerald
XRF Analyzer	Bruker Scientific	Petrie, Berkeley, Duke, UCL*	Lee Drake, Adam Gibson
Fiber optic Reflectance Spectroscopy	Equipoise Imaging	Berkeley†, UCL*	Bill Christens-Barry

*UCL Phantom surrogate papyrus samples †UC Berkeley Tebtunis Center s.n. cartonnage fragment

Table 1. Participating institutions and imaging techniques used during cartonnage imaging.

The integration of data and work processes from a variety of scientific tools, disciplines and institutions required storage, dissemination, and searchable access to data from instruments that provide output in different formats, some of which were unique to the research methods and disciplines (Emery et. al., 2004). While common standards and processes across institutions were encouraged, this was difficult with data and standards from technologies as diverse as nuclear synchrotrons and optical cameras. In addition, many contributors to this project volunteered their time and equipment for imaging and basic processing, but had limited time to spare from their day-to-day responsibilities – ranging from medical personnel preventing blindness to particle physicists studying elemental changes in bone formation.

Data Storage and Management

The approximately 300 Gb of data products– including images, individual reports, captured and processed data sets, analytical data and metadata– are now freely available online at https://www.ucl.ac.uk/dh/projects/deepimaging/data. This data set comprises a core content set of digital images, analytical data and technical reports on the imaging and analysis of mummy mask cartonnage and modern surrogates from the multiple imaging institutions. UCLDH established this project website to host the project information and data at the same for scholars, scientists and the public (UCL, 2017).

Collecting, organizing and hosting data with appropriate metadata from multiple institutions and systems around the globe proved to be a complex problem. This included providing access to and sharing of timely, complete, and relevant data during the project. This was due to both different data collection standards and the wide range of output from proprietary equipment. A key strength of this program was all institutions agreed to make all data freely available under Creative Commons license. This allowed the free exchange of all data for digital processing, analysis and research.

The data structures of the Archimedes and Galen Palimpsests and the University of Pennsylvania's OPenn served as models, but had to be adapted to include the various types of data sets for each image and data collection modality. To support scientific data integration, the team also used the Library of Congress CLASS-D data

model. Some adjustments were needed to previous flat file access protocols to make the data product more accessible to users and future researchers. As an example, the large captured multispectral data sets were put in separate folders from the processed images, with the former available for follow-on digital processing and research, and the latter available for immediate visualization of our findings produced with current processing tools.

The need for quality assurance to verify and validate the data proved important. Once the data was integrated, some type of feedback mechanism was needed to validate and check the data against other data in collaboration with the collector as part of collaborative research. This highlighted the value of the data in conjunction with other data, with feedback on the efficiency and quality of the data and its reproducibility as initially structured and standardized. This significantly improved data sharing and preservation across the research team.

Conclusions

Effective data management, integration and technical support are critical enablers in any broad digital research program to ensure data availability for follow-on research, even those (like this one) with a limited budget. The ability of imaging equipment to produce a standard data output with relative ease of use by the operator and researcher is important to the visualization, storage of and access to the data. Standardized procedures and data output better allow independent imaging of the same object with multiple technologies, with subsequent integration of data to leverage the strengths of each technology and technique.

References

Emery D., France, F.G., Toth, M.B. (2009) "Management of Spectral Imaging Archives for Scientific Preservation Studies", Archiving 2009, IS&T, May 4-7, 137-141

Mazza R. (2014) "Another Indiana Jones? Josh McDowell, mummy cartonnage and biblical papyri." Faces and Voices. 2014. https://facesandvoices.wordpress.com/2014/05/05/another-indiana-jones-josh-mcdowell-mummy-cartonnage-and-biblical-papyri/. (Accessed 22 Mar 2018).

UCL (2017) "Deep Imaging Mummy Cases, The Data" https://www.ucl.ac.uk/dh/projects/deepimaging/data, (Accessed 27 Feb 2018).

Towards A Digital Dissolution: The Challenges Of Mapping Revolutionary Change In Pre-modern Europe

Charlotte Tupman
c.tupman2@exeter.ac.uk
University of Exeter, United Kingdom

James Clark
j.g.clark@exeter.ac.uk
University of Exeter, United Kingdom

Richard Holding
r.j.holding@exeter.ac.uk
University of Exeter, United Kingdom

This work-in-progress paper offers for critical review the current challenges of an ambitious project to create a digital framework for interpreting the dissolution of monasteries in Europe. The most dramatic episode of the European Reformation (*c*.1517- *c*.1648), the state suppression of monasteries, the dispersal of their populations, the re-distribution of their property and the re-deployment of their infrastructure, represented the largest and furthest-reaching re-ordering of society, economy and culture before the Industrial Revolution (Chadwick, 2001; Youings, 1971). The scale, scope, pace and reach of the process make it perhaps the most formidable of all pre-modern territories for the data-driven researcher, and have ensured that narrative histories founded on conventional methods of data analysis have consistently failed to provide perspectives of adequate breadth, depth, and accuracy.

In respect of research data, the medieval monastery presents both the best and worst of all prospects. A world in microcosm, possessed of its own demographic, economic, social, cultural and environmental imprint, in principle there are multiple layers to its source-base. It also runs deep through time, passing any polity, dynasty, and even place of settlement to reach back to the remote beginnings of Christian-occupied Europe. Yet for these same reasons, the sources of the medieval monastery are also uniquely unstable. The self-containment of the monastery was such that while the form and function of its documentary record might be comparable one to another, it is never quite the same. A durable - but not always enduring - presence in a world that was chronically disturbed, the record underwent repeated and extended interruptions. The monastery invited the manipulation of those in power, and its records are susceptible to conscious distortion. Even well-preserved monastic records can confound the researcher.

The closure and re-constitution of these centuries-old institutions brings these data complexities into collision with the records of the state, city, commune and of private individuals at a moment when these constituencies were in transition to a post-medieval world. The bare historical record may give the impression of the dissolution as an event bounded by the dates of specific acts of state, but in fact its course and consequences were a collective experience which unfolded over several generations. This means that for effective interpretation, datasets should be defined not by the intrinsic criteria of a particular monastery but rather by those that can be related to the contexts in which it was situated, relating to a

range of organisational and social networks and to physical place and space. This requires drawing from the monastery's records data that they were not originally created to document. For example, a contextualised approach to data on the monastery's population profile demands not only a raw numeral but also a measure of its geographical origin, social status and generational mix, each in relation to other neighbourhood constituencies. Because the dissolution was experienced over *la longue durée*, a wide chronological frame is needed: only by capturing data from 1450 to 1650 can the process of dissolution be traced in real time. Given the inherent characteristics of the records, this can be no conventional time-frame bringing a strict linear order to each dataset. With unequal interruptions in every category of record, instead the timeline must be drawn between irregular census points derived from individual documents.

Presently, we are applying these principles to a single case-study, the English Benedictine abbey of Battle, in the county of Sussex, dissolved in 1538. A substantial foundation, holding territory across seven counties of England and Wales, overseeing diverse agricultural, commercial and industrial interests and governing a network of satellite churches and communities, Battle presents sufficient scale and complexity to guide, and test, our emerging methodology (Evans, 1941-2; Searle, 1974). We are creating datasets which aim to measure (1) every aspect of the monastery's presence in and imprint upon its neighbourhood in the period before its dissolution and (2) the pattern and pace of change in that presence and imprint as the monastery was suppressed. We have defined data categories to evaluate its dynamic role in its neighbourhood, providing a series of key performance indicators at those census points which can be established. Although these do not always directly reflect the categories of the monastic records, generally it has been possible for data to be anchored by a specific documentary reference. However, it is sometimes necessary to make use of proxies. For example, because the family origins of monks are rarely documented, surnames are taken as an index of origin and social position, and because the precise site and proportions of monastic buildings are not consistently documented we have adopted a 'best-guess' principle, utilising historic mapping, field-, excavation and environment surveys, realising the benefits of the Archaeology Data Service (www.archaeologydataservice.ac.uk) and Heritage Gateway (www.heritagegateway.org.uk).

Our paper explores how we are addressing these complexities and challenges in our sources by combining a webapp built on an open source XML database (xQuery-based eXist-db, http://exist-db.org/) with highly customisable mapping using jQuery and GoogleMaps API to create a digital framework for analysing the process of dissolution across Europe. The framework allows researchers to interpret its events and sources at levels from regional to site-specific, utilising a comparative approach to reveal and visualise patterns that have been largely obscured within this often chaotic set of sources. We are building the webapp to be redeployed by others: its source code and documentation will be released freely on GitHub, enabling others to reuse it for their own research aims. The complexity of the process of dissolution might suggest that certainty in interpretation is an impossible goal, but our work to date suggests that far from this being a deterrent to digital approaches, it instead raises important questions about how we describe our datasets and how we can represent with honesty and clarity the uncertainty, inconsistency and gaps in our sources. These are questions with which every digital humanities scholar must grapple, and having set out the solutions we have identified so far, we are keen to invite discussion on how we might resolve or improve our approach for the benefit of current and future projects encountering similar issues.

References

Bradshaw, B., (1974). *The dissolution of the religious orders in Ireland under Henry VIII*. Cambridge University Press: Cambridge

Chadwick, O. (2001). *The early reformation on the continent*. Oxford University Press: Oxford, pp. 151-180.

Evans, A. (1941). Battle Abbey at the Dissolution. *Huntington Library Quarterly*, 4:4, pp. 393-442; 6:1 (1942), pp. 53-101

Knowles, D. & Hadcock, R.N., (1971). *Medieval Religious Houses: England and Wales*. 2nd edition, Routledge & Kegan Paul: London

Searle, E. (1974). *Lordship and community: Battle Abbey and its banlieu, 1066-1538*. Pontifical Institute of Medieval Studies: Toronto

Youings, J. (1971). *The dissolution of the monasteries.* Routledge & Kegan Paul

An Archaeology of Americana: Recovering the Hemispheric Origins of Sabin's Bibliotheca Americana to Contest the Database's (National) Limits

Mary Lindsay Van Tine
mva@upenn.edu
University of Pennsylvania, United States of America

This long paper will offer an archeology of the Gale database *Sabin Americana, 1500-1926*, tracing its origins through an earlier microfilming project to Joseph Sabin's *Bibliotheca Americana*, a monumental 29-volume "Dictionary of works related to America" begun in 1868 and completed in 1937. While Bonnie Mak, Ian Gadd, and others have explo-

red the bibliographic roots of much-used digital resources like the ESTC and EBBO, the category of Americana has a distinct bibliographic tradition whose digital implications have not been examined. While many contemporary databases derive from earlier bibliographic projects organized by language or nation, "Americana" was for Sabin and his contemporaries a transnational and multilingual category that understood "America" as the entire Western Hemisphere. Sabin and other nineteenth-century bibliographers of "Americana" ultimately produced works with an implied teleological view of a New World history that began with "discovery" and culminated in the emergence of the United States; nevertheless, they conceived of the early history of the hemisphere as a shared one, and their work emerged from an extended scholarly network that encompassed not only the Anglophone but also the Hispanophone world.

While Gale's database borrows Sabin's name and title, it is otherwise strikingly vague on the exact nature of its relationship to the original print bibliography. A close examination reveals that, although the structuring logic of the database is not dissimilar to Sabin's alphabetic schema and indexing, its selection principles and framing radically redefine America as the United States. Unlike the original bibliography, the vast majority of the works included are in English, with few in Spanish and even fewer in indigenous languages. The search interface offers "subject" options that uncritically sort the entire span of New World history into U.S.-based periodizations: colonial era, early republic, antebellum, postbellum, and so on. These silent omissions both assume and reinforce the conflation of "America" and "United States." When a database that claims to be "drawn from Joseph Sabin's famed bibliography" and, like it, to "cove[r] four centuries of life in North, Central, and South America, and the West Indies," returns overwhelmingly English-language sources from the "colonial era," or fails to produce a single hit for one of the most prominent Mexican historians of the nineteenth century while returning dozens for his U.S. counterpart, the effect is not just inaccurate but deeply pernicious. I will argue that this dramatic shift is not so much a function of digital remediation as of a changed scholarly infrastructure that cannot accommodate the capaciousness of "Americana" in its earlier bibliographic sense. The logic of nineteenth-century Bibliotheca Americanas, I suggest, invites us to think otherwise, offering an alternate bibliographic framework that might inform the development of non-proprietary digital systems for bibliographic control.

I will conclude by considering my own work towards this end in the context of the Digital Bibliotheca Americana project. It assembles a freely-available dataset that re-centers indigenous and Spanish-language texts, offers insight into the contours of Americana at scale, and enables computational analysis of the material and conceptual relocation of "Americana" to the United States over the course of the nineteenth century.

Tweets of a Native Son: James Baldwin, #BlackLivesMatter, and Networks of Textual Recirculation

Melanie Walsh
melanie.walsh@wustl.edu
Washington University in St. Louis, United States of America

In the wake of Michael Brown's murder in Ferguson, Missouri, on August 9, 2014, and the non-indictment of police officer Darren Wilson on November 25, 2014, backlashing protests and riots took to the streets of Ferguson and to other major American cities across the country. They also took to the Twittersphere. A national conversation about police brutality and the American criminal justice system exploded on Twitter during this time period, eventually elevating the hashtag #Ferguson, tweeted over 27 million times, to the most frequent in Twitter's ten-year history, and the hashtag #BlackLivesMatter, tweeted over 12 million times, to third place (Sichynsky, 2016). First coined by Alicia Garza, Patrisse Cullors, and Opal Tometi in July 2013, the hashtag #BlackLivesMatter became a banner for a national protest movement and an index for conversations about the systematic devaluing and elimination of black life. Over the last five years, literary scholars and historians have noted that, within this massive social media movement, the novelist, essayist, and civil rights literary icon James Baldwin seemed to be often and increasingly invoked (Maxwell, 2016). The perceived frequency of Baldwin-related tweets has been pointed to by many as evidence of the Harlem-born author's 21st-century resurrection and recent political resonance (Glaude Jr., 2016; Robinson, 2017). Because tweets can be digitally archived and made computationally tractable, they can be collected, measured, and analyzed at scale, and they can offer a picture of Baldwin's social media reception that goes beyond perception and anecdotal evidence. This talk will share work-in-progress from my project *Tweets of a Native Son* (http://www.tweetsofanativeson.com/), which brings large-scale social media data and computational methods to bear on Baldwin's 21st-century remediation, recirculation, and reimagination. This talk will discuss the methods and progress made in the project thus far, argue that social media analysis might usefully contribute to a growing body of computationally-assisted scholarship focused on readership, reception, and textual circulation, and finally gesture to how such an approach might change our understanding of how texts are shared between communities of people, namely through its emphasis on networks.

Methods, Analysis, Initial Findings

First I "hydrated," that is, retrieved the full JSON information for, an archive of over 32 million tweets that were sent between June 1, 2014 and May 31, 2015 and that

mentioned Ferguson, Black Lives Matter, and 20 other black individuals who were killed by the police during this time period, which was first purchased from Twitter and shared by Deen Freelon, Charlton McIlwain, and Meredith D. Clark (Freelon, McIlwain, Clark, 2016). I next searched for all the tweets that mentioned "James Baldwin" by his first and last names using the Python and command-line tool "twarc" and the command-line JSON processor "jq," which returned 7,326 tweets and retweets. By using twarc utilities, a k-means clustering algorithm, and manual tagging, I then identified the most retweeted tweets in the archive and the text that appeared most often across all tweets in the archive, which revealed that the most frequent appeal to Baldwin during this time period was through quotation and overwhelmingly through the quotation of Baldwin's 1960s-era essays, radio interviews, and television appearances.

By studying the text of the most retweeted and most frequently cited tweets, and by tracing tweeted Baldwin quotations back to their literary and historical origins, my project argues that Baldwin's appeal as a #BlackLivesMatter muse comes, at least in part, from the remediation of much of his non-fictional work into YouTube videos and free online essays; from his aphorisms with deep roots in African American written and oral traditions; and from his sympathetic proximity to but never full embrace of black radicalism. Another goal of *Tweets of a Native Son*, however, is to let others explore, hypothesize, and learn about Baldwin's #BlackLivesMatter-related social media reception through a series of interactive data visualizations on the project's website. These interactive visualizations are meant to provide a perspective on Baldwin's living legacy, a refracted vision of Baldwin's life and career through those who actively called upon him in a moment of political and emotional urgency, a means by which others can come to their own conclusions about Baldwin's resurrection.

DH Reception Studies and Networked Reading

Tweets of a Native Son most broadly hopes to join and affirm recent digital humanities work that is trained on readership, reception, and textual circulation, such as Lincoln Mullen's *American's Public Bible* and Ryan Cordell and David Smith's *Viral* Texts, and to amplify Katherine Bode's call that the digital humanities better attend to and account for the ways in which literary texts "circulated and generated meaning together at particular times and places" (Mullen, 2016; Cordell and Smith, 2017; Bode, 2017). Like the 19th-century newspaper archives used by Mullen, Cordell, and Smith, social media archives offer a window into how texts travel, how texts are used and changed by individuals, and what these texts mean in context. Social media archives additionally offer massive amounts of (relatively) clean, recent data. Though of course with these advantages, they also present more ethical challenges, since this data is often tied to corporations and produced by still-living human beings whose consent, possible harm, and creative attribution must always be considered.

Finally, however, I believe that social media data might help us better theorize and make visible the networked structures of readership, reception, and textual circulation, because social media data, such as Twitter data, is often inherently networked in structure, recording retweets, replies, follower communities, hashtag communities, and more. This networked structure emphasizes the way that texts are not only engaged with by individuals but are shared between individuals, taking on social and communal meanings. For the particular case of Baldwin and #BlackLivesMatter in 2014-2015, the quotations of Baldwin's words were often recirculated as coalition- and community-building material, helping to forge connections between individuals across space, time, and American history. During the future stages of this project, I hope to employ network science and network visualization to better understand Baldwin's significance within #BlackLivesMatter.

References

Bode, K. (2017). The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History. *Modern Language Quarterly*, 78(1): 94.

Cordell, R. and Smith, D. (2017).Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines, http://viraltexts.org.

Freelon, D., McIlwain, C. D., and Clark, M. D. (2016). Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice. Center for Media and Social Impact.

Glaude Jr., E. S. (2016). James Baldwin and the Trap of Our History. *Time*.

Maxwell, W. J. (2016). Born-Again, Seen-Again James Baldwin: Post-Postracial Criticism and the Literary History of Black Lives Matter. *American Literary History*, 28(4).

Mullen, L. (2016). America's Public Bible: Biblical Quotations in U.S. Newspapers, http://americaspublicbible.org.

Robinson, Z. (2017). Ventriloquizing Black Feeling, Re-Voicing Black Life: Speaking Baldwin on the Internet. *Communities in Conversation: Digital Baldwin*, Rhodes College.

Sichynsky, T. (2016). These 10 Twitter Hashtags Changed the Way We Talk about Social Issues. *The Washington Post*.

Abundance and Access: Early Modern Political Letters in Contemporary and Digital Archives

Elizabeth Williamson
e.r.williamson@exeter.ac.uk
University of Exeter, United Kingdom

Letters stand as one of the most extensive sources of information on daily life in the early modern period and the study of epistolary culture(s) is a vital and growing area in Renaissance studies (see Daybell and Gordon, 2016; Daybell, 2012; Del Lungo Camiciotti and Pallotti, 2014). Access to such archives and collections is rapidly expanding – and changing – in the wake of mass digitization, online editions, OCR and federated search. In this paper I explore the extension of the narrative of archival history and epistolary provenance into the digital realm. Specifically, I compare the contextual afterlife of early modern letters in nascent state archives to their representation in the digital world, with particular emphasis on classification and metadata, surrogacy and access. Going beyond paralleled modern and early modern anxieties of information overload (the standard comparison of the print and digital revolution), this allows me to explore issues of access, search, and retrieval; control, preservation, and loss, then and now. This is an under-studied area ripe for discussion, and this paper aims to test these ideas in preparation for a wider study that connects the gathering, transmission, and preservation of political information in the early modern period to the digital life of these material primary sources and to our lives as digital researchers. There is a ready parallel to be found between the burgeoning administrative and institutional drive to preservation found in the early modern period – essentially the evolution of state archiving – and the informational anxieties of the internet age, where that largest of archives can offer everything and nothing, excess and restriction, results or dead ends. I explore tensions around archives facilitating both preservation and forgetting, which finds its apotheosis in the endless loss and abandonment of digital data, and in digital methods of retrieval as strict gatekeepers (a roulette of keyword search, privations of metadata, and dreams of text analysis).

I will use the concept of copia, fundamental to early modern humanism and classical pedagogy, to explore these twin pressures of abundance and lack, of meaningful quantity and meaningless repetition. Copia in the early modern period referred to the abundance of language, where mastery over the myriad ways of expressing a single idea gave students the rhetorical strategies to navigate the vast expanse of language. The incessant imitation of classical models, particularly concerning letter-writing, was encouraged not least by Erasmus in the wildly popular *De Copia*, and became a ubiquitous part of humanist education. This concept, of expertise and thus authority being created by sheer mass, by repetition, is particularly apposite when considering rhetoric and knowledge creation today. In fact, this abundance was framed as both knowledge and folly, particularly from the late sixteenth century, when the drive to systematizing information and rise of scientific method pushed against classical humanism (Francis Bacon offers a good example of a writer in this transitional time who both criticized copia and performed it in his criticism). Christine Hoffman, in her recent *Stupid Humanism: Folly as Competence in Early Modern and Twenty-First-Century Culture*, has also identified a productive parallel here, and links early modern rhetorical strategy and modern excess in online 'news' and social media (Hoffman, 2017). I will focus rather on early modern and digital archive creation in order to explore how access to these constituent building blocks of knowledge shapes our historiography and thus the world we construct. I will use copia's two semantic faces, abundance and copying, to firstly think through wider preoccupations with sheer tides of (often repeated) information acting as knowledge and secondly to consider our creation and reception of digital surrogates of primary sources. The early modern relationship to copia as copying is complex: on one hand it is intimately related to the massive growth of bureaucracy and paperwork, where the copy is the transmitter of authority, on the other hand the advent of print and the selfsame abundance of paperwork led to associations of degradation and inferior quality. If print could be alternately the thing itself or the degraded copy of the original, the digital surrogate today is held intermittently as the preserved, unquestioned work and as the flat representation that has lost both the materiality and authenticity of the original.

I will combine discursive reflection with concrete examples that draw parallels between early modern and modern concerns, to consider how the preoccupations and experiences of a particularly early modern growth of the archive and associate concern with amassing, preserving and accessing information inflects our understanding of the internet age, and vice versa. I draw parallels between search and preservation concerns today and amongst early keepers of the state paper office, demonstrating that long-held anxieties around access and information overload continue through into the digital archive. I place the digital archive in a history of indexing and cataloguing, which capitalizes on recent interest in early modern construction of metadata witnessed by Oxford University's 2017 conference 'The Book Index', for example. In both this history of information management and in our relationship to archiving today, what is kept, who has access, and how meaningful and stable that access is, are all essential questions. No less do we need to engage critically with the term access in an increasingly business and market-driven university sector. I point to the open-access philosophy as increasing the availability of digital primary sources, particularly around libraries and archives releasing high-quality digital images and the IIIF initiative making sharing images increasingly possible on a practical level, and to the untold opportunity for connecting resources in the abundant meta-archive promised by the LOD philosophy. From endorsed letters held in the labelled drawers of the Elizabethan Secretary of State's office, to authority files and shared standards, metadata has often been our key to texts. It both enables and restricts

our access, it channels our vision and helps construct our understanding of our sources. Understanding the stories of what is kept and the nuance of how it is described is vital to reveal the limits of this vision. We cannot forget that archives are variously permissive, proscriptive, and problematic, constructed through a history of loss, antiquarianism, colonialism, class and power structures: those with power leave records, those with power control them. With the risk of reproducing existing and long-standing power structures in our digital representations of the early modern world, we need to ask what we want from the modern digital archive, and who will be invited in?

How people create, access, and preserve (particularly political) information can tell us much about the power structures, value systems and personal concerns of both the writers and keepers of texts and the society at large. I argue that understanding the conditions of production and preservation of early modern letters is necessary to fully comprehend their use and meaning: the natural extension of this is that in order to fully engage with these texts we must attend to the methods and conditions of our own access in an increasingly digital scholarly environment. The digital has a natural and increasingly significant place in this textual history: we need to recognize and interrogate the continuous history or provenance that connects the first moment of textual creation to the most recent instance of representation and remediation in a digital space. Reflecting on this reveals that we can read in what is preserved and how it is described the power structures inherent in the society that creates the archive: this is absolutely true for the digital.

References

Daybell, J. (2012). *The Material Letter in Early Modern England: Manuscript Letters and the Culture and Practices of Letter-Writing, 1512-1635*. Basingstoke; New York: Palgrave Macmillan.

Daybell, J. and Gordon, A. (eds). (2016). *Cultures of Correspondence in Early Modern Britain*. Philadelphia: University of Pennsylvania Press.

Daybell, J. and Gordon, A. (eds). (2016). *Women and Epistolary Agency in Early Modern Culture, 1450-1690*. (Women and Gender in the Early Modern World). Burlington, VT: Ashgate.

Del Lungo Camiciotti, G. and Pallotti, D. (2014). Letter Writing in Early Modern Culture, 1500-1750. *Journal of Early Modern Studies*, 3 doi:10.13128/JEMS-2279-7149-3. http://www.fupress.net/index.php/bsfm-jems/issue/view/1023 (accessed 27 October 2017).

Erasmus, D. and Thompson, C. R. (1978). *Collected Works of Erasmus*. Toronto; London: University of Toronto Press.

Hoffmann, C. (2017). *Stupid Humanism: Folly as Competence in Early Modern and Twenty-First-Century Culture*. S.l.: Springer International Pu.

Balanceándonos entre la aserción de la identidad y el mantenimiento del anonimato: Usos sociales de la criptografía en la red

Gunnar Eyal Wolf Iszaevich
gwolf@gwolf.org
Instituto de Investigaciones Económicas, UNAM, Mexico

La criptografía por fin está detrás de prácticamente cualquier acción que emprendamos en Internet — Hemos llegado por fin a una adopción masiva del cifrado para la mayor parte de las comunicaciones en línea. Esto puede leerse desde ángulos muy distintos — Por un lado, nuestras comunicaciones están seguras del espionaje o modificación por parte de terceros. Por otro lado, está *firmada* –por nosotros y por nuestra contraparte– de forma que permite un *no repudio*.

La criptografía puede abordarse y estudiarse desde muy distintos ángulos. El ángulo matemático propone, desarrolla, valida (o refuta) los esquemas que se van presentando; abordar el tema desde la ingeniería en seguridad informática presenta diferentes aplicaciones, analiza modelos de amenaza, estandariza algoritmos en protocolos y formatos, etcétera. Puede hacerse también un análisis legal — Es bien sabido que hasta octubre del 2000, los Estados Unidos consideraban a la *criptografía fuerte* como municiones, y prohibían su *exportación*; afortunadamente sus legisladores reconocieron que los tiempos que corrían eran distintos y rectificaron dicha ley (Export Administration Regulations 15 CFR Part 730 et seq.), pero el tema respecto a quién y cómo debe tener derecho a mantener su privacidad incluso ante entidades de gobierno se mantiene vigente, con ejemplos como el del teléfono del asesino del ataque de San Bernardino (2016).

En esta intervención, el objetivo es hacer un análisis de la criptografía desde un punto de vista social: ¿Qué se entiende por *hacker*? ¿Por qué los *hackers* y la criptografía van tan de la mano? ¿Puede verse el trabajo de éstos ya sea en la comprensión social de la criptografía o en su avance técnico? E incluso yendo un paso más allá, ¿cómo han ido definiendo la criptografía los grupos *hackers* a los diferentes movimientos sociales en que encuentran cabida, tanto en países desarrollados como en la región latinoamericana?

Abordaremos casos como el *Chaos Computer Club* alemán, con más de 35 años de existencia y un congreso que atrae a más de 10,000 personas anualmente y la *Electronic Frontier Foundation*, fundado en 1990 y enfocado a defender legalmente los derechos de libertad de expresión en ambientes en línea, pero también como el *Rancho Electrónico* en México, *Vía Libre* en Argentina, *Patio Maravillas* en España.

Pero además, todos estos grupos, con sus características únicas y sus distintos grados de formalidad/in-

formalidad/aformalidad, se han ido engranando y retroalimentando. La *cultura hacker* es, ante todo, cultura. Los espacios (presenciales o electrónicos) de reunión de hackers son, necesariamente, espacios de creación cultural. Y mostraremos cómo todos estos grupos se han convertido en referentes de la creación cultural, de la generación de movimientos sociales — Estando, en todo momento, vinculados con sus principios creadores: Con la belleza técnica y elegancia algorítmica que posibilitan la existencia de la criptografía.

A White-Box Model for Detecting Author Nationality by Linguistic Differences in Spanish Novels

Albin Zehe
zehe@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Daniel Schlör
schloer@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Ulrike Henny-Krahmer
ulrike.henny@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Martin Becker
becker@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Andreas Hotho
hotho@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Introduction

Automatic nationality detection of authors writing in the same language (such as Spanish) can be used for many tasks, like author attribution, building large corpora to analyse nationality specific writing styles, or detecting outliers like exiled or bilingual authors. While machine learning provides many methods in this area, the corresponding results are usually not directly interpretable. However, in the Digital Humanities, explainable models are of special interest, as the analysis of selected features can help to confirm assumptions about differing writing styles among countries, or reveal novel insights into country-specific formulations.

In this work, we aim to bridge this gap: Our assumption is that nationality or country of origin of an author is strongly connected to their writing style. Thus, we first present a machine learning approach to automatically classifying literary texts regarding their author's nationality. We then provide an analysis of the most relevant features for this classification and show that they are well interpretable from a literary and linguistic standpoint.

Related Work

The problem of detecting regional linguistic differences is at the core of Digital Humanities, as it touches research questions in both traditional linguistics and modern computer science. In Spanish philology and linguistics, the analysis of different regional varieties has a long tradition (see for example Alvar 1969, Eberenz 1995, Noll 2001). There are well-known differences between the Spanish spoken and written in Spain itself and the variations used in the former colonies, for example in forms of address ("vosotros/ustedes" vs. just "ustedes", voseo) and articles (le/la vs. lo).[1] More recently, these differences have been investigated with quantitative methods, for example by applying Zeta to find distinctive words for novels from Spain and from Latin America, respectively (Schöch et al. 2018).

Model

Baseline SVM-Model for classifying author nationality

We assume that writers from different countries are distinguishable by a) their vocabulary and b) phrases that are more or less popular in different regions (cf. Section "Related Work"). Thus, we choose to use an n-gram model to represent our corpus in a computer readable way: First, we determine all word n-grams of length 1 to 4 in the corpus. Then, we select the 1000 most frequent n-grams of each length. We also tried selecting the 100 or 10000 most frequent n-grams, which led to slightly worse results. We represent a piece of text as tf*idf vectors of these n-grams (see Manning 2008).

We then train a linear SVM (see Steinwart 2008) to predict the nationality of an author given a piece of text. The linear SVM is known for good results in text classification (Joachims 1998) and - essential for interpretability - allows to inspect the importance of specific features.

Enhancing Feature Interpretability

When examining our classification model, we observed an over-representation of geographical entities (e.g., frequent locations like Buenos Aires) as well as names. To instead enforce linguistic properties, we replaced all uppercase tokens by distinct UNKNOWN-tokens (except at the beginning of a sentence). For example "¡Oh, María, María! ¡Cómo deseaba triunfar, conquistar Buenos Aires [...]", becomes "¡Oh, UNK_1, UNK_2! ¡Cómo deseaba triunfar, conquistar UNK_3 UNK_4 [...]". This ensures that n-grams with proper nouns will never be frequent enough to be used as a feature in our classification task.

1 http://lema.rae.es/dpd/?key=voseo, http://lema.rae.es/dpd/?key=loísmo&lema=loísmo

Augmenting Training Examples

The success of machine learning algorithms depends largely on the amount of training data. Thus, to increase the number of training samples, we split each novel into multiple segments of equal length, assigning each segment the same label as the entire novel.The cross validation split was performed before segmentation, ensuring that no novel was present in both training and test set. The classifier is then trained and evaluated on individual segments, resulting in a set of "votes" for the nationality of each novel in the test set. The nationality is then established by majority vote.

Corpus

We use a corpus composed of 100 novels from four Spanish-speaking countries, specifically Spain, Argentina, Cuba and Mexico, written in the 19th and early 20th century (Calvo Tello 2017, Henny-Krahmer 2017). Figure 1 shows the distribution over countries and the distributions over subgenres in the countries. All countries are represented by a roughly equal number of texts. We note that our corpus may have a bias towards a specific subgenre in some countries, which will later be addressed in the analysis of the features.

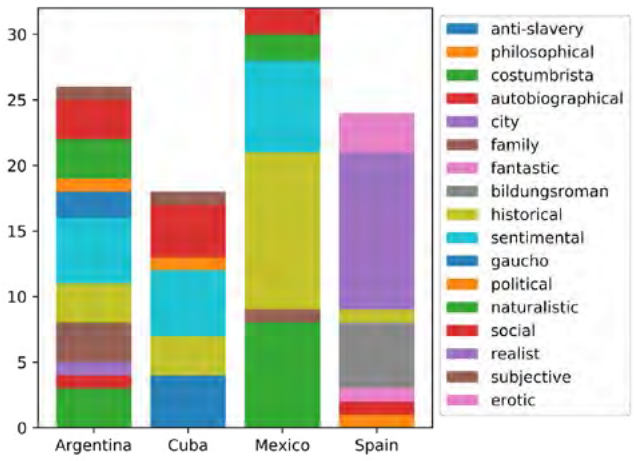


Figure 1: Distribution of countries and subgenres in our corpus

Experiments

We performed extensive experiments on the dataset to determine the accuracy of our approach. The main hyper-parameters of our model are the segment size s, determining how many words a segment contains, and the parameter C of the SVM. We performed parameter optimisation by grid search, choosing from s ∈ {100, 200, 500, 1000, 5000, 10000, 100000, ∞} and C ∈ {10^{-5}, 10^{-4}, …, 10^{5} }. The setting s = ∞ does not perform segmentation. We also varied the maximum length of n-grams: unigrams (n = 1) vs. n-grams of length 1 to 4. All scores reported below are weighted average F1-scores over 10-fold cross validation.

Generally, our model performed best when using only unigrams, removing uppercase tokens and splitting the novels into segments of length s = 1000 (see Table 1 for details).

	precision	recall	f1-score	support
0	0.800	1.000	0.889	24
1	0.923	0.667	0.774	18
2	0.824	0.875	0.848	32
3	1.000	0.885	0.939	26
avg / total	0.882	0.870	0.868	100

Table 1: Classification report for the best configuration, using only unigrams, segments of length s = 1000 and C = 10000

This can be explained by the small dataset: Unigrams are likely to occur in multiple samples even in a small corpus, while higher-order n-grams possibly only occur once and can therefore not be used for classification.

Figure 2 shows the results for varying s and C. Segments of a length around 1000 perform best, yielding F1-scores of up to 86.8 %. Very small segments fail to deliver satisfying results, while larger segments still provide reasonable classification accuracy. The value for C must be set high enough, but the specific value does not matter for C > 10.

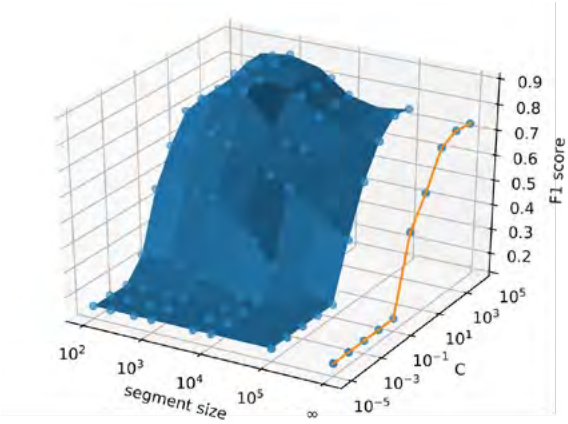


Figure 2: Weighted average F1-score depending on the segment size s and the cost parameter C of the SVM. The separated line denotes no segmentation (s = ∞). Only unigrams were used as features.

Using all n-grams of length 1 to 4 also delivered good accuracy (highest F1-score of 80.4% for C = 10000, s = 1000). Removing uppercase tokens had a positive effect when using unigrams, while it hardly influenced the accuracy using all n-grams.

A detailed view of all results can be found on GitHub.[2]

Feature Analysis

Using a linear SVM enables us to analyse the 10 n-grams that provide the strongest evidence for and against a country (according to internal weights). In the following, we focus on features that are weighted strongly in all or at least multiple folds of the cross validation.

Generally, we identify three feature groups: topical features, features related to the geographical setting and linguistic features. The presence of topical features can be explained by the bias in subgenres that is present in our corpus and is not necessarily representative. The geographical features seem to point to a tendency of the authors to base their stories in their respective home countries rather than other countries.

With regard to the different model variants, the model based on \emph{unigrams without removing uppercase tokens} tends to select names as its top-features such as the country itself or characteristic cities, for example "Madrid" for Spain. While these features are surely helpful for classification (yielding an F1-score of 81.7%), they are not particularly interesting for linguistic analysis. The features selected after removing uppercase tokens, on the other hand, seem more relevant from a linguistic viewpoint, while at the same time providing the best accuracy. Table 2 shows features that are among the highest weighted for more than 5 folds for each country in this setting.

Country	Unigrams	Comments
Spain	+ ello	linguistic (personal pronoun)
	+ seorito	linguistic (diminutive)
	+ duros	currency
	+ seores	linguistic/topical (noun)
	- pesos	currency
Cuba	+ esclavo/esclava	topical
	+ mulato	topical (ethnic group in Cuba)
	+ aadi	linguistic/narrative (verb, probably used to mark direct speech)
	- quiz	linguistic (adverb)
	- huerta	topical/linguistic (noun)
Mexico	+ hacienda	topical (haciendas are typical of Spanish colonies
	+ mexicano	
Argentina	+ entretanto	linguistic (temporal adverb)
	+ gaucho	topical
	+ misia	linguistic (form of address typical in South America)
	+ mate	topical (drink typical to Argentina)

Country	n-grams	Comments
Spain	+ se me figura que	linguistic (locution)
	+ de la huerta	topical (huerta is common in Spain)
Cuba	- de cuando en cuando	linguistic (temporal phrase)
Mexico	+ de/en la casa de	topical (probably due to a subgenre bias)
	+ al cabo de	linguistic (temporal phrase)
	+ de la hacienda	topical (typical of Spanish colonies)
	+ as es que	linguistic (locution)
	- al mismo tiempo	linguistic (temporal phrase)
	- la	linguistic (lesmo)
Argentina	+ en ese momento	linguistic (temporal phrase)
	+ se puso de pie	linguistic (verb)
	+ de vez en cuando	linguistic (temporal phrase)
	+ el hecho es que	linguistic (locution)
	+ al fin al cabo	linguistic (temporal phrase)
	- al cabo de un	linguistic (temporal phrase)

Table 3: N-grams with large weights assigned by the SVM. Features marked with + and - are signals for and against a country, respectively.

Discussion

Technical Aspects

We found that segmenting novels to augment the training data does improve results, but only if the segments are not too short and thus do not contain enough information to detect the author's nationality.

Removing uppercase tokens improves the classification accuracy and makes the selected features more interesting from a linguistic standpoint. We assume that otherwise proper nouns are picked up by the classifier as important clues on the training set, which fail to generalise to the test set.

Feature Interpretation

The words and phrases that our algorithms selects for differentiating between nationalities strongly resemble features that humans would consider given the same task. These include well-known linguistic differences (leísmo) as well as country-specific words (hacienda/huerta). However, it also finds phrases, such as temporal expressions, that are not very well known to be specific for some countries, but should be further investigated in future work. We also observe that authors in our corpus appear to have a strong tendency towards writing about their respective home countries, as evidenced by the selection of city or country names.

Conclusion and Future Work

We have presented a classifier that is able to distinguish between novels from different countries based on word n-grams. Our experiments show that this classifier is able to select features that are interpretable and reveal interesting insights into the language used in novels from different Spanish-speaking countries.

We note that our findings are only based on a limited dataset. However, the tools we have built enable us to replicate the experiments and confirm our findings as soon as larger collections of text become available.

Thus, our work is an important step towards combining machine learning with in-depth analysis and discovery of novel concepts in corpus-based linguistic studies through interpretable models.

In future work, we believe that replacing the majority vote over segments by more sophisticated methods can further improve our results. We also believe that incorporating linguistic information like parse-trees into our features can help to reveal more interesting insights into subtle linguistic differences between countries.

2 https://github.com/cligs/projects2018/tree/master/country-dh

References

Alvar, Manuel (1969). *Variedad y unidad del español: estudios lingüísticos desde la historia*. Editorial Prensa Española.

Calvo Tello, José (ed.) (2017). *Corpus of Spanish Novel from 1880-1940*. Würzburg: CLiGS. https://github.com/cligs/textbox/blob/master/es/novela-espanola.

Eberenz, Rolf (1995). "Norm und regionale Standards des Spanischen in Europa und Amerika". In: Oskar Müller, Dieter Nerius, Jürgen Schmidt-Radefeldt (eds.). *Sprachnormen und Sprachnormenwandel in gegenwärtigen europäischen Sprachen*. Rostock: Universität Rostock, 47-58.

Henny-Krahmer, Ulrike (ed.) (2017). *Collection of 19th Century Spanish-American Novels (1880-1916)*. Würzburg: CLiGS, 2017. https://github.com/cligs/textbox/master/spanish/novela-hispanoamericana/.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML*-98, 137--142.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press. ISBN: 0521865719, 9780521865715

Noll, Volker (2001). *Das amerikanische Spanisch: ein regionaler und historischer Überblick*. Tübingen: Niemeyer.

Schöch, C., Calvo, J., Zehe, A., Hotho, A. (2018). Burrows Zeta: Varianten und Evaluation. *DHd 2018*.

Siskind, Mariano (2010): "The Globalization of the Novel and the Novelization of the Global. A Critique of World Literature." In: *Comparative Literature* 62 (4), 336-360. https://doi.org/10.1215/00104124-2010-021

Steinwart, I., Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated. ISBN: 0387772413

Media Preservation between the Analog and Digital: Recovering and Recreating the Rio VideoWall

Gregory Zinman
gzinman3@gatech.edu
Georgia Institute of Technology, United States of America

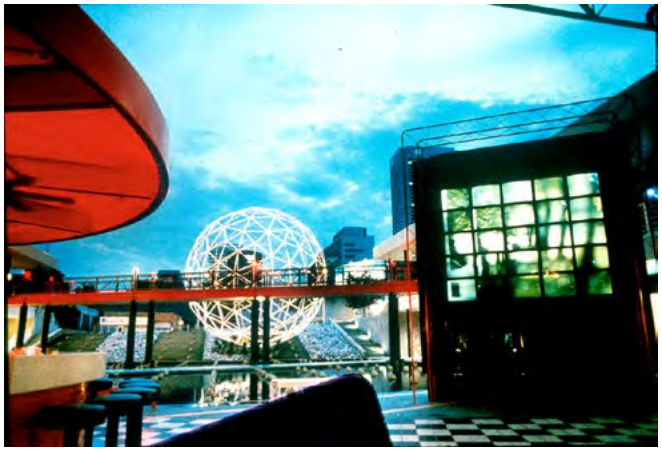


Figure 1: *The Rio VideoWall* (Dara Birnbaum, 1989), installed at the Rio Shopping Complex, Atlanta. Image courtesy of the Smithsonian American Art Museum, NEA Birnbaum collection.

"Media Preservation between the Analog and Digital" is a project that centers on the development of a digital recreation of pioneering video artist Dara Birnbaum's now-lost *Rio VideoWall* (1989), the first multi-screen artwork to be installed in a public setting in the United States. In its original instantiation, in downtown Atlanta, the work employed twenty-five identical 27" CRT monitors, stacked in a five-by-five grid, powered by eight LaserDisc players and proprietary computer code written specifically for the piece. Today, however, only a portion of the code remains; neither the CRT monitors, nor LaserDisc players, nor original computers, are in production, and to recreate the artwork—even in a lab setting—would involve a significant reimagining of the original piece. But there are additional considerations that extend beyond hardware and software: the *VideoWall* was installed in the Rio Shopping Complex, a mall in Atlanta's Old Fourth Ward, a historically African American neighborhood. The artist was attuned to this, and designed the artwork to combine scenes of the natural landscape that had been displaced by the mall with an unedited live-stream of CNN, an Atlanta-based company, all filtered through the moving silhouettes of mall patrons in real time. Neither the nature footage nor the CNN live-stream—let alone the mall patrons—presently exist (the mall was torn down in 2000), so a recreation of the artwork would need to identify footage that captures the spirit, if not the reality, of the piece. To do so

would therefore involve an analysis of the themes engaged by the original artwork: the legacy of segregation, the 24-hour media cycle, surveillance culture, the relationship between art and commerce, and the Anthropocene. This paper will provide a brief overview of the project, with an emphasis on the conceptual challenges it engages, describe the recovery work underway, and describe the current work and next steps toward the *VideoWall*'s ultimate recreation.

Recovering the Rio VideoWall: Conceptual Challenges

Recent seminars and symposia have explored the range of issues associated with doing digital art history (e.g. Harvard metaLab's "Beautiful Data," 2015, the Getty Foundation's "Art History in Digital Dimensions," 2016). At the same time, the field has devoted increasing attention to issues associated with digital preservation (e.g. the BitCurator initiative, or any number of conversations at the Digital Library Federation). And yet, with its hybrid analog/digital design, and its site-specific setting, the *Rio VideoWall* presents a unique case study for thinking through the additional conceptual challenges related to the preservation of public media art. For instance, Matthew Kirschenbaum has argued that every media artifact "leaves a "trace," by which he means a past that can be unearthed and understood. But what happens when the original artifact no longer exists, as is the case with the *VideoWall*, which was dismantled and discarded in 1999? And what is the "trace" that is left by public art that makes use of technology as a material support, as in the *VideoWall*'s CRTs and LaserDisc players, which are rarely understood in terms of their material properties? When considering the unique case of the *VideoWall*, additional questions arise: Can the artwork's original public setting can be reimagined in a creative way, perhaps online? Or should the *VideoWall* be rebuilt physically, and installed in the city of Atlanta, and if so, where? (The original site has long been built over). Or should a recreated *VideoWall* be incorporated into a museum's collection, so that it can be benefit from institutional resources? But what would be lost by limiting community access?

Recovering the Rio VideoWall: The Digital Archive

The first phase of recovering the Rio VideoWall, as well as planning for its recreation, involves the creation of a public-facing digital archive that documents the remaining materials associated with the *VideoWall's* design, construction, and eventual demise. The archive includes Birnbaum's original proposals and plans for the piece, her own 35mm slide documentation of the *VideoWall* and her recently-recovered video footage of the artwork's public opening, as well as correspondence, press clippings, and archival photographs of the site. The archive also features exhibits with demographic data and visualizations that illustrate the racial and economic makeup of the areas around the Rio Shopping Complex at the time of the mall's opening in 1987, at the time of its destruction in 2000, and today. Eventually, the site will also feature oral histories of Atlanta citizens who experienced the work. These histories will provide access to lived and felt experiences often absent from the histories provided by critics and scholars, that are nevertheless an essential component of accounting for the impact of public art within a community.

Recovering the Rio VideoWall: The Digital Recreation

Following the completion of the digital archive, the project team will begin the second phase of the project—reimagining the *VideoWall* in digital form. We are currently considering, and mocking up designs for, web-based, physical, and VR-based approaches to recreating the artwork. A web-based version might include visual overlays of the original site of the artwork on the current site. A physical recreation might involve a single, very large flat-screen display that could be separated into adjacent or tiling windows, drone footage of the chosen site, and touch-screen capabilities. A VR version of the work could be situated in a number of virtual locations, and would allow for diverse populations to interact with the artwork from around the globe. A workshop, currently being planned, will bring together scholars, conservators, designers, curators, and community organizers to discuss the computational and creative possibilities of each medium. In addition to describing the conceptual challenges associated with the piece, and highlighting certain key features of the digital archive, the paper presentation will present these mockups and solicit audience feedback in preparation for the final design and construction of the *VideoWall*.

The (Digital) Space Between: Notes on Art History and Machine Vision Learning

Benjamin Zweig
b-zweig@nga.gov
Center for Advanced Study in the Visual Arts, National Gallery of Art, United States of America

Machine vision learning and art historical practice are often poised as operations that are antithetical to one another (Spratt and Elgammal, 2014a, 2014b). A frequent criticism leveled by art historians against machine learning algorithms is that they do little that a trained art historian cannot do already (Bishop, 2017). A second criticism is that the results gleaned from machine vision

learning are heuristic exaggerations. And a third criticism is that computer scientists simply do not understand how to approach visual art, and in the process wrongly (albeit unintentionally) define the field of art history for a much larger audience than the one the humanities tend to generate. Much of the value placed on machine vision learning as it pertains to understanding artworks has been on its ability to sort, classify, and match images with similar ones through style and genre (Saleh and Elgammal, 2015, 2016), taxonomies that fail to reflect the current state of the history of art.

The above criticisms present a fair critique of the approach of machine vision learning to the history of art – but only to a certain degree. Such criticisms fail to recognize that art historians are precisely the ones who have the greatest stake in and the greatest potential for contributing to the questions raised by machine learning image analysis. Art historians simply ask different questions about artworks – questions of history, scale, tactility, surface, and representation – than the ones of which computer scientists are aware. One reason for this disjuncture is that art historians have often kept to themselves instead of engaging with other disciplines that are intensely interested in visual imagery.

Rather than simply critiquing and lamenting how computer and data scientists approach visual imagery, this short paper addresses a few "between points", as I call them, rather than intersections, where art historians can bring much critical insight into machine vision learning. For example, the issue of texture is a complex question in painting, for it can signify the texture of the paint, or the texture of the canvas weave, or how textured paint application is used to represent different physical textures, such as silk or fur. How could these distinctions be brought into machine vision learning? Another issue would be to see if a machine could identify when a painting was re-touched or repaired. Or one might compare how the descriptive terms generated by machine vision learning output correlate to the terms art historians would use when describing an object. The purpose of this paper is ultimately to pose some questions about how art historians and computer scientists might create a better dialogue in their respective practices.

References

Bishop, C. (2017). Against Digital Art History. https://humanitiesfutures.org/papers/digital-art-history/.

Spratt, E. and Elgammal, A. (2014a). Computational Beauty: Aesthetic Judgment at the Intersection of Art and Science. In Agapito, L. et al (eds), *Computer Vision – ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science*8925. Cham: Springer International Publishing, pp. 35–53.

Spratt, E. and Elgammal, A. (2014b).The Digital Humanities Unveiled: Perceptions Held by Art Historians and Computer Scientists about Computer Vision Technology. *Cornell University Library arXiv Physics Archive*. http://arxiv-web3.library.cornell.edu/abs/1411.6714v1.

Saleh, B. and Elgammal, A. (2016). Large-scale Classification of Fine-Art Paintings: Learning the Right Metric on the Right Feature. *International Journal of Digital Art History*, 2: 71-94.

Posters

World of the Khwe Bushmen: Accessing Khwe Cultural Heritage Data by Means of a Digital Ontology Based on Owlnotator

Giuseppe Abrami
abrami@em.uni-frankfurt.de
Goethe-University Frankfurt, Germany

Gertrude Boden
boden@em.uni-frankfurt.de
Goethe-University Frankfurt, Germany

Lisa Gleiß
zoi-m.gleiss@gmx.de
Goethe-University Frankfurt, Germany

Poster Abstract

The Khwe are a group of former hunter-gatherers living in Bwabwata National Park in northeast Namibia. They are one of the indigenous groups in Southern Africa known as "San" or "Bushmen". The documentation of their language and cultural heritage was a mission of Oswin Köhler (1911-1996), a German scholar in African Studies.

Between 1959 and 1992 he built up an integral collection of written vernacular texts, audio files, photographs, video files, ethnographic objects, dried plants and drawings from the Khwe, currently housed in the Oswin Köhler Archive at the Goethe University Frankfurt. As his main oeuvre on the Khwe, Köhler had planned an encyclopedia on ideally every aspect of Khwe culture in vernacular texts with German translations, titled "The World of the Khwe Bushmen [Die Welt der Kxoé-Buschleute]".

Four of twelve planned parts have been published in print so far (Kölher 1989, Köhler 1991, Köhler 1997). Köhler has supplemented, revised, split, merged and moved the texts for this encyclopedia from one part or from one section within a part to another over a time period of more than thirty years.

In order to identify and visualize these processes, a team of computer scientists and anthropologists/linguists has developed an OWL ontology for the semantic use of the Köhler encyclopedia. It maps the histories of individual texts and of their position in the overall structure of the encyclopedia but also the relations between Khwe terms, subject areas, text versions, versions of table of contents, footnotes to texts describing manipulations, codes for recurrent types of manipulations to the texts, object types, specific objects, video- and audio-files, photographs, drawings, people and places. It thus allows for a more holistic or integral understanding of Khwe concepts and cultural practices by presenting them in the multiple contexts where they occurred. At the same time it allows for retracing the formation of this cultural heritage documentation by revealing the impact of individual actors in changing and manipulating the documentation, with regards to content as well as numerically, e.g. the replacement of loan words for Khwe terms or attempts to standardize syntax. All this is done with the help of the so-called *OWLnotator* (Abrami et al. 2012).

OWLnotator, as part of the eHumanities Desktop (Jussen, Mehler, Ernst 2007), is a flexible annotation system for annotating inter- and intramedial relations in multimedia corpora and can be used as an annotation platform for any project. By using OWL ontologies as an annotation scheme, arbitrary annotation tasks can be defined.

For this purpose, OWLnotator provides a generic graphical web interface that displays the available classes and properties of the underlying ontology and allows linking to arbitrary resources. These resources are provided by the integration of OWLnotator into the so-called *ResourceManager*. ResourceManager is also part of the eHumanities Desktop and provides access to various types of resources as text documents, images, audio- and video-files, as well as their individual segments and more. In addition, the OWLnotator can also import data from CSV files, provided by the ResourceManager, and assign them to the corresponding ontologies. Furthermore, in this version of the OWLnotator *Blazegraph* (https://www.blazegraph.com/) is used as the new database backend. The poster presents the challenges of understanding and designing the multiple relations between individual items within the ontology, of formally describing and transforming existing data and databases to render them readable or automatically importable, and of visualising the items and relations in the ontology. As a result, the poster will display selected Khwe concepts with all their relations. The analytical potential of the ontology will be exemplified by presenting the results of a number of queries visualization with help of OWLnotator.

References

Abrami Giuseppe and Mehler Alexander and Pravida Dietmar (2015). *Fusing text and image data with the help of the OWLnotato*r. In Sakae Yamamoto, editor, Human Interface and the Management of Information. Information and Knowledge Design, volume 9172 of Lecture Notes in Computer Science, pages 261–272. Springer International Publishing

Jussen, Bernhard and Mehler, Alexander and Ernst, Alexandra (2007). *A corpus management system for historical semantics*. Sprache und Datenverarbeitung. International Journal for Language Data Processing, 31(1-2):81–89.

Köhler, Oswin (1989). *Die Welt der Kxoé-Buschleute im südlichen Afrika: eine Selbstdarstellung in ihrer eigenen Sprache*. 1. Die Kxoé-Buschleute und ihre ethnische Umgebung. D. Reimer, Berlin, Germany.

Köhler, Oswin (1991). *Die Welt der Kxoé-Buschleute im süudlichen Afrika: eine Selbstdarstellung in ihrer eigenen Sprache*. 2. Grundlagen des Lebens: Wasser,

Sammeln und Jagd, Bodenbau und Tierhaltung. D. Reimer, Berlin, Germany.
Köhler, Oswin (1997). *Die Welt der Kxoé-Buschleute im Südlichen Afrika: Die Welt der Kxoe-Buschleute im südlichen Afrika*. 3. Materialle Ausrüustung; Wohnplatz und Buschlager. D. Reimer, Berlin, Germany.

Design on View: Imagining Culture as a Digital Outcome

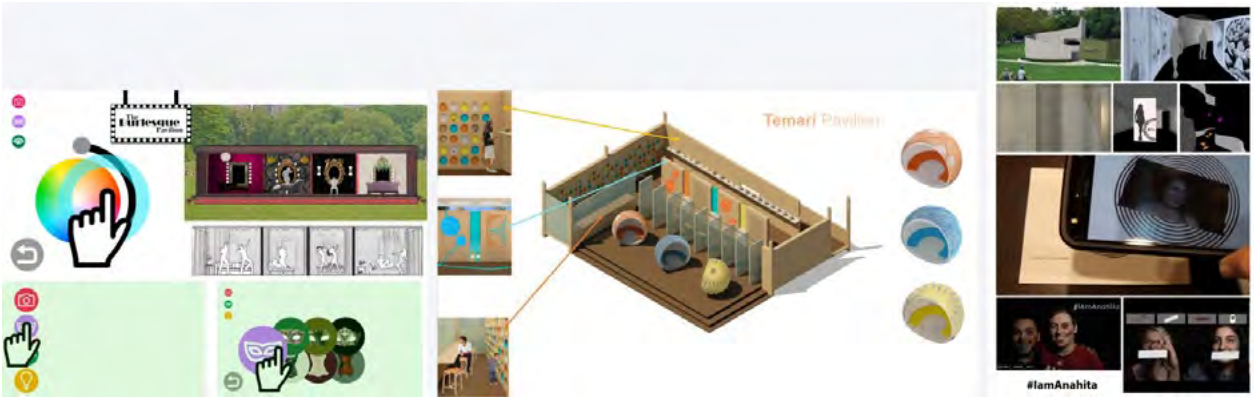
Ersin Altin
ersin.altin@njit.edu
New Jersey Institute of Technology, United States of America

Can **design** represent a culture/nation? Can the tools of digital design be used in collaboration with industrial and interior design to establish an interactive communication with culture? While design and **designwork** were seen as essential symbols of nation-based identity construction in most of the 20th century, today, the notion of design deliberately shies away from exposing its cultural/national implications because of global aspirations. Today's world, dominated by multinational corporations, with its imposition on self-centered identities seemingly curtains the close connection/flirtation of design to its cultural roots. The project that is developed as a collaborative design task at School of Art + Design at New Jersey Institute of Technology (NJIT) aims to question and build on the assumption that suggests a connection between design and culture/nation, with the emphasis on the fact that **nation** is also a social construction (Anderson, 1983).

This poster visualizes the results of the collaborative design project that I taught at NJIT in Fall 2016 and again in 2017. Throughout the semester students from different design fields were expected to work as a group on the design of a pavilion for the culture/nation of their selection that together with other teams formed an imaginary exposition center. Instead of superficial identifications, systematic research process and critical design concepts based on intellectual analysis of the findings determined a basis for the design project. By both researching and producing, teams aimed to create a digital tool that would be developed to investigate whether **designwork** can represent a culture/nation, subculture or simply a cultural issue. Three teams consist of three students from three different design fields worked on their pavilions that are imagined as interactive tools. These tools incorporating data processing software, motion capture, virtual and augmented reality establish vivid, interactive communication with the user. In doing so, instead of creating informative two-dimensional representations, projects aimed to involve users to explore their contribution to the dynamics of a culture. In other words, instead of imposing a **meaning**, pavilions ask users to build new meanings via their interactions both with the pavilion and with other users.

The poster documents three different design processes each of which produced its own interactive digital tools to communicate culture. One team envisioned a mobile pavilion for Burlesque culture that offered users to design their own shows. Augmented reality helped users/performers select and **put-on** a stage costume digitally. With a digital control panel performers were given a chance to adjust atmospheric effects such as light level and color, while physicality of the setting was conceived through a meticulous analysis of the Burlesque culture, such as heavily ornate historic furniture, wallpapers, textile, and decoration.

Second team created a digital crafting tool to educate visitors about Japanese Temari balls, which are toy balls made from embroidery may be used in handball games. Team tackled weaving as a craft with the question how and why weaving can be utilized as data analysis with an emphasis on its fabrication processes by using Japanese Temari balls as a case study. The pavilion encouraged visitors not only to learn about Temari tradition, but also share their experience with other users, who do not necessarily speak the same language or come from similar cultural backgrounds by transforming Temari making into a cultural activity that is virtually organized around a **ball game** / spectacle.



Burlesque Pavilion by Hideyoshi Azama, Emily Gutierrez, Tulio Squarcio (left); Temari Pavilion by Danielle Archibold, Wuraola Ogunnowo, Florencia Pozo (middle); Pavilion Anahita by Negaar Amirihormozaki, Albeirys Francisco-Parra, Nazifa Hamidullah (right)

The third team designed a pavilion that aimed to create a community by gathering people both physically in the space of the pavilion and virtually through social platforms such as Facebook, Twitter, Snapchat, and Instagram. The team problematized Iran's mandatory hijab law by connecting the issue to sexism in different parts of the world that creating a network on women's rights issues. Hijab's ban in some countries and its enforced use in others were carefully examined to generate a digital forum for different opinions on this specific issue.

This research was conducted to investigate culture's changing perceptions. Rather than attempting to redefine a preconceived notion of culture by simply incorporating modern technologies, digital tools, and social media, it aimed to reveal new interactive networks that culture forms with other notions and omit others when conventional relations needed replacement; for example, a new interconnectedness instead of nationality. Finally, this project highlighted areas that were defined by the conventional cultural tools and perceptions that are still relevant.

References

Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London and New York: Verso.

Introducing Polo: Exploring Topic Models as Database and Hypertext

Rafael Alvarado
ontoligent@gmail.com
University of Virginia, United States of America

Since the invention of Latent Dirichlet Allocation (Blei, et al. 2003) and early demonstrations of its utility for identifying lexical clusters in collections of historical and literary texts (Block and Newman 2006, Blevins 2010), topic models have become a mainstay of the digital humanities. However, the use of topic models within the field remains narrowly conceived, restricted largely, with some exceptions, to the discovery of topics and topic trends within corpora, even though the method has been extended significantly since first introduced. One reason for this conservativism may be that, like many methods drawn from data science, both the process and the output of topic model algorithms remain interpetively opaque to the humanists (and, arguably, to the computer scientist as well). Aside from the complexity of the math involved, a contributing factor to this opacity has been the limited way in which the results of topic models are presented to the user. One the one hand, the data provided by standard topic modeling tools (whether in Java, Python, or R) are often trapped in data files or shielded by objects that cannot be queried directly or visualized freely without the use of ad hoc programming or spreadsheet software. On the other hand, the outputs typically provided by these tools, such as top words per topic (often visualized as word clouds), show a highly restricted, decontextualized, and potentially distorted picture of the model (Schmidt 2013). Recently, various tools have emerged to fill this gap, such TOME (Klein et al. 2015), which is designed to allow scholars to explore topic models more fully. In this talk I will present Polo, a topic model browser developed at the Data Science Institute at the University of Virginia designed to present topic models to users in a direct, transparent, and complete manner, so that the representational quality of models may be explored, questions, and adjusted interactively. Built on top of MALLET, Gensim, and NLTK, Polo is a Python package that provides tools to both create topic models and to inspect them by combining the source corpus with all of the data produced by the core software into a single, normalized relational database (in SQLite). This database in turn forms the foundation of an interactive web application that effectively converts the output model with associated data and the source corpus into a single hypertext relating words, topics, and documents. A key design feature of Polo is that it employs the statistical properties of the model -- such as topic entropy in documents or mutual information among topics -- not simply as readouts on a dashboard but as navigational devices that allow the user to move from a reduced dimension, high-level perspective of a corpus to its source documents, and to move laterally through the network of topics and documents that compose the model. Using examples from both newspaper and journal collections, I will demonstrate how Polo enables scholars both to investigate implied cultural newtworks in these corpora and to explore the various ways in which topics may be said to convey meaning.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2002. "Latent Dirichlet Allocation." In *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani, 601–608. MIT Press.

Blevins, Cameron. 2010. "Topic Modeling Martha Ballard's Diary." *Cameron Blevins* (blog). April 1, 2010, http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/.

Klein, Lauren F., Jacob Eisenstein, and Iris Sun. 2015. "Exploratory Thematic Analysis for Digitized Archival Collections." *Digital Scholarship in the Humanities* 30 (suppl_1):i130–41.

Newman, David J., and Sharon Block. 2006. "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." *Journal of the American Society for Information Science and Technology* 57 (6):753–767.

Schmidt, Benjamin M. 2013. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities*. April 5, 2013.

El primer aliento. La expedición de los lingüistas Swadesh y Rendón en las ciencias computacionales (1956-1970)

Adriana Álvarez Sánchez
adralvsan@gmail.com
Universidad Nacional Autónoma de México, Mexico

El póster tiene como objetivos aportar conocimientos sobre los orígenes de las HD en América Latina y mostrar los primeros procesos computacionales realizados por los investigadores de la UNAM, Maurice Swadehs y Juan José Rendón, dedicados a la lingüística comparada de las lenguas mayas utilizando una computadora IBM 650 de alquiler.

Se abordan tres aspectos que serán ilustrados con documentos originales conservados en los archivos de la propia universidad: la gestión de recursos, el proceso técnico y los resultados de la investigación de estos lingüistas. Las cartas enviadas por Swadehs al director del Centro de Cálculo Electrónico de la UNAM para hacer uso de la computadora evidencian las gestiones realizadas por el lingüista, así como la descripción del proyecto. El proceso computacional ha quedado descrito en cartas pero también en los informes que este investigador hacía anualmente, así como en el archivo personal de Rendón, donde además he encontrado diagramas de flujo, matrices y otros documentos institucionales sobre el proceso realizado. Finalmente, parte de los resultados de la expedición de estos lingüistas en la aplicación de las ciencias computacionales al análisis de las lenguas quedó registrada en la revista *Estudios de Cultura Maya*. La lectura de esta publicación permite conocer también parte de los debates de aquella época acerca del uso de técnicas y tecnologías computacionales para el estudio de las lenguas. Ejemplo de ello fue el trabajo de los epigrafistas de la Sección Siberiana de la Academia de las Ciencias de la URSS sobre escritura maya antigua.

Los documentos consultados y reproducidos en el póster proceden de los Fondos Documentales Alfonso Caso del Instituto de Investigaciones Antropológicas de la UNAM, donde se encuentra el Fondo Juan José Rendón que consta de más de 100 cajas. En las Colecciones Especiales de la Biblioteca Juan Comas, Sección de Antropología del mismo Instituto se encuentran copias de los expedientes institucionales de ambos académicos.

La presente investigación se encuentra en un nivel avanzado de desarrollo y busca reconstruir la manera en la que los humanistas hicieron uso de los avances tecnológicos en sus disciplinas. Me interesa postular hipótesis acerca de si fueron las condiciones institucionales, el desinterés por metodologías de esta naturaleza o las propias corrientes de la lingüística, la razón por la cual no hubo continuidad en el desarrollo de estas investigaciones. Un proceso semejante se dio en los estudios históricos a nivel mundial, a raíz de la publicación de obras que dieron mayor peso a la estadística, dejando de lado la interpretación histórica.

Conocer la manera en la que las disciplinas humanísticas enfrentaron y/o aprovecharon los cambios tecnológicos de mediados del siglo XX contribuye a comprender el estado en el que hoy interactúan las distintas áreas del conocimiento, incluidas las ciencias computacionales.

Las aportaciones de este trabajo se concretan en ofrecer conocimientos sobre las primeras experiencias de aplicación de las ciencias computacionales que se llevaron a cabo en distintas latitudes y que forman parte de los orígenes de las HD. También se rescatan investigaciones olvidadas, incluso para la historia de la lingüística, que exploraron la relación entre la tecnología y el estudio de las sociedades. Se reconstruyen procesos computacionales del tercer cuarto del siglo XX aplicados al estudio de las lenguas indígenas y, finalmente, se ponen de manifiesto aspectos que hoy continúan conformando condicionantes para el desarrollo y sostenimiento de proyectos digitales en las universidades.

References

Barrera Vásquez, A. (1962). Investigación de la escritura de los antiguos mayas con máquinas calculadoras electrónicas: síntesis y glosa, *Estudios de cultura maya*, II: 319-342.

Beltrán, S. (1959). Carta del Ingeniero Sergio Beltrán al Dr. Maurice Swadesh, 5 de diciembre de 1959". *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 147-148.

Knorozov, Y. (1963). Ap.licación de las matemáticas al estudio lingüístico, *Estudios de Cultura Maya*, III: 169-185.

Rendón, J. J. (1967-68). Plan de trabajo de Juan José Rendón, 1967-68, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 05, Carácter C, núm. de registro 29, ff. 21-24.

Rendón, J .J. (1973). Epistolario, *Fondo Juan José Rendón*, caja 1.

Redón, J. J. (s.a.). Listas diagnósticas. Léxico-Estadística, *Fondo Juan José Rendón*, caja 12.

Redón, J. J. (1971) Reseña Breve introducción a la computación lingüística de Paul L. Garwin, *Anales de Antropología*. 8: 313-314.

Swadesh, M. (1958-60). Plan de Trabajo y Desarrollo, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 200-201.

Swadesh, M. (1960). Interrelaciones de las lenguas mayenses, *Anales del Instituto Nacional de Antropología e Historia*, XIII: 231-267.

Swadesh, M. (1961). Carta del Dr. Maurice Swadehs al Ing. Pablo Martínez del Río", 16 de febrero de 1961, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06,

Carácter C, núm. de registro 30, f. 113.
Swadehs, M. (1966). Carta del Dr. Maurice Swadehs al Dr. Miguel León Portilla", 29 de agosto de 1966, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 41-42.
Swadesh, M. Curriculum Vitae del Dr. Mauricio Swadesh Talnoper, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 99-100.
Swinggers, P. (2016). Tras las huellas de Mauricio Swadehs: en búsqueda de una lingüística total, *Revista de investigación lingüística*, 19: 107-130.

The Spatial Humanities Kit

Matt Applegate
mapplega@gmail.com
Molloy College, United States of America

Jamie Cohen
jamesncohen@gmail.com
Molloy College, United States of America

This poster session showcases "the spatial humanities kit": a combination of gear, open source code, and teaching materials for narrative GIS projects (http://spatialhumanitieskit.org/). The kit was derived and assembled from two international mapping projects executed by students and guided by faculty at Molloy College and Hofstra University. The kit includes the following gear and code, all of which will be available for faculty to interact with at DH 2018: an introduction to GeoJson with code also applicable for Mapbox, Open Street Map, and ArcGIS, an Insta 360 Camera, Snapchat Spectacles, two Garmin ETreX 20x GPS devices with preinstalled maps, a Samsung 360 Camera, a chicken foot tripod, Samsung Gear Oculus HMD, user cell phone cameras, a Skyroam Global Hotspot, a GoPuck Qualcomm Charge 3.0, and a GoPro Session.

Project Description & Framework:

The spatial humanities kit is a durable toolset designed to fit in a backpack. The gear and code that it features are meant to combine and enhance two approaches to GIS related work in the humanities. First, the combination of gear included in the kit is designed for their user to narrativize the spaces that they map. Following Jason Farman's approach to locative media, the gear's use is predicated on two concepts in particular, "site specificity" and "urban markup." Site specificity, as Farman defines it, pertains to "the unique qualities of a unique location that cannot be transferred onto another place," whereas urban markup refers to "the various ways that narrative gets attached to a specific place in a city." The spatial humanities kit is designed to capture both.

In the summer of 2016 students at Molloy College traveled to Northeastern Ireland and documented their trip under faculty guidance via the spatial humanities kit and an Omeka archive (http://molloymediaarchaeology.org). Students documented and narrativized their experience of urban and rural space, historical sites, and religious sites, combining the unique qualities of each location (GPS coordinates, landmarks, etc.) with a linear telling of their site specific experiences. In the summer of 2017, the project was refined and expanded to Hofstra University. Students used the kit under faculty guidance in Italy to research and report on social inequality, government corruption, recovery and revitalization, and media change in earthquake damaged L'Aquila, the Naples region of Scampìa, and the Roman town of Frascati (http://lhscmediaarchaeology.org).

Ultimately, both projects, especially in their map's function as an artifact, play with the spatial humanities use and function. Where our use of the kit has emphasized autoethnography, social good, and bringing accountability to historical narratives, the spatial humanities kit augments the discipline's preoccupation with space-time. Consider Ian Gregory's engagement with Doreen Massey's work in "Exploiting Time and Space: A Challenge for GIS in the Digital Humanities": "Time is needed to tell the story of how an individual place developed to become what it is now, however without space there is only one story and thus the risk that it is seen as the only possible story and the inevitable story." Thus far, the spatial humanities kit has expanded the narrative possibilities of humanities GIS projects by multiplying narratives about spaces that are mapped.

Interactive Experience:

Our proposed poster session will offer faculty the opportunity to learn what the spatial humanities kit is, how they can adopt it, and how students can operationalize it. In addition to the kit itself, we will offer faculty syllabi, access to the Molloy and Hofstra University projects, as well as the source code for our maps. Our goal is to maximize the use and function of the kit by making our work, and our student's work, more broadly available. Further, we aim to approach the interaction between tools for digital storytelling and approaches to spatial humanities differently by combining both toolsets with their attendant pedagogical applications.

References

Farman, Jason. *The Mobile Story: Narrative Practices with Locative Technologies.* Routledge, 2013.
Gregory, Ian. "Exploiting Time and Space: A Challenge for GIS in the Digital Humanities." *The Spatial Humanities*. Eds. David J. Bodenhamer, John Corrigan, and Trevor M. Harris. Indiana University Press, 2010

The Magnifying Glass and the Kaleidoscope. Analysing Scale in Digital History and Historiography

Florentina Armaselu
florentinaa@zoomimagine.com
University of Luxembourg, Luxembourg

Introduction

What is the meaning of scale in historical writings and migration narratives? Can digital tools and methods assist the detection of scale-related patterns in these categories of documents? May this enquiry be formalised into a system for scale analysis in texts? To address these questions, the paper combines theoretical background from historical, historiographical, linguistic and literary studies with digital tools and methods for text analysis and visualisation. The project is in an early phase; theoretical hypotheses and preliminary experiments are presented.

Methodology

Two types of corpora were considered: (1) historiographical - history writings mingling micro and global perspectives; (2) historical - migration narratives (autobiography). The first, in which variations of scale are clearly present, will serve to develop a prototype. The second, where representations of scale are more difficult to assess, will be used to test the approach.

Corpora

Although recent research in "global microhistory" (Trivellato, 2011) draws attention to the variable scale representation in history, the question of how this phenomenon is expressed through language in historians' discourse is less studied. Research enquiries may be related to: topics distribution pertaining to scale (local to global, micro to macro); "story" versus "study" distinctions (Kracauer, 2014: 122); epistemological explorations (Boudon, 1991). Corpus (1) samples: Brook (2009), Rothschild (2013), Wills (2001).

Corpus (2) is intended to East-West migration narratives, e.g. Kaminer (2011), Kassabova (2009), Verboczy (2017). Potential queries: representation of space and its scale-related particularities, e.g. the intimate, symbolic meaning, inspired by Bachelard (1957), of the old and new "home" (interior objects, house, street, city, country, continent) and its connections to geo-historical or cultural spaces, and a certain sense of belonging. Other elements could be considered: relations, names, events, time references.

Approach

The aim is to bridge "distant" and "close" reading, using zooming metaphor as an interpretative tool (Armaselu and Heuvel, 2017). Thus, a corpus/text can be explored via the hypothetical schema:

Level1: topic_X (obj_1, obj_2, …, obj_n)
Level2: topic_X.1 (obj_1.1, obj_1.2, …), topic X.2 (obj_2.1, obj_2.2, …), …
Level3: topic_X.1.1 (obj_1.1.1, obj_1.1.2, …), topic_X1.2, …, topic_X2.1, etc.

Where, 'obj_topic[.subtopic]' represents a whole/section/fragment of a document associated to a topic and a scale-related logic. The system will allow zooming-in/out the different topics, traversing the conceptual space, e.g. from general to specific, and accessing the corresponding objects. One of the challenges is that the levels hierarchy and the degree of granularity may not be unique but depend on different "perspectives". Corpus (1) can imply different viewpoints and objects grouped by topics on levels 1, 2: (a) world history – 17th, 18th century; (b) world history – trade routes, slavery; (c) world history – Europe, Asia, America. Some fragments generalise on world history, others discuss world trade routes between Europe, America and Asia, others narrow down to family history or paintings description. Like in a kaleidoscope, by rotating the device (changing the "magnifying-glass"), new patterns can emerge.

Proof of Concept (PoC)

The PoC phase (in progress) will test these hypotheses on corpus (1). Two experiments on Brook (2009) are presented below.

Figure 1 illustrates Paper Machines topics for each chapter. It is assumed that by combining these groupings with an analysis of the contexts where the corresponding words appear, e.g. co-occurrences, lexical chains, paths in a lexical-semantic hierarchy, a scale-related model of the text can be derived. Its levels may reflect how knowledge is organised, from synthesising, manipulating abstractions, through intermediate descriptions, to in-detail accounts referring to particular facts, persons, objects or quotations of sources.

Figure 2 shows a visualisation via Z-editor (Armaselu, 2010). The scalable layout in chapter 2 (created manually) is explored by zooming through the European hatters history in the fifteenth and sixteenth century, the opening of the beaver pelts Canadian supply and Champlain's fight with the Mohawks, the customs of wearing a hat and the rules of courtship in seventeenth century Netherlands, and, Vermeer's painting, *Officer and Laughing Girl,* illustrating these practices.

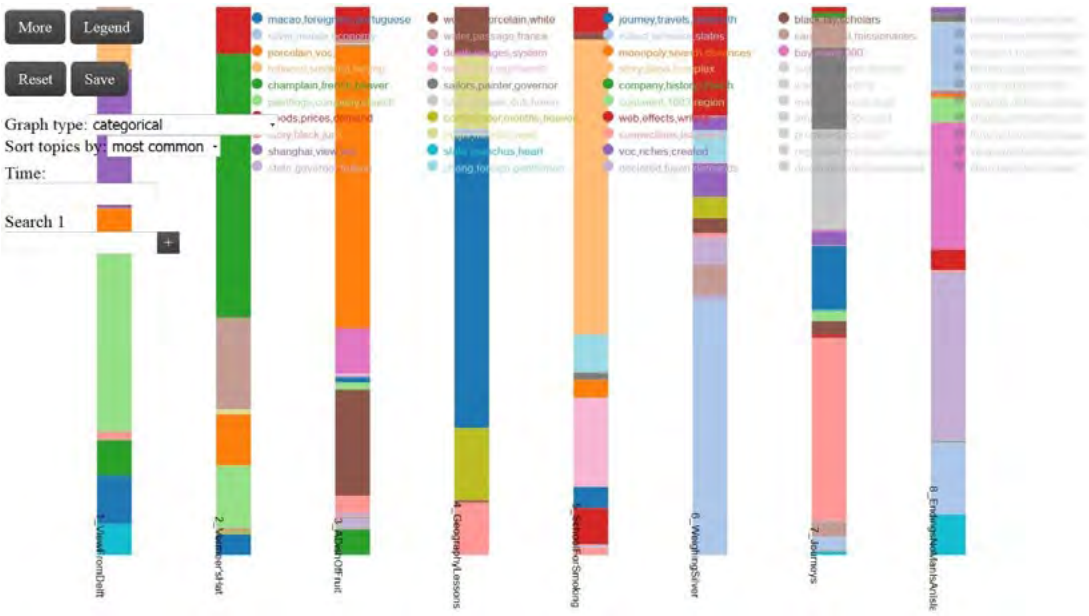


Fig. 1. *Vermeer's Hat*. Zotero - Paper Machines (topic modelling by subcollection/chapters)

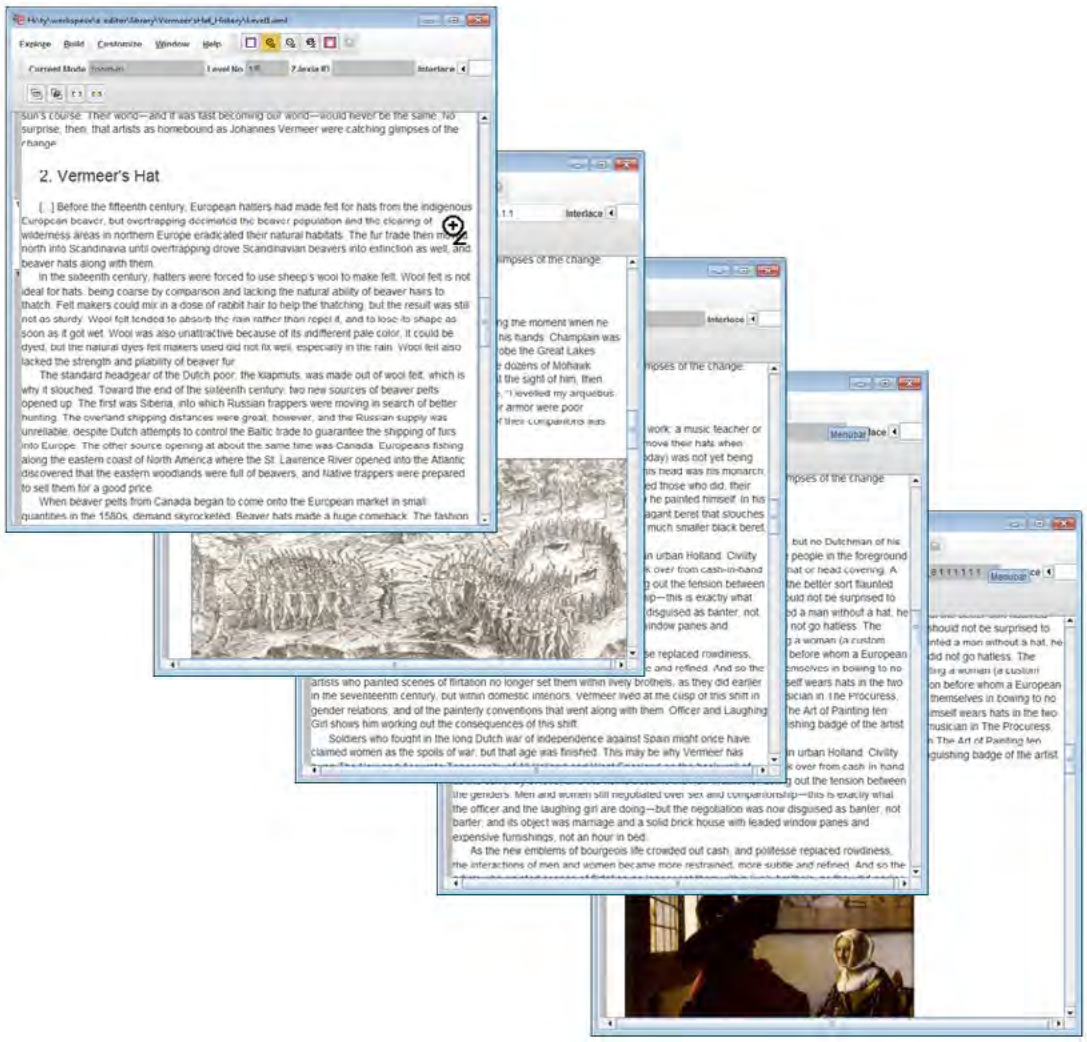


Fig. 2. *Vermeer's Hat*. Z-editor (zoomable text)

Tools/methods currently under testing: topic modelling (MALLET), textometry (TXM), lexical-semantic resources (WordNet), Named Entity Recognition (GATE), lexical chains and text structure (Morris and Hirst, 1991), visualisation (graphs, textual zooming). The PoC outcome will consist of insight into the advantages/limitations of these tools/methods in building a prototype for scale analysis.

Conclusion

The paper presents theoretical points and experiments for a system dedicated to scale analysis in historical/historiographical texts. By a combined approach, evoking the metaphors of the magnifying glass and the kaleidoscope, the system may allow both scale-related patterns detection and perspective change.

References

Armaselu (Vasilescu) F. (2010).Ph.D. Thesis, *Le livre sous la loupe : Nouvelles formes d'écriture électronique*, Papyrus, University of MontrealInstitutionalRepository.

Armaselu, F. and Heuvel, C. van den. (2017). "Metaphors in Digital Hermeneutics: Zooming through Literary, Didactic and Historical Representations of Imaginary and Existing Cities", In *Digital Humanities Quarterly (DHQ)*, Volume 11, Number 3.

Bachelard, G. (1957). *La poétique de l'espace*, PUF.

Boudon, P. (1991). *De l'architecture à l'épistémologie. La question de l'échelle*, PUF.

Brook, T. (2009).*Vermeer's Hat.The Seventh Century and the Dawn of the Global World*, Profile Books.

Kaminer, W. (2011). *Russian Disco,* Translated by Michael Hulse, Ebury Press.

Kassabova, K. (2009). *Street Without a Name: Childhood and Other Misadventures in Bulgaria*. New York: Skyhorse Publishing.

Kracauer, S. (2014). *History. The Last Things Before The Last,* Markus Wiener Publisher.

Morris, J. and Hirst, G. (1991). "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", In *Computational Linguistics,* Volume 17, Number 1, Association for Computational Linguistics.

Rothschild, E. (2013). *The Inner Life of Empires. An Eighteenth Century History*, Princeton University Press, 2011, paperback 2013.

Trivellato, F. (2011). "Is There a Future for Italian Microhistory in the Age of Global History?", *California Italian Studies*, 2(1).

Verboczy, A. (2017). *Rhapsody in Quebec. On the Path of an Immigrant Child.* Translated by Casey Roberts. Montréal: Baraka Books.

Wills Jr., J. E. (2001). 1688. *A Global History*, New York, London: W.W. Norton & Company.

Tools

GATE - General Architecture for Text Engineering, https://gate.ac.uk/.

MALLET - MAchine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/topics.php.

Paper Machines - http://papermachines.org/.

TXM – Textométrie project, http://textometrie.ens-lyon.fr/?lang=en.

Z-editor - http://www.zoomimagine.com.

WordNet - https://wordnet.princeton.edu/.

Encoding the Oldest Western Music

Allyn Waller
awalle18@g.holycross.edu
College of the Holy Cross, United States of America

Toni Armstrong
toarmstrong@clarku.edu
Clark University, United States of America

Nicholas Guarracino
nmguar18@g.holycross.edu
College of the Holy Cross, United States of America

Julia Spiegel
jrspie19@g.holycross.edu
College of the Holy Cross, United States of America

Hannah Nguyen
hnguye19@g.holycross.edu
College of the Holy Cross, United States of America

Marika Fox
marfox@clarku.edu
Clark University, United States of America

Problems of Encoding

This project describes a system for digitally encoding neumes and corresponding text in parallel aligned documents in order to create a digital, diplomatic edition of chant texts with neumes. Neumes are graphic marks denoting relative changes in pitch; they predate staff notation and are written above text. Each neume can be marked with performance variations, called episema or liquescence. They also include musical directions abbreviated in Latin, with 15 significative letters such as 't' for 'tenere' to indicate holding a note longer. There are at least a dozen styles of neumes, each of which has its own set of graphical symbols, like different fonts, to represent the same neumes.

A diplomatic edition of a neumed chant text must record the neumes as characters, not as absolute pitches.

It also must align neumes with text, as they are visually aligned by syllable in chant manuscripts.[1]

The 'Virgapes' system is based on a four-part encoding scheme for neumes that is flexible, extensible, and universal.[2] We have also developed a parallel document structure to align separate documents of text and neumes.

The 'Virgapes' System

In the Virgapes encoding, each neume is represented with a four-part code point. Each part is an integer standing for an aspect of the neume. The first integer denotes the number of pitches in the neume. The second integer is an arbitrarily assigned identifier within that group.[3] The system is flexible; it can expand to accommodate new or lesser known neumes. The third integer indicates the presence of episema, 1 for presence, 0 for absence. Likewise, the fourth notes liquescence in the same binary pattern.

For example: virga is a one-pitch neume, encoded as 1.1.0.0 in absence of episema or liquescence; pes is a neume of two ascending pitches, encoded (if liquescent) as 2.2.0.1.

The inclusion of episema and liquesence allows editors to note graphic marks indicating performance changes without imposing meaning. Our system also allows for specified searching: for all instances of virga or only instances of virga with episema, depending on the needs of analysis.

Parallel Aligned Documents

In addition to encoding neumes, we align transcriptions of neumes with transcriptions of texts. In a manuscript, this is done graphically: the neumes appear above the text.

In our digital editions, we create two parallel documents aligned by canonical citation using a Canonical Text Services (CTS) URN system to uniquely identify each passage.[4] With this, the two documents share a work hierarchy and a passage hierarchy. Consider the URN: 'urn:cts:chant:antiphonary.einsiedeln121.text:11.introit'. The CTS namespace is 'chant' for the domain of chant texts. The group is 'antiphonary' for the type of chant book. The specific work is Einsiedeln 121. The last section notes the version, text or neume.

The second portion of the URN system is a passage hierarchy, which subdivides the work hierarchy. A parallel would be the act, scene, and line in plays. It first identifies the feast day using numbers delineated in the *Antiphonale Missarum Sextuplex*.[5] Then, the subsection: introit, verse, etc, with further identifying numbers for graphically separated passages.

The URN system provides a citation scheme to align the texts; within the documents they are aligned by syllables, as each syllable must have at least one neume. This also provides a check for our encoding—there must be equal syllables in the text and neume document.

Digital encoding of neumes allows for advanced searching and analysis. With our two-part encoding solution, it is possible to search for repeated musical sequences, to determine if Zipf's law applies to neumes, or to analyze musical texture based on the neume:text ratio.

1 Among chant scholars, the most important digital resources are the manuscript databases of the Cantus Index network. (See http://cantus.uwaterloo.ca/) These datasets include information about the manuscripts themselves in addition to the encoding of text and music. Unlike the Cantus system, however, we encode staffless neumes without imposing interpreted equivalences to later musical notation on staffs.
Of the XML systems, the most significant is the Music Encoding Initiative (MEI) (http://music-encoding.org/documentation/3.0.0/neumes). It is largely inspired by work at Tübingen. Our system also allows encoding of basic neumes with extended properties (liquescence, episema) in a specified syntax, enabling us to take account or ignore these properties in computational manipulation. Neither the current XML schemes nor the Cantus Index allow these properties to be optional.

2 Called 'Virgapes' for the first one and two-note neumes, *virga* and *pes*.

3 These are available from our Github repository: https://github.com/HCMID/chant.

4 This system was developed as part of the CITE Architecture for the Homer Multitext Project 2010-18, and applied to this project: http://cite-architecture.github.io/ctsurn/.

Creating a Digital Edition of Ancient Mongolian Historical Documents

Biligsaikhan Batjargal
biligsaikhan@gmail.com
Research Organization of Science and Technology
Ritsumeikan University, Japan

Garmaabazar Khaltarkhuu
garmaabazar@gmail.com
Mongolia-Japan Center for Human Resources Development,
National University of Mongolia, Mongolia

Akira Maeda
amaeda@is.ritsumei.ac.jp
College of Information Science and Engineering
Ritsumeikan University, Japan

Introduction

In this poster, we introduce a digital edition of the Altan Tobchi, a Mongolian historical manuscript written in traditional Mongolian script. The Text Encoding Initiative (TEI) guidelines were adopted to encode the named entities. A web prototype was developed for digital humanities scholarship for utilizing digital representations of ancient Mongolian historical manuscripts as scholarly tools. The proposed prototype has the capability to display and search TEI encoded traditional Mongolian text

5 A standard chant reference work compiled by Dom Hesbert in the early 20th century. It contains transcriptions of six important sources of Gregorian chant.

and its transliteration in Latin letters along with the highlighted named entities and the scanned images of the source manuscript. This poster discusses how to develop a digital edition of Mongolian historical documents.

Mongolian manuscripts

Mongolian historical documents have been written in numerous scripts, i.e., the traditional Mongolian script, Square or Phags-pa script, Todo or Clear script, Soyombo script and Horizontal square script (Shagdarsuren, 2011). Among them, the traditional Mongolian script is the most popular and longest-surviving script for over 800 years. This research focuses on the traditional Mongolian script.

In 1946, Mongolia has made language reforms to eliminate a difference between written and spoken Mongolian language, and the Cyrillic script was adapted to Mongolian. The spelling of modern Mongolian was based on the pronunciations in the Khalkha dialect, the largest Mongol ethnic group (Sečenbagatur et al., 2005; Svatesson et al., 2005). Such a radical change separated the Mongolian people from their historical archives written in traditional Mongolian script. Reading traditional Mongolian documents by using literacy in modern Mongolian is not a simple task. Thus, a digital text representation that explains a given manuscript in a modern Mongolian is helpful for users who want to read, search and browse ancient Mongolian manuscripts.

Mongolian manuscripts in the digital age

To the best of our knowledge, there are a small number of digital texts of ancient Mongolian manuscripts. A few ancient Mongolian manuscripts including (1) 'Qad-un ündüsün-ü quri-yangyui altan tobči neretü sudur' (The Golden Summary: Short history of the Origins of the Khans) a.k.a. "Little" Altan Tobchi and (2) the 'Asarayči neretü-yin teüke' (The Story of Asragch) have been converted to digital texts and made publicly available (Batjargal et al., 2012).

Information processing of Mongolian manuscripts

Batjargal et al. have developed the traditional Mongolian script digital library (TMSDL) (Batjargal et al., 2012), which can be used to access and retrieve historical manuscripts written in traditional Mongolian script using a query in modern Mongolian (Cyrillic). Moreover, Batjargal et al. also proposed a named entity extraction method (Batjargal et al., 2016), which extracts proper nouns from digitized text of ancient Mongolian documents using Support Vector Machine with 0.6993, 0.5679 and 0.6268 of precision, recall and F-measure respectively. These researches have motivated us to create a digital edition that reflects ancient Mongolian historical manuscripts.

Digital edition of Mongolian manuscripts

We utilized Edition Visualization Technology (EVT) for creating a digital edition of Mongolian manuscripts, which is encoded according to the TEI XML schemas and guidelines (Del Turco et al., 2014). As shown in Figure 1 and Figure 2, all the personal names and place names (Figure 3) in the Altan Tobchi are highlighted by using the results of a named entity extraction method (Batjargal et al., 2016) and the named entities' indices obtained from the "Qad-un ündüsün quriyangyui altan tobči –Textological Study" (Choimaa, 2002). We made the following customizations in EVT to make it suitable for Mongolian manuscripts in traditional Mongolian script.

Figure 1. Image-to-text link and personal names' highlights

Parallel-text editions with transliteration

The proposed prototype can present scanned image-based editions with two edition levels: (1) diplomatic interpretative and (2) transliteration. Transliteration is helpful for those who are not familiar with a script of a certain language but understands that language. Transliteration in Latin letters of Mongolian historical documents is popular among scholars.

There is a limited recommendation to encode transliterations in TEI. Soualah and Hassoun (Soualah & Hassoun, 2012) proposed to implement transliteration by using a specific model, which uses the *<ref>* element with the *@xml:lang*, *@target*, and *@type* attributes. However, we consider transliteration as a separate edition and use it as parallel-text editions as shown in Figure 2.

Supporting the traditional Mongolian script

A unique feature of traditional Mongolian script is displaying vertically, from top to bottom, in columns advancing from left to right. Due to poor support for traditional Mongolian script at the EVT, we customized it to display the scanned images at the top and the corresponding text in traditional Mongolian script below with the direction top to bottom and left to right (Figure 1). We also set to dis-

play text in traditional Mongolian script on the left, and the corresponding transliteration in Latin letters on the right that can be used to compare them.

Additionally, we added a simple virtual keyboard composed of 22 traditional Mongolian letters and their corresponding Latin letters to help users to input a Mongolian keyword to benefit free-text search and keyword highlighting (Figure 4).

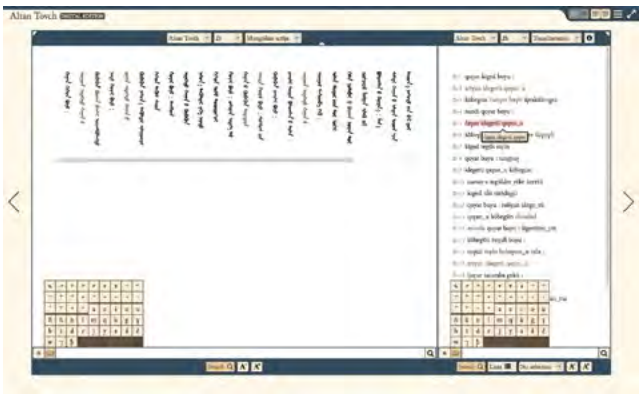


Figure 2. Parallel-text editions with transliteration

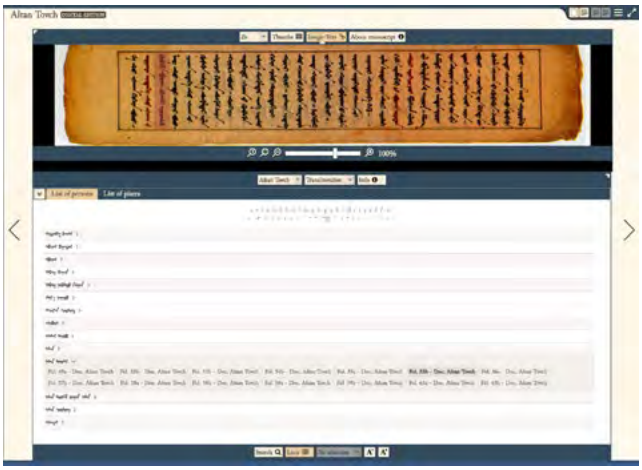


Figure 3. List of personal names in traditional Mongolian script

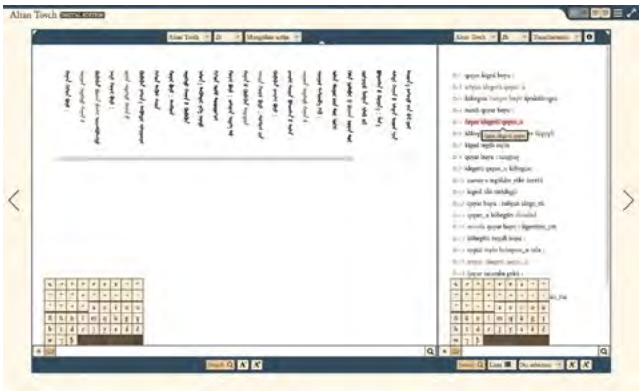


Figure 4. A simple virtual keyboard in parallel-text editions with transliteration

Summary and future directions

In this poster, we discussed our development of creating a digital edition (http://www.dl.is.ritsumei.ac.jp/AltanTovch/) of Mongolian historical manuscripts of the 13-16th century. The proposed method could be applied to other documents in Todo, Manchu, and Sibe, which are the derivative scripts of traditional Mongolian. We will further improve the proposed prototype by adding features to support critical editions.

We believe the proposed digital edition will enable users to search and browse ancient Mongolian manuscript with the highlights of historical figures and ancient place names.

References

Batjargal, B., Khaltarkhuu, G. and Maeda, A. (2016). *Named Entity Extraction from digitized texts of Mongolian Historical Documents in Traditional Mongolian Script, Conference Abstracts of Digital Humanities 2016*, pp. 734-735.

Batjargal, B., Khaltarkhuu, G., Kimura, F. and Maeda, A. (2012). Developing a Digital Library of Historical Records in Traditional Mongolian Script, *International Journal of Digital Library Systems*, 3(1): 33–53.

Choimaa, Sh. (2002). *Qad-un* ündüsün *quriyangyui altan tobči (Textological Study)*. vol. 1. Ulaanbaatar: Centre for Mongol Studies, National University of Mongolia, Urlakh Erdem, 2002. (in Mongolian).

Del Turco R. R., Buomprisco G., Pietro C. D., Kenny J., Masotti R., and Pugliese J. (2014) Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative*, Issue 8, DOI: 10.4000/jtei.1077.

Sečenbagatur Q., Tuyag-a B., and Ying U. (2005). *Monggul kelen-ü nutug-un ayalgun-u sinǰilel-ün uduridqal,* Hohhot: Öbür Monggul-un arad-un keblel-ün qoriy-a.

Shagdarsuren, T. (2011). *Study of Mongolian Scripts (Graphic Study of Grammatology)*, National University of Mongolia, Ulaanbaatar: Urlakh Erdem Khevleliin Gazar.

Soualah M. O., and Hassoun M. (2012). A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts. *Journal of the Text Encoding Initiative*, Issue 2, DOI: 10.4000/jtei.398.

Svantesson J., Tsendina A., Karlsson A., and Franzén V. (2005). The Phonology of Mongolian, New York: Oxford University Press.

Shedding Light on Indigenous Knowledge Concepts and World Perception through Visual Analysis

Alejandro Benito
abenito@usal.es
University of Salamanca, Spain

Amelie Dorn
amelie.dorn@oeaw.ac.at
Austrian Centre for Digital Humanities Austrian Academy of Sciences, Austria

Roberto Therón
theron@usal.es
University of Salamanca, Spain

Eveline Wandl-Vogt
eveline.wandl-vogt@oeaw.ac.at
Austrian Centre for Digital Humanities Austrian Academy of Sciences, Austria

Antonio Losada
alosada@usal.es
University of Salamanca, Spain

The way we conceptualise our world is dependent on various aspects, differing with culture, time and language, and may even be subject to change over the years [5,6]. In this paper, we introduce a visual analysis tool that supports the exploration of indigenous knowledge concepts of a historic language collection, the Database of Bavarian dialects in Austria (DBÖ, dboe@ema), originally and partially collected by means of systematic questionnaires in the area of the former Austro-Hungarian empire. The collection we focus on in this work consists of 109 (original-conceptual) and 9 (supplementary) questionnaires, designed between 1913 and 1920, with answers (about 5 million paper slips). Around 11.100 persons of regional importance with various professional backgrounds and different roles in the compilation process were involved for almost a century [further info c.f. 1,8].

Our tool results from a series of iterations [3] of a custom-made, agile and collaborative workflow inspired by work from other authors [4] that was especially designed for the Digital Humanities (DH). The workflow places data visualisation as the main dialogue facilitator between the different stakeholders participating in the project. By applying user-centered design [2] techniques such as design probes [7], we can direct the development of several micro-prototypes towards the answering of fine-grained research questions. This prototype comprises the results of a full iteration of this iterative and incremental software development cycle.

Attending to the technical aspect of our approach, we employ different distant reading techniques to provide the user with a realistic view of the contents of the questionnaire and with visual mechanisms to help her form a mental image of the cultural connections of the terms at the time the questionnaires were made.

Our visualization plays with lights, colours and shadows to display related concepts, a relationship that is obtained by analysing coincident terms in the questions: the more times two or more terms appear together, the more important they all look in the visualization. The main visual component of our pilot tool is an adjacency matrix tweaked to meet the needs of the multivariate analysis task at hand. This matrix represents one single questionnaire of the collection and its rows and columns the questions conforming it. Each cell is colored to show the number of different concepts two questions have in common (richer coincidences are coloured in darker colours), forming different visual patterns that inform the user about the general distribution and importance of the concepts across the questionnaire.

The main matrix view is escorted by two other views placed on its right and at the bottom respectively: The first one offers an overview of the individual concepts in the questionnaire attending to the number of times they appear, each one represented by a coloured circle. Less frequent (and therefore, less important in our approach) concepts are moved to the top of the visualization, whereas the more important ones are placed at the bottom. Whenever the user hovers over one element, the cells in which that concept appears are in turn highlighted in an effect that imitates refraction of light, allowing for a rapid identification of particularities in the exploration process. At the bottom, the specific concept associations can be found in a similar way. More populated associations appear bigger in the visualization, whereas the more common are placed to the left. We provide an example below related to the use of colour terms:

Although thematically restricted to a single questionnaire (Q53), colours occur in questions throughout the entire collection offering valuable insights on their connection to cultural concepts. Within a single questionnaire, concept patterns/groupings across questions are revealed (see Figure 1). Interestingly, in the case of Q53 the most frequently occurring colour term *bleich* (pale) groups across questions towards the end of the questionnaire.

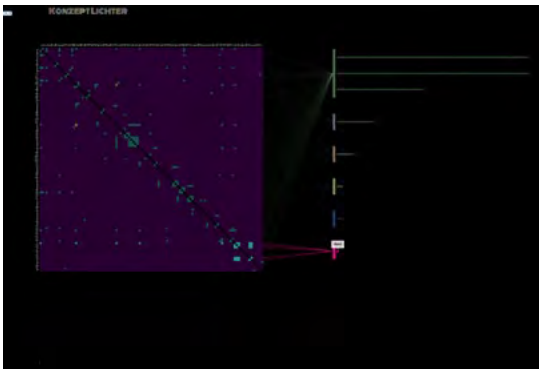


Figure 1: Visual distribution of 'bleich' (pale) grouped across questions in questionnaire 53.

Additionally, yellow (*gelb*) is the term/concept occurring most frequently across questions in questionnaire 85, thus playing an important role in the description of "The flora of our meadows / *Die Pflanzenwelt unserer Fluren* " (Q85) (see Fig. 2). Further, frequent collocations of colour terms in questions are revealed, which also shed light on the structuring of language and part of the conceptualisation of certain topics (see Fig. 3).

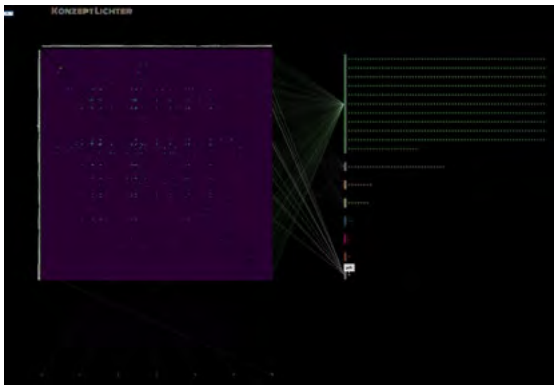


Figure 2: Distribution of 'gelb' (yellow) across questions in questionnaire 85.

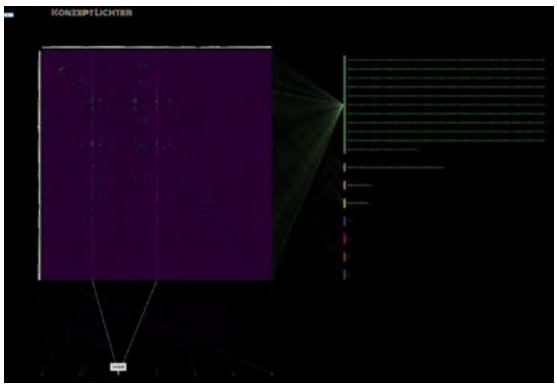


Figure 3: Visualisation of co-occurrence of terms 'rot-gelb' (red-yellow) across questions in questionnaire 85.

Note: Note: Preview of the prototype: https://concept-lights.herokuapp.com/ (Google Chrome only). Please share your remarks with us at explore@oeaw.ac.at. Thanks.

Data:

Datenbank der bairischen Mundarten in Österreich (DBÖ) | Database of Bavarian Dialects in Austria (DBÖ). Austrian Academy of Sciences: 11.2017.

Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema) | Database of Bavarian Dialects in Austria electronically mapped (dbo@ema). Ed. by Eveline Wandl-Vogt: Austrian Academy of Sciences: 2012 / 11.2017.

References

Abgaz, Yalemisew, et al.: "A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis." P*roceedings of W23 - 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, LREC 2018*, 21-29. http://lrec-conf.org/workshops/lrec2018/W23/index.html [last accessed: 26.04.2018]

Abras, C., Maloney-Krichmar, D. and Preece, J., 2004. User-centered design. Bainbridge, W. Encyclopedia of Human-Computer Interaction. *Thousand Oaks: Sage Publications*, 37(4), pp.445-456.

Benito, A., Therón, R., Losada, A., Wandl-Vogt, E. and Dorn, A., Exploring Lemma Interconnections in Historical Dictionaries. *2nd Workshop on Visualization for the Digital Humanities*. October 2017 - Phoenix, Arizona, USA.

Bernard, J., Daberkow, D., Fellner, D., Fischer, K., Koepler, O., Kohlhammer, J., Runnwerth, M., Ruppert, T., Schreck, T. and Sens, I., 2015. VisInfo: a digital library system for time series research data based on exploratory search—a user-centered design approach. *International Journal on Digital Libraries*, 16(1), pp.37-59.

'Concepts of the World': Publishing in Mexico's Indigenous Languages. https://publishingperspectives.com/2017/08/mexico-indigenous-language-publishers/ [last accessed: 26.04.2018]

De Beule, J. and De Vylder, B., 2005, January. Does language shape the way we conceptualize the world?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 27, No. 27).

Gaver, B., Dunne, T. and Pacenti, E., 1999. Design: cultural probes. *interactions*, 6(1), pp.21-29.

Wandl-Vogt, Eveline. "…wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (mit 10 Abbildungen)" P. Ernst (Ed.), Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20. – 23. September 2006. Wien: 2008. Praesens, (pp. 93–112).

The CLiGS Textbox

José Calvo Tello
jose.calvo@uni-wuerzburg.de
Universität Würzburg, Germany

Ulrike Henny-Krahmer
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Germany

Christof Schöch
schoech@uni-trier.de
Universität Trier, Germany

Katrin Betz
katrin.betz@uni-wuerzburg.de
Universität Würzburg, Germany

Introduction

This poster presents the textbox published by the CLiGS group both from the perspective of creating the textbox and of using it for research. The poster will highlight key aspects of the manner in which the collections of literary texts included in the textbox have been compiled, annotated and published. Furthermore it suggests several ways in which the text collections can be used for research in literary studies. This poster aims to showcase the unique XML-TEI-based collections we make available and to encourage their re-use by others.

What is the textbox?

The CLiGS textbox is dedicated to making collections of literary texts in Romance languages freely available. It currently contains novels, novellas and short stories published between 1830 and 1940 in France , Spain, Italy, Portugal, and Spanish-America as well as plays published between 1640 and 1680 in France with a total of 357 texts or about 14 million words. The texts are published in XML-TEI as well as in plain text versions and include detailed document-level metadata. All texts are in the public domain and the XML-TEI markup including the metadata is published with a Creative Commons Attribution license (CC-BY) or in case of the Italian novels with a NC-SA-BY. The text collections are curated and published using a public GitHub repository. In addition, main releases are automatically archived on Zenodo.org, a long-term data and publications archiving service for researchers across Europe managed by OpenAire and supported by CERN (see Nielsen, 2013). Each release receives a DOI (Digital Object Identifier), providing the unambiguous identification and long-term availability of the resource.

Text selection

The individual text collections were created with various usage scenarios in mind, and each collection has been compiled in a slightly different manner. For example, the two collections of Spanish novels, the *Corpus of Spanish Novels (1880-1940)* and the *Collection of 19th century Spanish-American Novels (1880-1916)*, have been prepared to be used for authorship attribution. Accordingly, the two collections have been balanced with regard to the number of texts from different authors. The poster will give an overview of the sub-collections of the textbox and also about the principles guiding their compilations.

File Formats

Independently of their original source format (e.g. html or EPUB), the texts are prepared (with Python scripts or XSLT) according to a common TEI schema established by the CLiGS group . In addition to that reference format, each collection is made available in a simple plain text format automatically derived from the XML-TEI version, containing only the text included in the body of the novels and plays (in particular, excluding prefaces, other paratext, or notes) and with external metadata provided in tabular format.

Moreover, the collections of French, Spanish, Spanish-American, and Portuguese novels as well as the Italian short stories are made available in a version combining basic structural markup (chapter and sentence divisions) with token-level linguistic annotation, including lemma, part-of-speech, morphology, and basic semantic annotation using Freeling (cf. Padró and Stanislovsky, 2012) and WordNet (see Figure 1). Finally, the collection of French plays is not only available in XML-TEI, but also in the "Zwischenformat" developed by the DLINA group (Kampkaspar et al., 2015).

```
<s>
    <w form="Temía" lemma="temer" tag="VMII3S0" ctag="VMI" pos="verb" type="main"
     mood="indicative" tense="imperfect" person="3" num="singular" wnsyn="01780202-v"
     wnlex="verb.emotion">Temía</w>
    <w form="un" lemma="uno" tag="DI0MS0" ctag="DI" pos="determiner" type="indefinite"
     gen="masculine" num="singular" wnsyn="xxx" wnlex="xxx">un</w>
    <w form="despertar" lemma="despertar" tag="NCMS000" ctag="NC" pos="noun" type="common"
     gen="masculine" num="singular" wnsyn="05678745-n" wnlex="noun.cognition">despertar</w>
    <w form="lúgubre" lemma="lúgubre" tag="AQ0CS00" ctag="AQ" pos="adjective"
     type="qualificative" gen="common" num="singular" wnsyn="xxx" wnlex="xxx">lúgubre</w>
    <w form="." lemma="." tag="Fp" ctag="Fp" pos="punctuation" type="period" wnsyn="xxx"
     wnlex="xxx">.</w>
</s>
```

Linguistic annotations in an XML format that is a minimal departure from the TEI standard to allow multiple token-level annotations

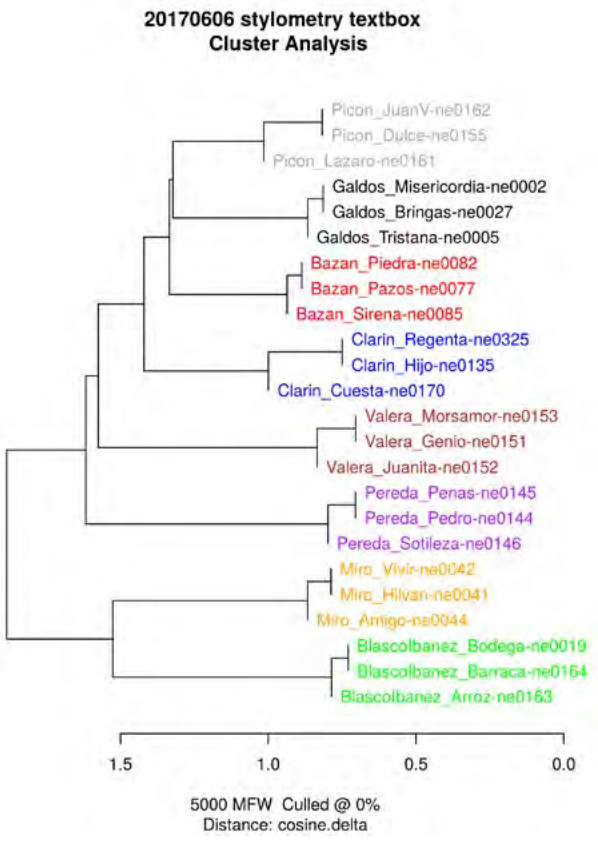
Metadata

Besides the administrative metadata like license, responsibility etc. the collections focus on descriptive metadata. There are four main areas about which information is documented: metadata concerning the authorship (VIAF, name, country, gender), metadata concerning the literary

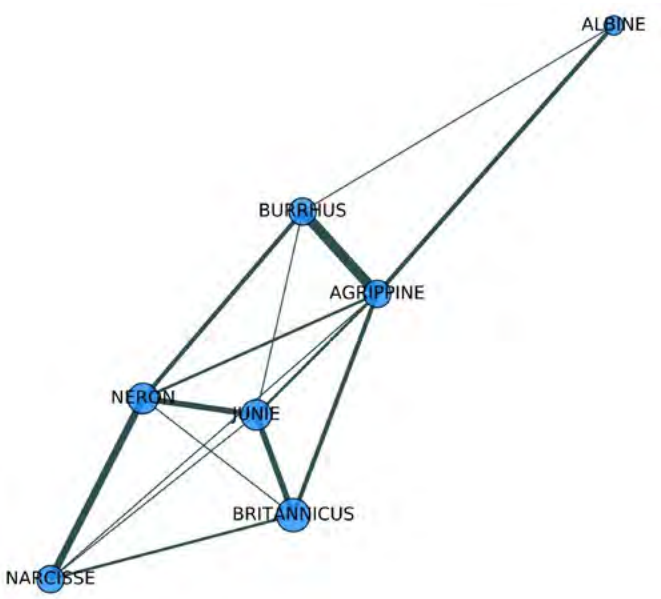
work and editions (VIAF or other identifier, extent of the texts, print and the digital source), and finally metadata concerning the genre: Since the main focus of the project is literary genre, a considerable part of the metadata is directly connected to it. Any reference to genre in the title of the work is collected as a genre label. Besides that, a hierarchical system is used, comprising supergenre (e.g. "narrative" or "drama"), genre (that is, novels or novellas), subgenre (the subtype of the novel, for example "adventure novel" or "political novel") and subsubgenre (optional, used for further differentiations like "war novel").

Usage Scenarios

There are many possible use cases for the textbox collections. The poster will demonstrate some results of these methods from the areas of authorship attribution (using the stylo package for R; Eder et al., 2016), network analysis (using NetworkX in Python), and topic modeling (using MALLET with "tmw" for Python). These scenarios are intended not only as examples of analyses conducted within the CLiGS group, but also as suggestions for potential users of the CliGS textbox, Figure 2 and 3 demonstrate some results for authorship attribution and network analysis.



Authorship attribution, results of cosine delta on the Corpus of Spanish Novels (cf. Smith and Alridge, 2011; Evert et al., 2017)



Character network based on number of words spoken in mutual presence (represented by the thickness of the lines), for Jean Racine's tragedy Britannicus (1669)

References

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry in R: A package for computational text analysis. In *The R Journal*, 16 (1): 1-15.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. In *Digital Scholarship in the Humanities*, 32 (suppl_2): ii4-ii16. doi: 10.1093/llc/fqx023 https://academic.oup.com/dsh/article/doi/10.1093/llc/fqx023/3865676/Understanding-and-explaining-Delta-measures-for (accessed April 26 2018).

Kampkaspar, D., Fischer, F. and Trilcke, P. (2015). Introducing Our 'Zwischenformat'. In *Network Analysis of Dramatic Texts*. https://dlina.github.io/Introducing-Our-Zwischenformat/ (accessed April 26 2018).

Nielsen, L. H. (2013). ZENODO – An innovative service for sharing all research outputs. In *Zenodo*. doi: 10.5281/zenodo.6815 http://doi.org/10.5281/zenodo.6815 (accessed April 26 2018).

Padró, L. and Stanislovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf (accessed April 26 2018): 2473-2479.

Smith, P. W. H. and Alridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. In *Journal of Quantitative Linguistics*, 18(1): 63-88. doi: 10.1080/09296174.2011.533591.

CITE Exchange Format (CEX): Simple, plain-text interchange of heterogenous datasets

Christopher William Blackwell
christopher.blackwell@furman.edu
Furman University, United States of America

Thomas Köntges
thomas.koentges@gmail.com
University of Leipzig

Neel Smith
nsmith@holycross.edu
The College of the Holy Cross, United States of America

Introduction: Sharing text libraries and data collections for teaching and research in the humanities

Source text collections and other complex datasets can be very difficult to share and reuse, and especially difficult to aggregate and disaggregate. CEX, CITE Exchange is a plain-text, self-documenting, technology-agnostic format for capturing citable texts, data collections, and arbitrary relationships among citable data at any level of granularity. CEX is based on the CITE/CTS architecture[1] and it positions itself as an alternative and complement to TEI XML and relational database schemas. TEI XML is a great archival format for storing textual information and metadata of individual editions. Managing and sharing text collections, however, can be cumbersome, especially if you only want to share a collection of excerpts based on hundreds of individual TEI XML files. When teaching text-heavy humanities disciplines, such as history, literature or classics, scholars are constantly faced with the problem of creating a source-text collections (that is, a corpus of excerpts of a bigger corpus that is deemed a representative sample able to answer a scholarly investigation), and the challenge of easily sharing this newly generated collection with students and colleagues. Based on current forms of data exchange, scholars and their students facing this task needed to have intimate knowledge of either database solutions like eXistDB[2] or of API-calls[3]. CEX circumvents this problem by simplifying the format of exchanging texts and related objects following the OCHO2 principles laid out in the CITE/CTS architecture[4].

Likewise, data collections (coins, geo-spatial data, manuscript folios, etc.) are efficiently served intact by relational databases. Extracting subsets, sharing datasets in whole or in part, and aggregating disparate collections with schemas can be very difficult. CEX, as an exchange format, simplifies this.

This paper is directed to two types of scholars: technology-savvy colleagues who want to discuss simple interchange formats for data-sets and colleagues who want to build, analyze, and exchange soure text collections with fellow researchers and students. The paper will introduce CEX, its design, utilities, and code libraries for creating, validating, and manipulating it, and examples of two types of end-user applications: applications that help to build CEX collections and applications that enable students and scholars to perform natural language processing on exchanged CEX collections. In the first part of the paper we will describe the format and structure of CEX, while the second part showcases sample applications.

The CEX format

CEX is based on clearly defined data models for texts and data collections. These data models define semantics of scholarly primitives. CITE and CTS URN citations capture the semantics of the objects they identify. CEX defines catalogs documenting repositories of texts and collections, and blocks of data capturing the data of the texts and collections themselves.

A CEX file is a plain text, UTF-8 file, containing blocks for distinct types of data. The CEX specification provides blocks for:

- Text Catalogs
- Textual Data
- Collection Catalogs
  - Collection Property Definitions
  - Collection Data
- Extensions to Collections, e.g. "Image Collections"
- Relations among citable objects
- Data models formally specifying further aggregation of primitives

Each text-block consists of a header line, followed by data records. Each line is a record, and fields within the lines are separated by a delimiting character ("#" is the default, but this is configurable).

Blocks are optional. A CEX file may contain only textual data, only collection data, or a combination of these. We will demonstrate using CEX files that contain millions of words of textual data and hundreds of thousands of data-records for collections.

In this paper, we will present these blocks, and the clearly defined abstract generic data models that they implement. ## Sample applications

We will demonstrate a sampling of utilities, services, and applications for creating, valdating, browsing, and analyzing scholarly data from CEX-formatted text files. All of these are openly licensed, with source code freely available on GitHub.

1 http://cite-architecture.github.io.
2 http://exist-db.org/exist/apps/homepage/index.html.
3 http://capitains.org.
4 http://cite-architecture.github.io.

Corpus building software: Brucheion and TEItoCEX

TEItoCEX[5] is a CTS TEI XML parser. It is written in Go and can be compiled to any operating system, including OS for Raspberry Pi [6]. It can be copied to the parent folder of multiple TEI CTS XML files. Once executed it will produce a valid CEX collection from all the XML files. **screenshot1**

Brucheion is a browser-based transcription desk written in Go and based on ICT2. **screenshot2** While the backend is based on the fast BoltDB, users can export and import CEX files to share transcriptions or feed them into analytical applications, such as (Meletē)ToPān v.0.3. **screenshot3 and screenshot4**

Microservices based on CEX: Citemicroservices written in Go and Scala

There are currently two implementations of CEX-based microservices. The first implementation we created is a prototype that helps the end-user retrieve textual nodes from CEX files via an HTTP API by simply using the barebone file system of the server as a backend.[7] It is written in Go and can be compiled to any operating system. The gitHub repository also includes executables for all major operating systems and documentation for usage.[8] A useful feature is that one can also retrieve textual nodes from remote CEX repositories by modifying config.json. Simplicity of implementation and usage was the driving force in developing this prototype. The work on this prototype spawned another implementation in Scala that could make use of all the features of the already existing Scala CTS/CITE library.[9]

Another implementation written in Go is currently under development. This implementation includes features of the Scala library and some additional natural language processing capabilities and that make use of BoltDB in its backend.[10]

Exploring CEX collections through Topic Modelling: (Meletē)ToPān v.0.3 and Metallō

(Meletē)ToPān is a GUI for topic modelling introduced at DH2016 in Krakow. It was written in R and is published on Zenodo[11] and Github[12]. Initially ToPān supported the predecessor of CEX, 82XF, which was developed in discussion at DH2016.[13] Since v.0.3, it also supports CEX. Using the CEX files for topic modelling gives users easy but precise control over which documents should be included in the analysis. It has become possible for any scholar to build a source text corpus with TEItoCEX and then feed it into ToPān. The researchers can then use Metallō[14], a commandline based topic model exploration tool, to have a closer look at the created result.

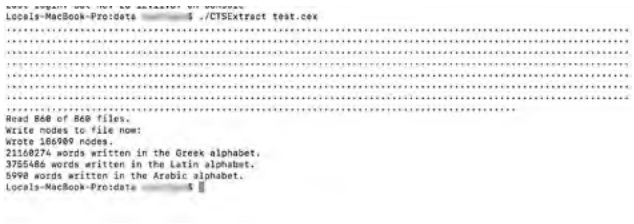
The paper will include a live demonstration of TEItoCEX, ToPān, and Metallō. Scholars visiting this talk will be able to use and build CEX collection for analysis, teaching, and research.

Browsing Integrated Text & Data Libraries with the CiteApp

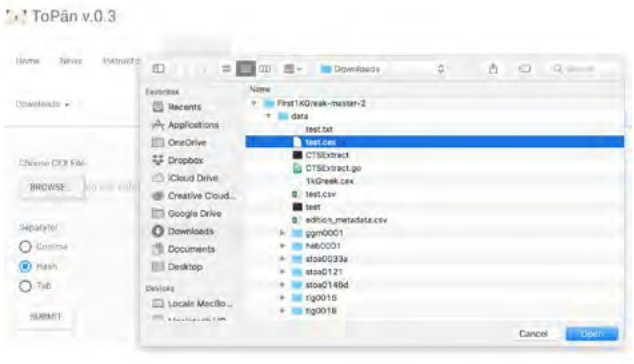
CiteApp is a "zero-config" web application written in ScalaJS that allows browsing and searching of libraries of text, collection data, and images. CiteApp itself consists of a single HTML file and can run locally or on a server. CiteApp reads data in the form of CEX files (with, optionally, an accompanying directory of image files). CiteApp allows browsing texts in a library, browsing parallel texts and translations, querying texts for strings, tokens, and NGrams, browsing data-collections and querying their records, and browsing collections of images, querying metadata, and interacting with zoomable views of high resolution images. Images may be local or hosted remotely.[15]

The audience for this talk will be able to take home a copy of CiteApp with sample libraries in CEX format of byzantine manuscript images and editions of the texts they contain.

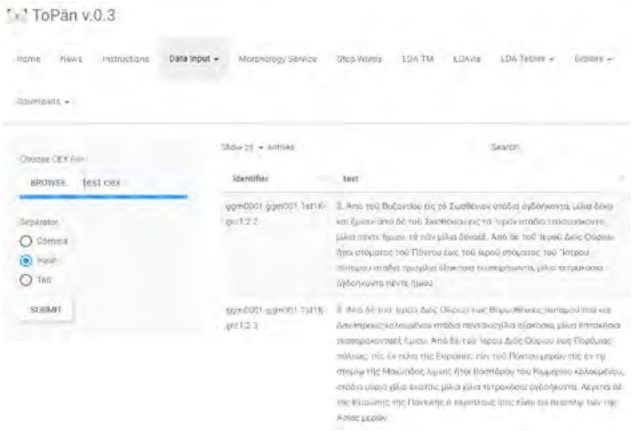
Screenshots of CEX-Capable Applications



TEItoCEX: Converting libraries of TEI documents to easily shared CEX files

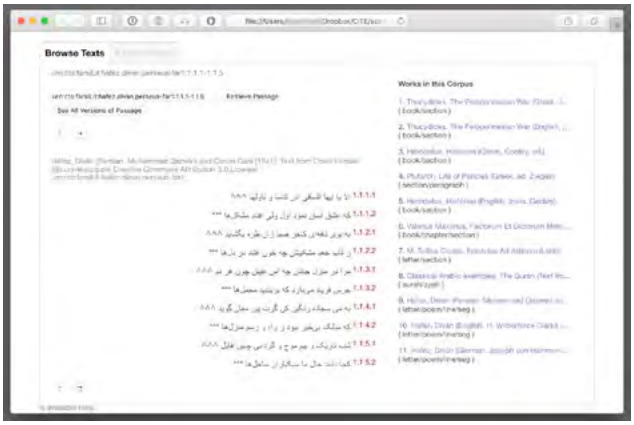


5 https://github.com/ThomasK81/TEItoCEX.
6 https://www.thepolyglotdeveloper.com/2017/04/cross-compiling-golang-applications-raspberry-pi/.
7 https://github.com/cite-architecture/citemicroservices.
8 https://github.com/cite-architecture/citemicroservices.
9 https://github.com/cite-architecture/scs-akka.
10 https://github.com/ThomasK81/Nucleus.
11 https://zenodo.org/record/556708#.WhhosLbMwll.
12 https://github.com/ThomasK81/ToPan.
13 http://neelsmith.github.io/2016/07/13/82xf_v2/.
14 https://github.com/ThomasK81/TopanTerminalBrowser.
15 https://github.com/cite-architecture/CITE-App.

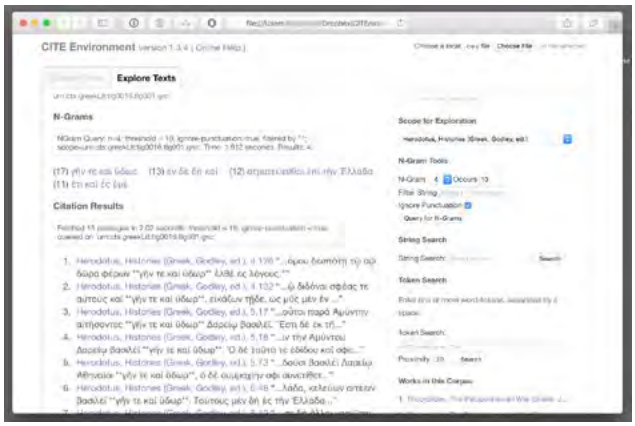


TEItoCEX
(Meletē)ToPān: Topic Modeling files in CEX format

Brucheion: Integrated Image and Textual data



Brucheion
CiteApp: Browsing a multilingual text library
CiteApp
CiteApp: Searching for NGramsTEItoCEX



CiteApp

References

Smith, D. Neel, and Gabriel Weaver. "Applying Domain Knowledge from Structured Citation Formats to Text and Data Mining: Examples Using the CITE Architecture." Text Mining Services, 2009, 129.

Blackwell, C., and D.N. Smith. "A Gentle Introduction to CTS & CITE URNs." Homer Multitext Project Documentation, November 2012. http://www.homermultitext.org/hmt-doc/guides/urn-gentle-intro.html.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. "What Is Text, Really?" ACM SIGDOC Asterisk Journal of Computer Documentation 21, no. 3 (August 1997): 1–24. https://doi.org/10.1145/264842.264843.

Digitizing Whiteness: Systemic Inequality in Community Digital Archives

Monica Kristin Blair
mkb4rf@virginia.edu
University of Virginia, United States of America

Introduction

In recent years, the digital revolution has transformed the idea of the archive. Once associated with grand library buildings filled with ancient books and artifacts, today scholars are making archives out of social media, metadata, and all things born digital. Traditional archives are also revolutionizing the way that users interact with their objects by taking digitization to the next level with techniques like linked data and photogrammetry.

However, archivists and scholars are not the only ones experimenting with digital curating. Online communities are making their own virtual collections. Librarians and digital humanists, including those at the United States Library of Congress, have encouraged and assisted people

in creating these "community digital archives" (LeFurgy). But how should scholars respond when these community digital archives are linked to institutions with extremely fraught histories of white supremacy, ableism, homophobia, transphobia, xenophobia, or sexism? This study explores that broad question by analyzing community digital archives created by alumni of historically segregated K-12 private schools in Virginia, USA to investigate the form, function, and ethics of studying community digital archives attached to historically prejudiced institutions.

Research questions

How does institutional inequality manifest itself in digital community archives?

How should scholars read community digital archives that are public, but may not intended for outside audiences?

What commonalities and differences exist between these repositories, traditional archives, and digital archives curated by scholars and archivists? How do those similarities and differences affect how scholars should interact with these communities?

What are the ethics of using these archives for scholarly research?

Literature

Scholars like Bergis Jules and Piia Varis have worked to define the ethics and best practices of archiving digital sources (Jules, 2016; Varis, 2014). Moreover, several digital projects such as DocNow and Take Back the Archives have modeled how scholars can engage with digital archiving methods to advance scholarly questions and social justice. Most of this scholarship has focused on archiving the experiences and activism of marginalized groups. That work is both vital and admirable. This study contributes to this literature by examining the opposite end of the spectrum. By looking at how white communities that have supported segregated education use community digital archives, I illuminate how these groups remake and reaffirm systemic inequality in the digital landscape. In the process, I also examine the ethics of analyzing and writing about digital communities that have white supremacist roots.

Case study

This study uses the historical subject of segregation academies as its basis. Segregation academies are private schools founded in the southern United States during the 1950s and 1960s in order to provide segregated education for white students whose families refused to comply with court-ordered school desegregation following the United States Supreme Court case Brown v. Board of Education.

White supremacy was at the heart of these schools' foundation, and this study examines how whiteness is reflected in the digital community archives of these schools' alumni pages on Facebook and Classmates.com. Both Classmates and Facebook are for-profit businesses, but it is former students, teachers, and administrators who post old photographs, pamphlets, yearbooks, and personal memories of their times at these intuitions on these websites. The memorabilia they gather and publish serves as an important window into the past, and their contemporary comments reveal the ways that white southerners navigate their personal ties to this history of white supremacy in the contemporary digital landscape.

Facebook and Classmates do not follow the practices of traditional archives as outlined by archivist Kate Theimer (Theimer, 2012); nonetheless, both platforms exhibit some of the classic characteristics of archives. They are repositories for institutional and personal histories. Donors contribute to these archives by providing digital copies of their personal papers, photographs, and yearbooks. The aim of these groups is to preserve the history of an institution and, in doing so, craft historical narratives about said institutions. Ultimately, the content, organization, and narratives on these websites are fundamentally shaped by the motive of the curators of these archives. The patrons who create these archives do so out of sentimentality about their former-schools. The web hosts, Classmates and Facebook, profit from this nostalgia, and thus have no incentive to challenge the whitewashed school histories their users promote. Thus, sanitized, color-blind versions of these schools histories prevail on these digital community archives, thereby erasing decades of systemic inequality and prejudice from view.

References

Documenting the Now. http://www.docnow.io (accessed 29 November 2017).

Jules, Burgis. (2016). Some Thoughts on Ethics and DocNow. *Medium*. https://news.docnow.io/some-thoughts-on-ethics-and-docnow-d19cfec427f2 (accessed 29 November 2017).

LeFurgy, Bill. (2013). Resources for Community Digital Archives. Library of Congress. https://blogs.loc.gov/thesignal/2013/06/10-resources-for-community-digital-archives (accessed 29 November 2017).

Take Back the Archive. University of Virginia. http://takeback.scholarslab.org (accessed 29 November 2017).

Theimer, Kate. (2012). Archives in Context and as Context. *Journal of Digital Humanities,* Vol. 1, No. 2, Spring 2012. http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/ (accessed 30 January 2018).

Varis, Piia. (2014). Digital Ethnography. *Tilburg Papers in Cultural Studies*. https://www.tilburguniversity.edu/upload/c428e18c-935f-4d12-8afb-652e19899a30_TPCS_104_Varis.pdf (accessed 30 January 2018).

How to create a Website and which Questions you have to answer first

Peggy Bockwinkel
peggy.bockwinkel@ilw.uni-stuttgart.de
University of Stuttgart, Germany

Michael Czechowski
mail@dailysh.it
University of Stuttgart, Germany

Initial situation and target group

Our target group are (digital) humanities scholars who want to present small, non-funded projects on their own website but have little or no experience with web design and development. Always the same questions have to be asked at the beginning of a website project to find out how complex or easy it is to implement the project. These circumstances are excellent conditions to offer a "how to"-flowchart with different applications as decision support. The poster is therefore designed as a decision flowchart and gives an overview of the possibilities and limitations of certain applications. All applications are differentiated according to whether static, dynamic or semi-dynamic websites can be built with them. Depending on which content should be presented and how, even beginners with little or no previous knowledge can create a website or get a feeling for when the support of a web developer is necessary. The questions that are answered in the flowchart concern...

1. … the complexity of the site (see 2.1. for details).
2. … the reason why the website is build, e.g. to publish a digital edition or to present research.
3. … the format of the texts to be published: Do they have a uniform format? Can they be read dynamically?
4. … the level of knowledge / technical affinity of the person creating the website, but also the capacity in terms of time and/or manpower and/or budget, i.e. for example whether it is possible to employ a student assistant or even a professional web developer.
5. … the scalability: is it possible to estimate how the project will develop, e.g. with regard to complexity? To what extent can the website be expanded and where are the limits, e.g. how far can you get with the individual technology stacks?
6. … hosting (see 2.2. for details).
7. … sustainability.

Questionnaire

Of all the relevant questions, two are presented here as examples. Nevertheless, the flowchart is made available on **github** so that as many scholars as possible can benefit from it.

What kind of data should be published?

If it is data that can be copied and pasted onto the website, a static website can be created with the website generator **jekyll**, **omeka** or the content management system **wordpress**. This can look like this, for example: http://www.germanliteratureglobal.com/

If it is data, that requires several tabs and contains recurring queries, it is a dynamic website. This type of website is also required if data is to be made available for downloading (a database is required for larger amounts of data). This can look like this, for example: http://www.berliner-intellektuelle.eu/

If your data is available in TEI format, the **TEIpublisher** or the **EVT** are good choices for publication. If you are moving in the dynamic area, the effort can quickly become very high and extensive knowledge of website development is necessary. In this case you should contact a web designer.

How can the website be published (Hosting)?

If you have chosen **jekyll**, hosting is very easy via **github** or **githubpages**. There are no additional costs. If you have access to a webserver, you can always use it. Often universities offer such server systems for its employees or even students.

Discussion

In the course of this project, we were confronted with various issues that all revolve around sustainability in the broadest sense:

What if the formats presented here are obsolete? This risk exists for any technology application. This project is meant to be a pragmatic guide. We cannot solve the problem, but we use tools that are freely available and extendable.

If a website should be sustainable, i. e. available in the long term, where should it be hosted to guarantee long-term accessibility? Again, there is no guarantee how long the services presented here are available. As far as the sustainability of humanities websites, i. e. cultural knowledge in any form is concerned, we see it as universities, libraries and archives duty to provide and maintain the corresponding infrastructure.

References

Baillot, A. (Ed.). (n.d.). *Briefe und Texte aus dem intellektuellen Berlin um 1800*. Retrieved April 27, 2018, from http://www.berliner-intellektuelle.eu/

Edition Visualization Technology. (2013). Retrieved April, 27, 2018, from http://evt.labcd.unipi.it/

GitHub. (2018). Retrieved April 27, 2018, from https://github.com/

GitHubPages. (2018). Retrieved April 27, 2018, from https://pages.github.com/
Jekyll. (2018). Retrieved April 27, 2018, from https://jekyllrb.com/
Omeka (n.d.). Retrieved April 27, 2018, from https://omeka.org/
Richter, S. (Ed.). (2017). Retrieved April 27, 2018, from http://www.germanliteratureglobal.com/index.php/Hauptseite
TEIpublisher. (n.d.). Retrieved April 27, 2018, from https://teipublisher.com/index.html
Wordpress (n.d.). Retrieved April 27, 2018, from https://wordpress.org/

La Aptitud para Encontrar Patrones y la Producción de Cine Suave (Soft Cinema)

Diego Bonilla
bonilla.diego@gmail.com
California State University, Sacramento, United States of Americas

La computadora e Internet han fomentado cambios significativos en la forma en la que las narrativas cinematográficas son construidas, teniendo un impacto no sólo en los medios digitales que serán recibidos en el propio computador, sino también en los medios análogos tradicionales (Buckland, 2009). El cine hiperliga (o *hyperlink cinema*) se refiere a películas en las que la narrativa no sigue un arco específico, presentando una historia de forma no lineal. El uso del término *hiperliga* proviene de los textos digitales en donde se pueden especificar ciertas palabras como referencias directas a otros textos. En el ámbito hipertextual de la red mundial de computadoras, los lectores desarrollan aptitudes de lectura diferentes a las de los lectores de libros impresos; por ejemplo, los lectores de hipertexto desarrollan la habilidad de encontrar patrones y conexiones en múltiples textos hiperligados (Landow, 1992). El acceso en masa a un medio de comunicación basado en procesos computacionales ayuda a redefinir las formas en la que se conceptualiza el contenido, cómo se lleva a cabo su autoría y cómo éste es recibido por las audiencias. Por lo tanto, la proliferación del cine hiperliga, en el cual los arcos narrativos tradicionales no son seguidos, se puede entender, en parte, como resultado de la adopción de la computadora como medio de comunicación (Buckland, 2009).

El término "cine suave," *o soft cinema,* es un compuesto de las palabras *software* y *cinema,* y se refiere al cine que está basado en principios computacionales (Manovich, 2005). El cine suave, de igual forma que ocurre con la lectura de hipertexto, se refiere a la capacidad de alterar la secuenciación de una narrativa cinematográfica a través de procesos computacionales. A diferencia del cine tradicional o "rígido," el cine suave presenta una narrativa en un gran número de secuencias diferentes determinadas por algoritmos.

Una característica importante del cine suave es que el vidente no tiene que interactuar de forma constante con el computador a lo largo de la presentación de la narrativa audiovisual. Como ocurre con el cine tradicional, el público recibe la obra de forma "pasiva." Esta característica es fundamental para separar a las narrativas del cine suave de las narrativas presentes en los videojuegos. Algunos ejemplos de cine suave son *A Space of Time* (Bonilla, 2003), *Soft Cinema* (Manovich, 2005) y *Accidental Occurrence* (Bonilla, 2017). En el caso de *A Space of Time*, los módulos narrativos se presentan a través de un algoritmo que determina la secuenciación y la longitud de cada versión de la película antes de que ésta sea presentada. *A Space of Time* es similar a *Soft Cinema* ya que en ésta última la narrativa se cuenta con un audio lineal y la secuencia de los elementos visuales es determinada por algoritmos. *Accidental Occurrence* también sigue una serie de algoritmos que establecen la secuenciación de cada versión antes de que la obra pueda ser vista.

Una diferencia sustancial entre *Accidental Occurrence* y las obras de cine suave citadas anteriormente es que ésta ofrece al vidente la capacidad de alterar la forma en la que los algoritmos re-editan la película. En otras palabras, la obra ofrece cierto nivel de interactividad al vidente/usuario al inicio de la experiencia cinematográfica: Se puede alterar la longitud de la película desde un mínimo de 6 minutos hasta un máximo de 70 minutos y se puede dar más énfasis a un personaje que a otro. Los dos tipos de variaciones llegan a ofrecer 9.11E+124 versiones diferentes de la obra; la experiencia de esta variabilidad evoca un sueño recurrente en el que se tiene la misma "vivencia" de una narrativa aunque ésta siempre se lleva a cabo de forma diferente.

La construcción de una narrativa que será presentada como cine suave conlleva un nivel alto de complejidad y requiere de una audiencia acostumbrada a arcos narrativos no tradicionales. Es decir, no sólo la obra debe de ser creada de tal forma que pueda variar, sino también se debe de contar con un público dispuesto a "solucionar" la película tal y como lo hace con el cine hiperliga. En otras palabras, los videntes que están acostumbrados a la no linealidad, ya sea por el uso habitual de la red mundial de computadoras o por frecuentar el cine hiperliga, son más capaces de apreciar las narrativas ofrecidas por el cine suave.

References

Bonilla, D. (2017). *Accidental Occurrence* [Program]. Retrieved November 23, 2017, from https://www.modular.film/ Programa para generar narrativas audiovisuales.

Bonilla, D. (2003). *A Space of Time (CDROM)* Hypergraphia, LLC. URL: http://www.diego.today/a-space-of-time
Buckland, W. (2009). *Puzzle films: complex storytelling in contemporary cinema*. Malden, MA: Wiley-Blackwell.
Landow, G. P. (1992). *Hypertext. The convergence of contemporary critical theory and technology* (p. 134). Baltimore: The John Hopkins University Press.
Manovich, L. (n.d.). Soft Cinema (project description). Retrieved November 23, 2017, from http://manovich.net/index.php/projects/soft-cinema
Manovich, L., & Kratky, A. (2005). *Soft cinema (DVD)*. Cambridge, MA: MIT Press.

Women's Faces and Women's Rights: A Contextual Analysis of Faces Appearing in Time Magazine

Kathleen Patricia Janet Brennan
kpjbrennan@gmail.com
SUNY Polytechnic Institute, United States of America

Vincent Berardi
berardi@chapman.edu
Chapman University, United States of America

Aisha Cornejo
corne129@mail.chapman.edu
Chapman University, United States of America

Carl Bennett
bennetca@sunyit.edu
SUNY Polytechnic Institute, United States of America

John Harlan
harlanj@sunyit.edu
SUNY Polytechnic Institute, United States of America

Ana Jofre
jofrea@sunyit.edu
SUNY Polytechnic Institute, United States of America

We are developing a methodology for exploring and finding meaning in large corpuses that contain images, such as archives of periodic publications. We focus this work on *Time* magazine, and in particular on images of faces in *Time*. We use computer vision analysis, combined with contextual research and methods from the humanities, to elucidate trends and patterns in the visual culture reflected by the publication. In particular, we are examining how representations of the human face have changed over time, and seeking relationships between the visual features we discover and their corresponding socio-political contexts. Specifically, we are interested in gaining insight about how the form and context of representations of women and ethnic minorities have changed over time. Our preliminary research focuses on the correlation between changes in facial representations in *Time* magazine and the Women's Liberation movement in the United States in the 1970s and 1980s. The main outcome of this project will be a meaningful and accessible web-based platform through which both researchers and the general public can explore the archives of *Time* magazine to discover insights into our cultural history. We expect that we will be able to apply our methodology to any periodical publication, but we chose *Time* because it stands as a record of the many pulses of U.S. and world politics and their intersections with American culture. We believe that because it is such a culturally important and ubiquitous publication much can be learned from these archives about how Americans perceived politics and culture throughout the twentieth and early twenty-first centuries.

Our methodology combines computational processes, such as computer vision analysis, with contextual research, such as the history of the magazine's production process, as well as the cultural and political climate in which each issue appears. A brief summary of our methodology is as follows. Using the entire *Time* magazine corpus (about 4800 issues spanning over 93 years), we are identifying and extracting every facial image within the corpus, and running computational analyses on the images to quantify their visual features (such as RGB pixel values). We are building a database of the images that includes their associated metadata (year, issue, page number), as well as the extracted visual feature data. Within this database, we are also including more detailed metadata for each image: the face's gender, race, the context in which the face appears (ad or feature story), whether or not the face is smiling, and whether it is an individual portrait or belongs to an image that contains more than one face. In parallel to building this database, we are developing timelines of significant contextual information, which includes a timeline of the evolution of printing technologies used by *Time* magazine, a timeline of culturally impactful geo-political events, a timeline of civil rights movements, and a timeline of women's movements. Our image database will be connected to our contextual timelines with visual analytics. The visualizations we create will be interactive, web-based, and open to the general public.

We present here compiled preliminary results using our methodology and samples from our private collections of *Time* magazine, along with a contextual timeline of the women's rights movement in the US. In the work presented here, we used human labor to extract face images from sample issues and to tag each face image with the metadata described above. We are using the data harvested through human labor to improve our facial recognition algorithms and to train new algorithms to identify gender, race, smiling, and context. There has been a great deal of interest in sentiment analysis and facial recognition across academia and the general public, and we feel

this project will allow us to examine how these complex categories interact with each other. For example, how do our understandings of race and gender impact how humans classify sentiment? How do these understandings impact algorithmic classifiers? This complexity is one of the primary motivations for developing a methodology that consciously moves back and forth between human and computer analysis.

The metadata extracted by human labor has been particularly insightful, especially when put into the context of our historical timelines. Specifically, we noticed an increase in the number of female faces in the 1970s, coincident with the many milestones in the Women's Rights movement. Interestingly, our preliminary data also suggests that as the number of women represented in the magazine increases, the proportion of women in advertisements decreases. Our poster will focus on a close examination of the data and sociopolitical context of 1965-1990 in order to fully explore this potential correlation. We will also discuss our methodology and include a few examples of our visualizations.

This project aims, not only to gain insights from an analysis of *Time* magazine and to make these insights publicly accessible, but also to establish novel methodologies for the visual analytics of large data sets, particularly of image-based corpuses, which we hope to use for years to come and to share with other researchers.

The ultimate goal of this project is to create a website with contextualized interactive visualizations based on the entire archive. Our initial approach was inspired by Manovich's Selfie-city and Photo-trails work, and by his team's use of direct visualization (Crockett, 2016), which we see as a way to engage broad audiences into complex corpuses. We also draw inspiration from *Robots Reading Vogue* (King and Leonard) and *Neural Neighbors* (Leonard), which are projects based in the Yale University library system. By exploring specific, humanities-based research questions in this early phase of our project we will be able to make meaning and better contextualize the interactive visualizations in the end.

References

Crockett, D. (2016). Direct visualization techniques for the analysis of image data: the slice histogram and the growing entourage plot. *International Journal for Digital Art History*, 0(2) doi:10.11588/dah.2016.2.33529. http://journals.ub.uni-heidelberg.de/index.php/dah/article/view/33529 (accessed 8 November 2016).

King, L., and Leonard, P. (2018). Robots Reading Vogue : Colormetric Space http://dh.library.yale.edu/projects/vogue/colormetricspace/ (accessed 5 January 2017a).

King, L., and Leonard, P. (2018). Robots Reading Vogue http://dh.library.yale.edu/projects/vogue/ (accessed 8 November 2016b).

Lauridsen, H. (2014). What's in Vogue? Tracing the evolution of fashion and culture in the media *Yale News* http://news.yale.edu/2014/09/05/what-s-vogue-tracing-evolution-fashion-and-culture-media (accessed 8 November 2016).

Manovich, L. (2011). Mondrian vs Rothko: footprints and evolution in style space http://lab.softwarestudies.com/2011/06/mondrian-vs-rothko-footprints-and.html (accessed 30 December 2016a).

Manovich, L. (2010). One million manga pages http://lab.softwarestudies.com/2010/11/one-million-manga-pages.html (accessed 30 December 2016b).

Manovich, L., Hochman, N., and Chow, J. (2013). Phototrails: Visualizing 2.3 M Instagram photos from 13 global cities http://lab.culturalanalytics.info/2016/04/phototrails-visualizing-23-m-instagram.html (accessed 30 December 2016).

Rushmeier, H., Pintus, R., Yang, Y., Wong, C. and Li, D. (2015). *Examples of challenges and opportunities in visual analysis in the digital humanities*. vol. 9394. pp. 939414-939414–19 doi:10.1117/12.2083342. http://dx.doi.org/10.1117/12.2083342 (accessed 5 January 2017).

softwarestudies.com (2009). *Timeline: 4535 Time Magazine Covers, 1923-2009*. Photo https://www.flickr.com/photos/culturevis/3951496507/ (accessed 30 December 2016).

Decolonialism and Formal Ontology: Self-critical Conceptual Modelling Practice

George Bruseker
bruseker@ics.forth.gr
Centre for Cultural Informatics,
Institute of Computer Science-FORTH, Greece

Anais Guillem
aguillem@ucmerced.edu
School of Social Sciences, Humanities and Arts, University of California Merced, United States of America

Digital humanists taking up the challenge of the decolonialist approach face, with regards to information management, the question of how to structure their data in a way which escapes the confines of the repressive episteme that they seek to challenge. And yet, the database and the data form have enormous potential to replicate and even intensify, in a new medium, the colonial intersection of knowledge and power. A data model operates, at least potentially, on its 'subject' as an authoritative power, disenfranchising the epistemological constellations of those it chooses to represent and submitting them to a colonial order of knowledge. Such subjugation can be argued to be represented in classic arrangements of knowledge like the 'tombstone' data explaining objects in museums

and archaeological collections. In such data models, the analyses that go along with the object and which tie these objects into a web of knowledge privilege the interpretation of the scholars who speak of and for the object. It is often the agency of the 'discovering' or 'gifting' agent that is most associated to the object over/above the cultures, groups and individuals for whom the object was a living part of life and practice. (saywhatnathan, 2017) The digital humanist would work on a corpus of well-formatted data in order to build up a new knowledge, contesting colonial representations, but the epistemic, ethical and pragmatic challenge comes together here: what can be the form of this representation and how to conceptualize and maintain it, without re-introducing imperialistic paradigms?

In this question, the theoretical and practical interests of decolonialism and the discipline of knowledge engineering / formal ontology overlap and have the chance for a fruitful methodological dialogue. By decolonialist thought we intend the theoretical tendency building up from the post-colonialism of Said (1979) on to the work of Mignolo (2011), Borgstede (2010), and Smith (2012) amongst others. This movement looks to challenge the identification drawn between scientific practices originally developed in the West, meaning the traditions of European scholarship, and a universalist objectivity. The critique is undertaken in order to identify and recognize limits of the Western, scientific project, towards the end of opening a space for the self-articulation of suppressed modes of discourse, so that they can reach expression and be understood as autonomous spaces of potential truth disclosing, as elaborated under non-dominant conceptual paradigms. By knowledge engineering and formal ontology, we intend methods proposed since the 1990s (Gruber, 1995; Guarino, 2997; Smith, 2003) as a means to make better data structures within information systems by engaging in an interdisciplinary practice to build these latter through a disciplined dialogue between computer science, philosophy and the domain practitioners concerned. Established and well known applications of this method are known in the areas of linguistics with DOLCE and cultural heritage CIDOC CRM as described in Gangemi et al. (2002) and Doerr (2003).

The general aim of adopting a formal ontological approach in a research discipline or community is to serve as a means to robustly model data and create more accurate digital representations in a way that creates community consensus around the generic representational form. Within the digital humanities, formal ontology is an important tool to solve the long term data integration and data provenance problems that are correlate to the creation of ever greater datasets by scholars. A formal ontology offers much that the decolonialist digital humanist would need in their toolset. Can it, however, meet their epistemic and ethical requirements?

Here we would argue that decolonialist thinking and well founded formal ontological thinking share fundamental theoretical commitments which are mutually beneficial. The potential for enrichment is two-way, offering a path forward for an information structure suitable to decolonialist studies but also providing to formal ontology research an important control point. In particular, what binds together these two approaches is a shared commitment to a radical and critical approach to known epistemic structures. Both are committed to a self-reflexive critique which does not accept the given epistemic prejudice of the 'form' of scientificity but rather aims to critique it in order to understand a wider form. This is expressed in a radical empiricism in the sense of a deliberate openness to understanding from practice rather than from the formalisms of science. What is to be modelled is not what is said but what is done. This commitment on the part of formal ontology leaves the final model of information representation always open to modification. The work of the decolonialist scholar brings material that can continually challenge the prejudice in a model and cause its redesign. On the other hand, the open ended design of the formal ontological model which does not follow the logic of the data form but of an open ended graph of knowledge, allows for the representation of multiple perspectives and the multi-participation of objects in different epistemological formations.

An illustration of this self-reflexive and openly critical practice in action can be taken from the modelling of 'discovery' activities in the CIDOC CRM extension, CRMsci. (Doerr et al., 2017) Classic data representation and inbuilt cultural prejudice would offer the 'intuitive' category of 'discovery' to describe scientific observation activities such as ethnography, archaeology, botany and so on. Such categorizations, however, are one-sided and privilege the 'discoverer' while decentering and subjecting the 'discovered'. Extensive, long-term dialogue and conversation over this issue, led to the elaboration of a general class of the ontology called 'Encounter'. 'Encounter' avoids one-sidedness of representation and the implication that something comes to be known through the encounter event. It shifts the representation to a third party point-of-view, and allows modelling the fact that some group met some thing. This encounter finds an object and may produce new knowledge, for the group that has initiated an encounter activity, but not as such.

The intersection of decolonialist thought and knowledge engineering in the practice of digital humanism offers the opportunity to lift the tombstone off cultural knowledge and open it to expression and contention with the dominant episteme by means of the construction of open graphs of knowledge that empower the representation, reconstruction and expression of suppressed knowledge by the actors from whom it originates.

References

Borgstede, G., Cipolla, C. N., Gullapalli, P., Lilley, I., Jiménez, J. R. P., Patterson, T. C., Preucel, R. W., et al. (2010). *Archaeology and the Postcolonial Critique*. (Ed.) Liebmann, M. & Rizvi, U. Z. Reprint edition. AltaMira Press.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3): 75.

Doerr, M., Kritsotaki, A., Rousakis, Y., Hiebel, G. and Theodoridou, M. (2017). *Definition of the CRMsci: An Extension of CIDOC-CRM to Support Scientific Observation*. Technical Report Crete: ICS-FORTH.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. and Schneider, L. (2002). Sweetening Ontologies with DOLCE. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. (Lecture Notes in Computer Science). Springer, Berlin, Heidelberg, pp. 166–81 doi:10.1007/3-540-45810-7_18. https://link.springer.com/chapter/10.1007/3-540-45810-7_18 (accessed 25 April 2018).

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. *International Journal of Human-Computer Studies*, 43(5): 907–928.

Guarino, N. (1997). Understanding, building and using ontologies. *International Journal of Human-Computer Studies*, 46(2): 293–310.

Mignolo, W. D. (2011). *The Darker Side of Western Modernity: Global Futures, Decolonial Options*. edition. Duke University Press Books.

Said, E. W. (1979). *Orientalism*. 1st Vintage Books ed edition. New York: Vintage.

saywhatnathan (2017). Maker unknown and the decentring First Nations People *Archival Decolonist [-O-]* https://archivaldecolonist.com/2017/07/21/maker-unknown-and-the-decentring-first-nations-people/ (accessed 25 April 2018).

Smith, B. (2003). Ontology. In Floridi, L. (ed), *Blackwell Guide to the Philosophy of Computing and Information*. Oxford: Blackwell, pp. 155–166.

Smith, L. T. (2012). *Decolonizing Methodologies: Research and Indigenous Peoples*. 2 edition. London: Zed Books.

Rules against the Machine: Building Bridges from Text to Metadata

José Calvo Tello
jose.calvo@uni-wuerzburg.de
University of Würzburg, Germany

Introduction

Digital literary studies advance in their research, requiring more specific metadata about literary phenomena: narrator (Hoover 2004), characters (Kastorp et al. 2015), place and period, etcetera. This metadata can be used to explain results in tasks like authorship attribution or genre detection, or to evaluate digital methods (Calvo Tello 2017). What could be the most efficient way to start annotating this information in corpora of thousand of texts in languages, genres and historical periods for which many NLP tools are not trained for? In this proposal, the aim is to identify specific literary metadata about entire texts with methods that are either language-independent or easily adaptable for humanists.

Two Ways from Text to Metadata

The two approaches to classify unlabeled samples applied here are rule-based classification and supervised machine learning. In rule-based classification (Witten et al. 2011), domain experts define formalised rules that correctly classify the samples. For example a rule based on a single token can be defined for each class to predict whether a text is written in third person (83% of the corpus) or first person using tokens for the two values are the Spanish words *dije* ('I said') and *dijo* ('he said'), and the rule:

1. if *dijo* appears 90% more than *dije*, the novel is written in third person
2. if *dijo* appears less, in first person

The results of applying this rule can be presented as a confusion matrix:

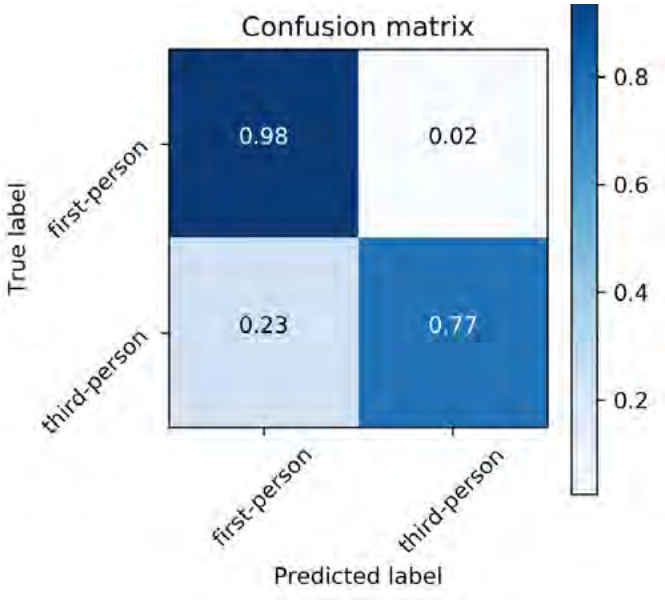


Fig 1. Confusion Matrix of rule-based results about narrator

For supervised methods (Müller and Guido 2016; VanderPlas 2016), we need labeled samples to train and

evaluate the method. In the following table, the different classifiers and document-representations achieve different accuracy scores:

	raw	relative	tfidf	zscores
SVC	0.90	0.83	0.83	0.88
KNN	0.83	0.88	0.81	0.81
RF	0.88	0.88	0.86	0.90
DT	0.84	0.83	0.84	0.82
LR	0.88	0.83	0.83	0.17
BN	0.72	0.72	0.72	0.82
GN	0.72	0.80	0.80	0.81

Fig 2. Accuracy (F1-score) for narrator

Corpus and Metadata

The data is part of the *Corpus of Spanish Novels of the Silver Age (1880-1939)* (used in Calvo Tello et al. 2017), with 350 novels in XML-TEI by 58 authors. Each text has been annotated manually with metadata and its degree of certainty has been assigned. 262 texts with either high or medium certainty have been used to create a gold-standard with the following classes:

1. protagonist.gender
2. protagonist.age
3. protagonist.socLevel
4. setting.type
5. setting.continent
6. setting.country
7. setting.name
8. narrator
9. representation
10. time.period
11. end

Modelisation and Methods

The scripts have been written in Python (available on GitHub) (https://github.com/cligs/projects2018/tree/master/text2metadata-dh). The features have been represented as different document models (Kestemont et al. 2016):

- raw frequencies
- relative frequencies
- tf-idf
- z-scores

Different classify algorithms (cross validation, 10 folds) and amount of Most Frequent Words have been evaluated. For each class a single token was used to represent each class value and a ratio was assigned for the default class value (see repository in GitHub for rules). Both approaches were compared to a "most populated class" baseline, quite high in many cases.

Results

The results of both approaches are as following:

Class	F1 baseline	F1 Rule	F1 Cross Mean	F1 Cross Std	Algorithm	Model	MFW	Winner
end	0.60	0.54	0.60	0.02	LR	tfidf	100	Baseline
narrator	0.83	0.80	**0.91**	0.04	**RF**	**tfidf**	**1000**	**ML**
protagonist.age	0.55	0.25	0.55	0.01	LR	tfidf	100	Baseline
protagonist.gender	0.80	0.68	0.80	0.01	BN	tfidf	100	Baseline
protagonist.socLevel	0.63	0.49	0.64	0.07	SVC	zscores	5000	Baseline
representation	0.88	0.80	0.88	0.01	LR	tfidf	100	Baseline
setting.continent	0.95	0.94	0.96	0.01	SVC	zscores	5000	Baseline
setting.continent.binar	0.95	0.95	0.95	0.19	LR	zscores	500	Baseline
setting.country	0.93	0.38	0.94	0.01	SVC	zscores	1000	Baseline
setting.country.binar	0.87	0.47	0.88	0.03	SVC	zscores	1000	Baseline
setting.name	0.64	**0.85**	0.71	0.02	SVC	zscores	1000	**Rule**
setting.type	0.48	0.46	**0.71**	0.05	**SVC**	**zscores**	**5000**	**ML**
time.period	0.95	0.95	0.97	0.01	BN	zscores	5000	Baseline

Fig 3. Results

In many cases the baselines are higher than the results of both approaches. The rule outperformed the baseline in the case of name of the setting with very good results. In two cases (narrator and setting's type), Machine Learning is the most successful approach and its F1 is statistically higher than the baseline (one sample t-test, = 5%). The algorithms Supported Vector Machines, Logistic Regression and Random Forest are most successful, while tf-idf and speacilly z-scores got the best results, the last one a data representation "highly uncommon in other applications" different from stylometry (Kestemont et al, 2016).

Conclusions

In this proposal I have used simple rules and simple features in order to detect relatively complex literary metadata in many cases with high baselines. While Machine Learning showed a statistically significant improvement in detection for two classes (type of setting and narrator), rules worked better for the name of the setting. This is a promising point to continue researching in order to annotate the rest of the corpus.

References

Calvo Tello, J. (2017). What does Delta see inside the Author?: Evaluating Stylometric Clusters with Literary Metadata. III Congreso de La Sociedad Internacional Humanidades Digitales Hispánicas: Sociedades, Políticas, Saberes. Málaga: *HDH*, pp. 153–61 <http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf>.

Calvo Tello, J., Schlör, D., Henny-Krahmer, U. and Schöch, C. (2017). Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels. Montréal: *ADHO*, pp. 181–83 <https://dh2017.adho.org/abstracts/037/037.pdf>.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4): 453–75.

Kastorp, F., Kestemont, M., Schöch, C. and Bosch, A. Van den (2015). *The Love Equation: Computational Modeling of Romantic Relationships in French Classical Drama. Sixth International Workshop on Computational Models of Narrative*. Atlanta, GA, USA. <https://zenodo.org/record/18343>.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, 63: 86–96 <http://dx.doi.org/10.1016/j.eswa.2016.06.029>.

Müller, A. C. and Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientist*. Beijing: O'Reilly.

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. First edition. Beijing Boston Farnham: O'Reilly.

Witten, I., Frank, E. and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. San Francisco: Morgan Kaufmann.

Prospectiva de la arquitectura en el siglo XXI. La arquitectura en entornos digitales

Luis David Cardona Jiménez
lcardona@ucm.edu.co
Universidad Católica de Manizales, Universidad de Caldass, Colombia

Desde la irrupción y aplicación de los adelantos de los medios digitales en los procesos de diseño y creación en arquitectura, los ámbitos académicos y profesionales han encontrado alternativas, posturas y desarrollos para la adopción de la tecnología digital, la cual, ha propiciado impactos en las formas de trabajo y en los abordajes del problema arquitectónico definiendo interesantes posibilidades en la imaginación y creación arquitectónica.

Se ha establecido por varios autores como, Carpo, Piccon, Frazer, Menges Cache, entre otros, que estos procesos de cambio en la imaginación y procesos de diseño y creación del espacio arquitectónico se dieron a principios de la década de los 90's con la aparición de los primeros programas comerciales de CAD (Computer-Aided Design) los cuales ofrecieron oportunidades de transformación y manejo en la proyectación geométrica de espacios y edificios.

Es importante mencionar que la tradición del pensamiento arquitectónico desde sus orígenes se ha basado en la geometría euclidiana y en los sólidos platónicos, prisma, cilindro, cubo, pirámide, esfera, son figuras que se encuentran en las arquitecturas de todas las civilizaciones antiguas, las cuales se podían identificar claramente como arquetipos únicos y aislados (Fernández-Álvarez, 2014)

El giro post-digital

El termino post-digital es relativamente reciente y aún en construcción, sin embargo, una postura es no entenderlo como "después" de lo digital o lo "anti" digital, más bien se debe pensar como en la relación y el dominio de lo humano sobre lo tecnológico. El término "post-digital" apunta la llamar la atención sobre "una actitud que se preocupa más por ser humano que por ser digital" (Zreik & Gareus, 2012)

Reiteradamente se ha mencionado a los años 90 como el momento en el que se evidencia la aplicación de medios y tecnólogas digitales en arquitectura. De acuerdo con Buchanan (1992), el reposicionamiento de nuevas ideas y planteamientos desencadenados por las pociones e intenciones de interpretación de nuevas preguntas y prácticas en todo al diseño incorporando tecnologías alternativas con entornos de simulación buscando productos y materiales innovadoras.

La incorporación de medios digitales al pensamiento, imaginación y creación arquitectónica expresadas en la representación y visualización del espacio arquitectónico. Aquí cabe mencionar de nuevo a Buchanan (1992) el cual, sin ser arquitecto, pero con una consciencia y formación en diseño, nombra la arquitectura deconstructivista como una de las iniciativas arriesgadas y agresivas que contribuirían a recuperar el significado que trasmite la obra arquitectónica.

El futuro de la arquitectura, el Siglo XXI

La arquitectura como disciplina de tradicional arraigada a principios y postulados casi inmutables, viene des-

de hace tres décadas presentado cambios en la forma de abordar el problema del diseño del espacio habitable representado en la ciudad, edificios o viviendas. Con el constante cambio tecnológico, el fortalecimiento de las relaciones humanas a través de los digital, la arquitectura empieza a responder acertadamente a las demandas de nuevas formas de abordar la trasformación de la realidad.

En un mundo hiperconectado, con una producción diaria de datos incalculables, el Big Data se convierte en una herramienta que permite crear plataformas de trabajo colaborativo, fortaleciendo la relación entre usuarios y diseñadores.

Phil Bernstein, arquitecto y profesor de Yale University visualiza el futuro de la arquitectura a través avatares para el análisis de comportamiento de usuarios en entornos construidos virtualmente desde la visualización del Big Data.

> Big Data ya está transformando la forma en que los arquitectos diseñan edificios. Cambiando las potencias Big Data y la realidad virtual, se avanzará en la práctica arquitectónica a pasos agigantados (Phil Bernstein)

Esta visión prospectiva de la arquitectura es emergente y esta en proceso de convergencia. Hoy, aunque incipiente, en Latinoamérica se empieza a mostrar un interés por avanzar en el entendimiento y aplicación de conceptos de investigación e innovación a través de las posibilidades que las tecnologías digitales ofrecen para el desarrollo de una arquitectura que responda a las expectativas del mundo en constante proceso de cambio.

References

Alexenberg, M. (2011). *The Future of Art in a Postdigital Age: From Hellenistic to Hebraic Consciousness.* Chicago: Intellect Ltd.

Allen, S. (2009). *"Velocidades terminales" en La digitalización toma el mando.* Barceolna: Gustavo Gili.

Amado, R. (2007). La arquitectura como interfaz. En *Arte, Arquitectura y Sociedad_ Digital* (págs. 107-109). Barcelona: Universitat de Barcelona/ESARQ UIC.

Baltazar, A. (2009). *Cyberarchitecture: the virtualisation of architecture beyond, Tesis doctoral.* University College London, UCL.: The Bartlett School of Architecture.

Buchanan, N. (1992). Wicked Problems in Design Thinking paper. *Design Issues*, 5-21.

Carpo, M. (2013). *The Digital Turn in Architecture 1992 - 2012.* Chichester, UK: John Wiley & Sons Ltd.

Cross, N. (1982). Designerly Ways Of Knowing paper. *Design Studies*, 221-227.

Fernández-Álvarez, Á. J. (2014). Riding the cloud. Information and architectural representation in the post-digital age. *EGE: Revista de Expresión Gráfica en la Edificación,* , 159-166.

Ortega, L. (2009). *La digitalización toma el mando.* Barcelona: Gustavo Gili.

Períes, L. (2016). *Estereotomía y topología en arquitectura.* Buenas Aires: Editorial de la Universidad Nacional de Córdoba.

Piccon, A. (2010). *Digital Culture in Architecture.* Basilea: Birkhäuser.

Sandoval Vizcaíno, M. (2014). Herramientas de diseño y arqutiectura, la relación intrínseca entre herramientas y diseño . *Revista Legado de Arquitectura y Diseño*, 39-56.

Zreik , K., & Gareus, R. (2012). *PostDigital Art - Proceedings of the 3rd Computer Art* . Paris: Europia Productions.

Visualizando Dados Bibliográficos: o Uso do VOSviewer como Ferramenta de Análise Bibliométrica de Palavras-Chave na Produção das Humanidades Digitais

Renan Marinho de Castro
renan.castro@fgv.br
Fundação Getulio Vargas, Brazil

Ricardo Medeiros Pimenta
ricardopimenta@ibict.br
Instituto Brasileiro de Informação em Ciência e Tecnologia, Brazil

O objetivo dessa pesquisa é mapear, através da identificação de termos de palavras-chave, quais as principais atividades presentes nas humanidades digitais construindo e visualizando mapas bibliométricos oriundos de uma revisão de literatura deste tema. Dessa forma é proposta uma análise desses dados a partir da utilização do software VOSviewer para construção de redes de relacionamento dos termos provenientes das bases: Web Of Science (WoS) e Scopus. Assim, foram gerados grafos de palavras-chave baseados nos termos atribuídos à literatura registrada nessas duas bases de dados. Buscamos dessa forma combinar essas duas análises a partir da construção de dois mapas distintos e possibilitar seu cotejamento.

Partindo dessa proposta, elaboramos uma expressão de busca[1] para dar conta de recuperar a publicação sobre *digital humanities* em inglês, espanhol e português nas bases de dados eleitas para esta revisão. Adotou-se como padrão a opção de filtro que contemplasse o '*abstract*', sendo o campo 'resumo' escolhido como foco da recuperação por apresentar maiores concentrações de termos relacionados à indicação temática dos documentos. Os resultados reportados pelas buscas foram expor-

1 A expressão de busca aplica às bases selecionadas pode ser representada pela *string* ((((((„Digital Humanities")) OR (("Humanidades Digitales")) OR (("Humanidades Digitais"))))))

tados no formato compatível com o VOSviewer, no caso da Web Of Science, 'Tab Delimited (Win)' e no caso da Scopus, o formato 'CSV'. Foram recuperados na Web Of Science 1067 documentos e, na Scopus, 1575.

De posse dos arquivos extraídos, utilizamos do recurso de criação de grafos baseados em co-ocorrência de palavras-chave. Essa análise oferece as opções *'Author´s keywords'* e *'Keywords Plus'*, por isso elegemos a opção *'all keywords'* que engloba essas duas modalidades, além do método de *full counting* que atribui o mesmo peso para cada link em co-ocorrência. Na WoS foram totalizadas 2826 palavras-chave com exigência mínima de 8 ocorrências para integrar a análise, essa filtragem resultou em 38 núcleos conectados. No caso da Scopus também elegemos a opção *'all keywords'* para contemplar as palavras-chave atribuídas pelos próprios autores (*Author´s Keywords*) além da opção *'index keywords'*, cuja atribuição é proveniente da base. Foram, assim, identificadas 5195 palavras-chave e a nota de corte elevada à recorrência mínima de 15 vezes. Essa configuração produziu um grafo com 64 (após desambiguação: 61) termos com núcleos de conexão entre si. Este grafo também considerou o método *'full counting'*.

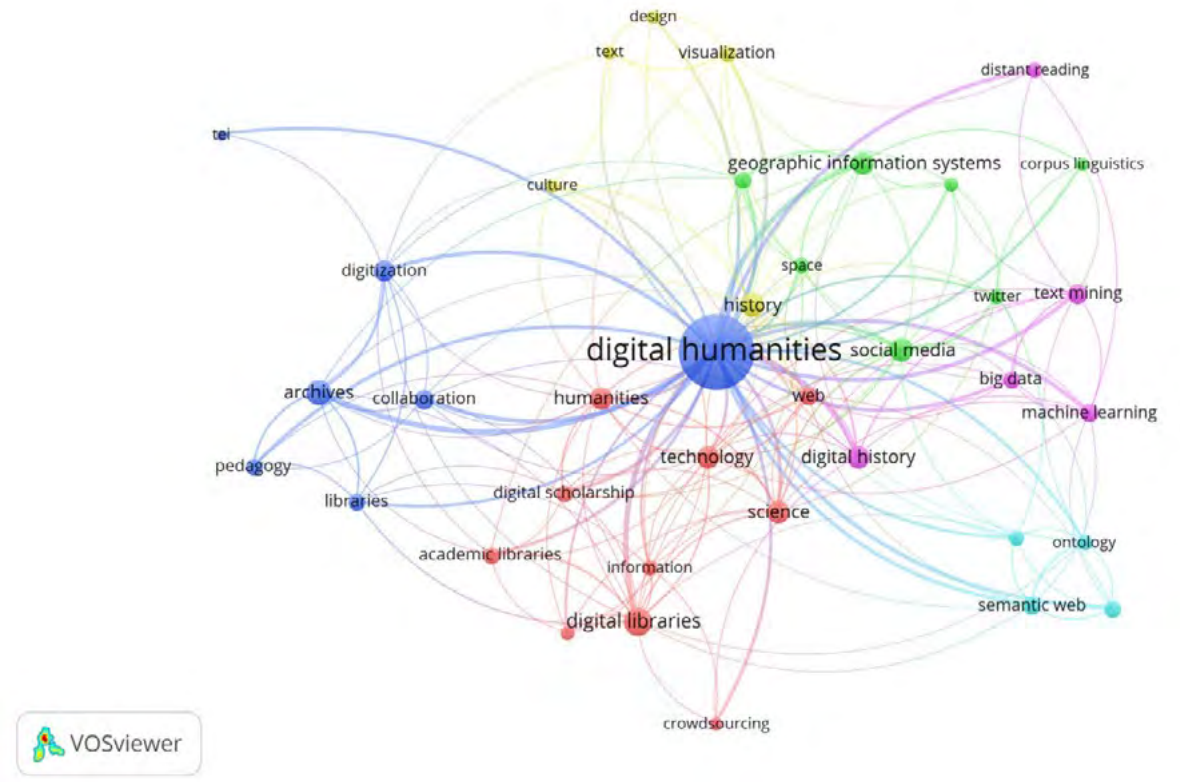


Figura 1 Grafo de palavras-chave da produção registrada na Web Of Science com ‚nós' calculados segundo seu link total de força.

Na sequência produzimos dois mapas baseando-se nas respectivas fontes de literatura e, baseados nestas, geramos, além dos dois mapas, a mineração dos termos mais recorrentes que servem de base para construção do grafo. O grafo respectivo à WoS possui 6 clusters compostos por 10 termos no de maior tamanho e 4 no menor. A análise de clusters permite identificar que dentre estes há clusters estreitamente relacionados às bibliotecas digitais e à ciência da informação (C.I.) (por exemplo termos como *information* e *technology*), como no caso do cluster 1. Também há um cluster relacionado às técnicas de visualização (cluster 2). O cluster 3 volta a apresentar termos relacionados à C.I. como *archive, digitization e libraries*. Já o cluster 5 volta-se às técnicas das humanidades digitais como *text mining* e *machine learning*.

O grafo com dados da Scopus também possui 6 clusters tendo no maior deles 16 termos e, no menor, 7. Também é possível perceber a recorrência de um cluster voltado às técnicas de visualização (cluster 4: *visualization, data visualization e gis*) bem como a reverberação da presença da C.I. com os termos *digital libraries, digital archives e digital collections* (cluster 2). Outras técnicas das humanidades digitais reincidem como *data mining* e *text mining* (cluster 3), além de outros termos relacionados à ciência da informação: *archives, libraries* e *digitization* (cluster 5). Vale destacar que o termo com maior peso foi *digital libraries* tanto na WoS como na Scopus.

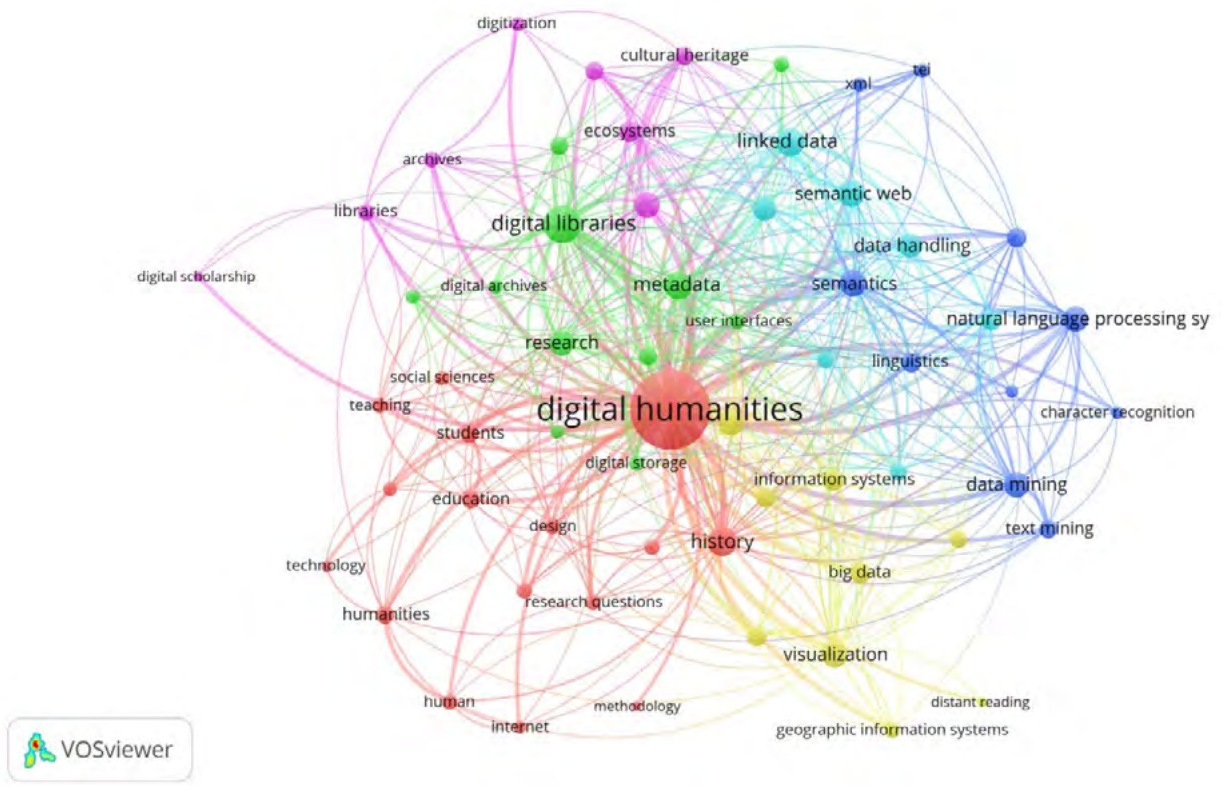


Figura 2 Grafo de palavras-chave da produção registrada na Scopus com ‚nós' calculados segundo seu link total de força.

Dessa forma os mapas permitem visualizar termos e conceitos mais presentes na literatura e, consequentemente, possibilitam a clarificação da relação entre eles. Apesar da grande rede de relacionamento que os mapas exibem é possível, mesmo interpretando apenas os clusters criados, contemplar, por exemplo, as áreas principais que interagem para formar a ideia de humanidades digitais na literatura. Além disso, sobretudo, o cotejamento dos grafos provenientes de cada repositório de literatura permite corroborar quais termos se consolidam através de sua reincidência nos mapas.

References

DACOS, Martin. (2011). Manifesto das Humanidades Digitais. *ThatCamp Paris*, [S.l.] 26 mar. 2011. Disponível em: <https://tcp.hypotheses.org/497> Acesso em 10 out. 2016.

ECK, Nees Jan Van; WALTMAN, Ludo. (2016) *VOSviewer Manual*. Disponível em http://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.5.pdf Acesso em 10 de julho de 2017

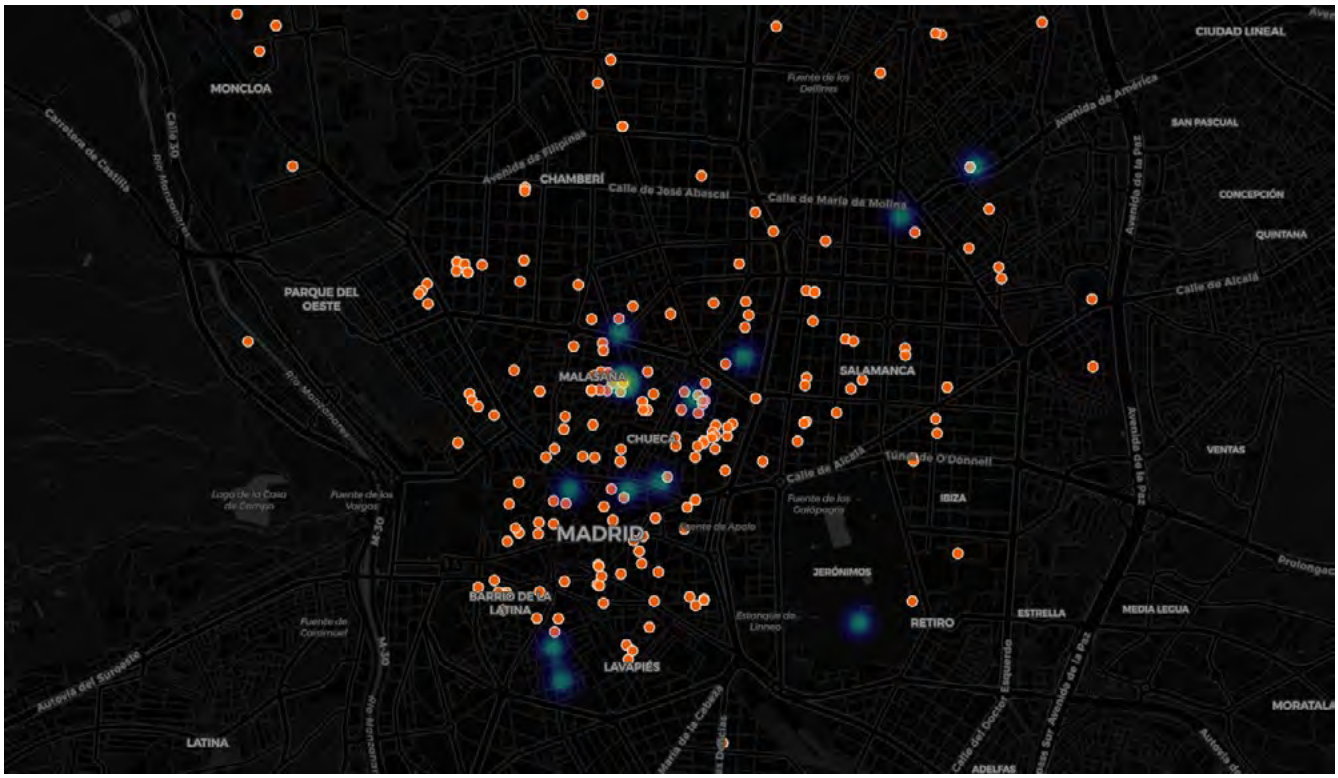
KOLTAY, Tibor. (2016) Library and information science and the digital humanities: perceived and real strengths and weaknesses. *Journal of documentation*, 72(4), pp. 781-792.

TANG, Muh-Chyun; CHENG, Yun Jen; CHEN, Kuang Hua. (2017) A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses, *Scientometrics*, v.113, n.2, pp.985-1008, nov. 2017.

Mapping the Movida: Re-Imagining Counterculture in Post-Franco Spain (1975-1992)

Vanessa Ceia
vanessa.ceia@mcgill.ca
McGill University, Canada

Mapping the Movida is an open web archive and geo-spatial project that visualizes the cultural and creative hubs and networks of the *Movida madrileña*, a sociological phenomenon and cultural renaissance that emerged in the first decade of Spanish democracy (roughly 1976-1986), most notably in central Madrid. Among the Movida's most well-known artists are filmmakers Pedro Almodóvar and Iván Zulueta, photographers Alberto García-Alix and Ouka Leele, illustrators El Hortelano, Nazario, and Ceesepe, poet Eduardo Haro Ibars, novelist Eduardo Mendicutti, fashion designers Jesús del Pozo, Manuel Piña and Agatha Ruiz de la Prada, and musicians Ana Curra and Alaska (Olvido Gara), among many others. One of the most striking characteristics of those who have been historicized as so-called "artists of the Movida" is that



they are, with few exceptions (Almodóvar, Ruiz de la Prada, Curra, Alaska), men of upper middle-class upbringing. Additionally, when examining the canonized geographies of the Movida—that is, the cultural hubs and culturally productive spaces of the period—we find that these canonical artists are primarily centered around Madrid's core neighborhoods, such as Malasaña, and, in few instances, the affluent north-central sector of the Spanish capital. In the canonized Movida, peripheral, and often working-class neighborhoods are largely excluded from the cultural map and countercultural histories of this period.

This project is a scholarly response to the limited scope of artists—mostly male and professionally active in central Madrid—historically associated with the Movida in mainstream press and scholarship. In its mission to bring to light and build "BRIDGES/PUENTES" with uncharted human geographies of the period, Mapping the Movida aims to: (1) re-create the Madrid of the Movida using a range of visual, textual and spatial media, data, and thick (Presner, Shepard, Kawano 2014) and deep mapping technologies that document the Madrid of the past; (2) visualize creative networks and cultural hubs of the Movida through various cultural and critical lenses—including mainstream Spanish media outlets (*El País*, *ABC*, *El Mundo*), scholarly articles, and subcultural publications from the period (*La Luna de Madrid*, *El Víbora*, *Ozono*, *Madrid Me Mata*)—to reveal how each lens represents the Movida in different, divergent, and/or similar ways and "provoke negotiation between insiders and outsiders, experts and contributors, over what is represented and how„ (Bodenhamer, Corrigan, Harris 2015: 4); (3) create a public archive and searchable database of Movida events and artists' documented movements in Madrid during the Movida; and, perhaps most importantly, (4) de-colonize the geographies of the Movida by revealing spaces, artists, and socio-economic groups that problematize the cultural and spatial canon of the period.

This poster, grounded in archival research from Brown University's Revistas de la Movida Collection, will exhibit the methodology and tools (Carto, Esri Story Maps) that have been used, the archival and theoretical concerns that have arisen, and the revelations that have been made during the various stages of project development. It will also demonstrate how Mapping the Movida's marriage of archival research and technology questions and queers the scope of what has been historicized and canonized as the "culture of the Movida" over the last nearly 40 years. At stake in this project is our understanding of the cultural and human geographies of Madrid during this period as well as our knowledge of artists and cultural products that have rarely, if at all, been studied and imagined within the corpus of so-called Movida artists and texts.

References

Bodenhamer, D.J. and J. Corrigan, T.M. Harris. (2015). *Deep Maps and Spatial Narratives*. Bloomington: Indiana University Press.

Presner, T. and D. Shepard, Y. Kawano. (2014). *HyperCities: Thick Mapping in the Digital Humanities*. Cambridge: Harvard University Press.

Intellectual History and Computing: Modeling and Simulating the World of the Korean Yangban

Javier Cha
javiercha@gmail.com
Seoul National University, Korea

This poster presentation demonstrates the use of computational methods to discover hidden collectives and communities from Korean historical data. The overarching question is derived from the intellectual history of early modern Korea, which was defined by the coalescence of several schools of Neo-Confucian thought and literary movements. Such developments took place at a time of increasing localization of population, material resources, state institutions, and culture. In the existing body of research, the connections between the material and ideational aspects of the yangban aristocracy have been unclear, owing in large part to the undue attention given to a small number of famous personalities, source materials, and locations. Can this skewed picture be redrawn from the bottom-up, through a more balanced and fuller use empirical data? Fortunately for social scientifically-minded historians of Korea, the government of South Korea has aggressively funded the digitization of cultural heritage. Access to this "big data" has allowed me to embark on a critique of existing reified generalities with large-scale data analysis. This kind of data also demands a new type of research concerning social, cultural, and historical entities which may not yet have been identified and therefore not yet been given a label. The data are drawn from two sources: (1) 50,000 civil service examination degree holders and their extended kin and (2) 198 million Sinitic characters of writing extracted from 1200 collected works. The pilot run has already revealed a surprising assemblage of *yangban* aristocrats interconnected via complex ties of patronage and marriage. As the method gets refined, and more data gets added and cleaned, I expect to discover other hidden entities and groupings. Finally, I will explain the theoretical and philosophical implications of historical entity discovery through computing by engaging with the works of social scientists and philosophers such as Gilles Deleuze, Manuel DeLanda, Norbert Elias, Zhuangzi, and Su Shi.

In addition to sharing this digital project's historical and philosophical contributions to East Asian Studies, I will share my experience with the uses of software tools to address key issues in early modern Korean history. Computational history entails the processing of digitized or born-digital sources using software packages and algorithms designed for use in another discipline or industry. Moreover, historians of East Asia may need to consider the support for Unicode encoding or rare Sinitic characters. I will explain the strategies I developed to collate genealogical data and scrape a large amount of text with the aid of a macro program. Thereafter, I will discuss my adaptation of Cytoscape, a network visualization platform designed for bioinformatics, to analyze the robust ties of marriage that contributed to the self-perpetuation and regional division of the early modern Korean *yangban* aristocrats. A highlight of this demonstration will be my linking of multiple data sources and the subsequent extraction of a subnetwork (~300 nodes) from a large network (~20,000 nodes). The marriage networks and subnetworks will be compared against the patterns of localization discovered through spatial data and text analysis. The presentation will consist of large-format prints as well as digital media shown on a monitor or a projection screen (which I will bring with me).

More Than "Nice to Have": TEI-to-Linked Data Conversion

Constance Crompton
constance.crompton@uottawa.ca
University of Ottawa, Canada

Michelle Schwartz
michelle.schwartz@ryerson.ca
Ryerson University, Canada

For developers of TEI-based projects, linked data is often much-desired but nonessential, an added output that would be nice to have, but that is not critical to ultimate success of the project. The recent catalyzation of interest in linked open data in the context of TEI (including the revitalization of ADHO's LOD SIG and the TEI's Ontologies SIG) is, however, a promising sign of our field's engagement with linked data, and our readiness to join international efforts to produce and publish linked data (Huber et al.; Pattuelli et al.; Lehmann et al.; Shadbolt et al.; Hellmann et al). Currently linked data only makes up 1% of the web, and much of that 1% is used for commercial rather than scholarly purposes (Simpson and Brown). The conversion of existing digital humanities data into linked data offers humanities scholars an opportunity to intervene in the semantic web as it is being built. It allows the power of the semantic web to be harnessed for more than just commercial purposes, and offers rich and readily accessible information about the research topic of the liberal arts: the human record. The underlying assumption of the semantic web is the same as the underlying assumption of humanities research—we can never assume ourselves to be in a full state of knowledge; there is always new information that may come to light. The creation and exposure of linked data from the vast number of existing authoritative TEI projects could enable scholars to embrace linked cultural data at scale. But what is the path to success? Our poster reflects on the technical and

institutional challenges to linked data creation, and proposes a workflow and toolset for the creation of linked data from TEI.

Despite calls in the digital humanities for TEI-linked data compatibility (Simpson and Brown, Ciotti and Tomasi), scholars have yet to develop best practices for creating linked data from richly encoded TEI resources. For many projects, the production of linked data is an ancillary goal, one that would be gratifying to achieve, but one that is secondary to the encoding itself, or only necessary to facilitate aggregation. We propose the development of XSLT-backed tools to convert and connect otherwise incommensurable data sets. The tools will require human checks, since mapping the unique usages of hierarchical elements by TEI-based projects onto existing ontologies—including CIDOC-CRM, FOAF, SKOS, schema, dcterms, and others—is hardly one-to-one. Furthermore, the historical primary source material that the TEI permits encoders to so diligently represent requires significant contextualization, since the conditions of its production were often underpinned by historical worldviews that today may be read as racist, sexist, ableist, or homophobic. Without machine and human-readable contextualization, historic intents, biases, and worldview may be reified by the inferencing that linked data permits. The ideal outcome would instead be an understanding, without valourization, of those worldviews. We are testing our tools and workflow against data sets that present exactly these challenges. We are working with four sample TEI-based data sets representing four hundred years of Atlantic cultural production, including manuscripts, books, periodicals, biographies, art works, legislation, places, and events, representing 45,000 entities. The data spans four hundred years, two regions (Europe and the Americas), five religions, three languages, all with particular historical-contextual specificity. The upcoming phases of our work will involve testing the tools against more diverse TEI sets. We are especially interested in the poster format, as we are keen to solicit feedback from peers on the balance between granularity and generality in the representation of people, places, time, and cultural production as linked data.

References

Ciotti, F., Tomasi, F. (2016). Formal Ontologies, Linked Data, and TEI Semantics. *Journal of the Text Encoding Initiative*. https://doi.org/10.4000/jtei.1480

Hellmann, S., et al. (2014). Knowledge Base Creation, Enrichment and Repair, in: Auer, S., Bryl, V., Tramp, S. (Eds.), *Linked Open Data — Creating Knowledge Out of Interlinked Data, Lecture Notes in Computer Science*. Springer International Publishing, pp. 45–69.

Huber, J., Sztyler, T., Noessner, J., Murdock, J., Allen, C., Niepert, M. (2014). LODE: Linking Digital Humanities Content to the Web of Data. I*EEE/ACM Joint Conference on Digital Libraries*. http://arxiv.org/abs/1406.0216.

Pattuelli, M.C., Miller, M., Lange, L., Fitzell, S., Li-Madeo, C. (2013). Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project. *The Code4Lib Journal*.

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., Schraefel, M.C. (2012). Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems* 27, 16–24. https://doi.org/10.1109/MIS.2012.23

Simpson, J., Brown, S. (2014). Inference and Linking of the Humanist's Semantic Web, in: Implementing New Knowledge Environments. Presented at the *Building Partnerships to Transform Scholarly Publishing*, Whistler, BC.

Animating Text Newcastle University

James Cummings
james.cummings@newcastle.ac.uk
Newcastle University, United Kingdom

Tiago Sousa Garcia
tiago.sousa-garcia@newcastle.ac.uk
Newcastle University, United Kingdom

Animating Text Newcastle University

This *DH 2018* poster will provide an introduction to a new kind of digital humanities research network and the pilot projects it is building. Animating Text Newcastle University (ATNU) is a three year interdepartmental research project exploring new frontiers at the cross-roads between traditional scholarly textual editing, digital editing, digital humanities and computer science. It is a collaboration between humanities researchers and computing scientists that is exploring research questions raised by pre-1860 editing projects. The poster at *DH 2018* will introduce the ATNU network, the successes and failures of the project so far, and the individual pilot projects it has undertaken.

ATNU connects original historical research from across Newcastle University from the School of English Literature, Language and Linguistics, the School of Arts and Cultures, the School of Modern Languages, and the School of History, Classics and Archaeology, with the transformational research of the Digital Institute. The intention is to share expertise and intellectual resources and to work to deliver ambitious, future-facing research that will nurture future large-scale collaborative projects. The network is hosting invited expert workshops, visiting speakers, and undertaking pilot digital projects informed by editing challenges. It is hoped that this will not only increase familiarity with DH methodologies and technologies inside the institution but foster partnerships outside it.

Why pre-1860 texts?

In these earlier periods the characteristics of manuscript and the printed book (and their relationship with one another) are fundamentally distinct from how they are in the period from the late nineteenth century to the present. Yet the ways in which pre-1860 texts are re-presented in current print and digital editions often fails to recover their vital, distinctive contexts (the relations between authors, copyists, printers, publishers and booksellers), and the way the printed page is meant to facilitate particular experiences. ATNU is contributing to a vital debate not just about the history of the text and the future of the book, but also about the place of historically-focussed editorial scholarship in the story of the humanities and its digital future.

Funding Streams and Resistance to Failure

A frustrating aspect of many research projects is the tendency to promote their successes and ignore failures. These projects may produce excellent outputs which benefit the humanities, but in discussing their projects they often count the hits and ignore the misses. It is completely understandable when highlighting the success of their projects to those who funded them. However, ATNU is fortunate in being slightly different: it is funded by Newcastle University's Research Investment Fund specifically to bolster digital humanities research at the institution. Part of the ATNU mission is the development of additional grant applications for cutting edge projects that specifically have their basis in more risky blue skies thinking. Moreover, in order to develop these funding bids ATNU is undertaking a series of pilot projects but because these are funded internally they are allowed to be more experimental. They do not have to be successes -- failure is indeed an option! Where the pilot projects succeed they will go on to be the base for external funding bids, but where these projects are less successful, their failures can be publicly documented and projects can be re-oriented towards more successful techniques.

Pilot Projects

The network's pilot projects are in three categories: "Manuscripts and Print", "Performance", and "Translation". The projects in each of these have a set of shared interests, methodologies, and an overlap of possible technological solutions.

- **Manuscript and Print:** the projects in this area investigate topics such as scholarly digital editing, the process of collaborative editing, the presentation of editions, and the handling of variation across multiple versions. The first pilot is a prototype digital edition of the Sarum Hymnal involving text, image, and music encoding.
- **Performance:** many texts have a life beyond the page, and include acoustic and visual experiences. ATNU is exploring how best to represent and enable these performative and interactive dimensions. One pilot in this looks at a visual, interactively animated, view of James Harrington's early modern proposal for reforming voting systems, another experiments with the acoustic effect of punctuation in early modern texts.
- **Translation:** investigating pre-modern texts and their translations, how these entities relate, and developing tools for researchers comparing texts in translation. A pilot under this theme is examining the concept of the social translation.

The poster will provide more details about the network and its pilot projects.

Una Investigación a Explotar : Los Cristianos de Alá, Siglos XVI y XVII

Marianne Delacourt
marianne.delacourt@univ-tlse2.fr
Université Toulouse Jean Jaurès, France

Véronique Fabre
veronique.fabre@univ-tlse2.fr
Maison des Sciences de l'Homme et de la Société de Toulouse/ CNRS, France

En los siglos XVI y XVII, el Mediterráneo fue el reto geopolitico entre la Monarquia Española y el Imperio Otoman. Entre batallas e incursiones, muchos cristianos fueron reducidos a la esclavitud por los Berberiscos. Unos,para suavizar sus condiciones de vida o por fuerza, se convertieron al Islam, y fueron llamados *Renegados* Ellosfueron *puentes* entre las dos civilisaciones y religiones.

Cuando regresaban a la vida cristiana, fueron juzgados por la Inquisición.

Bartolomé Bennassar, historiador frances, hizo, al fin de los 80, fichas de papel sobre más de 1550 renegados, basadas en las fuentes de los archivos de la Inquisición.

Nuestro proyecto es digitalizar esas fichas y crear una base de datos, albergada en la plataforma francesa de humanidades digitales del CNRS : *HUMA-NUM*

El poster que queremos presentar da cuenta del método y de las etapas de un proyecto de *Humanidades digitales* entre dos instituciones que no suelen trabajar juntas.

La numerisacion fue bastante facil... construir la base de datos que permitara interrogar a las fichas es muchas mas dificil.

El Profesor Bennassar habia preparado fichas dactilograficas con datos fijos tipo nombre, lugar de nacimeinto, condiciones de renegacion, etc...Pero, leyendo el archivo de los procesos, añadia muchas informaciones

manuscritas que vienen enriquecer el perfil de vida del renegado, pero que no son «normalizadas».

Es decir que para construir la base de datos teniamos que ser pertinente con las estructuras de interrogacion y decidir cual informaciones adicional teniamos que tomar en cuenta para aclarar unos datos biograficos del renegado y sucitar el interes del investigador .

Entonces, pedimos a historiadores de validar cada etapa de la elaboracion de la base de datos.

Esperamos que esos datos seran explotados: que los datos geograficos un dia sean explotados por un logiciel de visualisacion y analisis de redes, que la informacion «se casa con uzansas de moros», permiten investigar sobre la vida intima de los renegados, etc…

Asi que como lo dice el titulo del poster : « los critianos de Ala, una investigacion a explotar»

The Iowa Canon of Greek and Latin Authors and Works

Paul Dilley
paul-dilley@uiowa.edu
University of Iowa, United States of America

This poster will introduce the Iowa Canon of Greek and Latin Authors and Works, which aims to be the most comprehensive list of classical texts from the origins of Greek and Latin literature through the end of the Antiquity (the 6th century CE), and associated metadata, made available for researchers through an innovative online interface. The Iowa Canons are affiliated with the Big Ancient Mediterranean Project, for which I am a co-PI with Sarah Bond, with lead developer Ryan Horne, which seeks to provide an interface for the coordinated exploration of linked textual, geospatial, and network data relating to the ancient world. Both BAM's interface and the Iowa Canons are in development; a beta-version of the Iowa Canon of Latin Authors and Works is available at http://bam.lib.uiowa.edu/iclaw/. The Iowa Latin Canon currently stands at over 5,400 works; a more extensive version, paired with the Iowa Canon of Greek Authors and Works, which currently includes over 9,000 entries, will be published in May 2016. I have been assisted in data collection by students in my graduate seminars on distant reading, as well as undergraduate and graduate research assistants.

The goal of both Iowa Canons is to integrate existing canons of Greek and Latin Literature, especially the Perseus Catalog, the Thesaurus Linguae Graecae (TLG) Canon, the Packard Humanities Institute (PHI) Classical Latin Texts, the Brepols Library of Latin Texts (LLT-A), and other resources such as the Clavis Apocryphorum; to increase their granularity and the amount of associated metadata; and to make this data collection searchable in an interface that integrates Greek and Latin texts, which none of the previous Canons do. None of the existing Canons include lost works, and they group fragmentary works under a single entry (e.g. "Fragmenta"), with no functionality to search for individual titles within it, which sometimes number in the hundreds. The Iowa Canons, in contrast, will include all known lost or fragmentary works, and include additional metadata, such as time and place of composition, genre (using the same "in-house" classification system for both Greek and Latin texts), meter (if poetic), and Christian/non-Christian content. Finally, the Iowa Canons will cross-reference each work to existing canons (when possible), as well as to the Perseus Catalog, which will provide stable reference urns for Greek and Latin works, a project with which we are collaborating.

The Iowa Canon of Greek and Latin Authors and Works will make this data available to users through an interface, which will provide faceted search of available metadata, for example, by selecting all works of a particular genre, in a specified time period and/or location. The results of the search are displayed geospatially, with circles around all locations with relevant works, their diameters proportionate to the number of "hits" in that location. Clicking on the circles reveals those "hits." When combined with the extensive records of lost and fragmentary titles, this search functionality will greatly facilitate research into Greek and Latin literary history beyond the usual focus on canonical works, which will themselves be contextualized. Jockers has described this sort of research metadata as the "lowest hanging fruit of literary history" (Jockers 2013: 35); his work, as well as Franco Moretti's (Moretti 2009), have explored the possibilities of this approach for studying certain genres of 19th and 20th century literature in English, which is of course far more extensive than surviving ancient Greek and Latin literature. But the cumulative metadata that will be accessible through the Iowa Canons will offer a unique picture of an entire literary field, with over 60 genres, as it developed over centuries, and in several languages. The poster will be of interest not only to digital classicists, but to literary scholars working in other languages and eras, from whom I will solicit feedback about its functionality, as well as its potential for distant reading.

References

Jockers, Matthew, *Macroanalysis: Digital Methods and Literary History* (University of Illinois Press, 2013)

Moretti, Franco, "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740-1850)," *Critical Inquiry* 36 (2009): 134-58.

Packard Humanities Institute Latin Author List: http://latin.packhum.org/browse

Perseus Catalog: http://catalog.perseus.org/

Thesaurus Linguae Graecae: http://stephanus.tlg.uci.edu/index.php

Digital Storytelling: Engaging Our Community and The Humanities

Ruben Duran
ruben.duran@hccs.edu
Houston Community Collegeo, United States of America

Charlotte Hamilton
charlotte.hamilton@hccs.edu
Houston Community Collegeo, United States of America

Houston Community College is one of the leading two-year colleges in the United States incorporating digital storytelling into the curriculum while reaching out to the community to achieve the history of the diverse communities with vibrant background that provides such a rich tapestry that makes Houston the city it is today.

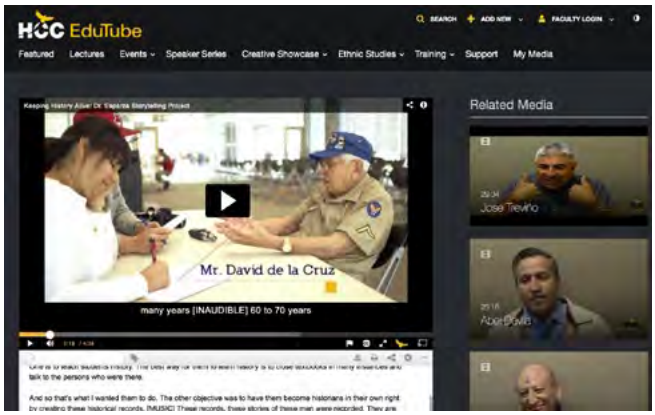
Working with the Center for Digital Storytelling we have trained our faculty and staff to incorporate these initiatives into the instructional curriculum.

Digital Storytelling supports projects that bring ideas and insights of the humanities to life for general audiences. Our past projects engage humanities scholarship to analyze significant themes in disciplines such as history, literature, and art history. Our projects support and encourage activities that involve members from the many Houston cultural communities through collaboration with humanities scholars and students. We have also invited contributions from the community in the development and delivery of humanities programming.

These presentations provide video examples of the following initiatives:

History of Latino war veterans from Korean and Vietnam Wars

Students from a Mexican American history class interviewed veterans from the Korean and Vietnam Wars. The veterans expressed pride in their contributions to the war, some of them for the first time since returning home from their deployments many years ago. Students shared their excitement while developing insight into history through stories not contained in their textbook.

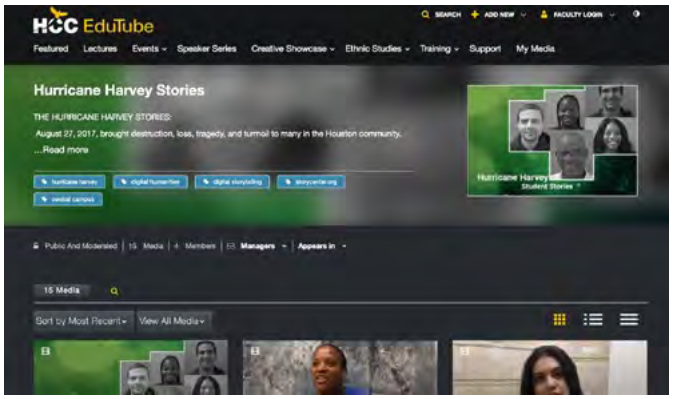


MECA grant for outreach to K-12 students

MECA is a community-based non-profit organization committed to the development of under-served youth and adults through arts and cultural programming, academic excellence, support services, and community building. Under the tutorage of library trainers, using staff and equipment resources from our institution, students produced short videos of their interviews with members of their community. This project fostered in the students a better understanding of the significant contributions of their community peers, and it helped them to develop discipline, self-esteem, and increased cultural pride.

Harvey Listening Stations provide support through student and staff stories

Following the disastrous hurricane that stuck Houston in August, 2017, we provided equipment and the opportunity for staff and counselors to capture stories reflecting the impact on individual students of the devastation of the flooding. The objective was both to allow students to tell their stories, as well as to determine whether there were specific actions we could implement to provide support for their continued success in their educational activities.



Current project is the collaboration with HHA 2018 Year of the Woman in Houston

Working through the Texas Historical Association and the Houston Historical Alliance we are providing class assignments that include developing short videos and scholarly research on women in the greater Houston community who affected or influenced the history of our region and the state. Based on established guidelines this initiative allows our students to complete the project as a class assignment in history, government or other disciplines. The women include pilots, activists, oil magnates, storytellers, scientists, ranchers, daughters, and mothers who have made significant contribution to the richness of our diverse communities. The digital stories should include notable women, as well as lesser known figures. These videos will be hosted on our Media Space as a reference tool and

will be eligible for selection by a peer jury for inclusion in the online *Handbook of Texas Women*.

The poster session showcases tools from our storytelling arsenal that includes the Listening Stations and iPads displaying referencing videos from our initiatives. All the projects were developed using WeVideo, a collaborative cloud editing application that serves as the online video editor that makes it easy to capture, create, view and share the stories.

The stories are shared in Edutube, HCC's media community tube. The Learning Station was designed as a public kiosk for people to share stories with a listener, hold a conversation, or be part of an interview. The included app automates the upload and delivery of files to participating organizations and the participants. The apps integrates the metadata collection, registration, release and transcription processes, making it a state-of-the-art tool for gathering primary source material for documentary projects.

References

Center for Digital Storytelling. Storycenter.org
WeVideo. Wevideo .com
Edutube. https://edutube.hccs.edu

Text Mining Methods to Solve Organic Chemistry Problems, or Topic Modeling Applied to Chemical Molecules

Maciej Eder
maciejeder@gmail.com
Pedagogical University in Kraków,
Institute of Polish Language, Poland

Jan Winkowski
jan.winkowski@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences), Poland

Michał Woźniak
michal.wozniak@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences), Poland

Rafał L. Górski
rafal.gorski@ijp.pan.pl
Pedagogical University in Kraków,
Institute of Polish Language, Poland

Bartosz Grzybowski
nanogrzybowski@gmail.com
Pedagogical University in Kraków,
Institute of Polish Language, Poland; Ulsan National Institute of Science and Technology, Korea

Introduction

The Renaissance Humanism was probably the last moment in the history of ideas when the development of exact sciences was shaped according to the intellectual paradigms of the humanities (the Liberal Arts, to be precise). After the advent of the Scientific Revolution in the 17th century – with its empiricism, experimental reasoning, mathematical apparatus, and so forth – the exact sciences became the point of reference for all the other disciplines, in terms of scientific inference and its methodology. The imbalance between the humanities and the sciences has been growing ever since. Nowadays, statistical analysis is routinely applied in social sciences, cognitive linguistics tries to take advantage of the fMRI technology, text analysis studies are overwhelmed by numerous machine-learning techniques, ranging from hierarchical cluster analysis to Support Vector Machines classification and Deep Learning. The exact sciences have affected the humanities to a considerable extent, but at the same time they continue to be rather resistant to any methodological inspirations coming from the "soft" scholarship. This study is an example of such a reversed influence, since we propose to apply text mining methods to study chemical molecules. Arguably, the phrase "If an atom is a letter, then a molecule is a word", even if popular in chemistry, sounds rather naïve for anyone who has some expertise in linguistics. Nonetheless, despite a shallow similarity between language structures and organic chemistry at first glance, the methodology developed in text mining proves very promising as a way to discover internal molecule structures.

The problem

One of the biggest issues in contemporary organic chemistry is an enormous number of different molecules

and their fragments that play role in chemical reactions. To cut a long story short: any reaction involves certain changes in molecules' structures, which usually means that certain bonds are disjoined, and particular atoms change their positions within each molecule. On theoretical grounds, these changes can be predicted and/or controlled. In practice, however, predicting optimal bond cuts requires high-level expert knowledge, due to the extreme complexity of the problem, or an enormous computer power to run brute-force combinatoric algorithms. This is, however, still far beyond our capabilities, because completing a task that involves testing billions of billions of combinations would require decades if not centuries. For that reason, the big question at stake is how to optimize the entire process of identifying relevant molecule substructures (Ruddigkeit et al., 2012).

Splitting complex chemical molecules into "meaningful" substructures is the first problem to be overcome. In this context, "meaningful" means groups of atoms that are local centers of reactions. The nature of bonds between atoms is very well understood since the first half of the 20th century. However, it is still unclear why certain clusters of atoms tend to keep together while reprehend some other groups. Being one of the most crucial issues in organic chemistry, this question has been approached using different methods, which are aimed at finding repetitive fragments of molecules. It can be assumed that methods derived from text mining can be adopted to (partially) solve the task.

Chemical "words"

Let us assume that a molecule is a sentence (with some obvious caveats in mind, non-linearity of molecules being the most important one). If so, then a list of known molecules can be considered a corpus. Quite striking is the fact that a commonly used convention of describing chemical structures (referred to as SMILES) uses sequences of characters, what makes any comparisons to corpora even more natural. E.g., caffeine is coded as follows: CN-1C=NC2=C1C(=O)N(C(=O)N2C)C.

To make the language–chemistry parallel complete, one has to define "words" as well, keeping in mind that there are no explicit substructure boundaries in molecules. To this end, we adopt the idea of Cadeddu et al. (2014), who compared a few thousands of molecules pairwise, in order to extract their maximum common substructures, with the belief that they represent chemical "words"; this step was followed by a term frequency–inverse document frequency (tf/idf) heuristic.

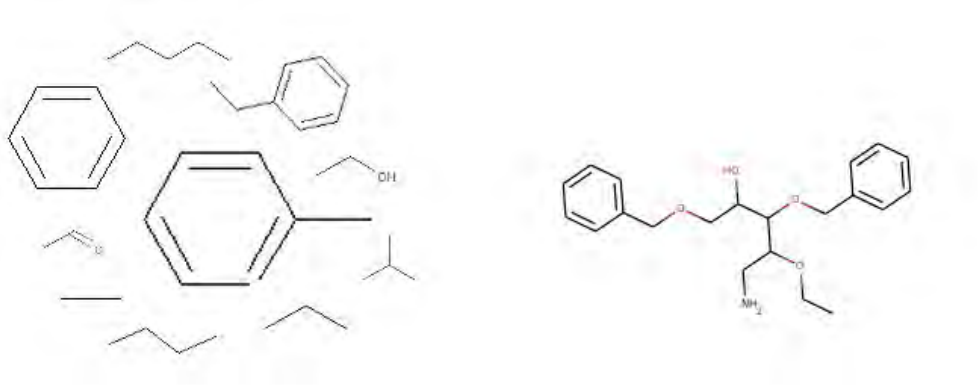


Fig. 1: Chemical "words" defined as maximum common substructures shared by chemical molecules: 10 most frequent chemical "function words" (left), and an example of an unfrequent "content word" (right).

Using the above idea of extracting "words", we picked randomly 50,000 reactions from the Reaxys database (www.reaxys.com), and computed the pairwise comparison, resulting in a corpus of >800,000 word types and 2.5 * 10^9 tokens. Interestingly enough, the chemical "words" share the characteristics of a typical natural language, e.g. they follow the Zipf's law, but they also exhibit the behavior of function and content words in their relation to frequency (see Fig. 1). Moreover, the chemical "words" can be subject to time-proven text mining methods such as keywords analysis, as has been demonstrated in our previous study (Woźniak et al., 2018).

Topic modeling

In order to identify any relations between chemical "words", we analyzed our corpus using topic modeling (Blei et al., 2003), a technique that attracted a good share of attention in Digital Humanities, but has never been popular beyond text-centric applications. Topic modeling belongs to a group of distributional semantics methods, which are based on a general assumption that the meaning of a word is defined by its lexical context (Firth, 1962). In its extended form, the distributional hypothesis says that the degree of semantic similarity between words can

be modeled as a function of the degree of overlap among their linguistic contexts (Miller and Charles, 1991; Baroni and Lenci, 2010). Topic modeling, usually computed via the LDA algorithm (Blei et al., 2003) assumes the "bag-of-words" type of context, which means that the sequence of words in a sentence is irrelevant. This feature allows for computing chemical "words", which, essentially, do not follow any linear sequence.

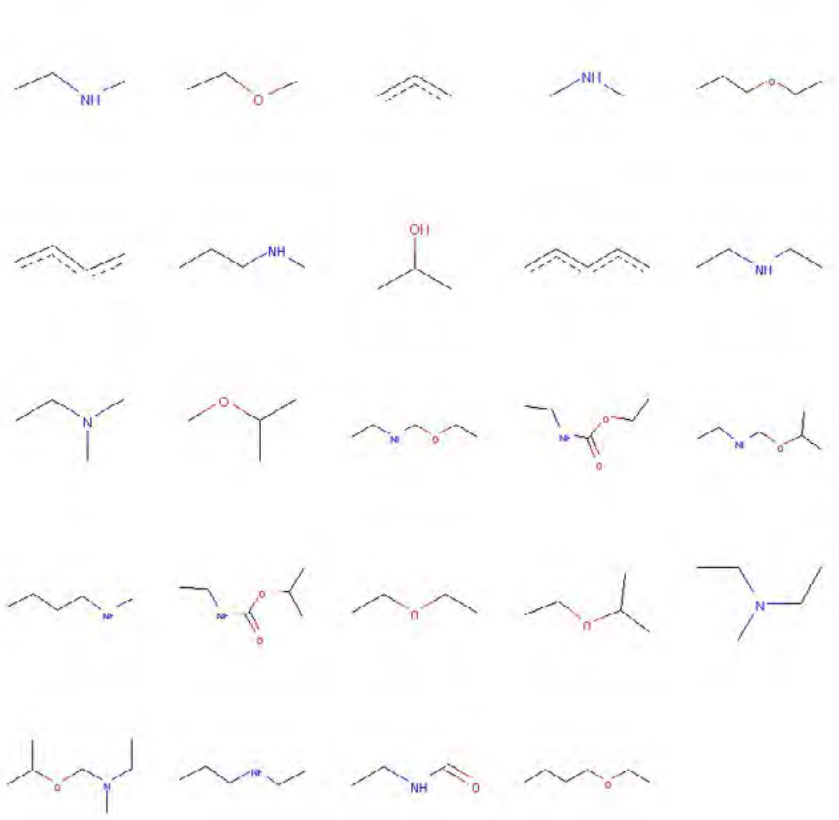


Fig. 2: Topic 47 extracted from the corpus of chemical "words".

We trained a few models ranging from 50 to 200 topics, using the LDA technique. Therefore, we were able to substantially reduce the enormous number of >800,000 "word" types into a small number of word constellations (topics) that contain meaningful information about co-occurring chemical fragments. One of the topics is shown in Fig. 2. Among the 24 most distinctive "words" one can recognize some amines, fragments of aromatic rings, fragments containg carboxyl functional groups, and so on. Inspected by trained practitioners in organic chemistry, the topics revealed several collocations that seemed meaningful, and could not have been identified in the original (raw) collection of molecules. Despite the intuitive interpretation via close-reading, however, such an outcome inevitably leads to a more serious question, namely if one can define *meaning* in organic chemistry, in the context of distributional semantics.

Classification

Interesting as they are, the chemical topics cannot solve any real-life problem *per se*, even if they seem to be meaningful from the naked eye's perspective (note that the same holds for topic modeling based on texts). Specifically, one cannot discover any general structure of, say, natural products by manual inspection of their prominent topics, nor can one predict if a given substance is likely to be toxic. There is a plethora of similar classification (or prediction) tasks where topics might prove useful, provided that the analysis goes beyond the close-reading perspective. If the topics' proportions are indeed significantly different across the corpus – i.e. if they really keep some information about semantic differentiation between the molecules – they should be applicable as a set of input features for machine-learning classification.

To test this hypothesis, we designed a controlled experiment on a (somewhat artificial) problem of classifying molecules as potential drugs. Again, we used the same Reaxys database to extract relevant training material: 1,800 known drugs and a similar number of known non-drugs. Our two-class supervised setup involved a simple neural network (implemented via Keras with Tensorflow backend), the input layer being the most probable topics for each chemical molecule. The final results varied de-

pending on a topic model used for prediction, nevertheless they turned out to be fairly optimistic. The best accuracy was: 0.7851 (the model for 200 topics), the worst: 0.7135 (the model for 50 topics). Even if preliminary, these results suggest that some semantic information can be indeed extracted from chemical corpora using text mining algorithms.

Acknowledgements

This research is part of project UMO-2014/12/W/ST5/00592, supported by Poland's National Science Centre.

References

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4): 673–721.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Cadeddu, A., Wylie, E. K., Jurczak, J., Wampler-Doty, M. and Grzybowski, B. (2014). Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angewandte Chemie*, 126(31): 8246–50.

Firth, J. R. (1962). A synopsis of linguistic theory 1930-55. In Firth, J. R., *Studies in Linguistic Analysis*. Oxford: Blackwell.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1–28.

Ruddigkeit, L., Deursen, R. van, Blum, L. C. and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11): 2864–75.

Woźniak, M., Wołos, A., Modrzyk, U., Górski, R. L., Winkowski, J., Bajczyk, M., Szymkuć, S., Grzybowski, B. and Eder, M. (2018). Linguistic measures of chemical diversity and the 'keywords' of molecular collections. *Scientific Reports*, 8: forthcoming.

Studying Performing Arts Across Borders: Towards a European Performing Arts Dataverse (EPAD)

Thunnis van Oort
t.vanoort@uva.nl
University of Amsterdam, The Netherlands

Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, The Netherlands

EPAD is an emerging network of scholars investigating the history of European performing arts (theatre, music, cinema) using digital methods and (shared) datasets. EPAD builds on the infrastructure and expertise collected in existing projects at the CREATE research program at the University of Amsterdam with a data-driven approach to the history of cinema, theatre and music performances, functioning as point of departure from where more extensive European cross-sectorial cooperation can develop.

Cultural performances in theatre, music and film have contributed vividly to the formation of individual and social identities in the European past. Cinemas, theatres and concert halls are places par excellence to examine how modern notions of identity like nation, class or gender were forged in a collective, 'live' appropriation of ideas, images and experiences (Balme, 2014; Furnée, 2012).

Traditionally, scholarship in music, theatre and film history has prioritized the study of the artwork over its consumption. Since the 1980s, the prevalent text-oriented perspectives have been complemented by a substream of historiography contextualizing the distribution and reception of performing arts (Allen & Gomery 1985; Booth 1991; Fischer-Lichte 1997; Gerhard 1992; Johnson 1995; Staiger, 1992; Weber 1975 Wollenberg and McVeigh 2004). This research tradition is dominated by qualitative approaches often based on distinct case studies, the results of which have proven hard to compare or generalize beyond the local scale (Biltereyst et al. 2018; Cowgill and Rushton 2006; Maltby 2006; Müller 2014). More advanced digital methods and larger datasets can push the research agenda beyond the prevailing particularism by providing wider comparative frameworks and new levels of generalization. Upscaling the scope yields largely uncharted possibilities for transnational perspectives on the relations between cultural consumption and the formation of shared identities (Balme 2015; Charle 2008; Garncarz 2015; Hall-Witt, 2007; Sedgwick 2000). Furthermore, EPAD's interdisciplinarity promises rare insights in the extent to which audiences of theatre, music and film overlapped and shared socio-cultural characteristics (Engelen et al. 2017; Furnée 2017; Röttger 2017).

Current historiography on the consumption of performing arts is predominantly conceived in local or national frameworks, often limited to the discipline-specific object music, theatre or film. In relative isolation, European musicologists, film and theatre scholars are confronting similar historical questions and methodological and technical issues. Joining forces opens up an agenda of transnational and cross-sectoral comparative research, that does justice to the capacity to travel across geographical, social and medium boundaries that is so characteristic of the performing arts. Moreover, data-driven historical audience research has the capacity for significant revisions of established cultural canons or genre hierarchies (Blom and Van Marion 2017; Garncarz 2015; Nieuwkerk 2017; Weber and Newark forthcoming).

Dozens of performing arts databases are scattered across Europe (Baptist et al. forthcoming). These multiform online data collections contain a variety of information on programming, and/or the venues, locations, people and organisations involved in theatrical presentation. Aggregated and combined with socio-economic data, these data can generate new insights in the social meanings of the cultural exchanges in European theatres and concert halls, for instance by delineating taste patterns and (other) socio-spatial audience characteristics.

To realize a data-driven history of the performing arts we need to join forces. The EPAD network strives to open up an exchange of expertise, data and technical know-how. To develop this research agenda, collaborating scholars need to find solutions in three (interlocking) domains:

1) address methodological-ontological questions. To facilitate comparative research into the socio-cultural dynamics of performing arts audiences, we need to reflect on the definitions of the objects of study. What exactly constitutes a performance, a venue? Can we agree on shared ontologies for structuring our data?
2) develop and refine a theoretical-historiographical framework for comparative, transnational and interdisciplinary research into the performing arts that addresses the relation between cultural consumption and social identity formation.
3) confront technical-infrastructural issues: outline the conditions for data interoperability. How can existing facilities and tools best be utilized for creating a virtual research infrastructure for comparative transnational research on the history of performing art cultures? We aim to build upon the CLARIAH infrastructure and tools for harmonizing and querying socio-economic datasets based on a linked data approach (CLARIAH Structured Data Hub). The work involves developing ontologies, shared data models and thesauri containing internationally shared terms for performing arts data, as well as building the actual infrastructure within the context of the European Digital Research Infrastructure for the Arts and Humanities DARIAH.

References

Allen, R. and D. Gomery. *Film History: Theory and Practice*. New York, 1985.

Balme, C. *The Theatrical Public Sphere*. Cambridge, 2014.

Balme, C. 'The Bandmann Circuit: Theatrical Networks in the First Age of Globalization,' *Theatre Research International* vol. 40 no. 1 (2015) pp. 19-36.

Baptist, V., T. van Oort and J. Noordegraaf. 'Mapping European Performing Arts Databases: An Inventory of Online Historical Data Projects,' in: N. Leonhardt ed. *The Routledge Companion to Digital Humanities in Theatre and Performance*. Abingdon, forthcoming.

Biltereyst, D., T. van Oort and P. Meers, 'Comparing Historical Cinema Cultures: Reflections on New Cinema History and Comparison with a Cross-National Case Study on Antwerp and Rotterdam,' in: R. Maltby, D. Biltereyst and P. Meers eds. *The Routledge Companion to New Cinema History*. Abingdon, 2018 (in press).

Blom, F. and O. van Marion. 'Lope de Vega and the Conquest of Spanish Theater in the Netherlands,' *Prolope. Anuario Lope de Vega. Texto, literature, cultura* no. 23 (2017) pp. 155-177.

Booth, M. *Theatre in the Victorian Age*. London, 1991.

Charle, C. *Théâtres en capitales. Naissance de la société du spectacle à Paris, Berlin, Londres et Vienne, 1860-1914*. Paris, 2008.

Cowgill, R. and J. Rushton. *Europe, Empire, and Spectacle in Nineteenth-Century British Music. Music in 19th-Century Britain*. Aldershot, 2006.

Engelen, L., R. Vande Winkel and L. Van de Vijver eds. Spektakelcultuur in de Lage Landen. Special Issue *Tijdschrift voor mediageschiedenis* vol. 20 no. 2 (2017).

Fischer-Lichte, E. *Die Entdeckung des Zuschauers. Paradigmenwechsel auf dem Theater des 20. Jahrhunderts*. Tübingen and Basel, 1997.

Furnée, J. *Plaatsen van beschaafd vertier. Standsbesef en stedelijke cultuur in Den Haag, 1850-1890*. Amsterdam, 2012.

Furnée, J. 'Cultuurliefhebbers. Sociale structuren en persoonlijke voorkeuren,' Inaugural Lecture Radboud University, Nijmegen (24 March 2017).

Garncarz, J. *Wechselnde Vorlieben:* Über *die Filmpräferenzen der Europäer, 1896-1939*. Frankfurt and Basel, 2015.

Gerhard, A. *Die Verstädterung der Oper: Paris und das Musiktheater des 19. Jahrhunderts*. Stuttgart, 1992.

Hall-Witt, J. *Fashionable Acts: Opera and Elite Culture in London, 1780-1880*. Hanover, 2007.

Johnson, J. *Listening in Paris: A Cultural History*. Berkeley, 1995.

Maltby, R. 'On the Prospect of Writing Cinema History from Below', *Tijdschrift voor mediageschiedenis* vol. 9 no. 2 (2006), pp. 85–7.

Müller, S. *Das Publikum macht die Musik. Musikleben in Berlin, London und Wien im 19. Jahrhundert*. Göttingen, 2014.

Nieuwkerk, M. van. 'The Felix Meritis Concert Program Database. Work-in-progress in Research and Data Curation,' Paper at CREATE ACHI Conference, October 2016, Amsterdam.

Röttger, K. 'Technologies of Spectacle and "The Birth of the Modern World": A Proposal for an Interconnected Historiographic Approach to Spectacular Culture,' *Tijdschrift voor mediageschiedenis* vol. 20 no. 2 (2017) pp. 4-29.

Sedgwick, J. *Popular Filmgoing in 1930s Britain: A Choice of Pleasures*. Exeter, 2000.

Staiger, J. *Interpreting Films: Studies in the Historical Reception of American Cinema*. Princeton, 1992.

Weber, W. *Music and the Middle Class. The Social Structure of Concert Life in London, Paris and Vienna*. London, 1975.
Weber, W, and C. Newark, eds. *The Oxford Handbook of the Operatic Canon*. Oxford, forthcoming.
Wollenberg, S. and S. McVeigh eds. *Concert Life in Eighteenth-Century Britain*. Aldershot, 2004.

The Archive as Collaborative Learning Space

Natalia Ermolaev
nataliae@princeton.edu
Princeton University, United States of America

Mark Saccomano
mss2221@columbia.edu
Columbia University, United States of America

In 2014, the Columbia University Rare Book & Manuscript Library (RBML) acquired a unique archival collection. The Serge Prokofiev Archive, which contains materials related to the twentieth-century Russian composer Sergei Prokofiev (1891-1953), contains more than 17,500 diverse items: music manuscripts, letters, financial documents, scores, concert programs, notebooks, monographs, articles, journals, photographs, audio and visual recordings, and ephemera in original, photocopy, and digital formats. The archive was first established in 1994 at Goldsmiths College, London (Mann, 2008). In the twenty years that it grew, a complex, intricate, and item-level descriptive apparatus evolved alongside. By the time the collection came to Columbia, the archival items were accompanied by hundreds of metadata files in formats such as spreadsheets, Word documents, text files, PDF, Endnote databases, Access database, MARC records, and various XML encodings. Our poster describes how we — an archivists and digital humanities researcher — curated, explored, and analyzed, this dense and diverse body of data.

Our first steps were to satisfy the immediate need of funders and stakeholders: making records of the Prokofiev Archive publicly available through the finding aid on the RBML website. Though the goal was clearly defined — records in XML using Columbia's EAD (Encoded Archival Description) schema and style guide — the process was complex. Records from Goldsmiths differed in both structure and content depending on the item catalogued. For example, data about books was captured in EndNote and MARC, while information about music manuscripts was kept in Excel spreadsheets, and correspondence records were in an Access database. We worked to transform all data into XML, and then ran customized XSLT transformations to generate standard EAD. However, what we gained in standardization we lost in information richness: this custom EAD schema didn't allow the encoding elements at the level of granularity we had in the original records. Significant scholarly information was lost. In addition, the conventional finding aid interface limits the user's options for exploring a large archival collections: content is presented in blocks of narrative, long lists of items, and search and browse organized by series and sub-series that does not allow for easy cross-collection discovery.

Thus, our next task was to find alternatives for the analysis and representation of the Serge Prokofiev Archive. We decided to pivot our approach, and moving away of EAD, transformed structured XML into a series of CSV files that could be manipulated with various data analysis and visualization tools. Not surprisingly, both the processes and our results deepened our understanding of the archive and of Prokofiev's work and legacy: an alluvial chart of the music manuscript series, for example, showed patterns in the way Prokofiev used multiple languages for different types of annotations and markings as he wrote his scores; a map using location data of Prokofiev's letters revealed his correspondence with Russian-American composers who had emigrated to China; a network graph using metadata about the secondary literature on Prokofiev (books, journal articles) showed surprising connections between editors and authors in Soviet and Western publications.

Our experience demonstrated the value of creative engagement with archival data; through experimentation and play, the Serge Prokofiev Archive became a site of collaborative research and learning. Our work was guided by two important conceptual shifts in the library and archives profession: one is the "Collections as Data" movement, which encourages reframing the digital object as data (Padilla, 2016), and the second is the move away from locating value exclusively in the *objects* of a collection to the impact collections have on *people* and *communities*. In Kate Theimer's notion of "archives as platform," tools and technologies help users interact with archives in creative ways that add value to their lives and experiences. Work that takes place "behind the scenes" (Theimer, 2014) by archivists and their collaborators helps define the archive as a dynamic cross-disciplinary learning space.

References

Noëlle Mann, "The Serge Prokofiev Archive in London - A Complex Story," *Fontes Artis Musicae*,Vol. 55, No. 3 (July-September 2008), pp. 543-547.
Thomas Padilla, "On a Collections as Data Imperative," conference report, Collections as Data: Stewardship and Use Models to Enhance Access, Library of Congress, Washington, DC, September 27, 2016,
Kate Theimer, "The Future of Archives is Participatory: Archives as Platform, or A New Mission for Archives," April 3, 2014. http://archivesnext.com/?p=3700&cpage=1#comment-4180873

Tensiones entre el archivo de escritor físico y el digital: hacia una aproximación teórica

Leonardo Ariel Escobar
leonardo.ariel11@gmail.com),
Universidad Autónoma del Estado de Morelos, Mexico

En mi investigación doctoral exploro el "archivo de escritor" como un artefacto que cambia la manera en que los lectores se relacionan con los textos de determinado escritor, entendiendo como dispositivo la: "disposición de una serie de prácticas y de mecanismos (conjuntamente lingüísticos y no lingüísticos, jurídicos [...]) con el objetivo de hacer frente a una urgencia y de conseguir un efecto" (Agamben). En este orden de ideas, una cita a Derrida es acertada y ahí radica lo arcóntico del archivo: "No solo aseguran la seguridad física del depósito y del soporte sino que también se les concede el derecho y la competencia hermenéuticos. Tienen el poder de *interpretar* los archivos" (Derrida 1997 10).

En la tesis se ha tomado como materia de estudio el archivo del escritor Gabriel García Márquez, repartido entre la Universidad de Texas en Austin y la Biblioteca Nacional de Colombia, aunque el segundo no sea muy numeroso y se trate solo de cierto material bibliográfico especifico (La Nación, 2014).

El estudio se emprende en medio de un panorama teórico que no resulta muy numeroso respecto a las definiciones del archivo de escritor. Se sabe que el archivo de escritor puede tener diferentes significados, no obstante, debemos aclarar que se tomará en la siguiente de sus acepciones dentro de esta propuesta:

> un conjunto organizado de documentos, de cualquier fecha, carácter, forma y soporte material, generados o reunidos de manera arbitraria por un escritor a lo largo de su existencia, en el ejercicio de sus actividades personales o profesionales, conservados por su creador o por sus sucesores para sus propias necesidades o bien remitidos a una institución archivística para su preservación permanente (Goldchluk y Pené 13).

Hoy en día una de las principales preguntas que se realizan a la hora de postularse a una beca de estancia en un archivo físico de escritor es justificar el porqué es obligatoria la consulta del archivo del escritor en físico, aunque se encuentre gran parte de dicho legado en forma digital (Harry Ransom Center, 2017). La idea del presente escrito es precisamente observar qué tensiones se encuentran presentes entre una y otra forma del artefacto, ya que aunque se pudiera decir que son equiparables y que equivalen a lo mismo, están lejos de cumplir una misma función en común, debido a que sustancialmente pienso que funcionan de maneras distintas. Encontrar una aproximación teórica en torno a estas tensiones es precisamente el fin de este escrito. Interesa explorar estas cuestiones porque el archivo no es un artefacto inocente, sino que: "[...] se constituye como el espacio físico que resguarda los documentos, pasando por su institucionalidad arcóntica que ejerce su poder de custodia y autoridad hermenéutica legitimadora [...]" (Nava 96).

Hay que decir que en muchas ocasiones las opciones digitales son tomadas como las más amenas, precisamente por su disposición pública y su libertad, no obstante, a través de este escrito pienso que esto debe verse con sumo cuidado:

Con el advenimiento de las tecnologías vinculadas a la información y la comunicación, y la generación de espacios virtuales donde se pueden almacenar y consultar volúmenes considerables de documentos, [...]. Entra en escena el concepto de domiciliación, definido por Derrida (1997) como el lugar donde los documentos residen de modo permanente, transitando el camino institucional que va de lo privado a lo público. Esta domiciliación implica algo más que una simple noción espacial, es el reconocimiento de ese espacio dentro de una dimensión jurídica que le asigna determinadas características específicas" (Goldchluk y Pené 14).

Así que la domiciliación de los documentos se convertiría en un primer escollo de esta problemática. Esta se hace patente sobre todo cuando se decide aquello que se digitaliza y se pone en público y qué se deja en privado, resguardado a la parte física del archivo. El domicilio se apropia de la materia de los documentos, y ésta sería una primera tensión.

En un segundo momento la domiciliación que se aúna a la desterritolización, porque aquello que se posee está localizado y resguardado y solo se consulta con permiso institucional. La segunda de las tensiones que se presentan entre una y otra forma del archivo pienso que va por el lado de la desterritolización de las literaturas, precisamente porque opera un dispositivo, es decir, una conjunción entre el poder y la institucionalidad (Agamben). En últimas es un ejercicio de poder el que determina qué nación se apropia de un archivo. Dicha desterritolización no se presenta únicamente con nuestras literaturas, también pasa lo mismo con otras literaturas, por ejemplo, sobre los diversos ejemplares literarios del dadaísmo francés (Iowa University), de tal manera que la labor arcóntica de los archivos estadounidenses ha estado presente desde hace algún tiempo, y va en aumento constante.

Se conoce que un archivo de escritor está básicamente poblado de documentos y es obvio que el presente debate también va en la vía de las tensiones y la actualización obvia de dicho concepto. Se puede decir en cierta media que los verdaderos documentos se encuentran en la versión física y que muchas veces los archivos digitales se limitan a ser solo una muestra. Esto es notorio en la descripción que se puede leer en la página del Ransom Center sobre el archivo de García Márquez: de más de 1000 documentos guardados solo 33 están para la consulta pública en línea. De esta manera queda difícil emprender una labor como la que propone Foucault:

> ahora bien, por una mutación que no data ciertamente de hoy, pero que no está indudablemente terminada aún, la historia ha cambiado de posición respecto del documento: se atribuye como tarea primordial, no el interpretarlo, ni tampoco determinar si es veraz y cuál sea su valor expresivo, sino trabajarlo desde el interior y elaborarlo. […] (Foucault 9-10).

Así, ¿cómo es esto posible si los archivos no se poseen? Es la pregunta que queda en el aire para nuestra propia tradición crítica.

En tercer lugar, encuentro que los países que no tienen en su poder los archivos de sus escritores tienen menos opciones de poder proceder a ediciones críticas de sus literaturas que tengan en cuenta el modelo de la genética textual, puesto que dichos manuscritos y demás son tenidos en cuenta meramente como materia para especialistas que se puedan desplazar hasta estos lugares de consulta, la mayoría de las veces más accesible para aquellos que se encuentren dentro del ámbito lingüístico al que pertenece el archivo. Un ejemplo claro de esto es el documento de las galeradas corregidas de la versión de conmemoración que hizo la RAE en el año 2007 de *Cien años de soledad*, si no fuera por estas galeradas que reposan en Austin, entonces no sabríamos los cambios (casi imperceptibles) que tuvo la novela en su edición revisada, lo cual sería una tarea titánica de comparación de ediciones (Harry Ransom Center, 2017). Se piensa así que el acceso a los archivos físicos da mayor opción a cierta actualización editorial de la obra.

Para señalar la última de las tensiones, citamos de nuevo a Foucault: "reconstituir, a partir de lo que dicen esos documentos – y a veces a medias palabras- el pasado del que emanan y que ahora ha quedado desvanecido muy atrás de ellos; el documento seguía tratándose como el lenguaje de una voz reducida ahora al silencio: su frágil rastro, pero afortundamente descrifrable (Foucault, 9).

Dicha idea de Foucault está muy conectada precisamente con la idea de iterabilidad de Derrida, es decir, poder reconstruir el enunciado del emisor aunque no se cuente con su presencia:

La posibilidad de repetir, y en consecuencia, de identificar las marcas está implícita en todo código, hace de éste una clave comunicable, transmisible, descifrable, repetible por un tercero, por tanto por todo usuario posible en general. Toda escritura debe, pues, para ser lo que es, poder funcionar en la ausencia radical de todo destinatario empíricamente determinado en general (Derrida 1998 364).

Obviamente si no se tiene un acceso físico a los archivos, la capacidad de su iterabilidad se desvanece, sobre todo en lo que tiene que ver con la genética de los textos. Al ordenar y definir qué se da al público y qué se conserva privado se está resguardando de cierta forma la capacidad de iterabilidad que podría tener tal documento, en ese orden de ideas, su capacidad de iterable se disminuye. ¿Hasta qué punto son más iterables aquellas obras que se ponen en público y en digital y aquellas a las que se les guarda con más celo?

Finalmente, lo que se quiere lograr con esta aproximación es observar qué contrastes existen entre ambas formas de presentación del archivo de escritor y las diversas tensiones que se producen entre una forma y otra de presentación de los archivos, a pesar de su supuesto carácter de equiparabilidad.

References

Agamben, Giorgio. "¿Qué es un dispositivo?". *Arte y pensamiento*. Universidad Internacional de Andalucía. Web. 20 de noviembre de 2017. http://ayp.unia.es/r08/IMG/pdf/agamben-dispositivo.pdf

"Colombia: polémica por la venta del archivo personal de Gabriel García Márquez a la Universidad de Texas". La Nación. 24 de noviembre de 2014. Web. http://www.lanacion.com.ar/1746618-colombia-polemica-por-la-venta-del-archivo-personal-de-gabriel-garcia-marquez-a-la-universidad-de-texas

Derrida, Jacques. "Firma, acontecimiento, contexto". Márgenes de la filosofía, Madrid, Cátedra, 1998, pp. 347-372. Impreso.

Derrida, Jacques. *Mal de archivo: una impresión freudiana*. Madrid: Trotta, 1997. Impreso.

Foucault, Michel. *La arqueología del saber*. Buenos Aires: Siglo XXI Editores, 2002. Impreso.

Goldchluk, Graciela y Pené, Mónica Gabriela. "Archivos de escritura, génesis literaria y teoría del archivo". Repositorio institucional. Universidad Nacional de la Plata. Web. 20 de noviembre de 2017. http://www.memoria.fahce.unlp.edu.ar/trab_eventos/ev.772/ev.772.pdf

Harry Ramson Center. "2018–2019 Research Fellowships Application Instructions". Universidad de Texas en Austin. Web. 20 de noviembre de 2017. http://www.hrc.utexas.edu/research/fellowships/application/

Harry Ramson Center. "Gabriel García Márquez: Un Inventario de sus documentos en el Harry Ransom Center". Universidad de Texas en Austin. Web. 20 de noviembre de 2017. http://norman.hrc.utexas.edu/fasearch/findingAid.cfm?eadid=01084

Iowa University. "Digital library". Web. 20 de noviembre de 2017. http://digital.lib.uiowa.edu/

Nava Murcia, Ricardo. "El mal de archivo en la escritura de la historia". *Historia y grafía*, núm. 38, enero-junio 2012, pp. 95-126. Impreso.

Using Linked Open Data To Enrich Concept Searching In Large Text Corpora

Christine Fernsebner Eslao
eslao@fas.harvard.edu
Harvard Library, United States of America

Stephen Osadetz
osadetz@fas.harvard.edu
Harvard University, United States of America

This poster presents the library metadata aspects of a web-based text mining application for sifting corpora of unstructured text in order to find particular passages that deal with a concept of interest. In addition to overcoming the limitations of vendor-supplied search platforms, which tend to be based on simple keyword searches that place the burden of interpreting, refining, and iterating on search results on the laborious grunt work of scholarly users (De Bolla, 2013), this tool demonstrates the utility of reconciling named entities with external structured data to refine its results and to enrich its output for use in research, visualizations, and secondary analytic tools by leveraging demographic (Hwang, 2015), temporal, and geographic data from the linked open data cloud. This necessitates the creation of entity resolution workflows with both automated matching tools and practices for manual reconciliation and maintenance, exploring a variety of open-source tools including OpenRefine (Van Hooland, 2014; Hwang, 2017), Python, and Mix'n'Match (Knoblock, 2017) and contributing to the development of "functional requirements for how [library] systems use and maintain these identifiers and associated data" (Folsom, 2017) by metadata librarians and researchers and "the complexities inherent in managing both locally-created and externally-assigned identifiers" in the context of library infrastructure (Tarver, 2017). Our goal is to integrate a tool catering to advanced researchers into library discovery platforms by "[exploring] partnerships with external entities to create game changing discovery" (Wones, 2017) and leveraging those users' domain expertise to "interrogate corpora of resources directly ... to discover new patterns that exist across the literature, perform their own ranking of relevance against particular parameters, and find new pathways for discovery more efficiently than could be enabled through existing information portals" (MIT Libraries, 2016). The process is as follows:

1. Combine vendor metadata for large corpora with bibliographic metadata from Harvard Library collections
2. Reconcile authors, including persons and organizations, in those metadata resources, with external URIs, including those of ISNI (International Standard Name Identifier), Wikidata, and Geonames entities, generating batches of new entities in external resources at scale as needed (Mika, 2017)
3. Integrate data from external URIs into a text mining tool for sifting large corpora to drive filters and enrich data extracted from that tool
4. Work with library technology staff and metadata librarians to facilitate retrieval of rare materials in Harvard Library collections, as well as their electronic reproductions, based on results of text mining tool and integration of URIs in library metadata
5. Export resulting data to produce visualizations and secondary analytic tools

Through this process, we hope to enable the serendipitous discovery (Bourg, 2017) of relevant but unknown works in library collections: traditional reading of the "great unread" (Cohen, 1999) facilitated by distant reading (Moretti, 2013). Our poster includes: an explanation of the linked data principles underlying the metadata aspects of the text mining tool, our entity reconciliation workflow, implications for library metadata and name authority practices in support of digital research projects, and an example of combined and enriched metadata for a work of eighteenth century literature, and an example of an iterative concept search and its output presented both as a static flowchart on the poster as well as an interactive prototype on a laptop.

References

Bourg, Chris. (2017). *Serendipity as prick* https://chrisbourg.wordpress.com/2017/02/11/serendipity-as-prick/ (accessed 18 November 2017).

Cohen, Margaret. (1999). T*he Sentimental Education of the Novel*. Princeton, N.J.: Princeton University Press.

De Bolla, Peter. (2013). *The architecture of concepts : The historical formation of human rights* (First ed.). New York: Fordham University Press.

Folsom, Steven. (2017). New Models Require New Action Plans: Implementing Linked Data within the PCC. PCC (Project for Cooperative Cataloging) *Strategic Planning Meeting Keynote*, 1 November 2017. https://docs.google.com/presentation/d/11DHY-Ry24F4aQjYbPsVmInO2ovcb_pujBPIwJBQ0RrAQ/edit?usp=sharing (accessed 27 November 2017).

Hwang, Karen. (2015). *Enriching the Linked Jazz Name List with Gender Information* https://linkedjazz.org/enriching-the-linked-jazz-name-list-with-gender-information/ (accessed 1 November 2017).

Hwang, Karen. (2017). *Using OpenRefine to Reconcile Name Entities* http://mnylc.org/fellows/2017/03/17/using-openrefine-to-reconcile-name-entities/ (accessed 10 October 2017)

Knoblock, C.A., et al. (2017). Lessons Learned in Building Linked Data for the American Art Collaborative. In: d'Amato C., et al. (eds) *The Semantic Web – ISWC 2017 : 16th International Semantic Web Conference*, Vienna, Austria, October 21-25, 2017, Proceedings, Part II (Lecture Notes in Computer Science, 10588). Cham: Springer International Publishing : Imprint: Springer.

Mika, Katie. (2016). *The Role of Librarians in Wikidata and Wikicite*. https://library.mcz.harvard.edu/blog/role-librarians-wikidata-and%C2%A0wikicite (accessed 1 November 2017).

MIT Libraries, *Ad Hoc Task Force on the Future of Libraries*. (2016). Institute-Wide Task Force on the

Future of Libraries—Preliminary Report https://future-of-libraries.mit.edu/sites/default/files/Future-Libraries-PrelimReport-Final.pdf (accessed 25 November 2017).
Moretti, Franco. (2013). *Distant Reading*. London: Verso.
Tarver, Hannah, & Phillips, Mark. (2017). *Identifier Usage and Maintenance in the UNT Libraries' Digital Collections* http://dcevents.dublincore.org/IntConf/dc-2016/paper/download/458/546 (accessed 27 November 2017).
Van Hooland, Seth, & Verborgh, Ruben. (2014). *Linked data for libraries : How to clean, link and publish your metadata*. Chicago, IL: Neal-Schuman.
Wones, Suzanne. (2017). *Harvard Library Digital Strategy, Version 1.0*. http://projects.iq.harvard.edu/files/overseers/files/vc_3_hl_digital_strategy_v2.pdf (accessed 10 November 2017).

Pontes into the Curriculum: Introducing DH pedagogy through global partnerships

Pamela Espinosa de los Monteros
espinosadelosmonteros.1@osu.edu
Ohio State University Libraries, United States of America

Joshua Sadvari
sadvari.1@osu.edu
Ohio State University Libraries, United States of America

Maria Scheid
scheid.31@osu.edu
Ohio State University Libraries, United States of America

This poster proposes a discussion on the challenges and lessons learned in the integration of digital humanities pedagogy into a traditional graduate foreign language course through a heterogeneous collaboration among global DH scholars and North American experts in geographic information systems (GIS), copyright, digital humanities, and area studies. Significant barriers of entry exist for humanities students and faculty attempting to introduce DH into their departments and classrooms. Uneven institutional infrastructure and programmatic presence of DH at universities leave faculty with the dilemma of simultaneously learning DH methods themselves and integrating DH pedagogy into the curriculum for their students. As such, DH methods can present real and perceived psychological and cultural barriers (Battershill and Shawna, 2017) that surpass the digital and technology competencies of students or faculty. Humanities departments recognize the value of DH methods, research, and the need to develop the next generation of DH scholars, but may lack in-house expertise to design the initial curriculum.

Partnership with the community of DH scholars, and the research library, may assist faculty member to overcome technological infrastructure and subject expertise lacking in their own departments. The proposed poster and case study will highlight a team based approach to introduce DH research and DH curriculum from the Lusophone world. Three DH methods were introduced including text-analysis, GIS, and text-encoding and transcription. Each method was paired with course content, DH literature, mediated exploratory assignments, and current DH research by scholars in the field. Sessions were team taught in workshop and lecture settings to provide students with both experimental learning models and theoretical background. DH curriculum was customized to meet the subject content of the Portuguese literature course and taught in both English and Portuguese.

The most significant and time-intensive DH assignment students completed during the course was the collaborative creation of an ArcGIS Story Map on the African diaspora of Lisbon. With the advent of web-based mapping platforms, user-friendly on-ramps exist for humanities scholars to integrate geovisualization and location-based storytelling into their research (Presner & Shepard, 2016), and this assignment was designed for students to recognize the utility of such a platform for their own work. After a brief introduction to some key GIS concepts and a hands-on tutorial, students collaboratively identified images and text associated with course topics to overlay points on a georeferenced historical map of Lisbon. In this way, students combined a growing knowledge of course subject matter, copyright considerations when identifying and incorporating suitable content, and newly-developed digital mapping skills to create an end product that differed from the more traditional written paper to which they might be accustomed. In collaboration with the faculty instructor, adjustments were made throughout the project to accommodate humanities students' varying levels of technical and information literacy proficiencies in the classroom.

In this poster, we will address challenges faced by the team to blend and balance traditional and DH pedagogy, multilingual limitations of existing DH tools, and design of an exploratory assignment with specific disciplinary content. A focal point of the poster will be the role of each participant and timeline for the project's implementation. By sharing our experiences in developing this introductory intervention, we hope to explore with attendees the ways in which DH methods, tools, and dispositions can be introduced into traditional foreign language humanities courses. This poster will outline lessons learned and promote discussion on unique challenges in curriculum design, collaborative instruction, and delivery of GIS DH instruction for a foreign language course.

References

Presner, T., & Shepard, D. (2016). Mapping the geospatial turn. In S. Schreibman, R. Siemens, & J. Unsworth

(Eds.), *A new companion to digital humanities*. Chichester, UK: John Wiley & Sons, Ltd., pp 201-212.
Battershill, C., & Ross, S. (2017). *Using digital humanities in the classroom: a practical introduction for teachers, lecturers, and students*. London: Blomsbury Academic.

Milpaís: una wiki semántica para recuperar, compartir y construir colaborativamente las relaciones entre plantas, seres humanos, comunidades y entornos

María Juana Espinosa Menéndez
mj.espinosam@uniandes.edu.co
Universidad de los Andes, Colombia

Camilo Martinez
gemartin@uniandes.edu.co
Universidad de los Andes, Colombia

Milpaís, proyecto de tesis para la maestría en Humanidades Digitales de la Universidad de los Andes de Colombia, nace como iniciativa del colectivo Savias y Sabias quienes en sus trabajos con comunidades expertas y no, han encontrado la necesidad de apropiarse de herramientas digitales que permitan democratizar el acceso al conocimiento experto sobre plantas, visibilizar el conocimiento tradicional y local y sobre todo defender este saber en tanto bien común (Bollier, 2016; Zuluaga Ramírez, 1994). Con especial énfasis discutimos los aspectos éticos y legales que tuvimos que sopesar al formular este trabajo en Humanidades Digitales sobre conocimientos tradicionales en el contexto global y en particular en el caso colombiano (Organización Mundial de la Propiedad Intelectual (OMPI), 2017; Gómez Madrigal, 2013). Al respecto, debimos considerar estrategias para prevenir la expropiación indebida de conocimientos mediante la definición de la catalogación y la visibilización en la wiki de los territorios, las personas y comunidades que cuidan, siembran y trabajan con las plantas. Este mapeo permite construir elementos probatorios de la pertenencia cultural de conocimientos colectivos circunscritos a territorios.

El *prototipo se desarrolla a partir de una wiki semántica del software libre Media Wiki* (*semantic-mediawiki.org*, 2018) para la gestión del conocimiento etnobotánico de comunidades que usan, defienden y comparten saberes sobre las plantas. Diseñada a partir de una reflexión ética y legal de lo que implica documentar, catalogar y difundir conocimientos tradicionales y locales, el prototipo cuida de no exponer contenidos susceptibles de expropiación indebida tales como componentes, fórmulas, técnicas y rituales. Es así que en este prototipo nos interesa conectar qué lugares, qué personas (comunidades) y de qué maneras se construyen las relaciones con las plantas, entendidas estas como uno de los bienes comunes que sostienen y equilibran entornos como el cuerpo y el medio ambiente (Lafuente, 2007).

En términos técnicos, la SMW permite estructurar una Base de Datos Relacional (BDR) mediante el uso de plantillas que integran notación semántica y vocabularios controlados. Para esta wiki utilizamos el estándar de metadatos FOAF(*The FOAF Project*, 2018) y un conjunto de metadatos propios y vocabularios controlados alimentados de diversas fuentes de catalogación etnobotánica (Royal Museum From Central Africa, 2017; *BRIT - Native American Ethnobotany Database*, 2003). Igualmente, se ha tenido y se tendrá en cuenta la información que colaboradores y posibles usuarios han reportado necesaria. Para recuperar la información y que se integre la notación semántica, el prototipo implementa los formularios de Semantic Media Wiki (*Page Forms - MediaWiki*, 2018). Estos formularios permiten a los colaboradores/creadores de la wiki ingresar la información mediante una interfaz amable sin necesidad de hacer notación semántica manual. Una vez se ingresa la información la SMW permite recuperar información relacional (qué personas son amigos de una planta, qué plantas sirven a las personas para hacer artesanía, qué comunidades resguardan una semilla en particular, quiénes y dónde hay médicos tradicionales, yerbateras, investigadores, médicos alópatas que trabajan con plantas, etc.) así como visualizar datos tales como los geográficos.

Finalmente, el prototipo tiene una fase piloto anterior a la implementación (2018-II) en la cual empezamos a trabajar la campaña de difusión "Adopta una planta y cultiva su conocimiento en la web". Dicha estrategia se enmarca en el trabajo que se adelanta con comunidades potencialmente usuarias en zonas aledañas y urbanas de Bogotá y en la cual se llevó a cabo un primer rastreo sobre la información que consideran importante documentar, compartir y defender. Así, la herramienta digital se dispondrá al servicio de procesos educativos con comunidades que quieran intercambiar saberes, investigar y defender el conocimiento tradicional y local sobre las plantas.

References

Biblioteca digital de la medicina tradicional mexicana (2009). Available at: http://medicinatradicionalmexicana.unam.mx/index.php (Accessed: 27 October, 2017).
Bollier, D (2016). *Pensar desde los comunes*. Madrid: Traficantes de Sueños.
BRIT - Native American Ethnobotany Database (2003). Available at: http://naeb.brit.org/ (Accessed: 27 March 2017).
Gómez Madrigal, L. S. (2013). *Protección de la tradición. Los derechos no tradicionales de la propiedad intelectual. Comité intergubernamental de recursos genéti-*

cos, conocimientos tradicionales y folclore de la OMPI. *Revista La Propiedad Inmaterial*. Available at: http://revistas.uexternado.edu.co/index.php/propin/article/view/3581/3798 (Accessed: 2 February 2017).
Lafuente, A. (2007). Los cuatro entornos del procomún. *Archipiélago*: *Cuadernos de crítica de la cultura*, (77), pp. 15–22. Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=2500491 (Accessed: 11 September 2017).
Organización Mundial de la Propiedad Intelectual (OMPI) (2017). *Guía para la catalogación de conocimientos tradicionales*, *OMPI*. Available at: http://www.wipo.int/edocs/pubdocs/es/wipo_pub_1049.pdf (Accessed: 8 May 2017).
Page Forms - MediaWiki (2018). Available at: https://www.mediawiki.org/wiki/Extension:Page_Forms (Accessed: 27 April 2018).
Royal Museum From Central Africa (2017). *Prelude*, *Medicinal Plants Data Base*. Available at: http://www.africamuseum.be/collections/external/prelude/plant_collection (Accessed: 27 April 2018).
The FOAF Project (2018). Available at: http://www.foaf-project.org/ (Accessed: 5 February 2018).
Semantic-mediawiki.org (2018). *Semantic MediaWiki*. Available at: https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki (Accessed: 5 March 2018).
Zuluaga Ramírez, G. (1994). *El aprendizaje de las plantas*: *en la senda de un conocimiento olvidado*. Bogotá: Seguros Bolívar.

Cataloging History: Revisualizing the 1853 New York Crystal Palace

Steven Lubar
lubar@brown.edu
Brown University, United States of America

Emily Esten
emily_esten@brown.edu
Brown University, United States of America

Steffani Gomez
steffani_gomez@alumni.brown.edu
Brown University, United States of America

Brian Croxall
brian.croxall@brown.edu
Brown University, United States of America

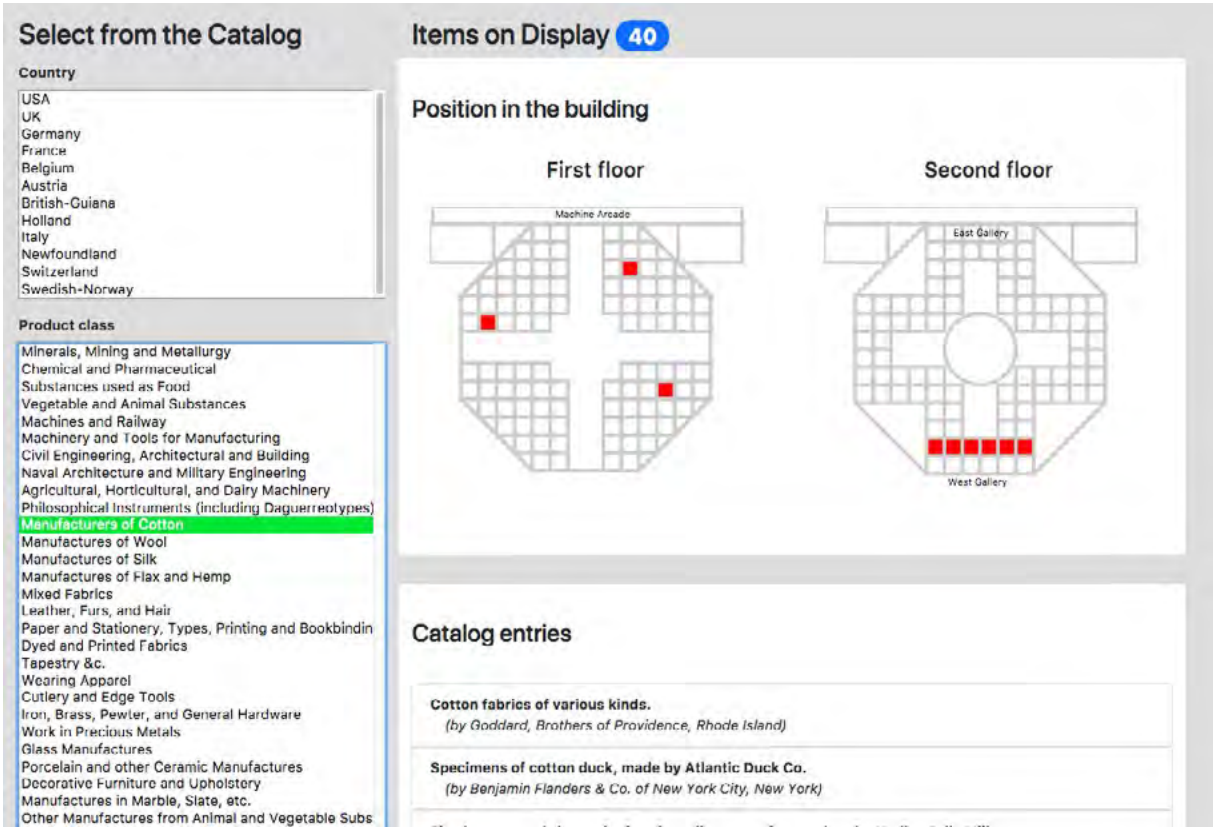
Patrick Rashleigh
patrick_rashleigh@brown.edu
Brown University, United States of America

The 1853 New York Crystal Palace, also known as the Exhibition of the Industry of All Nations, was the center of America's first World's Fair. Modeled on the Great Exhibition held at the Crystal Palace in London, the exhibition sought to "draw forth such a representation of the world's industry and resources as would enable us to measure the strength and value of our own, while it indicated new aims for our enterprise and skill." While the exhibition burnt down by the end of the decade, it survives through catalogues documenting its success and breadth.

This poster considers the catalogues from the 1853 New York Crystal Palace exhibition in physical and digital forms: as book, file, and database. Originating with the Museum Wormianum and the Louvre, catalogs were published by early museums to visualize and document their work for the world. The New York Crystal Palace exhibition generated multiple print publications to record the museum-like experience through databases and narratives. In the digital era, however, these traditional exhibition catalogs can be used in new ways. As both a physical and digital object, the cataloging forms encourage and allow relationships among user, data, and experience to come to fruition. This project created a dataset and subsequent database from a digitized copy of the New York Crystal Palace catalog to explore the artifacts documented inside. Explored through digital tools such as OpenRefine, Tableau, Palladio, D3.js, and Google Fusion, the Crystal Palace catalogs aid us in viewing catalogs, and their modern database descendants, more generally. How can looking at the Crystal Palace through digital tools let us see not only what others have seen, but also to see things better, see things differently?

Databases, by their nature, lend themselves to exploration – the 1853 Crystal Palace exhibition catalog is no exception. Creating a database from the catalog frees the information inside. The curators of the exhibition thought hard about the best way to arrange the *Official Catalog*, but once they decided, it couldn't be changed. Now, the data is alive and fluid. It can be analyzed, represented in new ways. It can be searched, sorted, faceted, mapped, and turned into networked and nodes. Revisualized and revitalized in digital scholarship, this database and subsequent visualizations offer non-narrative perspectives to the exhibition's construction and imagined possibilities. Contributing both to historical and museological scholarship, *Cataloging History* considers histories of technology through the catalog to understand exhibition construction and the ways in which we can reconstruct it in the digital age.

Ultimately, visualizations of any kind open up the catalog to a new form of interrogation. But they also ask us to reimagine the relationships connections embedded inside. What if we wanted to re-curate the Crystal Exhibition by role, showing off inventors and agents in the major divisions? What if we wanted to use this as an opportunity to tease out specific class categories? What if we wanted to reorganize the catalog by class first, instead of country? Visualizations offer us the ability to think through both



the construction of the exhibition and the catalog, while also allowing us the chance to reconstruct its data to new ends. We can use the database and visualizations as way to look into the past, evolving this nineteenth-century exhibition along with new forms oft he catalogue.

Developed in cooperative collaboration with representatives of Brown's Center for Digital Scholarship, *Cataloging History* examines the ways in which traditional museum data can be mobilized to reimagine historical spaces. Using the catalogue as a piece of technology for understanding the past, it also opened a new dialogue for thinking about catalogues of the future. Building on conversations around "collections as data," this project uses a historical example to pose to both scholars and museums about how cultural heritage may work to be more readily open to computation. How does digital humanties help us unpack historical collections? These visualizations highlight how digital tools can unearth relationships among data to better understand what was there. *Cataloging History* challenges us to think more deeply about what information is contained in a catalog, about what remains when an exhibition is gone, and about how datasets and tools like these promote the evolution of exhibitions.

References

Croxall, B., Esten, E., Gomez, S., Lubar, S., and Rashleigh, P. (2017). 1853 New York Crystal Palace accessed April 25, 2018. http://cds.library.brown.edu/projects/crystalpalace/.

Esten, E. (2017). "Visualizing the Crystal Palace." https://github.com/sheishistoric/Visualizing-the-Crystal-Palace.

Lubar, S. (2017). "A brief history of American museum catalogs to 1860." https://medium.com/@lubar/cataloging-history-eac876941db6.

Lubar, S. and Esten, E. (2017). "Catalog as Book, File, and Database." https://medium.com/@lubar/catalog-as-book-file-and-database-ac954096152e.

Lubar, S. (2017). "The New York Crystal Palace Catalogs." https://medium.com/@lubar/the-new-york-crystal-palace-catalogs-b09d1f2bd20e.

Lubar, S. and Esten, E. (2017). "Revisualizing the Crystal Palace." https://medium.com/@lubar/revisualizing-the-crystal-palace-d239e50d9e12.

New York Exhibition of the Industry of All Nations, New York, N.Y. (1853). *Official Catalogue of the New-York Exhibition of the Industry of All Nations. 1853*. New York: G.P. Putnam & Co.

Crowdsourcing Community Wellness: Coding a Mobile App For Health and Education

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Michael Thompson
michael.thompson@bucknell.edu
Bucknell University, United States of America

Jacob Mendelowitz
jpm061@bucknell.edu
Bucknell University, United States of America

Caroline Whitman
alw001@bucknell.edu
Bucknell University, United States of America

Shaunna Barnhart
sb060@bucknell.edu
Bucknell University, United States of America

In response to the widely reported increase in obesity and related health problems in the US, a team of faculty, staff, and students at Bucknell University have authored a mobile app that incentivizes exercise through the use of crowdsourced public-facing humanities content of local interest. ReadySetFit, available on both Apple and Android phones, is a completely student-coded app that leverages a Google Maps platform and the Google My Maps application. (http://www.readysetfitapp.org) The user can select from a set of walking paths that have been created using the Google My Maps app, which contain points of interest that present cultural/historical information to the user as he or she approaches the physical location of each point. Once a user has reached all points of interest, or manually clicked a button to finish a workout, the distance covered is saved to the handheld device and can be reviewed at a later date.

Key components of the success of ReadySetFit have been the ease of use and the localized and crowdsourced nature of the information provided. Griffiths and Barbour (2016) argue that the creation of "smart cities" greatly enhances the sense of place among local citizens. Our university collaboration with a local civic group (The Improved Milton Experience) in the post-industrial central Pennsylvania town of Milton has engaged in local history through crowd-sourcing content for specific points of interest while incentivizing citizens to walk around the town. Users receive rewards and discounts at local shops when they earn "Milton Bucks" by walking on set paths in the borough.

Furthermore, partnering with the statewide system of parks (DCNR) and its "Think Outside" higher-education partnership program has promoted the app to a wide user-base who are already visiting the parks but who want to know about the history and environment through which they are walking. (http://www.dcnr.pa.gov/Education/ThinkOutside/Pages/default.aspx) Newly launched to the public, ReadySetFit has shown potential to overcome the major obstacle to maintaining an exercise routine--incentive (Harris and Roushanzamir 2014: Conroy et al) 2014. The app's incentive is multi-dimensional: engaging with new and interesting place-based content in realtime, collecting completed pathways, obtaining fitness levels for financial rewards through local business partnerships, and contributing to the creation of new pathways. Through crowdsourcing content, user participation promotes both individual wellness and community buy-in. The place-based content that is provided to the user is created by members of the community and fosters active engagement in creating a sense of place (Lepofsky and Fraser, 2003). The poster presentation will demonstrate the app itself and also show the process undergone by the students in terms of technology and content development. We will also demonstrate the path creation-guidelines that have been shared with local organizations and can be adopted for creating pathways anywhere in the world with cellular data connectivity.

References

Conroy, David E., Chih-Hsiang Yang, Jaclyn P. Maher, "Behavior Change Techniques in Top-Ranked Mobile Apps for Physical Activity", In American Journal of Preventive Medicine, Volume 46, Issue 6, 2014, Pages 649-652, ISSN 0749-3797, https://doi.org/10.1016/j.amepre.2014.01.010.

Griffiths Mary and Kim Barbour. "'Imagine If Our Cities Talked to Us': Questions about the Making of 'responsive' Places and Urban Publics." In Making Publics, Making Places, 27-48. South Australia: University of Adelaide Press, 2016, http://www.jstor.org/stable/10.20851/j.ctt1t304qd.8

Harris,Felicia and Elli Lester Roushanzamir. "#Blackgirlsrun: Promoting Health and Wellness Outcomes Using Social Media." Fire!!! 3, no. 1 (2014): 160-89. doi:10.5323/fire.3.1.0160.

Leipert, Beverly D., Belinda Leach, and Wilfreda E. Thurston, eds. Rural Women's Health. Toronto; Buffalo; London: University of Toronto Press, 2012. http://www.jstor.org/stable/10.3138/j.ctt2tv021.

Lepofsky, Jonathan, and James C. Fraser. "Building Community Citizens: Claiming the Right to Place-making in the City." Urban Studies 40, no. 1 (2003): 127-42. http://www.jstor.org/stable/43084177

Bad Brujas Only: Digital Presence, Embodied Protest, and Online Witchcraft

Amanda Kelan Figueroa
browngirlsmuseumblog@gmail.com
Harvard University, United States of America

Ravon Ruffin
ravonruffin@gmail.com
National Museum of African American History and Culture, United States of America

Over a year before Donald Trump took the presidency in 2016, a group of self-identified Cuban brujas, latina practitioners of witchcraft and/or indigenous rituals, led by Yeni Sleidi released an online video titled "Brujas Hex Trump." From this platform on YouTube, the video called for fellow witches in both the digital and physical realms to intervene in the presidential candidate's campaign through a type of activism not previously considered political — ritual, witchcraft, hexing. Since this initial call via YouTube, monthly hexes have continued among Brooklyn-based latinas, organized via social media.

The organization of social justice activism through interpersonal networks on Facebook, Twitter, Instagram, and other platforms is not an unusual phenomenon for marginalized communities, evidenced by such movements as #BlackLivesMatter and #MeToo. However, the model of integration between online and offline practices demonstrated in social media witchcraft or brujería communities is worthy of note, as a reclamation of the female-identified body and indigeneity in this current political climate. Witchcraft in its traditional forms would seem to be the antithesis of digital media due to its emphasis on materiality, embodied presence, and physically-enacted rituals. However, these networked communities of digital brujas transcend this divide, as a politicized tool for empowerment, and decolonization of history and the female body.

Groups like Brujas Hex Trump capitalize on the ability of personal practices to have overarching political impacts, while an organization called Witch Cabinet creates workshops and digital courses for femme- and queer-identified people to learn self-care through magic ritual, and social media astrologer Danielle Ayoka gives horoscopes and personalized tarot readings via Twitter and Instagram. The intersection of the embodied rituals of witchcraft and the digital space of social media appear to be irreconcilable, and for this reason, digital expressions of witchcraft and magic are widely considered to be cheap, commercialized, or inconsequential. However, we will examine these points of apparent conflict, between medium and message that occur in these examples of witchcraft, in order to demonstrate a method for seeing the social media space as a mediator, and not an obstacle, for these practices.

Using a theoretical frame based on Chela Sandoval's work on dissident coalition building, Chon Noriega's understanding of museological power structures, and the investments of black digital studies in a radical black archival practice, we build from embodied theories of ethnic studies and art ecosystems to find a method for considering race and ritual in the digital sphere. Undertaken primarily by women of color, digital witchcraft is successful at translating online presence into embodied action in ways that perhaps offer strategies for other social, institutional, and cultural communities and their activism. By studying the ways in which these digital witchcraft communities make use of social media platforms in order to bridge these divides, sometimes by using them against their designed purposes, the potential of digital activism, and its implications for studies of chicana and black feminism, indigenous studies, and other branches of ethnic studies, digital or otherwise, can be considered.

La geopólitica de las humanidades digitales: un caso de estudio de DH2017 Montreal

José Pino-Díaz
jpinod@uma.es
Universidad de Málaga, Spain

Domenico Fiormonte
domenico.fiormonte@uniroma3.it
Università Roma Tre, Italy

Resumen: Las discusiones y reflexiones sobre los desequilibrios culturales, políticos, lingüísticos y de género en las Humanidades Digitales se han concentrado sobre aspectos generales y, con pocas excepciones (Dacos, 2016; Fiormonte 2017b; Grandjean, 2014; Weingart 2014; Weingart and Eichmann-Kalwara 2017), no han ofrecido un análisis de datos y casos concretos. Nuestra propuesta intenta aportar una contribución al debate a través del análisis de las 420 colaboraciones del congreso DH2017 de Montreal. Los datos archivados nos permitieron realizar mapas de colaboración entre países y entre países y centros académicos o de investigación; así como mapas de temas de investigación (palabras clave) y de redes de autores. El resultado es una imagen real de lo que son hoy en día las Humanidades Digitales a nivel global, y donde parece confirmado el papel hegemónico del Norte global, y sobre todo de los países anglosajones, en la comunidad internacional.

El conjunto de comunicaciones presentadas en el último congreso global Digital Humanities 2017 (DH2017), celebrado, del 8 al 9 de agosto de 2017 en Montreal (Canadá), constituye, hasta la fecha, el corpus de conocimiento cooperativo más actual sobre Humanidades Digitales. Este corpus documental constituye el elemento más actual y necesario para indagar en las relaciones de asociación que se establecen entre países y centros (centros de investigación, universidades, etc.). A partir de la afiliación de los autores de los documentos, el estudio multidisciplinar de la colaboración en la investigación se ha abordado comúnmente desde diversos campos científicos: Historia de la Ciencia, Filosofía, Documentación, Bibliometría o Sociología (González Alcaide y Gómez Ferri 2014).

Este trabajo, planteado desde la óptica de la geopolítica del conocimiento (Adriansen 2016 y 2017; Canaga-

rajah 2002; de Sousa Santos 2010; Graham et al. 2011; Fiormonte 2017a; Mignolo 2011)[1], evidencia los desequilibrios culturales, institucionales o políticos existentes en el ecosistema de las Humanidades Digitales. El estudio de las relaciones de coautoría establecidas en las comunicaciones presentadas al congreso DH2017, pone de manifiesto las desigualdades entre países y centros.

La información disponible de cada registro, accesible on-line en el sitio web del congreso[2], consta de: nombre, apellidos y correo electrónico de los autores; título, resumen y palabras clave de la comunicación; y, centro y país de procedencia. A partir de la información proporcionada por la web del congreso se ha elaborado un archivo de texto en formato *Research Information System* (RIS), base de conocimiento y partida para realizar los análisis de asociación.

VOSviewer[3] (Van Eck y Waltman 2010, 2011 y 2014), herramienta desarrollada en la Universidad de Leiden, facilita el análisis de la colaboración científica mediante la visualización y la clasificación de las redes bibliométricas (Waltman, Van Eck y Noyons 2010) de autores, palabras clave, países o centros, existentes, pero no explícitas, en archivos RIS. El método de normalización de los enlaces elegido ha sido el del "valor de asociación".

Se han analizado las relaciones de asociación entre países, entre países y centros, entre palabras clave y entre autores, y se han obtenido tres tipos de redes y mapas de calor, según se haga proporcional el tamaño de los nodos a los valores de las ocurrencias, al número total de enlaces o al peso total de los enlaces. Veánse los siguientes ejemplos:

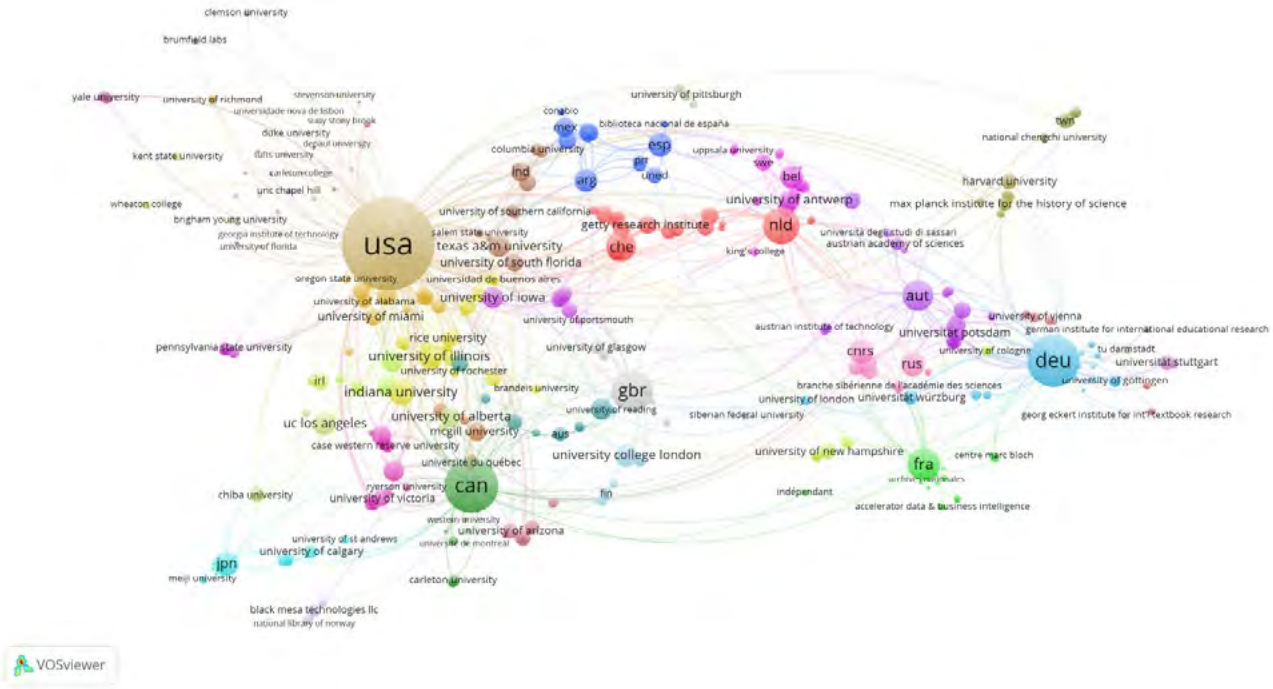


Figura 1.- Mapa de relaciones de asociación entre países y centros. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x0,5).

1 En el ámbito académico ver el reciente proyecto http://knowledgegap.org/

2 https://dh2017.adho.org/program/abstracts/

3 http://www.vosviewer.com/

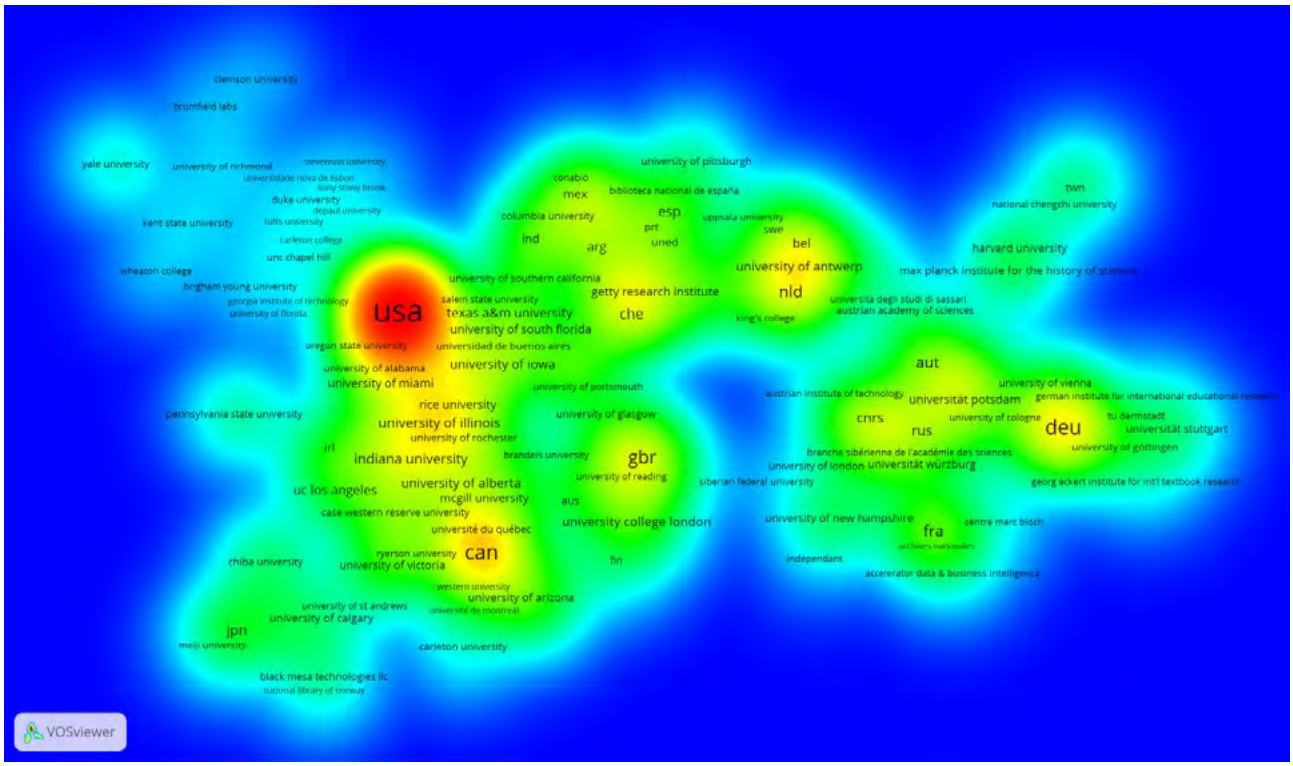


Figura 2.- Mapa de calor de las relaciones de asociación entre países y centros. Sólo aparecen los nodos conectados "total link strength"; factor de variación de tamaño x0,5).

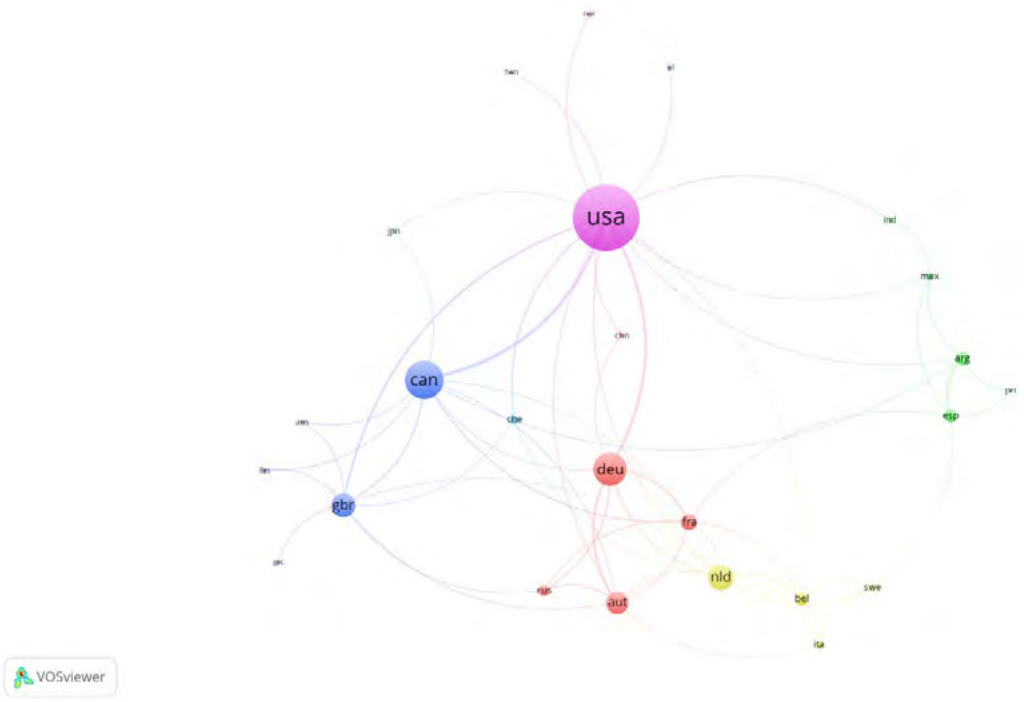


Figura 3.- Mapa de relaciones de asociación entre países. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x1).

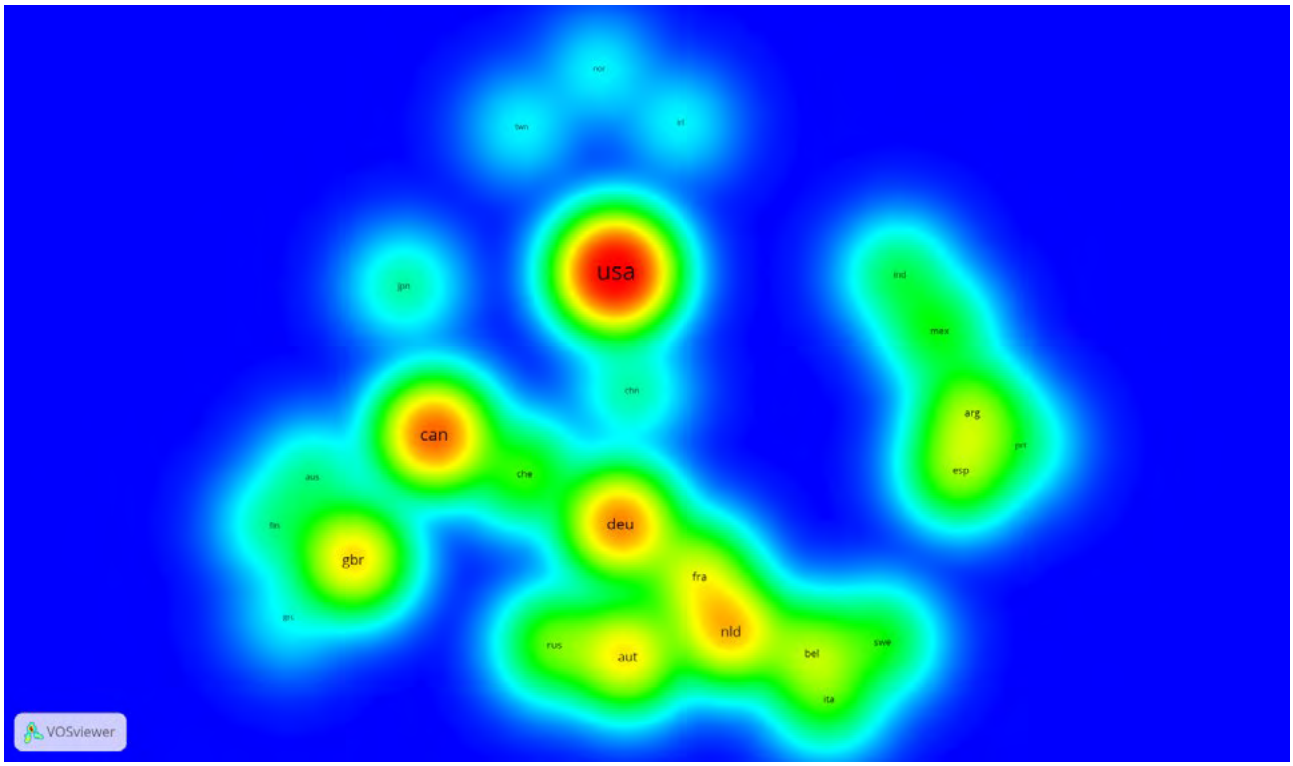


Figura 4.- Mapa de calor de las relaciones de asociación entre países. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x1).

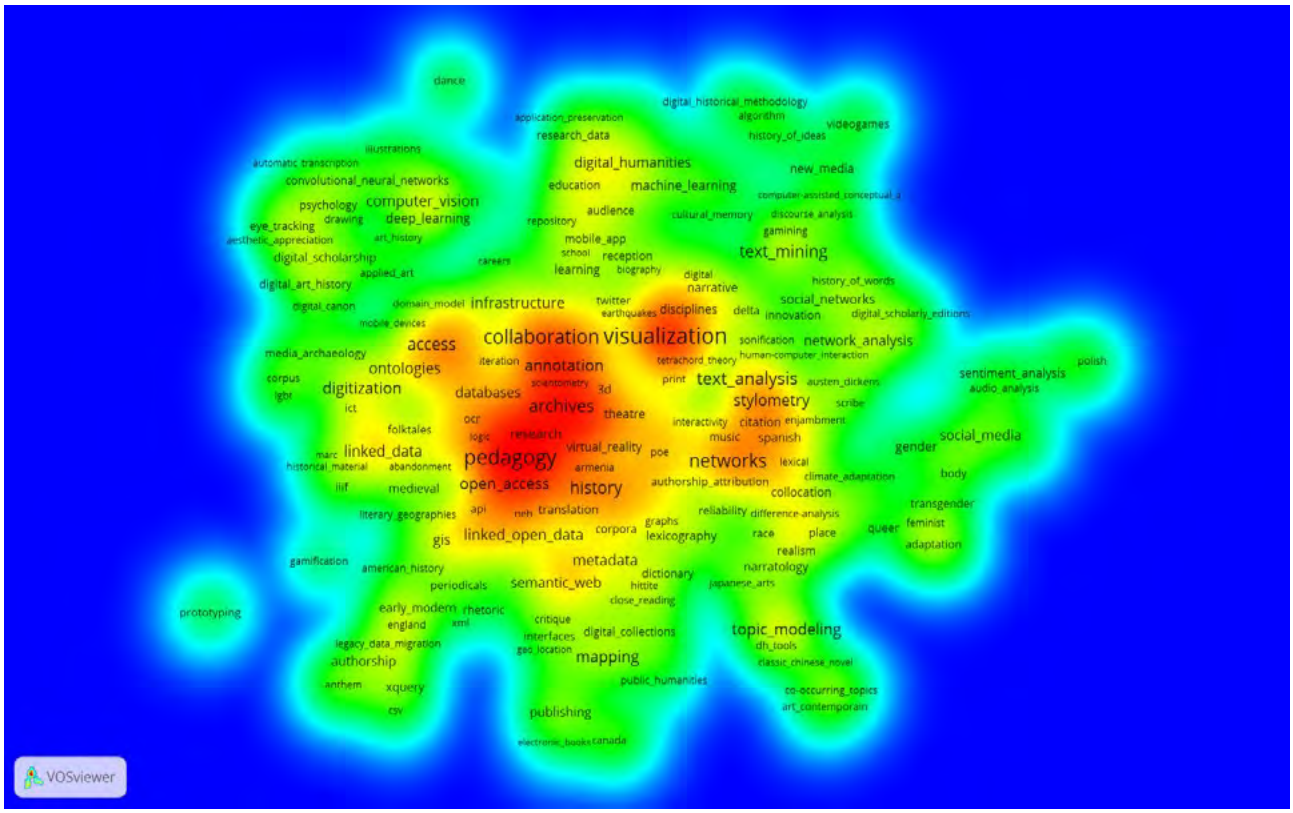


Figura 5.- Mapa de relaciones de co-ocurrencia entre palabras clave. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x0,5).

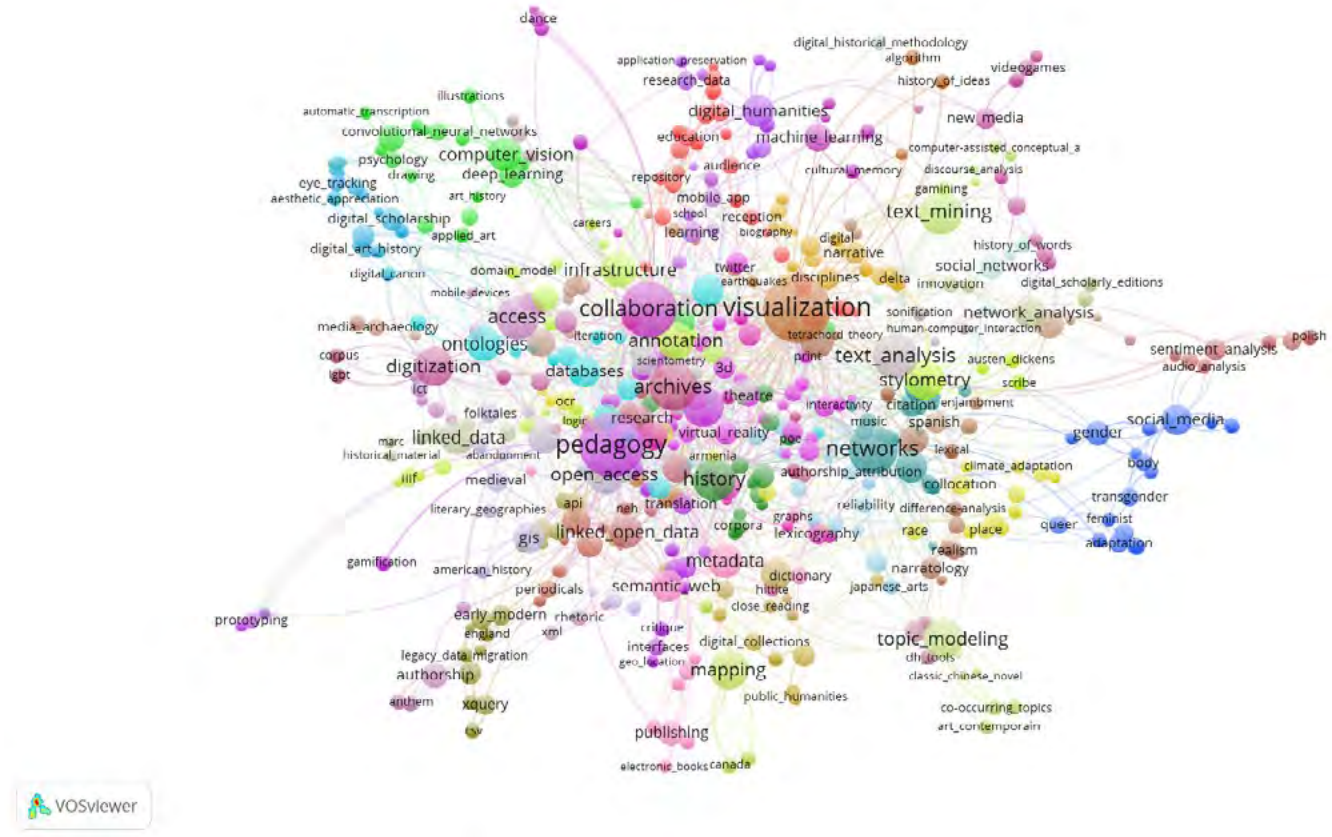


Figura 6.- Mapa de calor de las relaciones de co-ocurrencia entre palabras clave. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x0,5).

References

Adriansen, H. K. (2017). The power and politics of knowledge: what African universities need to do. *The Conversation*, September 28 2017. https://theconversation.com/the-power-and-politics-of-knowledge-what-african-universities-need-to-do-84233

Adriansen, H. K. (2016). Global Academic Collaboration: A New Form of Colonisation? *The Conversation*, July 8 2016. https://theconversation.com/global-academic-collaboration-a-new-form-of-colonisation-61382

Alcaide, G. G., & Ferri, J. G. (2014). La colaboración científica: principales líneas de investigación y retos de futuro. *Revista española de Documentación Científica*, *37*(4), 062.

Canagarajah, A. S. (2002). *A Geopolitics of Academic Writing*. Pittsburgh: University of Pittsburgh Press.

Dacos, M. (2016). La stratégie du sauna finlandais: Les frontières des Digital Humanities. *Digital Studies/Le champ numérique*. http://doi.org/10.16995/dscn.41

de Sousa Santos, B. (2010). *Descolonizar el saber, reinventar el poder*. Montevideo: Extensión, Universidad de la República-Ediciones Trilce.

Fiormonte, D. (2017a). Digital Humanities and the Geopolitics of Knowledge. *Digital Studies/Le champ numérique*, 7(1). http://doi.org/10.16995/dscn.274

Fiormonte, D. (2017b). Lingue, codici, rappresentanza. Margini delle Digital Humanities. *Filologia digitale: problemi e prospettive* (pp. 113-140). Accademia Nazionale dei Lincei. Contributi del Centro Linceo Interdisciplinare "Beniamino Segre", 135. Roma: Bardi.

Graham, et al, M. (2011). Visualizing the uneven geographies of knowledge production and circulation. *Global Higher Education*, 14.9. https://globalhighered.wordpress.com/2011/09/14/visualizing-the-uneven-geographies-of-knowledge-production-and-circulation/.

Grandjean, M. (2014). Le rêve du multilinguisme dans la science : l'exemple (malheureux) du colloque #DH2014. http://www.martingrandjean.ch/multilinguisme-dans-la-science-dh2014/

Mignolo, W. (2011). *The darker side of Western modernity*. Durham (N.C.): Duke University Press Books.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *arXiv:1109.2058 [cs]*. http://arxiv.org/abs/1109.2058

van Eck, N. J., & Waltman, L. (2014). Visualizing Bibliometric Networks. En Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice* (pp. 285-320). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10377-8_13

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629-635. https://doi.org/10.1016/j.joi.2010.07.002

Weingert, S. B. (2014). Submission to DH2015. http://scottbot.net/submissions-to-digital-humanities-2015-pt-1/

Weingart, S. B., & Eichmann-Kalwara, N. (2017). What's Under the Big Tent?: A Study of ADHO Conference Abstracts. *Digital Studies/Le champ numérique*, 7(1), 6. DOI: http://doi.org/10.16995/dscn.284

Using Topic Modelling to Explore Authors' Research Fields in a Corpus of Historical Scientific English

Stefan Fischer
stefan.fischer@uni-saarland.de
Universität des Saarlandes, Germany

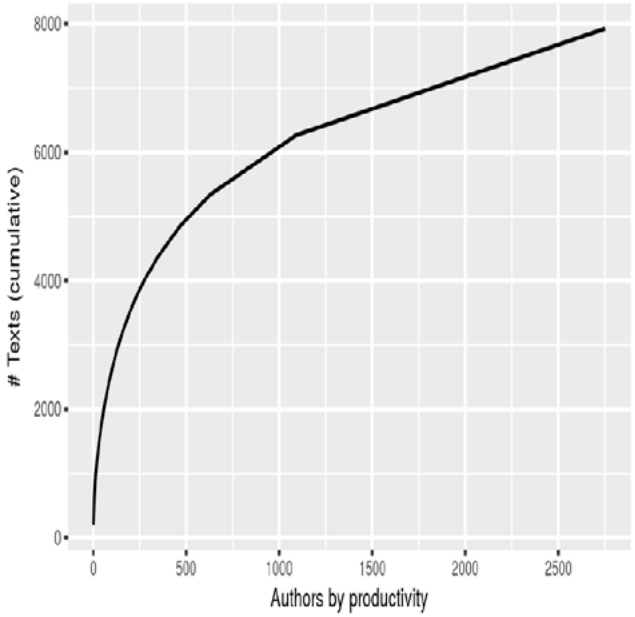
Jörg Knappen
j.knappen@mx.uni-saarland.de
Universität des Saarlandes, Germany

Elke Teich
e.teich@mx.uni-saarland.de
Universität des Saarlandes, Germany

Introduction

In the digital humanities, topic models are a widely applied text mining method (Meeks and Weingart, 2012). While their use for mining literary texts is not entirely straightforward (Schmidt, 2012), there is ample evidence for their use on factual text (e.g. Au Yeung and Jatowt, 2011; Thompson et al., 2016). We present an approach for exploring the research fields of selected authors in a corpus of late modern scientific English by topic modelling, looking at the topics assigned to an author's texts over the author's lifetime. Areas of applications we target are history of science, where we may be interested in the evolution of scientific disciplines over time (Thompson et al., 2016; Fankhauser et al., 2016), or diachronic linguistics, where we may be interested in the formation of languages for specific purposes (LSP) or specific scientific "styles" (cf. Bazerman, 1988; Degaetano-Ortlieb and Teich, 2016).

We use the *Royal Society Corpus* (RSC, Kermes et al., 2016), which is based on the first two centuries (1665–1869) of the *Philosophical Transactions* and the *Proceedings* of the *Royal Society of London*. The corpus contains 9,779 texts (32 million tokens) and is available at https://fedora.clarin-d.uni-saarland.de/rsc/. As we are interested in the development of individual authors, we focus on the single-author texts (81%) of the corpus. In total, 2,752 names are annotated in the single-author papers, but the activity of authors varies. Figure 1 shows that a small group of authors wrote a large portion of the texts. In fact, the twelve authors used for our analysis wrote 11% of the single-author articles.



Productivity of writers of single-author papers

Approach

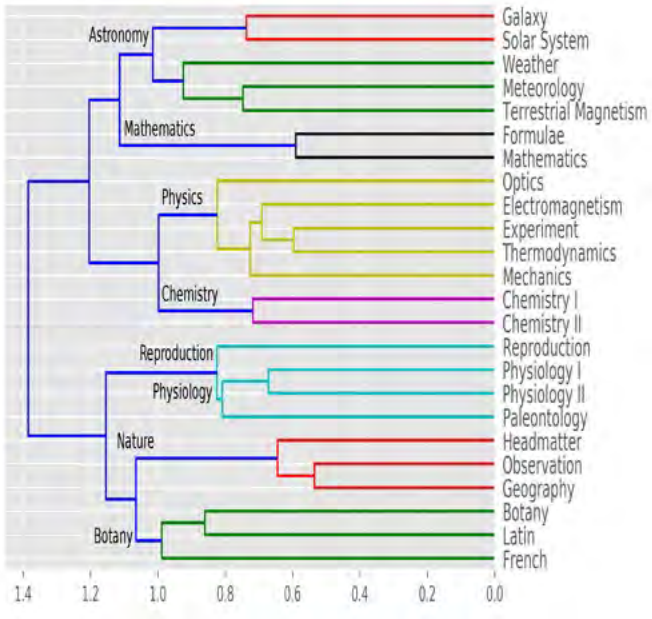
The topic modelling approach we take as a basis is Latent Dirichlet Allocation (LDA, Blei et al., 2003). LDA assumes that corpora contain a number of recurring topics and it treats texts as bags of words. Topics, which can be regarded as groups of semantically related words, are represented as probability distributions over words and each text is treated as a mixture of topics. Typically, topics are displayed as lists of the most probable words and labels are assigned manually. We also considered author-topic models (Rosen-Zvi et al., 2010) but their author-topic matrix implies that authors' topics are fixed over time.

As disciplines were not part of the original metadata of the RSC, we applied topic modelling to approximate disciplines. Using MALLET (McCallum, 2002), we built a model with 24 topics, which are shown along with their characteristic words in Figure 2.

Label	Words	%
Botany	plant leaves plants tab tree foliis folio seeds flowers bark seed species le...	2.0
Chemistry I	water acid grains quantity salt iron solution air experiments found lime col...	6.8
Chemistry II	acid water solution hydrogen oxygen obtained action salt cent alcohol gave s...	5.9
Electromagnetism	wire iron electricity experiments current experiment made end electric coppe...	4.3
Experiment	author present general subject state results nature similar case place great...	3.6
Formulae	cos sin oo tan ab sine axis ac io nt cd aa log vi cc arc al bc ef	4.3
French	la les le des en dans du par qui une qu il ou pour ce je sur au ne	4.2
Galaxy	stars distance position star obs equatorial diff small vf double magnitudes ...	1.1
Geography	sea water great miles found north part time river south side earth land east...	5.3
Headmatter	years year society time royal life age great number made letter part work pu...	0.3
Latin	quae quam sed ab sit vero hoc sunt ac qui esse etiam autem pro erit inter qu...	4.0
Mathematics	equation equations series number form terms values case equal order point cu...	0.7
Mechanics	force motion equal point surface velocity axis line plane body direction ang...	0.5
Meteorology	observations time hours tide water station hill height diurnal made stations...	4.0
Observation	made great found parts part make time small water body account long nature m...	5.5
Optics	light rays glass eye spectrum red lines colour colours surface blue white le...	1.3
Paleontology	bone part bones teeth surface upper lower side anterior length posterior jaw...	4.7
Physiology I	blood time animal day urine parts hours heart found food part days quantity ...	19.1
Physiology II	fibres nerves nerve part muscles vessels side muscular posterior anterior le...	2.8
Reproduction	cells form species surface structure cell membrane found part shell animal s...	1.6
Solar System	sun time observations made moon distance observed observation telescope limb...	7.8
Terrestrial Magnetism	needle magnetic ship observations direct force compass north made dip erebus...	6.7
Thermodynamics	air water heat temperature experiments tube experiment gas time made mercury...	0.1
Weather	rain cloudy ditto fair wind weather clear sw day fine ne cy se rn winds apri...	3.2

Topic labels and top words

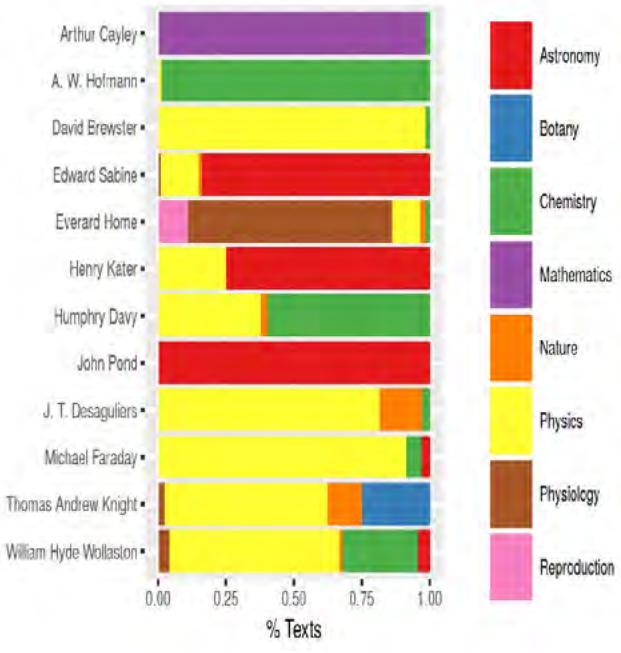
Following the approach of Fankhauser et al. (2016), we clustered the topics using Jensen–Shannon divergence. Figure 3 shows the resulting topic hierarchy. Based on this clustering, we identified broader research areas, which we marked on the branches of the dendrogram.



Hierarchical clustering of the 24 topics

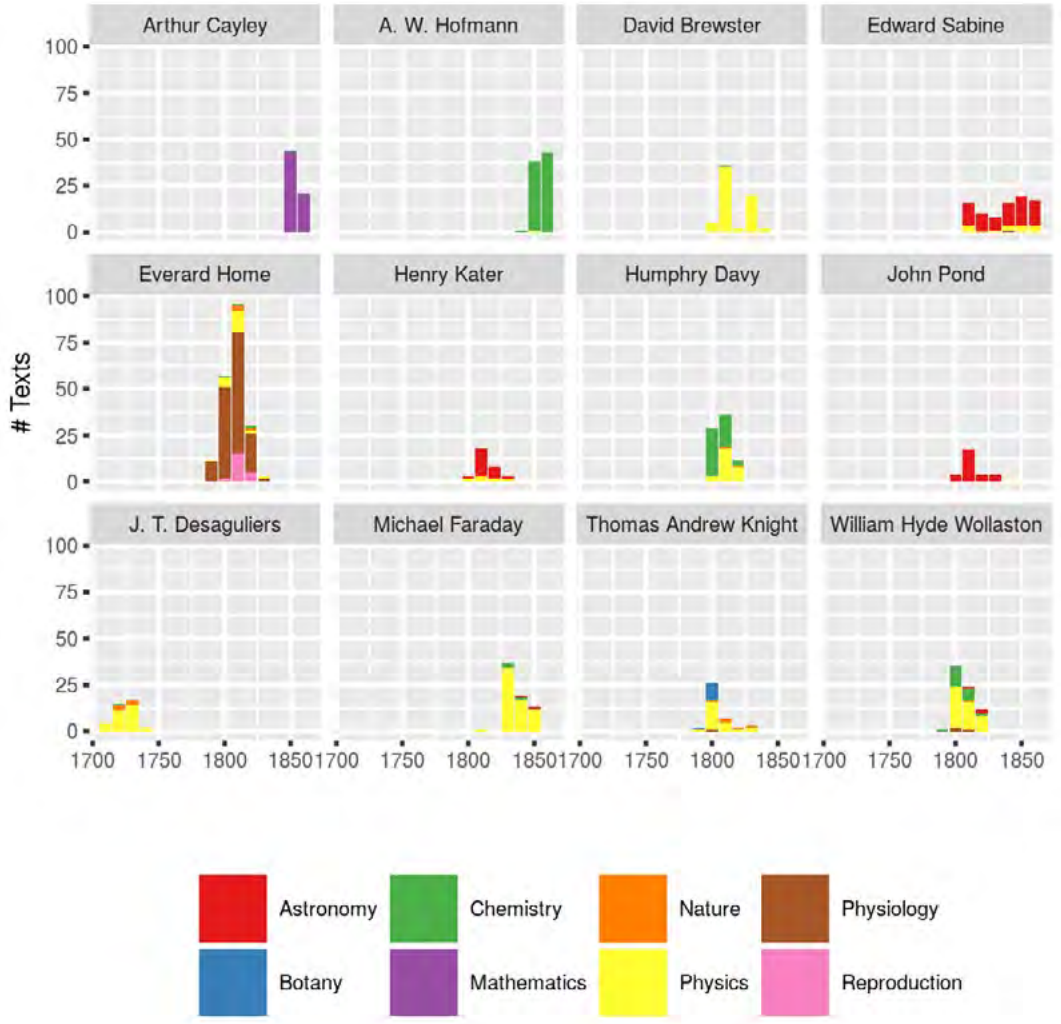
Results

Using these broader categories, we explore whether individual authors stayed in the same area or shifted their focus during their time of scientific production. For this purpose, we selected the most prolific authors (29–198 articles) in the RSC and tracked their topics over time (see Figures 4 and 5). We excluded names if we could not identify the author in the *Virtual International Authority File* or if publication years did not match the author's lifetime.



Comparison of topics of most prolific authors

Figure 4 shows the topics used by twelve authors during their career. We can see two groups of authors. Authors like *Arthur Cayley* dedicated their life to a single research area whereas *Humphry Davy* worked on two topics or in an interdisciplinary area. Figure 5 shows the development of the same authors over time. Overall, the authors' interests did not change dramatically over their professional life. However, one can identify a peak of productivity for most authors.



Development of individual authors over time

Conclusion

We proposed to use topic modelling as a method of exploring the development of the scientific orientation of individual authors over time. Taking topic as an approximation of discipline, our approach can be used to explore the contribution of a particular author to a given discipline over time or find authors with potentially interesting production profiles (e.g. authors shifting topics). In our future work, we will improve our models (e.g. avoid potential confusion of namesakes) by better metadata on the authors which we will obtain from the Royal Society.

Acknowledgement

We acknowledge the support of DFG (Deutsche Forschungsgemeinschaft) through the Cluster of Excellence *Multimodal Computing and Interaction* (MMCI).

References

Au Yeung, C. and Jatowt, A. (2011). Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. (CIKM '11). Glasgow, Scotland, UK: ACM, pp. 1231–1240.

Bazerman, C. (1988). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison: University of Wisconsin Press.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Degaetano-Ortlieb, S. and Teich, E. (2016). Information-based Modeling of Diachronic Linguistic Change: from Typicality to Productivity. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin, Germany: Association for Computational Linguistics, pp. 165–173.

Fankhauser, P., Knappen, J. and Teich, E. (2016). Topical Diversification over Time in the Royal Society Corpus. *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 496–500.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and Teich, E. (2016). The Royal Society Corpus: From Uncharted Data to Corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu (accessed 1 April 2018).

Meeks, E. and Weingart, S. B. (2012). The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities*, 2(1): 1–6.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P. and Steyvers, M. (2010). Learning Author-topic Models from Text Corpora. *ACM Transactions on Information Systems (TOIS)*, 28(1): 4:1–4:38.

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities*, 2(1): 49–65.

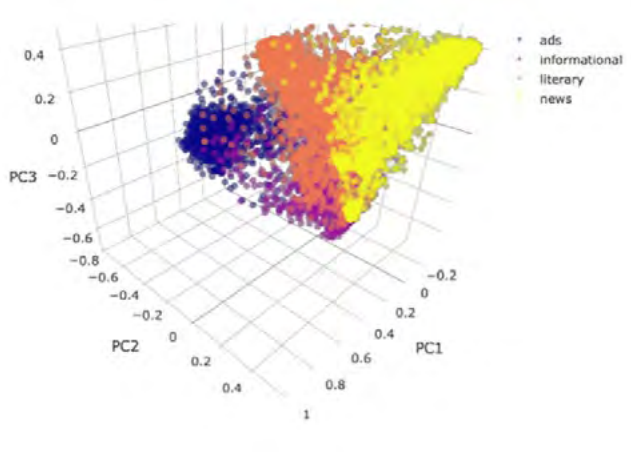
Thompson, P., Batista-Navarro, R. T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M. and Ananiadou, S. (2016). Text Mining the History of Medicine. *PLOS ONE*, 11(1): 1–33.

Stranger Genres: Computationally Classifying Reprinted Nineteenth Century Newspaper Texts

Jonathan D. Fitzgerald
fitzgerald.jo@husky.neu.edu
Northeastern University, United States of America

Ryan Cordell
r.cordell@northeastern.edu
Northeastern University, United States of America

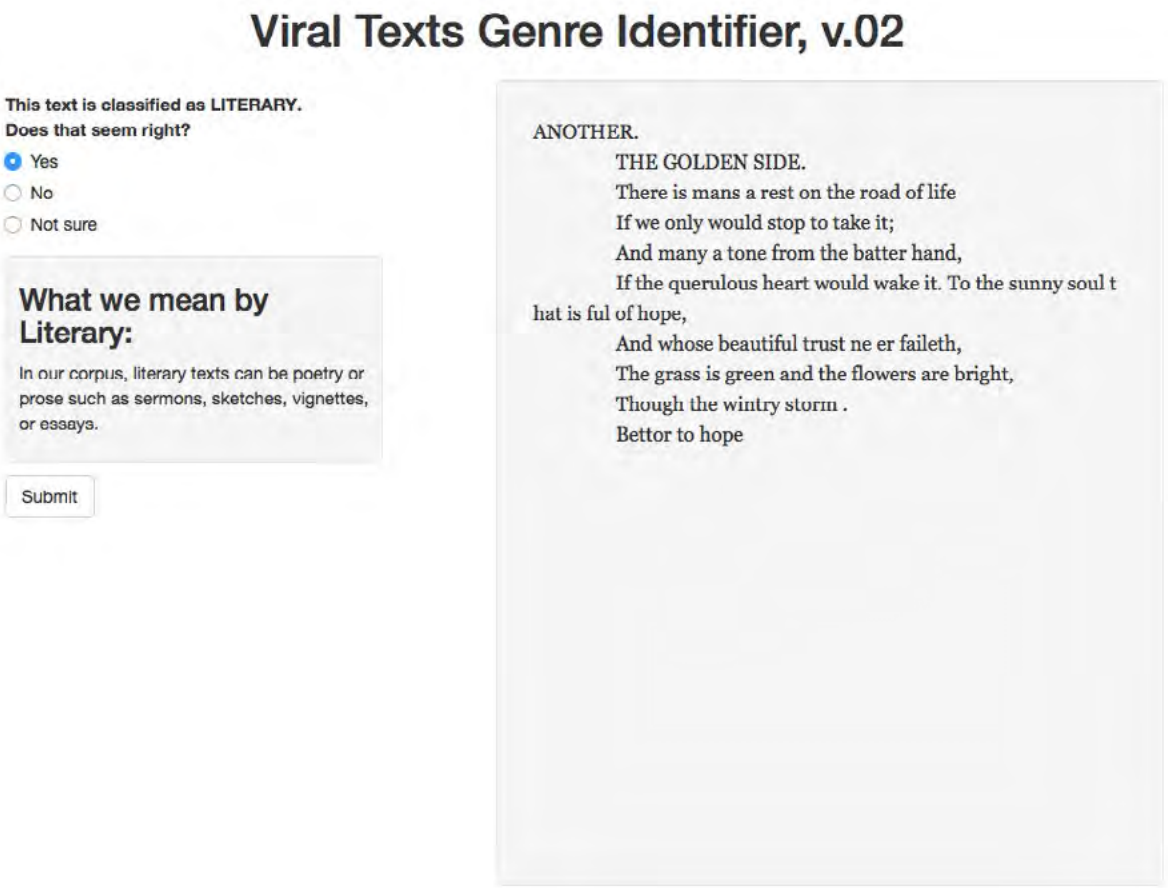
Since its inception in 2012, the *Viral Texts Project* has identified several millions of reprinted texts from corpora of nineteenth-century newspapers. The project began with the aim of isolating texts worthy of closer academic scrutiny from the "big data" of scanned newspapers, but the project's derived data is itself now so big that it cannot be effectively studied through browsing and reading alone. This poster describes our efforts to theorize and implement one solution to this challenge, through computational classification that identifies reprinted texts by genre. The poster will also share a prototype crowd-sourcing experiment that creates a bridge between computational research and various publics by encouraging scholars, students, journalists, and others to explore the strange genres of the nineteenth-century newspaper while enhancing our ground-truth data for training improved classifiers. Following other scholars who affirm the importance of human judgment in computational text analysis (Underwood, 2017; Klein, 2014; Long and So, 2015), our classification method employs unsupervised and supervised modelling: topic modeling and principal component analysis to group similar texts within a training set and generalized linear modelling to sort additional texts from the larger corpus. When the PCA data are visualized in three dimensional space, they cluster around four centers, which, upon closer inspection, can be described as four discrete but overlapping genres: news, advertisements, informational pieces, and literary pieces. Our GLM-based classifier—trained on data derived from PCA and confirmed by human readers—has been successful at finding and identifying thousands of previously unclassified texts in each of these genres.



These early experiments are helping our team more effectively isolate particular genres of texts for deeper literary-historical study, but these experiments are perhaps more valuable for the ways they are helping us reconsider our notions of genre itself in nineteenth century newspapers. Genres, as noted by other scholars who use computational methods to classify texts by genre (Schöch, 2017), are highly complex and fluid through time. In an effort to avoid presentist or anachronistic readings of genre, we

dispense with conceptions of journalistic genres drawn from twentieth- and twenty-first-century newspapers, and attend instead to the much more complex reality of the nineteenth century newspaper. For example, among the texts found in the "literary" category, we've identified many examples of what we name "vignettes"—short prose pieces that are a hybrid of fact and fiction, moral lesson and humorous anecdote. Vignettes of this kind are quite remote from contemporary journalistic genres, and yet we theorize that vignettes encapsulate the hybrid nineteenth century periodical press.

To make our classification efforts accessible to wider publics—and following other scholars who have done likewise in recent years (Beals, 2017; Mullen, 2016)—we have created a crowd-sourcing web application. This app, "The Amazing Generic Automaton," creates accessible paths into our work by allowing users to read a text alongside its most probable genre according to our classifier, asking users to determine whether our classifier has correctly identified the genre. If a user agrees with the classification, she simply clicks "Yes" to confirm; if, however, the genre does not appear to describe the text, the user may select "No" and a list of other genres, listed in the order of their probability as determined by the classifier, appear. The user can then select another genre, or instead choose "other," with a prompt to specify how she might classify the text. The results are saved as CSVs, which, when combined, constitute a new training set for *Viral Texts*. This app, in addition to confirming some of our classification efforts and providing a larger set of ground-truth data, fulfills a major goal of our work: it makes relatively complex computational work more accessible, thus adding a public face to our scholarship. For other humanities scholars less familiar with computational approaches, this app helps them see classification not as a "binary" decision, but instead as a constellation of overlapping generic probabilities.



The poster we propose will outline our process, describe what we're learning about genre in the nineteenth-century periodical press, present early results and visualizations, and offer conference attendees the opportunity to try out "The Amazing Generic Automaton." We expect our presentation will lead to meaningful conversations about innovative approaches to genre classification, the nature of literary genre situated in specific historical periods, and the benefits of creating bridges between complex computational text analysis work and the public.

References

Beals, M. H. (2017). Scissors-and-Paste-O-Meter Officially Launched for 1800-1900 http://mhbeals.com/scissors-and-paste-o-meter-officially-launched-for-1800-1900/ (accessed 28 November 2017).

Klein, L. F. (2014). Talk at Digital Humanities 2014 *Lauren F. Klein* http://lklein.com/2014/07/talk-at-digital-humanities-2014/ (accessed 28 November 2017).

Long, H. and So, R. J. (2015). Literary Pattern Recognition: Modernism between Close Reading and Machine Learning. *Critical Inquiry*, 42(2): 235–67 doi:10.1086/684353.

Mullen, L. (2016). America's Public Bible: Biblical Quotations in U.S. Newspapers http://americaspublicbible.org/ (accessed 17 April 2018).

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, 11(2) http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Underwood, T. (2017). We're probably due for another discussion of Stanley Fish *The Stone and the Shell* https://tedunderwood.com/2017/07/13/were-probably-due-for-another-discussion-of-stanley-fish/ (accessed 28 November 2017).

Humanities Commons: Collaboration and Collective Action for the Common Good

Kathleen Fitzpatrick
kfitz@msu.edu
Michigan State University, United States of America

Humanities Commons is an open-source, open-access not-for-profit social network and scholarly communication platform founded by the Modern Language Association and supported by a collective of scholarly organizations. Scholars and practitioners across the humanities and around the globe can create a professional profile, discuss common interests with colleagues, develop new publications, and share their work with other scholars and with the world.

Humanities Commons grew out of the MLA's experiences with its January 2013 launch of MLA Commons; the earlier platform was designed to serve the needs of MLA members by providing a range of types of open, networked communication. Early adopters, however, exhibited a strong desire to collaborate with scholars in fields other than those represented by the MLA. At the same time, the MLA was approached by several other ACLS member societies seeking similar networked communication solutions. Further, increasing concerns among scholars about the future disposition of commercial scholarly networks, given the sale of both Mendeley and SSRN to Elsevier and the problematic profit models being developed by ResearchGate and Academia.edu, revealed a strong desire for a sustainable not-for-profit alternative.

Given its successful prior work in the area, the MLA was well-positioned to explore the development of a federated platform that might be jointly supported by multiple scholarly societies, bringing together proprietary membership-oriented spaces with a range of fully public functions. With the support of the Andrew W. Mellon Foundation, the MLA met with a group of societies to discuss the possibility and then designed a pilot project to test the technical assumptions behind the federated network. Working with three partner organizations — the Association for Jewish Studies; the Association for Slavic, East European, and Eurasian Studies; and the College Art Association — the MLA launched Humanities Commons in beta in December 2016.

The network currently comprises four primary functions:

- a profile system permitting humanities practitioners to create a professional presence in a non-for-profit online space where they can easily connect with others in their fields;
- an open-access repository that allows members to archive and share the many products of their work, and to notify other members of their availability;
- a community platform, permitting members to join groups, share ideas, and discuss common interests;
- a publishing platform, permitting individuals or groups to create articles, books, teaching materials, Web sites, and blogs, to make their research public and to seek feedback on work in progress.

The network is built on the Commons In A Box (CBOX) platform, developed by the CUNY Graduate Center; CBOX is in turn based on WordPress and BuddyPress, which bring together a flexible publishing engine with rich social networking capabilities. The network's repository system is Fedora/Solr-based, with a WordPress front-end, developed in collaboration with the Columbia University Libraries and with the support of the National Endowment for the Humanities. Additionally, the network uses a federated authentication and identity management system, primarily based on COmanage and other Internet2-based systems, that communicates with the membership databases of participating scholarly organizations, thus allowing members to access all the organizations to which they belong through a single sign-on mechanism.

As of mid-April 2018, Humanities Commons has over 13,500 members who are actively developing their professional profiles. In order for the network to thrive, however, it must develop in a sustainable fashion. The planning and development for Humanities Commons were

undertaken by the MLA as a service to the profession, as well as to its sister societies, with the goal of providing an open-source, scholar-governed alternative to the available commercial services. That development was partially supported through grant funding, as noted above, but grant funding is not a business model; funders expect a project such as this to develop a sustainability plan to ensure its future. Humanities Commons is thus working toward collective action by and shared services for scholarly societies and other kinds of scholarly organizations who want to work together to provide a rich scholarly communication infrastructure for their members and for the profession at large.

This poster presentation will include an active demo of Humanities Commons as well as discussions of its platform, its community, its sustainability plan, and its development roadmap. We want to encourage members of the ADHO community to join the network and connect with one another across the conference and throughout the year. We also want to invite active participation among ADHO members and constituent organizations in the network's development process.

Making DH-Course Together

Dinara Gagarina
dinara@psu.ru
Perm State University, Russian Federation

The project involves the development of the MA-course "Concepts and Approaches of Digital Humanities", which is one of the basic courses in the MA-program "Digital Technologies in Sociocultural and Art Practices." Currently, the approbation of the experimental methodology in one of the universities is underway. The total volume of the discipline is three credits. The course is placed in the first semester, after which courses in specific areas of the DH are studied.

The goal of the course is to review the existing concepts of Digital Humanities, approaches to defining the subject and methods of Digital Humanities, theoretical and methodological foundations of using IT in various humanities, to get an idea of the relevant directions, tools, and projects.

The relevance of the project is primarily due to the growing importance of Digital Humanities as an interdisciplinary area. The developed MA program and this course as its part, allow combining the knowledge and methods accumulated in the application of IT in separate humanities disciplines, and train specialists of a new type. This association also takes place at the organizational level and facilitates the interaction of faculties, the implementation of joint projects.

Our approach to constructing the course is to actively involve students not only in research activities and projects, but also in the discussion and formation of the structure of the course, and then in filling it. The means of implementing this approach is the dynamic creation of the course site during the entire training period. We want to involve students in the practice of shaping the course, the constant retention of the focus of attention.

At the same time, we solve several problems and obtain a number of didactic opportunities and advantages.

Firstly, there is a problem of the lack of educational literature on DH in general, mainly prevailing literature and pedagogical developments in selected areas of DH. There is a language problem, for example, the only one reader by DH in Russian is released (Digital Humanities: A Reader, 2017). We use various types of materials - video, MOOC-courses, web resources, and actively read scientific literature.

Secondly, it is important for us to show a common landscape and a multitude of DH-directions on a scale. We work with masters who have different backgrounds in the bachelor's degree: historians, culture and art studies, philologists, PR, etc. There are also students in the group without humanities education at the bachelor's level.

Third, we want to combine teaching with the formation of a set of important DH skills: designing and retrieving information for web resources and databases, working with maps and timelines, corpora of texts. We use the students' conscious and active approach to learning and suggest that they become co-authors of the training course, first discussing possible course structures with them, and then jointly creating a special course site (Gagarina, 2017).

All links, texts related to the work of students during the course, are posted on the special site of the course. The entire group has access to viewing and commenting on all sections of the site and editing their own materials. Students learn to work in a team.

In the first lesson after the opening remarks, students are divided into pairs and offer their own model of the course. Since we are now conducting an experiment, we can already confidently say that students see DH with a skewing in its background or experience. During the discussion of these models, we jointly design a common framework and plan. For the convenience of combining the models, I suggest that students make a structure of 3-5 modules with a possible division into topics within the module. At this stage, one can clearly see the feature - students see as the first block the definition and history of DH, as 2-3 modules, most of them suggest considering directions within DH (computer linguistics and digital history, for example). Then many students propose to do their project. Almost no one talks about the consideration of common approaches, the general DH methodology, the DH infrastructure, and the classification of DH by technologies and tools.

The skill of the teacher is at this stage in combining your pre-designed curriculum and the vision of students.

It is quite possible to do this by shifting the emphasis somewhat.

At the conference, we plan to present the results of the experiment and the project as a whole.

The study is supported by Vladimir Potanin Foundation.

References

Gagarina D. (2018). *Studying Digital Humanities,* https://dhumanities.ru/.

Digital Humanities: A Reader (2017). Ed. M. Terras, J. Nyhan, E. Vanhoutte, I. Kizhner, Krasnoyarsk, 352 p. (in Russian).

Standing in Between. Digital Archive of Manuel Mosquera Garcés.

Maria Paula Garcia Mosquera
mpgarcia10@brown.edu
Brown University, United States of America

How have the life and achievements of preeminent Afro-Colombians been depicted in digital spaces? Which aspects of their lives have been highlighted in those efforts? What do those projects talk about the way these peoples have been remembered? Starting from these questions, *Standing in Between. The Digital Archive of Manuel Mosquera Garcés* is an initiative aiming to deepen in the history of Afro-Colombian politicians and intellectuals from the mid 20th century by creating an extended (digital) narrative of Manuel Mosquera Garcés.

Born in the Pacific coast in 1907, Mosquera Garcés was among the first Afro-Colombians to reach prominence in the Colombian government between the 1940s and 1970s. A leader of the Conservative party, Mosquera Garcés was part of a generation of politicians coming from the periphery who actively worked towards the inclusion of their home region into national dynamics. His story, however, has been blurred within the historical narratives of the country. Mosquera Garcés' legacy does not easily fit into the dominant narratives typical of a Colombia's official and centralized history (white, conservative, wealthy, eager to replicate Western and Catholic values), nor the mainstream narratives of the Afro-Colombians (black, liberal, underprivileged, eager to claim their African roots). His story in sharp contrast against those narratives, as he was a conservative politician from a marginalized region of the country who believed profoundly in Catholic principles. Additionally, he was black, a lawyer and a passionate reader of intellectuals of the Western tradition. He worked in Bogotá (capital city) while he was standing for his people in Chocó. The project is designed as a digital repository that will publicly display— for the first time — Mosquera Garcés' personal archive, along with additional documents related to his work, contextualizing the whole set as a curated collection.

Based on Kim Gallon's work on the "politics of recovery"(1) and the ways historiographical reinterpretations could be considered political enterprises to restore the "humanity" of black people as historical, political, and intellectual agents, *Standing in Between* will seek to restore the historical role and agency of Afro-Colombians in the digital domain. Connected to Liliana Ángulo's artwork "A case of reparation,"(2) which *liberates* archival sources to reveal historical erasures of the Botanical Expedition, the project is guided by the importance of offering sources to generate analysis with an extensive level of historical detail. Indeed among different local blogs and websites, including *Historia Personajes Afrocolombianos*, *Enamórate del Chocó*, and *República de Colores*, Mosquera Garcés has been included as a historical Afro-Colombian figure. In the form of biographies and informative articles, these private initiatives are rooted in an urgency to present the legacy of Afro-Colombians in order to incorporate these stories as part of the national identity and historical discourse. The University of Vanderbilt has published part of the correspondence of Manuel Zapata Olivella (black novelist) and historical documents of the Pacific Coast, while on a local level the appearance of digital initiatives and archives is still an emerging process.

Standing in Between aims to join these efforts examining Mosquera Garcés's archive, which was preserved by his family but until now it has not been scholarly reviewed, by considering three lenses that influenced his academic and political life: religion, language, and race. Archival material is diverse, and includes photographs, sound archives, bibliographic documents, and correspondence dating from the 1920s to the 1970s. Due to Mosquera's involvement in several periodical publications, as well as his work in the government in different capacities, the privately preserved documents do not offer a complete body of documentation of his political and scholarly life. In order to provide a more comprehensive context, the project has carried archival work in several public archives and libraries, to broadly identify his political agenda and academic interests. The archival work paid special attention to content reflecting his religious thought and conservative partisanship.

The initial work done on the digitization and cataloging of these materials, and the preliminary findings of curating this archive, will be presented in this poster. Additionally, in this early stage of the project, the design of a timeline will be displayed as a way of visualizing the connections between Mosquera Garcés and his generation of peers in his native Chocó poster, all of whom were bridging the gap between the center and the periphery through their participation in the national government. This first visual tool will add references to the collection, other digital projects on Afro-Colombians, and oral histories conducted for this Project.

References

Gallon, K. (2016). Making a Case for the Black Digital Humanities. In her article, Dr. Gallon In Gold M. & Klein L. (Eds.), *Debates in the Digital Humanities 2016* (pp. 42-49). Minneapolis; London: University of Minnesota Press. Retrieved from http://www.jstor.org/stable/10.5749/j.ctt1cn6thb.7

Ángulo, Liliana. (2015). *Un caso de reparación. Un proyecto de reparación histórica y humanidades digitales*. http://uncasodereparacion.altervista.org/?doing_wp_cron=1524636377.0468459129333496093750 (accessed 20 April 2018).

Research Environment for Ancient Documents (READ)

Andrew Glass
asg@uw.edu
Microsoft Corp., University of Washington

Stephen White
stephenawhite57@gmail.com
Stephen White - Italy

Ian McCrabb
ian@prakas.org
University of Sydney, Prakas, Foundation, Australia

The Research Environment for Ancient Documents (READ) is an integrated Open Source web platform for epigraphical and manuscript research. It may be configured as the underlying engine for a text repository or as a complementary research toolset to an existing repository. The defining innovation of this software is the atomization of text into orthographic subunits (as opposed to lines or words). This enables mapping across all layers of textual analysis, from factual data (the location of a character on a surface) through contestable (the transcription of a character) to the purely interpretive (a semantic annotation). This data architecture enables:

- The integration of physical, textual, and interpretive aspects of research
- The transformation of conventional editing practice into optimized workflows
- Granular attribution of components of a text, which allows for alternative interpretations and flexible collaboration

This poster outlines the workflows and outputs supported by version 1.1 (2017) of READ. The first release is optimized for use with Indic languages using ak ara-based writing systems (abugida). We will demonstrate the platform using documents in Gāndhārī language. We will also demonstrate the ability to generalize READ to support other languages and writing systems, e.g., Aramaic, Chinese, English, Italian, and Mayan.

The core workflows of the READ are:

1. **Creating a new item for study and inputting a text transcription**. A researcher creates a new item in READ and adds basic metadata and enters a transcription of the item in free text. Once entered, the researcher can immediately access two types of reports: a wordlist generated from the text; and alternate presentations of the text edition (diplomatic, reconstructed, and hybrid). These reports are available via READ's web interface, as well as the following downloadable export formats: HTML export, RTF, TEI (EpiDoc).
2. **Uploading images of the source text and linking to the text transcription.** READ provides tools to mark segment boundaries around the graphical units of the writing system depicted in the images. These segments are then automatically linked to the transcription entered in step 1. At this point the researcher can view the edition side by side with the image using synchronized scrolling provided by READ's web interface. In addition, the researcher can access a paleographic report generated from the image segments using the linked transcription. The TEI (EpiDoc) export includes the image as Facsimile element. All image segments can be exported as distinct files for paleographic processing using external tools.
3. **Creating a text glossary by adding lexicographical data to the generated wordlist.** The researcher uses tools provided by READ to add lexicographical data to the wordlist that was generated in step 1. At this point, a glossary can be generated and exported (HTML, RTF) or viewed in READ's web interface. Also, the edition viewer in READ's web interface integrates glossary data in flyouts associated with each word in the edition.
4. **Completing the glossary.** The researcher views the glossary created in step 3 in the READ's web interface and adds compound analysis to any compounds occurring in the text. At this point, glossary generation includes cross-reference entries for compound members.
5. **Annotating the edition.** The researcher uses annotation tools provided by READ to add footnotes and tags to the edition. The researcher can add text-structural information as well as textual parallels, translation, and alternate transliteration forms. These annotations can be viewed in the web interface, as footnotes in exported RTF and HTML output.
6. **Cubing the edition.** A researcher can integrate alternate editions of the same text using tools provided by READ. Any alternate editions so integrated, will be linked to the same image added in Step 2. Alternate editions can be viewed side-by-side in READ's web

interface to support comparison between alternate editions of a text.

7. **Sharing the research.** READ has been designed as a collaborative tool from the outset. Researchers can choose to share visibility and editing rights to any of the elements in their work. Work can also be published in mutable and immutable forms via the READ viewer interface, as well full text editions in TEI, exported HTML, and RTF that can be opened in common word processing and desktop publishing software applications.

The READ project began in 2013 and has been funded by Ludwig-Maximilians Universität, Munich, Germany; the University of Washington, Seattle, USA; Université de Lausanne, Switzerland; University of Sydney, Australia; and Prakaś Foundation, Sydney, Australia.

Manifold Scholarship: Hybrid Publishing in a Print/Digital Era

Matthew K. Gold
mgold@gc.cuny.edu
Graduate Center, City University of New York, United States of America

Jojo Karlin
jojo.karlin@gmail.com
Graduate Center,
City University of New York, United States of America

Zach Davis
zach@castironcoding.com
Cast Iron Coding, United States of America

This poster will present the Manifold Scholarship project (http://manifold.umn.org), an open-source scholarly communication and book publishing platform funded by the Andrew W. Mellon Foundation. Created by the University of Minnesota Press, The GC Digital Scholarship Lab, and Cast Iron Coding, Manifold aims to present the scholarly monograph in a new networked and iterative form that still has strong ties to print.

Manifold editions bridge the space between static print and ebook forms and custom web-based projects that are individually designed and programmed to meet the unique and specific needs of a particular scholar. Manifold editions present a multi-dimensional version of the book as we know it—a base text upon which a set of media and user-interaction layers can be added along with an archive space for related research materials. The reading experience offers a set of standard characteristics and constraints so readers who read and interact with one Manifold edition know how to interact with another, no matter the publisher.

Manifold offers a potentially powerful platform for publishers who hope to offer web-based editions of their books at scale. It can ingest ePubs, the format used most often by scholarly presses in their production practices, but it can also ingest Google docs, markdown files, and Microsoft Word docs. It is thus useful not only for scholarly presses, but also for individual DH practitioners who wish to publish their work in an attractive, responsive format with options for annotating and highlighting works. Future development on the platform will enable it to be used in classrooms by groups of students, who might comment together on OER materials that have been published on a Manifold instance.

This poster will explain what Manifold is, how it works, how it integrates into existing university-press publishing workflows, and how others may begun using it on their own for a variety of publishing and pedagogical needs.

Legal Deposit Web Archives and the Digital Humanities: A Universe of Lost Opportunity?

Paul Gooding
p.gooding@uea.ac.uk
University of East Anglia, United Kingdom

Melissa Terras
m.terras@ed.ac.uk
University of Edinburgh, United Kingdom

Linda Berube
l.berube@uea.ac.uk
University of East Anglia, United Kingdom

Introduction

Legal deposit libraries have archived the web for over a decade. Several nations, supported by legal deposit regulations, have introduced comprehensive national domain web crawling, an essential part of the national library remit to collect, preserve and make accessible a nation's intellectual and cultural heritage (Brazier, 2016). Scholars have traditionally been the chief beneficiaries of legal deposit collections: in the case of web archives, the potential for research extends to contemporary materials, and to Digital Humanities text and data mining approaches. To date, however, little work has evaluated whether legal deposit regulations support computational approaches to research using national web archive data (Brügger, 2012; Hockx-Yu, 2014; Black, 2016).

This paper examines the impact of electronic legal deposit (ELD) in the United Kingdom, particularly how the 2013 regulations influence innovative scholarship using the Legal Deposit UK Web Archive. As the first major case

study to analyse the implementation of ELD, it will address the following key research questions:

- Is legal deposit, a concept defined and refined for print materials, the most suitable vehicle for supporting DH research using web archives?
- How does the current framing of ELD affect digital innovation in the UK library sector?
- How does the current information ecology, including not for-profit archives, influence the relationship between DH researchers and legal deposit libraries?

Research Context

The British Library began harvesting the UK web domain under legal deposit in 2013. The UK Web Archive had, by 2017, grown to 500Tb. However, UK legal deposit regulations, based on a centuries-old model of reading room access to deposited materials, affect the archive's significant potential for research: in practice, researchers can only access the full range of UK websites within the walls of selected institutions. DH scholars, though, require access to textual corpora and metadata in addition to interfaces for discovery and reading (Gooding, 2012). Winters argues that "it is the portability of data, its separability from an easy-to-use but necessarily limiting interface, which underpins much of the exciting work in the Digital Humanities" (2017: 246). Restricted deposit library access requires researchers to look elsewhere for portable web data: by undertaking their own web crawls, or by utilising datasets from *Common Crawl* (http://commoncrawl.org/) and the *Internet Archive* (https://archive.org). Both organisations provide vital services to researchers, and both innovate in areas that would traditionally fall under the deposit libraries' purview. They support their mission by exploring the boundaries of copyright, including exceptions for non-commercial text and data mining (Intellectual Property Office, 2014). This contrast between risk-enabled independent organisations and deposit libraries, described by interviewees as risk averse, challenges library/DH collaboration models such as *BL Labs* (http://labs.bl.uk) and Library of Congress Labs (https://labs.loc.gov).

Methodology

This paper analyses the impact of the UK regulatory environment upon DH reuse of the Legal Deposit UK Web Archive. It presents a quantitative analysis of information seeking behaviour, supported by insights from 30 interviews with UK legal deposit library practitioners. Quantitative datasets consisted of Google Analytics reports, and web logs of UK web archive usage, which were analysed in SPSS and Excel. These datasets allowed us to identify broad patterns of information-seeking behaviour.

Practitioner interviews were hand-coded to three levels in Nvivo: initial coding, to provide the foundations for higher level analysis; focused coding, to further refine the data; and axial coding, using the convergence of ideas as a basis for exploring the research questions (Hahn, 2008). This analysis will inform two further research phases: a broader quantitative analysis of UK ELD collections; and qualitative analysis of the ways that the research community, and DH researchers, use ELD collections.

Conclusion

This paper provides a vital case study of how legal deposit regulations can influence library/DH collaboration. It argues that UK ELD regulations use a print-era view of national collections to interpret digital preservation and access. A lack of media specificity, combined with a more cautious approach to text and data mining than allowed under UK copyright, restricts DH research: first, by limiting opportunities for innovative computational research; and second by excluding lab-based library/DH collaborative models. As web preservation activities become concentrated in a small group of key organisations, current regulations disadvantage libraries in comparison to not-for-profits, whose vital work is supported by an ability to take risks denied to legal deposit libraries. The UK's approach to national domain web archiving represents a lost opportunity for computational scholarship, requiring us to rethink legal deposit in light of the differing affordances of born-digital archives.

References

Black, M. L. (2016). The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research. *International Journal of Humanities and Arts Computing*, 10(1): 95–109.

Brazier, C. (2016). Great Libraries? Good Libraries? Digital Collection Development and What it Means for Our Great Research Collections. In Baker, D. and Evans, W. (eds), *Digital Information Strategies: From Applications and Content to Libraries and People*. Waltham, MA: Chandos Publishing, pp. 41–56.

Brügger, N. (2012). Web History and the Web as a Historical Source. *Studies in Contemporary History*, 2 http://www.zeithistorische-forschungen.de/site/40209295/default.aspx (accessed 9 January 2017).

Gooding, P. (2012). Mass Digitization and the Garbage Dump: The Conflicting Needs of Quantitative and Qualitative Methods. *Literary and Linguistic Computing* doi:10.1093/llc/fqs054. http://llc.oxfordjournals.org/content/early/2012/12/22/llc.fqs054.abstract (accessed 30 July 2013).

Hahn, C. (2008). *Doing Qualitative Research Using Your Computer: A Practical Guide*. London: Sage Publications Ltd.

Hockx-Yu, H. (2014). Access and Scholarly Use of Web Archives. *Alexandria*, 25(1/2): 113–27.

Intellectual Property Office (2014). Exceptions to Copyright: Research *UK Government* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.
Winters, J. (2017). Coda: Web Archives for Humanities Research - Some Reflections. *The Web as History*. London: UCL Press, pp. 238–48.

Crafting History: Using a Linked Data Approach to Support the Development of Historical Narratives of Critical Events

Karen F. Gracy
kgracy@kent.edu
Kent State University, United States of America

This poster will present a progress report on a project that aims to explore how historians and other humanities scholars can most effectively access and use the data hidden in the silos of digital archival collections to craft narratives about significant developments and critical junctures in historical events, using Linked Data and event-based description. This project has two objectives: 1) to investigate the efficacy of an event-based model of description that will facilitate search across archival inventories and textual documents found in archival collections, and, 2) to develop and test a software tool that will allow scholars to more easily discover and use these hidden nuggets of information about events, and facilitate the construction of explanatory narratives about historical phenomena.

Linked Data and Event-Based Description

In the last two decades, the number of documents, photographs, and other archival material available in open digital archives worldwide has increased dramatically. Yet, these valuable sources of information are often hard to discover, due to long-standing practices in how archival materials are described and cataloged. Archival collections represent a tremendous source of untapped data, which is not discoverable without significant effort on the part of the researcher. Linked Data represents a new approach to information access that goes beyond simple tagging and indexing of documents using a predefined set of topics. Rather, it relies on semantically structured data embedded within the collection inventories, or even in the documents themselves, to interlink related information and make it searchable through semantic queries.

This particular project focuses on the difficulties of finding information on historical events in archival collections. Events are a special form of named entities, as they serve as a nexus point that marks a relationship between specific agents, places, and points in time (Gracy, 2015; Hyvönen, Lindquist, Törnroos, and Mäkelä, 2012). Thus, they act as gathering mechanisms for records of actions and are crucial aspects of archival information systems. To explore the concept of event-based description, the research team for the project has chosen the May 4, 1970 tragedy (during which four students were killed by members of the Ohio National Guard during a Vietnam War demonstration and nine others were injured) as our test case, as it has special resonance for our location at Kent State University. Kent State and other academic institutions have significant archival holdings and other information resources related to this event.

Usefulness of the Event-Based Model for Historical Research

This project employs archival finding aids and selected archival materials to create historical event vocabularies and ontologies, while creating and testing an event-based model that encompasses spatio-temporal dimensions and agents associated with events. The event vocabularies and ontologies are used as the basis for identifying and encoding information about persons, organizations, places, and topics. The event-based description model will be used as the basis for designing an information service that facilitates the linking of historical documents and archival descriptions related to an event, and will also help to link those materials and descriptions to other relevant published and archival sources.

Upon completion of the initial design of the event-based model (which is already underway), the project investigators will develop and test a prototype tool for event information discovery and use which can be used by scholars, students, and others interested in building historical narratives using archival material and related resources. Narrative building, which is the methodological stock in trade for many historians and humanities scholars, relies on the careful accumulation of data via the examination of documents relating to the topic under investigation (Barthes, 1977; White, 1984). This tool will also allow the investigators to test the validity of the event-based model as a suitable approach for facilitating information discovery for archival materials. This project proposes the process of historical research may be aided by a web-based tool designed to help with the discovery, collation, annotation, and sequencing of relevant information, and aims to build a web-based software with that functionality. The investigators propose that this project will have positive outcomes for digital history and humanities work, as it will empower humanities researchers to build complex historical narratives from various primary and secondary sources.

This poster will provide a progress report on the following activities: 1) Testing the event model with semantic metadata drawn from the May 4 Collection, which is an

archival collection from the Kent State University Libraries; 2) Developing and refining a web-based tool to assist historians and cultural heritage scholars in building and testing hypothetical narratives based on the linking of event information from various sources.

References

Barthes, R. (1977). Introduction to the Structural Analysis of Narrative. In Heath, S. (trans), *Image, Music, Text*. New York: Hill & Wang, pp. 79-124.

Gracy, K.F. (2015). Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges, *Archival Science*, 15: 239-254. doi: 10.1007/s10502-014-9216-2

Hyvönen E., Lindquist T., Törnroos J., & Mäkelä E. (2012). History on the Semantic Web as Linked Data—An Event Gazetteer and Timeline for the World War I. Proceedings of *CIDOC 2012, Enriching Cultural Heritage*, 10-14 June 2012, Helsinki, Finland. Retrieved from http://www.cidoc2012.fi/en/File/1609/hyvonen.pdf.

White, H. (1984). The Question of Narrative in Contemporary Historical Theory, *History and Theory* 23(1): 1-33.

Prosopografía de la Revolución Mexicana: Actualización de la Obra de Françoise Xavier Guerra

Martha Lucía Granados-Riveros
luciagranadosriveros@gmail.com
Escuela Nacional de Antropología e Historia, Mexico

Diego Montesinos
diegomontesinos@ciencias.unam.mx
Facultad de Ciencias UNAM

En 1985 se publicó el libro *México: del antiguo régimen a la revolución* del historiador Françoise Xavier Guerra, una referencia fundamental para el estudio de la Revolución mexicana. La obra revela las relaciones y tensiones entre la sociedad tradicional, un sistema heredado de la colonia y el Estado moderno proveniente en gran medida de los ideales liberales de la revolución francesa.

El trabajo de Xavier Guerra se inscribe en una amplia tradición de la investigación prosopográfica que se extiende desde el siglo XIX (Verboven, Carlier y Dumolyn, 2007) hasta el relativamente reciente uso de herramientas computacionales para el manejo de bases de datos (Blust, 1989; Keats-Rohan, 2010). El cuerpo biográfico de su investigación está compuesto por más de siete mil actores sociales entre los que figuran individuos y colectividades, con aproximadamente cien mil datos asociados a los movimientos políticos. Para su análisis se construyó una base que sistematizó los datos en más de cincuenta categorías que codifican dos tipos de sucesos; aquellos personales como fecha de nacimiento, muerte y ascendencia familiar y aquellos sucesos relacionados con la vida política y social del actor como participación en batallas o los cargos públicos ocupados. Los sucesos se organizaron en módulos independientes, lo cual permitió enriquecer la base de datos con la captura de nuevos módulos para personajes ya establecidos.

Dicha base de datos fue almacenada originalmente en tres cintas magnéticas de las cuales no se refieren más detalles, realizar nuevos análisis resulta inviable ya que el único medio en que está disponible actualmente es el impreso en los anexos de la obra señalada. El objetivo de este trabajo es la digitalización de la base de datos de Xavier Guerra, que permita la reproducción de los análisis del autor, así como la generación de nuevo conocimiento a partir del cruce de variables.

Con ese fin, se creó un programa en Python que ocupa Tesseract, una biblioteca de reconocimiento de caracteres. Debido a la estructura modular de la base de datos, los renglones, columnas y espacios en blanco son significativos. Por lo tanto, se realizó un pre-procesamiento de las imágenes, para detectar la estructura espacial del texto, de manera que Tesseract procesará pedazos de texto organizados. En esta etapa se ocupó el framework OpenCV y la biblioteca Pytesseract. Posteriormente, el programa organizó la información en un esquema de base de datos dentro de un archivo SQL.

En este póster presentamos el código desarrollado para la recuperación y organización de la base de datos, el funcionamiento de la base mediante algunas réplicas de los análisis que realizó Xavier-Guerra en su obra, así como el resultado de queries inéditos y por último el diseño inicial de la página que permita interactuar con los datos, de modo que los usuarios puedan consultar al sistema en términos de tiempo, geografía, compromisos políticos y relaciones de parentesco o sociales.

References

Blust, N. (1989). Prosopography and the computer: problems and possibilities, en Denley, P. (ed.) *History and computing* . no.2. Manchester, UK: Manchester University Press, pp. 12–18.

Keats-Rohan, K. (2010). Prosopography and Computing: a Marriage Made in Heaven?, *History and Computing*, 12(1), pp. 1–11.

Verboven, K., Carlier, M. y Dumolyn, J. (2007). A Short Manual to the Art of Prosopography, en Keats-Rohan, K. (ed.) *Prosopography Approaches and Applications. A Handbook*. Oxford: University of Oxford, pp. 35–69.

Developing Digital Methods to Map Museum "Soft Power"

Natalia Grincheva
natalia.grincheva@unimelb.edu.au
University of Melbourne, Australia

The project aims to employ Geographical Information Technologies to develop a pilot version of the digital mapping system "Museum Soft Power Map." It explores key factors in the time-space development of museum capacities to contribute to local creative economy by attracting tourism and generating economic activity. In collaboration with the Australian Centre for the Moving Image (ACMI), the project creates a dynamic digital map to visualize a growing in time geographic diversity of the Centre's collections, programming, audiences and partnerships. It reveals what factors affect the development of the ACMI's global brand recognition and influence its capacity to attract larger visitation and revenue.

Contemporary museums, as important actors in the international arena (Sylvester, 2009), increasingly serve as vital economic players helping their cities to compete for talent, tourism, and investment (Towse and Handka, 2013; Vivant, 2011; Werner, 2005). Though Nye's (2004) concept of "soft power" has been recently employed to discuss museum contribution to place branding, urban regeneration and tourism development (Lord and Blankenberg, 2015), there is a significant gap in the academic knowledge on what exact museum resources and activities accumulate "soft power" and how they affect the development of institutional global brand recognition in time and space.

The project tests a theoretical hypothesis that representing, promoting and celebrating cultural diversity help contemporary cultural institutions to attract larger global media attention, increase international visibility and appeal to more diverse audiences and partners (Nye, 2004, La Porte, 2012). The project traces a historic development of the ACMI's global brand that is based on the institutional vision to "be the leading global museum of the moving image" (ACMI, 2016). With diverse collection of hundreds foreign language films, representing a wide variety of cultures across the globe, ACMI runs a dozen of international tours and projects annually to strengthen its "reputation for world class exhibition experiences" (ACMI, 2016). In a close collaboration with ACMI, the project develops a customized Geographic Information System (GIS) that maps a growing international profile and visibility of the ACMI's collections, curatorial expertise and activities through time. The main goal of this digital mapping tool is to explore how attention to diversity on the level of collection acquisitions and a strategic focus on international outreach in its programming help the museum to accumulate institutional "soft power," measured through increase in its audienceship and selfearned income in Melbourne and other hosting cities.

A young, dynamic and ambitious institution, ACMI in 15 years of its existence, managed to develop a large audience reaching in 2016 1.5 million visitors to the Federation Square museum and 500 thousand attendants of its international exhibitions in six countries (ACMI, 2016). With 22% international visitors, ACMI generates $11.5 million through tickets sales and program services annually. As a partner in the project, ACMI is eager to provide its historical institutional records and digital expertise to develop the GIS software which traces and measure the development of its "soft power" in time and space.

The project employs museum records in the last 15 years in collection acquisitions and strategic programming to map and visualise a growing geographic diversity of the museum cultural resources and activities to explore how this international exposure affects audience development. The GIS software operates as a combination of deep mapping layers, each representing a different dimension of museum capitals tied to a specific location on the globe. Resources or Cultural Layer exposes a diversity and scope of museums' collections and main exhibits, highlighting geographic areas of their origins. Outputs or Social Layer maps complex museum "ecosystems" by visualizing museum social resources and telling stories about their engagements with constituencies, partners and audiences on the local and global levels. Impacts or Economic Layer builds on the metric of economic effects, measured through ticket sales at home and abroad, local and international program service revenue, membership dues as well as income received through museum shop, restaurant, and renting. The GIS processes the input data from three dimensions of museum capitals to map, visualize and draw correlations among cultural assets, social outcomes and economic impacts.

Combining and building on recent findings in academic scholarship on deep mapping (Bodenhamer et al., 2010; Gibson et al., 2010; Abrams et al., 2008) and museum evaluations (Jacobsen, 2016), the project designs a GIS system that advances a rapidly developing field of cultural mapping. The major outcome of this project is a research platform that can make a contribution both to applied knowledge and to academic scholarship. On the practical level, this research system can improve ACMI proactive management in global PR and programming. The digital map reveals geographic areas of missed opportunities by exposing locations where ACMI has a low or no cultural affiliations. Also, the system helps to identify "hot spots" of social density in terms of visitation and social activities, as well as to explore if stronger institutional efforts to target specific locations can result in a higher economic return on institutional investments. In academic terms, such a digital mapping tool advances the digital humanities scholarship by developing computational methods to explore cultural institutions and their impacts upon audiences. It combines quantitative and qualitative traditions within cultural mapping to reveal how collections, curatorial expertise and in-

ternational programming strategies can generate museum "soft power."

My poster presentation at the conference will present the first stage of the mapping system development. The first stage is focused on mapping ACMI collections and calculating collection appeal power index to different countries. The demo version of the application is available here: http://victoriasoftware.com/demo.html Integrating content analysis of the multicultural and multilingual collections with cultural analytics data, representing different countries around the globe, the online map shows where ACMI can have a stronger appeal with its offerings and holdings.

ACMI has unbelievably rich and diverse collections. It has 200 thousand original items and more than 40 thousand titles. The majority of the collections are accessible online through the online collection search system which currently allows to search through 41.713 titles. 70% of films are produced outside Australia not only in the US and UK but also in France, Germany, Japan, China or New Zealand. There are movies in around 50 different languages which are spoken in more than 230 countries around the world. For example, extensive collections in English that originate from Australia, New Zealand, Canada, the USA, the UK and other countries provide a potential content access to people from a hundred countries, while films in French could reach people in 38 countries.

To understand the potential appeal power of the ACMI collection to people from different countries, I considered two main types of criteria: collections characteristics and social demographic statistics. Collections criteria indicate how many items where produced in a certain country and how many films in the collection are in the language/s spoken in this country. Social demographic criteria bring to light such nuances as immigration statistics in Melbourne, annual tourism rate, ancestry data and internet penetration rate which affects the collection access and discoverability online. I calculated the collection appeal power index as a weight some of all subsidence's across two key criteria. The demo app available online (http://victoriasoftware.com/demo.html) demonstrates the Appeal Power Index that ranges from 0 to 1 and is visualized by the intensity of the blue color applied to different countries. When you click different countries, the app indicates how many movies from the ACMI collections were produced in this country, how many movies in the collections are in the spoken languages of this country as well as highlights secondary factors like tourism, ancestry and immigration from this country which increases the probability of the collection exposure and visibility among people of this geographic area.

References

Abrams, J. and Hall, P. (2008). *Else/Where: Mapping New Cartographies of Networks and Territories*. Minnesota: University of Minnesota Press.

Australian Centre for the Moving Image (ACMI). 2016. *Annual Report 2015-16*. http://bit.ly/2vEEfNN

Bodenhamer, D., Corrigan, J. and Harris, T. (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington: Indiana University.

Gibson, C., Brennan-Horley, C. and Warren, A. (2010). Geographic Information Technologies for cultural research: cultural mapping and the prospects of colliding epistemologies. *Cultural Trends* 19 (4): 325–348.

Jacobsen, J. (2016). *Measuring Museum Impact and Performance*. Rowman & Littlefield.

La Porte, T. (2012). The Legitimacy and Effectiveness of Non State Actors and the Public Diplomacy Concept. In *Public Diplomacy Theory and Conceptual Issues*, ed. International Studies Association, ISA Annual Convention.

Lord, G. D. and Blankenberg, N. (2015). *Museums, Cities and Soft Power*. Rowman & Littlefield Publishers. AAM Press.

Nye, J. (2004). Soft Power: T*he Means to Success in World Politics*. New York: Public Affairs.

Sylvester, C. (2009). *Art/Museums: International Relations Where We Least Expect It*. London Paradigm Publishers.

Towse, R. and Handka, C. (2013). *Handbook on the Digital Creative Economy*. Routledge.

Vivant, E. (2011). Who brands whom? *Town Planning Review 82* (1): 99-115.

Werner, P. (2005). *Museum, Inc: Inside the Global Art World*. Chicago: Prickly Paradigm Press

Brecht Beats Shakespeare! A Card-Game Intervention Revolving Around the Network Analysis of European Drama

Angelika Hechtl
angelika.hechtl@gmail.com
Vienna University of Economics and Business, Austria

Frank Fischer
ffischer@hse.ru
Higher School of Economics, Russian Federation

Anika Schultz
anika.schultz@hu-berlin.de
Humboldt University of Berlin, Germany

Christopher Kittel
contact@christopherkittel.eu
University of Graz, Austria

Elisa Beshero-Bondar
ebbondar@gmail.com
University of Pittsburgh, United States of America

Steffen Martus
steffen.martus@hu-berlin.de
Humboldt University of Berlin, Germany

Peer Trilcke
trilcke@uni-potsdam.de
University of Potsdam, Germany

Jana Wolf
jana_a_wolf@hotmail.com
University of Potsdam, Germany

Ingo Börner
Ingoboerner86@gmail.com
University of Vienna, Austria

Daniil Skorinkin
dskorinkin@hse.ru
Higher School of Economics, Russian Federation

Tatiana Orlova
taorkon.tootta@gmail.com
Higher School of Economics, Russian Federation

Carsten Milling
cmil@hashtable.de
Higher School of Economics, Russian Federation

Christine Ivanovic
christine.ivanovic@univie.ac.at
University of Vienna, Austria

This poster offers a playful introduction to network analysis as a means to study and compare dramatic texts. Its more serious purpose is a didactic intervention in the now well-established methods of literary network analysis, which are not always applied with sufficient reflection. The calculation of complex network metrics is often not followed by a leap to meaningful interpretation. What does it really mean, for example, that the average path length of the social network extracted from Shakespeare's *Hamlet* is 1.69 and the density of the same network is 0.34? However, when we look at these values in relation to the corresponding values of other dramatic texts, such network statistics become much more meaningful.

In order to cultivate comparative sensitivity in the context of literary network analysis, we build on a gamification approach. Unlike other experiments in this direction – such as the Android and web app *Play(s)* presented at the DHd2016 (Göbel/Meiners 2016), which encouraged the playful correction and enrichment of literary TEI corpora – we produce a true card game that invites players to explore network-analysis data in a new way.

The poster format is applied in two ways: On the one hand, the poster is a data visualisation based on a minimal canon of European drama. On the other hand, it is a card game that playfully acquaints audiences with the meaning of basic network metrics. This approach is not new in the arts and humanities and reaches back to card games like *Plattenbauten. Berliner Betonerzeugnisse* (Mangold et al. 2001), where technical data of different types of prefabricated concrete buildings had to be compared (cf. Richter 2006).

Our drama card game serves to instruct players in literary history, quantitative approaches, and network theory, based on a collection of 32 dramas ranging from the ancient Greeks to the modern age. Instead of a lexicon-like description of such a collection, the descriptive instrument here consists of visual and quantitative values that produce comparability – a type of card game known to English speakers as *Top Trumps* – see https://en.wikipedia.org/wiki/Top_Trumps –, or as *Supertrumpf* in the German context.

Each card presents a visualisation of a social network extracted from one of the 32 plays (very much along the lines of Fischer et al. 2016 and Fischer et al. 2018). Additional information on the cards consists of metadata (author, title, subtitle, year of publication/premiere) and static and dynamic network data (network size, network density, clustering coefficient, average path length, maximum degree incl. the name of the corresponding character, number of scenes). The front card contains an introduction to the project and its background as well as short definitions of network-theoretical terminology.

The poster is generated with the all-in-one drama analysis script *dramavis*, which has received a corresponding function in the new version 0.4 (Kittel/Fischer 2017). The collection of 32 plays used for the conference poster is in no way meant to be definitive or canonical, but is intended to present a diverse collection of plays from the history of European drama that feature comparably interesting social network data. Our collection ranges from antiquity (Aeschylus, Euripides, Sophokles, Aristophanes) to modern times (Marlowe, Shakespeare, Ben Jonson, Calderón de la Barca, Racine, Molière, Aphra Behn, Goldoni, Goethe, Mitford, Victor Hugo, Pushkin, Gogol, Grabbe, Ibsen, Strindberg, Schnitzler, Chekhov, Lasker-Schüler, Shaw, Pirandello, García Lorca, Brecht, and others).

The *dramavis* tool can be fed with a customisable canon file to create your own deck of cards.

References

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., Trilcke, P. (2016): *Distant-Reading Showcase. 200 Years of Literary Network Data at a Glance. Proceedings of DHd2016*, Leipzig. DOI: https://dx.doi.org/10.6084/m9.figshare.3101203.v1

Fischer, F., Kittel, C., Milling, C., Schultz, A., Trilcke, P., Wolf, J. (2018): Dramenquartett. Eine didaktische Intervention. *Proceedings of DHd2018*, Cologne. DOI: https://doi.org/10.6084/m9.figshare.5926363.v1

Göbel, M., Meiners, H.-L. (2016): Play(s): Crowdbasierte Anreicherung eines literarischen Volltext-Korpus. *Proceedings of DHd2016*, Leipzig.

Kittel, C., Fischer, F. (2017): dramavis v0.4. On Github, 2017. *Repo*: https://github.com/lehkost/dramavis
Mangold, C. et al. (2001): *Plattenbauten*. Berliner Betonerzeugnisse. Ein Quartettspiel. Berlin.
Richter, P. (2006): *Der Plattenbau als Krisengebiet. Die architektonische und politische Transformation industriell errichteter Wohngebäude aus der DDR am Beispiel der Stadt Leinefelde*. Hamburg, Univ., Diss. URL: http://ediss.sub.uni-hamburg.de/volltexte/2006/3041/

Visualizando una Aproximación Narratológica sobre la Producción y Utilización de los Recursos Online de Museos de Arte.

María Isabel Hidalgo Urbaneja
m.hidalgo-urbaneja.1@research.gla.ac.uk
University of Glasgow, United Kingdom

Publicaciones y exposiciones online, así como otros recursos interactivos, se encuentran entre los recursos online más utilizados por museos de arte en todo el mundo para transmitir historias vinculadas a obras de arte y colecciones. Los formatos tradicionalmente utilizados por museos para contar la historia del arte están siendo reconceptualizados a través de las cualidades y funcionalidades que nos ofrece el medio digital. El proceso experimental que aquí se expone nace con el objetivo de visualizar las particularidades que definen las narrativas generadas en los recursos online de museos de arte. Una selección de seis recursos online representativos de las tipologías más comunes producidos por museos de los Estados Unidos, España y Reino Unido* han sido la base para, por un lado, recabar datos sobre la perspectiva de los productores, y por otro, la de los usuarios especializados—una audiencia de perfil académico/investigador en el área de la historia del arte. Los datos se obtuvieron a través de dos métodos: entrevistas con los productores involucrados en la creación de los recursos online seleccionados, y en el caso de los usuarios, a través del protocolo conocido como "pensamiento en voz alta" (thinking aloud protocol) que ayuda a capturar información relevante a los procesos de navegación de los recursos online. Ambos procedimientos fueron grabados y transcritos para un facilitar el análisis posterior. Estos datos que se codificaron y analizaron desde una perspectiva narratológica permitiendo la observación de elementos configurantes de las narrativas: autoría, recepción como lectoespectador, estructuración, espacialidad, cronología.

Aunque en la investigación doctoral que da origen a los datos utilizados en esta propuesta se siguió una metodología cualitativa de corte más tradicional, en este póster se expone una aproximación experimental basada en la visualización de los códigos extraídos de las transcripciones. La visualización ofrece una visión complementaria al análisis inicial de los datos, orientado a presentar resultados de forma discursiva. De acuerdo con esta premisa, el póster compararía las posibilidades de análisis y presentación de las visualizaciones con el formato discursivo. Un análisis visual de los códigos revela aspectos cuantitativos de los datos, así como las conexiones entre los códigos de forma más explícita. En un cierto sentido, se presenta un resumen visual o vista general. La visualización de datos puede ayudar en la identificación de aspectos que habían sido obviados tras el empleo de la metodología más tradicional, y potencialmente, puede ofrecer nuevas conclusiones en la investigación.

La modalidad de visualización que se emplea en esta propuesta ha sido diseñada partiendo de diferentes metodologías y herramientas de visualización de datos. En primer lugar, toma como punto de partida en el uso de diagramas como herramienta de análisis narratológico (Ryan, 2007). Aunque el diseño de la metodología usada para visualizar datos en este póster emplea específicamente el procedimiento conocido como "map analysis" (Carley, 1993), éste permite la comparación de textos en base a los códigos extraídos y las relaciones entre ellos. Por otro lado, el trabajo de Luther (2017) propone un modelo y herramienta de visualización centrado en la representación de aspectos cualitativos y cuantitativos, éste fue desarrollado con el objetivo de estudiar aspectos de temática socio-histórico artística. No obstante, como herramientas se han elegido Gephi y d3.js ya que permiten representar la frecuencia de los códigos e interrelaciones. Las visualizaciones de este póster representan por separado los datos tanto de productores como de usuarios especializados de los recursos online, permitiendo comparar los seis recursos online. Las visualizaciones permiten estudiar las diferencias y similitudes existentes entre las perspectivas de los creadores, desde un punto de vista referente a la autoría, y las perspectivas de la audiencia especializada, como lectoespectadores de los recursos online. Conclusiones derivadas del proceso de visualización de datos serán argumentadas en el póster. Las visualizaciones se conciben como generadoras de discusiones además de ser una representación de la investigación llevada a cabo, éstas podrán ser consultadas en http://m-hidalgo.com.

Este trabajo es también resultado del proyecto de I+D: „HAR2014-51915-P. Catálogos artísticos: Gnoseología, epistemologías y redes de conocimiento. Análisis crítico y computacional".

*Los estudios de son recursos digitales de las siguientes instituciones: Museo Nacional del Prado, Museo Centro de Arte Contemporáneo Reina Sofía, National Gallery, Londres, National Gallery of Art, Washington DC, Metropolitan Museum of Art y MoMA.

References

Carley, K. M. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology*, 23, 75–126. http://www.casos.cs.cmu.edu/publications/papers/carley_1993_codingchoices.PDF

Drucker, J., (2014). *Graphesis. Visual Forms of Knowledge Production*. Cambridge, MA.: Harvard University Press.

Flick, U., (2010). *An Introduction to Qualitative Research*. London: Sage Publications.

Gee, K. (2001). The ergonomics of hypertext narrative: usability testing as a tool for evaluation and redesign. *ACM J. Comput. Doc.* 25, 1 (February 2001), 3-16. DOI=http://dx.doi.org/10.1145/383948.383950

Luther, A. (2017). The Entity Mapper: A Data Visualization Tool for Qualitative Research Methods. *Leonardo*, Volume 50, Issue 3, June 2017. Cambridge: MIT Press, p.268-271. Doi: 10.1162/LEON_a_01148. Abstract available at: http://www.mitpressjournals.org/doi/abs/10.1162/LEON_a_01148

Mann, L. (2016). Online scholarly catalogues: Data and insights from OSCI. *MW2016: Museums and the Web 2016*. Consulted November 26, 2017. http://mw2016.museumsandtheweb.com/paper/online-scholarly-catalogues-data-and-insights-from-osci/

Ryan, M. (2007). Diagramming narratives. *Semiotica*. 165: 1.4, 11-40.

Warwick, C. (2013). Studying users in digital humanities. Terras, Melissa; Nyhan, Julianne, and Vanhoutte, Edward, eds., *Defining Digital Humanities. A Reader*. London: Routledge https://blogs.ucl.ac.uk/dh-in-practice/chapter-1/

Transatlantic knowledge production and conveyance in community-engaged public history: German History in Documents and Images/ Deutsche Geschichte in Dokumenten und Bildern

Matthew Hiebert
hiebert@ghi-dc.org
German Historical Institute Washington DC, United States of America

Simone Lässig
laessigs@ghi-dc.org
German Historical Institute Washington DC, United States of America

This poster presents the technical redesign of the web resource *German History in Documents and Images/ Deutsche Geschichte in Dokumenten und Bildern* (GHDI) as a transatlantic knowledge production and conveyance model for community-engaged public history. It is a multilingual project led and based at the German Historical Institute Washington (GHI) in partnership with DARIAH-DE, the Max Weber Foundation, and the University of Southern California. It was awarded a three-year development grant from the German Research Foundation/ Deutsche Forschungsgemeinschaft (DFG) in 2017. We display the project's theoretical foundations and aims, the resulting technical design, and report on the proof-of-concept phase and first-year of development.

GHDI was first conceived in 2002 by a group of academic historians who sought to make a large collection of German historical documents openly available online in German and English translation. GHDI would consist of ten chronological volumes to cover German history from 1500 to 2009, each of which includes an introduction and a selection of historical documents, images, and maps, accompanied by interpretations. The site currently contains 1,784 German documents (along with an equal number of English translations), 2,374 images, and 55 maps (for a total of 16,068 pages), with content being expanded in the revamp. The project has developed a large and diverse international community of users, registering approximately 100,000 unique visitors a month.

The reconceptualization and revamp of the GHDI includes the encoding of original and new materials in TEI P5, Dublin Core metadata for all content, a site-wide co-created bibliography, and a scholarly annotation system. The integration of, and project development contributions to, *Scalar*—a robust open-source authoring, editing, and publishing platform with support for RDF content—allows users to navigate content in diverse ways and along various critical historiographical paths, challenging "master narrative" approaches to German history. The Scalar adapters developed by the project will link a number of important German archives to English-speaking scholarly communities for the first time, and the GHDI platform will ultimately allow users to use and "mix" this and other content to produce their own and collaborative scholarly outputs.

Data resources of the project are being described using Dublin Core metadata vocabulary. Sources with annotations or other semantic enhancement adhere to TEI (Text Coding Initiative) P5 using the DTA base format. Linked-open data representations are being be stored in RDF-XML. Using Scalar's built-in API, all content will be made available directly via URL-based requests in RDF-XML. This is also the technical basis for user content "remixing" and user publication facilitation being developed within the GHDI environment. Authority control for personal names and other entities, both in consumption and publication, will be assured through GND and similarly broadly accepted standards. Resources suitable for language analyses tools conform to Component MetaData Infrastructure (CDMI) as prescribed by CLARIN-DE data

centers. Geographic data is being encoded in GeoJSON. All data will be published to prioritize permissiveness of use under Creative Commons licensing.

A Tool to Visualize Data on Scientific Performance in the Czech Republic

Radim Hladík
radim.hladik@fulbrightmail.org
Institute of Philosophy of the Czech Academy of Sciences, Czech Republic; National Institute of Informatics, Japan

The poster introduces a project to develop a visualization application for a unique data source on Czech sciences. Information Register of R&D Results (RIV) is the Czech Republic's inventory of the outputs of basic and applied research since 1992. Although it is potentially an important source of data for analyses of various aspects of the intellectual organization and publication culture in Czech sciences, this particular data source has earned itself a pejorative nickname – "a coffee grinder" – for its central role in purely mechanistic science evaluation in the country.

By employing text-mining technique that are standard in the digital humanities and by getting inspiration from visualization platforms such as *Voyant Tools* (Sinclair and Rockwell 2012), the project aims to contribute to the shift in the Czech narrative of science evaluation from the exclusively bibliometric perspective to a more diverse one. For example, the hope is that the visual display of the plethora of topics that are discussed in the research outputs registered in RIV will implicitly criticize the myopic vision in which all disciplines are leveled to the singular measure of the number of publications. The latter system is not only intellectually dubious, but it has had documented adverse effects on the quality of research results. Crucially, it stimulates institutions as well as individuals to prioritize quantity over quality (Good et al. 2015; Grančay, Vveinhardt, and Šumilo 2017).

The ill-fated usage of the RIV data to mold nationwide fiscal policies for scientific research reminds us that data analytics is not necessarily a neutral enterprise. A proper treatment of the data is a matter that confronts a data analyst with questions on the borderline of ethics. Although it is perfectly feasible in technical terms, we wish to discourage users from attempts to track individuals researchers; instead we offer features that display institutional or disciplinary dimensions of the data (see Figure 1). Furthermore, the web application will provide a module to visualize textual information from the register. Textual strings, such as abstracts and keywords, have been part and parcel of the recorded entries, but have only served thus far as mere search terms. Meanwhile, the utility of textual data has been demonstrated in studies that strive to map the intellectual organization and relationships within and between disciplines (Leydesdroff 1989; Moody 2004).

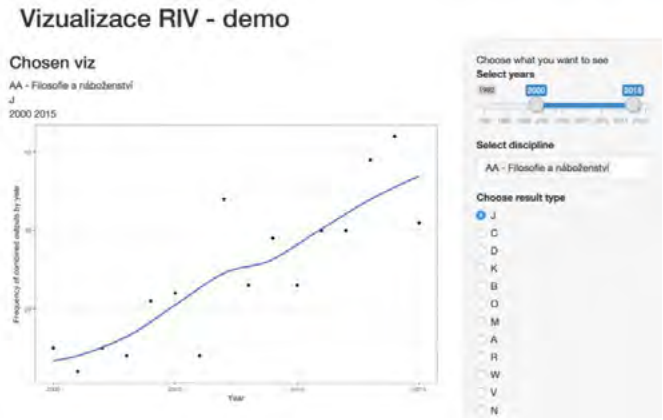


Figure 1. Using RIVVIZ to visualize a trend in the publication frequency of research outputs in the "J" (journal) category of the Information Register of R&D Results for the discipline "Philosophy and Religion" [note: the data are only a sample used in the development version]

The target group of the application are the researchers themselves. Namely, the textual module is intended to serve their needs by providing an overview of the trending topics in research or to identify institutions working on similar problems. The specialist user subgroup is envisaged to come from the fields focusing on social and other studies of science. The accessibility of visualized data and the simplicity of the interface can also attract journalists or other members of the public. The prospective users are also likely to be recruit from among the stakeholders in scientific policy-making and management who may wish to gain quick insights into the quantitatively assessed rates of output per research institutions or funding bodies.

The RIVVIZ application is developed in the R language and deployed on the R Server platform using the standard Shiny library. The data are imported from the publicly available repository of the Czech Research, Development and Innovation Information System. The internal setup is also fairly straightforward, relying predominately on the Tidyverse collection of packages, with ggplot2 library being the primary engine for visualization tasks. The underlying principles of the "grammar of graphics"(Wickham 2009) are particularly suitable for programming a user-oriented environment that allows for a control over a wide range of visualization parameters.

Giving the users more choices should help to make them more engaged with the application, although there is a trade-off between user-friendliness and complexity. Reasonable defaults can partially alleviate this dilemma. The user engagement will be important for the future application development (Galey and Ruecker 2010).

In the case of visualization schemes, locking users in a single – no matter how aesthetically pleasing – perspective is problematic. The apparent self-explanatory style and transparent communication of images may draw attention away from the complex and multifaceted nature of the data by making some of their aspects more easily accessible than others (Drucker 2011).

References

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* (DHQ), 5(1).

Galey A. and Ruecker, S. (2010). How a Prototype Argues. *Literary and Linguistic Computing*, 25 (4): 405-424.

Good, B., Vermeulen, N., Tiefenthaler, B. and Arnold, E. (2015). Counting Quality? The Czech Performance-Based Research Funding System. *Research Evaluation* 24 (2): 91–105.

Grančay, M., Vveinhardt, J. and Šumilo, Ē. (2017). Publish or Perish: How Central and Eastern European Economists Have Dealt with the Ever-Increasing Academic Publishing Requirements 2000–2015. *Scientometrics* 111 (3): 1813– 37.

Leydesdroff, L. (1989). Words and Co-Words as Indicators of Intellectual Organization. *Research Policy* 18 (4): 209–223.

Moody, J. (2004). The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. *American Sociological Review* 69 (2): 213–238.

Sinclair, S., Rockwell, G. and the Voyant Tools Team. (2012). *Voyant Tools* (web application).

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Dordrecht: Springer.

Augmenting the University: Using Augmented Reality to Excavate University Spaces

Christian Howard
ch4zs@virginia.edu
University of Virginia, United States of America

Monica Blair
mkb4rf@virginia.edu
University of Virginia, United States of America

Spyros Simotas
ss4ws@virginia.edu
University of Virginia, United States of America

Ankita Chakrabarti
ac4ze@virginia.edu
University of Virginia, United States of America

Torie Clark
vrc7de@virginia.edu
University of Virginia, United States of America

Tanner Greene
tjg6ph@virginia.edu
University of Virginia, United States of America

Project Website: http://reveal.scholarslab.org/

Introduction

Using augmented reality (AR) applications, our project, titled *UVA Reveal: Augmenting the University*, challenges the surface of our perceptions of objects and places. Our project specifically uses the University of Virginia (UVA), a large public state university, as its target. UVA is a southern historic campus with an enrollment of 22,000 students; given its history and recent spotlight in the news, UVA's campus is ripe for the historical inquiry and narrative intervention that our project proposes. In augmenting UVA's campus, we hope to expose the historical, cultural, (inter)national, (trans)sexual, and (dis)ability-related "archeology" of objects, places, and events.

Background of the Project

Augmented reality applications are becoming increasingly prevalent in society (witness Pokémon Go) and in the academy. For instance, a DH project titled *The Whole Story* uses an app that allows users to build AR statues of women and place them in the spatial landscape for others to see. By putting women back in the narrative, the app challenges the unequal statuary landscape and its implication that men are the makers of history. The digital spaces created by AR thus assume an openness and mobility that is lacking in physical space, which may be controlled or limited by socio-economic and political reasons. Nonetheless, these spatial boundaries can seemingly be circumvented in digital spaces,[4] and users can move rapidly across zones that they would be unable to otherwise. *UVA Reveal* is thus designed to explore how real spaces can be experienced through changing, mobile technologies that enable spatial and temporal augmentation.

The objects of our investigation include both buildings and documents at or connected to UVA, especially documents from the special collections library. In particular, we are attempting to renegotiate UVA's narratives about race, gender, and disability. For instance, a prominent mural on our campus depicts troubling scenes, including sexual harassment. We intend to use AR to highlight how women and other minorities are shown in this

4 We recognize that the same can be said about digital spaces, i.e. firewalls, paying services, language barriers, profile/password credentials, profiles set to private, digital literacy, etc. Our project, however, is open-source and freely available to the public.

mural by directing attention to them and challenging the patriarchal gaze.

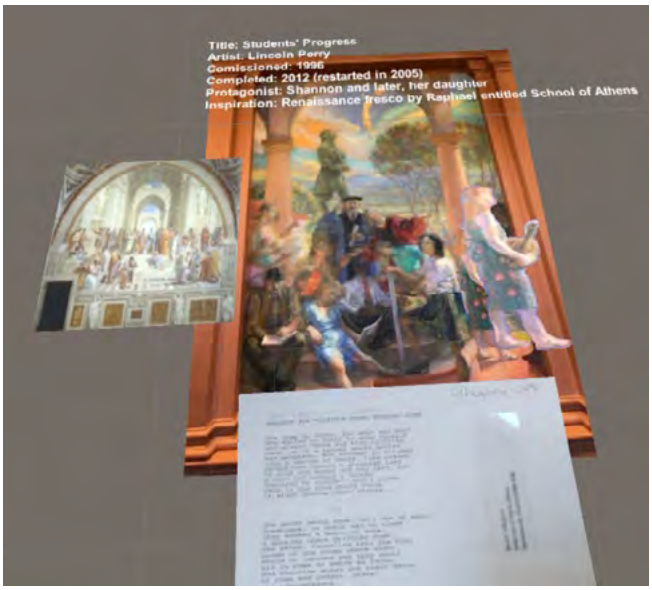


Image 1: Example of the augmentation of a prominent mural on UVA's campus as viewed through the Unity editor.

Theory

The spatial historian Richard White has claimed: "Visualization and spatial history… is a means of doing research: it generates questions that might otherwise go unasked, it reveals historical relations that might otherwise go unnoticed, and it undermines, or substantiates, stories upon which we build our own versions of the past." Similarly, our project neither contests nor reinforces the university's archive; rather, we supplement our archival research with broader research beyond the university's purview. As such, *UVA Reveal* enables viewers to make their own judgments about certain places and objects on UVA's campus by bringing those items to viewers' attention.

Methodology

UVA Reveal will have two primary instantiations: a web-based version and an app. The web version will clearly layout both our research methods and findings. Specifically, as we engaged with Special Collections, we realized that our project could have benefitted from a more directed search experience. To that end, we created a search function using UVA library data. Given a database with a sample of Special Collections holdings, the user may research a topic (narrowly defined for the scope of our project) using multiple keywords that relate to that topic; this cross-search exposes links between thematic data otherwise unavailable. We are using the d3 library to visualize the resulting data. This search function is integrated into our website.

The second version of the project will explicitly draw upon AR technology. In particular, our project uses Unity to layer 3D models on images – including university buildings and physical objects – that will enable the viewer to experience the virtual layering of time upon an object. Unity is easily exportable to Android, iOS, and HoloLens platforms. Our users will thus be able to engage the AR experience through their personal devices.

Our team is committed to open access. Thus, we are using GitHub to manage our content and ensure that our work process is openly accessible.

Conclusion

Through research in Special Collections, we plan to unearth the many historical layers upon which UVA is built. Ultimately, we hope to use AR to allow users to experience these limited-access spaces and objects in new ways that prompt critical reflection on the structure, culture, mission, and history of the university.

References

"Ambient Literature – This Is Your Part of the Story." *Ambient Literature*, UWE Bristol, Bath Spa University, the University of Birmingham, and Calvium Ltd., June 2016. <ambientlit.com/>.

E Silva, Adriana de Souza. "Mobile Narratives: Reading and Writing Urban Space with Location-Based Technologies," *Comparative Textual Media*. Ed. Katherine Hayles and Jessica Pressman. Minneapolis: University of Minnesota Press, 2013. 33-52.

White, Richard. "What Is Spatial History?" *Spatial History Project*, 1 Feb. 2010. <web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29>.

"The Whole Story." *The Whole Story*, 2017. <thewholestoryproject.com/>.

An Easy-to-use Data Analysis and Visualization Tool for Studying Chinese Buddhist Literature

Jen-Jou Hung
jenjou.hung@gmail.com
Dharma Drum Institute of Liberal Arts, Taiwan

In the field of Chinese ancient texts digitalization, the digitization of Buddhist scriptures has been regarded as a relatively complete and fruitful collection. The Chinese Buddhist Electronic Text Association (CBETA) has made the Chinese electronic Tripi□aka collection widely available for many years and provided a resourceful platform for the studies on Chinese Buddhist texts. As of the 2016

version(CBETA 2016), more than 210 million Chinese characters are freely and publicly available in digital form through the efforts of the CBETA.

The digital age that we have now entered has provided us with tools which help us in conducting surveys of Buddhist texts at a scale larger than before, and The text analysis techniques has been proofed as useful in many Buddhist literature research studies (Hung 2010, Bingenheimer 2017). It is with this goal in mind, our team made use of these new tools of the digital age to create a digital research environment which tailored to the needs of research in the field of Buddhist studies (and beyond). In order to achieve these goals, we established the CBETA Research Platform (http://cbeta-rp.dila.edu.tw/?lang=en). This research platform provides high-quality digital content from the CBETA corpus, combines with relevant reference materials based on the latest findings. Additionally, we implemented tools for quantitative analysis with the ultimate goal of creating a digital research platform which will assist scholars in their study of Chinese Buddhist texts or the underlying Indian origins.

CBETA Research Platform

The system architecture of CBETA Research Platform is shown in fig. 1. We have integrated the full text of CBETA corpus with Tripitaka catalogue, bibliographic databases, Buddhist dictionaries and authority databases of person and places to form the backend database of CBETA Research Platform. We then create tools to assist researchers in reading, searching and analyzing Buddhist literature.

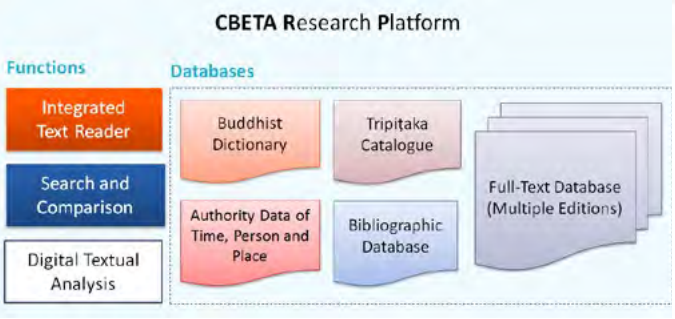


Fig 1. the system architecture of CBETA Research Platform

Concordance Search and Analysis

Concordance Search and Analysis is the first quantitative analysis tool implemented in CBETA Research platform[5]. It is a tool for gaining deeper insight into the search results from CBETA corpus. It allows user to aggregate search results from different dimensions (by Text Category, by Date and Dynasty, by Authors and Translators), and compare the results of multiple search terms.

Start a New Analysis

Concordance Search and Analysis will first require user to enter the keywords they want to compare and specify the search scope.

Fig 2. the start page of Concordance Search and Analysis system

Data

The system retrieves the complete search results and stores the search results for different keywords in the system cache at the same time. On data page, users can examine the complete list of the matches, and delete unwanted records from the result set.

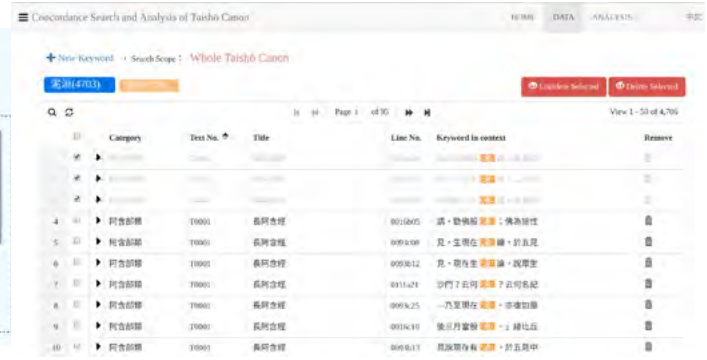


Fig 3. the data page of Concordance Search and Analysis system

Analysis

The System allows user to aggregate search results from different dimensions: by Text Category, by Date and Dynasty, by Authors and Translators, and compare the result of multiple search terms. Fig 4, 5 and 6, show the analysis results of two Synonyms: 泥洹(ní huán)and 涅槃(niè pán) form above-mentioned three different dimensions

5 Besides to Concordance Search and Analysis, CBETA Research Platform has provided an user-friendly reading interface (called as CBETA Online Reader, http://CBETAOnline.dila.edu.tw) for accessing texts and reference materials from backend database.

Fig 4 The statistics keywords in different text categories

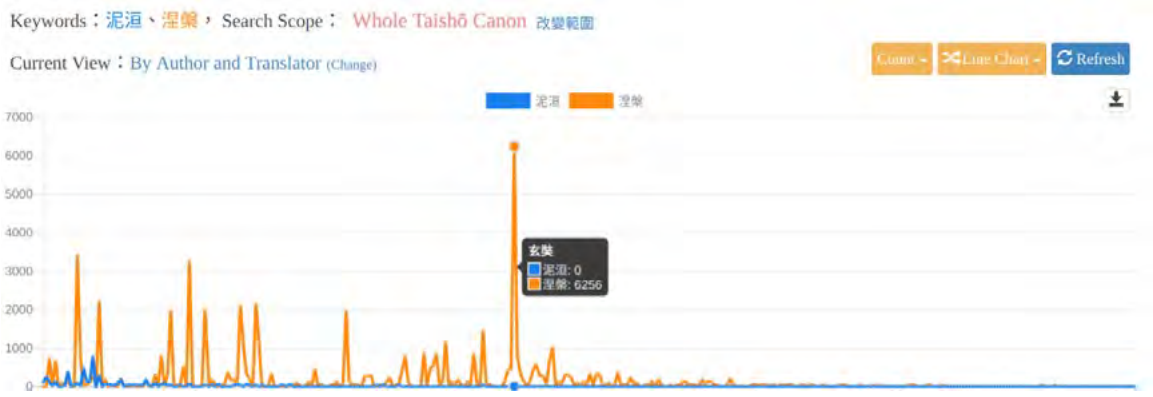


Fig. 5: The statistics of keywords with different translators

Fig. 6 The statistics of keywords in different dynasties

The system offers several statistical range settings. Thus, users are able to observe a wider usage of keyword from large-scale view, and at the same time, to trace a particular phenomenon back to the source text for identification and further research.

Fig. 7. statistics of keywords in different texts from Eastern-Jin Dynasty (C.E. 317 -420)

Fig. 8 statistics of keywords in different fascicles of 長阿含經(Dīrghāgama).

If we click the points represented the fascicle 3 of 長阿含經(Dīrghāgama) in the Fig.8, we will see sentences that actually contain keywords in the text.

Fig 9. sentences that actually contain keywords in fascicle 3 of Dīrghāgama.

In addition, the system also provides the „prefix and suffix analysis“ feature, allowing users to quickly access the statistics of a character before and after the keyword.

Fig 10. prefix and suffix analysis of keywords

In addition, in the spatial analysis function, we use a GIS system to display the location of the text containing the keywords, which allows users to compare the use of keywords geographically.

Fig 11. the spatial analysis of keywords

References

Bingenheimer, M.,Hung, J., and Hsieh, C. (2017) Stylometric Analysis of Chinese Buddhist texts – Do different Chinese translations of the Gandavyūha reflect stylistic features that are typical for their age? *Journal of the Japanese Association for Digital Humanities*, 2(1): 1-30

CBETA. (2016) *CBETA Chinese Electronic Tripitaka Collection*, Available at: http://www.cbeta.org/cbreader/help/index_e.htm (Accessed: 11 July 2017)

Hung, J., Bingenheimer, M., and Wiles, S. (2010) Quantitative Evidence for a Hypothesis regarding the Attribution of early Buddhist Translations *Literary and Linguistic Computing,* 25(1):119-134

'This, reader, is no fiction': Examining the Rhetorical Uses of Direct Address Across the Nineteenth- and Twentieth-Century Novel

Gabrielle Kirilloff
gkirilloff@gmail.com
University of Nebraska-Lincoln, United States of America

Though directly addressing the reader in fiction is often associated with cloying sentimentality, many different forms of direct address are employed across nineteenth- and twentieth-century novels. Upton Sinclair's use of address in *The Jungle* engulfs the reader in a tactile, fictional world, "your knife is slippery, and you are toiling like mad, when somebody happens to speak to you, or you strike a bone" (12). While Harriet Beecher Stowe's *Uncle Tom's Cabin* uses address to implicate the reader in systems of oppression, "And now, men and women of America, is this [slavery] a thing to be trifled with, apologized for, and passed over in silence?" (578). What is fascinating about address, is not only that it can be put to such a variety of purposes, but that these purposes are often antithetical, and have drastically different effects on real readers.

This project seeks to answer questions about the historical usage of address by employing computational methods to detect and extract instances of address from a corpus of 2,000 nineteenth- and early twentieth-century novels.[1] I examine how the frequency of address changes over time and among different groups of authors (such as female authors and African-American authors). In order to detect address I utilize a pattern matching approach that uses regular expressions to match sentences outside of dialogue that contain certain keywords, such as "reader," "you," and "this story." To remove dialogue from the corpus, I developed a pattern matching approach that eliminates quotations. This method accounts for various typographical inconsistencies, including missing quotation marks, embedded quotations, and quotations that extend across paragraphs. In order to learn more about the different types of address that authors have employed, I then used the Stanford Dependency Parser on the sentences extracted from each novel. The Parser is a tool that provides a representation of grammatical relations

1 The corpus is 70% male and 30% female authored; 70% American and 30% British. The texts come from freely available sources. The texts were written between 1800 and 1923.

between words in a sentence. This allowed me to examine the adjectives used to describe the reader or the verbs the reader performs in moments of address. In addition, I performed sentiment analysis on the sentences extracted from each novel using the Syuzhet package in R in order to track the emotional valence of address.

The results from the study indicate the prevalence of address across literary periods. Notably, the mean number of sentences containing address in each novel remains steady over time. Of the 2,000 novels examined, 1,864 contain address, with each novel on average containing 49 instances of address. These results are unexpected given the hypothesis put forward by Garett Stewart in *Dear Reader*: "outlawed in modernism, address went underground [at the beginning of the twentieth-century]" (33). The frequency of address and its prevalence across time push against the critical association (noted by Robyn Warhol in *Gendered Interventions*) of address with mid-nineteenth-century Victorian sentimentality. While the frequency of address remains relatively constant, the form of address radically fluctuates: authors decreasingly use "reader" to address their public in favor of addressing readers as "you."

Address is also correlated with author gender: male authors address their readers more frequently than female authors. Overall, address authored my female writers has a more "positive" emotional valence than address authored by male writers. In addition, male authors are more likely than female authors to use the word "reader" (rather than "you") in moments of address. Although there are notable exceptions, the distribution of "you" and "reader" maintains its correlation with author gender across time and nationality. These results intersect with Robyn Warhol's argument that female authors, more so than male authors, employ the intimate and personal "you" to foster a sense of connection with their readers in order to evoke sympathy for social causes.

References

Stowe, Harriet Beecher. (2009). *Uncle Tom's Cabin or Life Among the Lowly*. Cambridge, MA: Harvard University Press.

Sinclair, Upton. (2005). *The Jungle*. Boston, MA: Bedford/St. Martin's.

Stewart, Garett. (1996). *Dear Reader: The Conscripted Audience in Nineteenth-Century British Fiction.* Baltimore, MD: Johns Hopkins University Press.

Warhol, Robyn. (1989). *Gendered Interventions: Narrative discourse in the Victorian Novel*. New Brunswick, NJ: Rutgers University Press.

Reimagining Elizabeth Palmer Peabody's Lost "Mural Charts"

Alexandra Beall
abeall3@gatech.edu
Georgia Institute of Technology, United States of America

Courtney Allen
callen71@gatech.edu
Georgia Institute of Technology, United States of America

Angela Vujic
av.vujic@gmail.com
MIT, United States of America

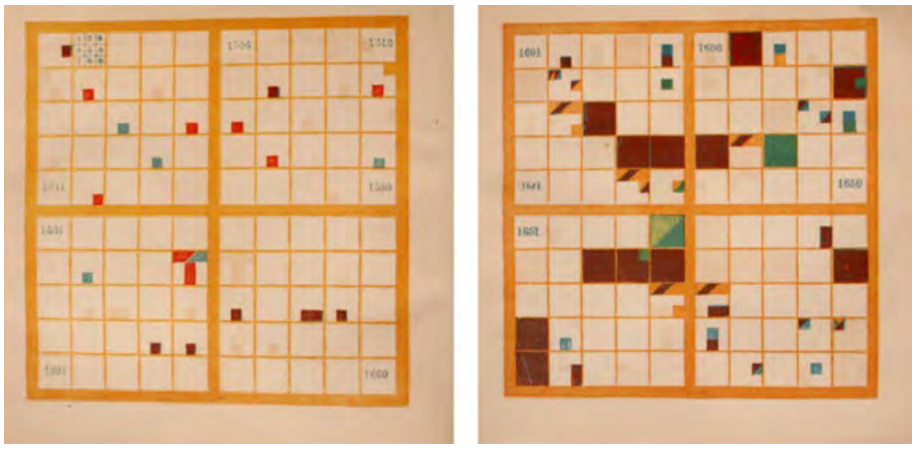
Lauren F. Klein
lauren.klein@lmc.gatech.edu
Georgia Institute of Technology, United States of America

Introduction and Overview

Writing to a friend in 1850, editor and educator Elizabeth Palmer Peabody (1804-1894) complained:

> Just now I am aching from the fatigue of making Charts for the Schools who will take the book… Every school must have a mural chart—& there is but one way of making them (until they can be made by ten thousands) & that is by stencilling… I can do one a day. But I must sell them cheap… To day I worked 15 hours—only sitting down to take my meals—& so I have done all week—so much fatigue stupefies one—but as soon as it is adopted in a few towns I shall be able to hire someone to do this drudgery for me.

In these lines, Peabody provides some of the only extant documentation of her "mural charts"-- large-scale versions of the pedagogical charts that she designed to accompany her U.S. history textbook, *A Chronological History of the United States* (1865). Peabody's textbook promoted data visualization as a pedagogical method. Her visualization scheme involved translating significant historical events into shape and color, and arranging them on a grid (see figures 1). Students could then use the grid as a visual mnemonic, inscribing each century of U.S. history into their memories.



Left: Significant events of the 16th century United States. Right: Significant events of the 17th century United States.

The Mural Chart Project

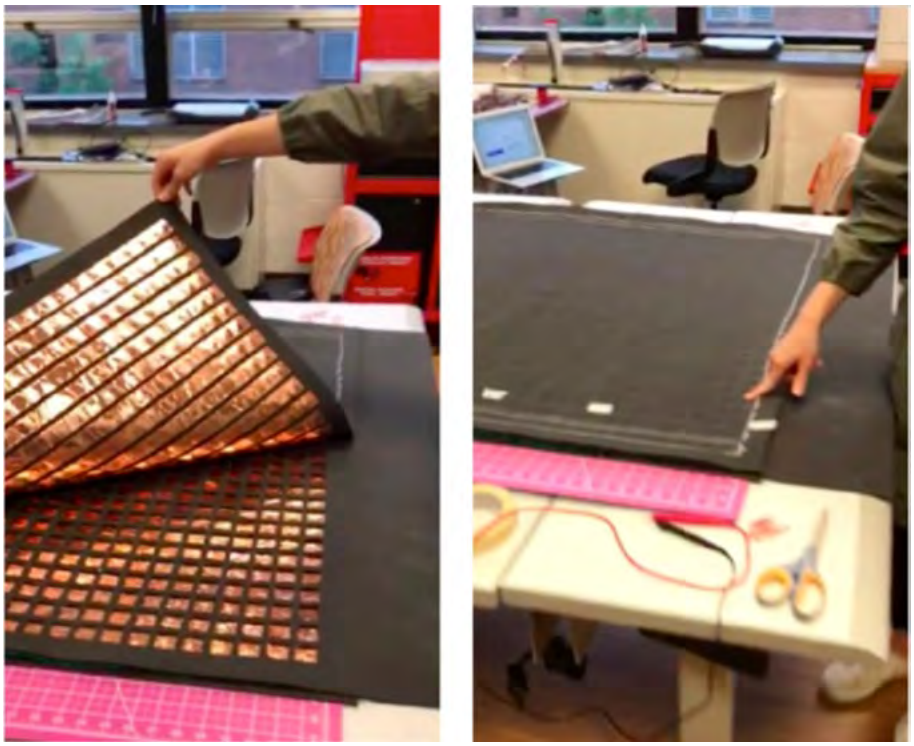
The project team has explored Peabody's visualization scheme in detail (e.g. Klein et al., 2017). But the "mural charts" that she describes in her 1850 letter have not been preserved. Scholarship describes how Peabody would lay the mural charts out on the classroom floor, inviting students to sit around the charts and discuss the colors and shapes that they perceived (Ronda 1999). We were captivated by how, in this particular use, the mural charts seemed to anticipate a form of embodied, experiential learning. We were also taken with the experiential aspects of making the charts-- the "fatigue" and the "drudgery"-- that Peabody describes in her letter. We thus embarked upon a project to recreate Peabody's lost mural charts using physical computing materials, amplifying the embodied and interactive aspects of interpreting the charts that are documented in these archival fragments, and attending to the additional experiential aspects of our own chart-making process. In doing so, we bring together historical fabrication work (e.g. Sayers 2015) with feminist making (e.g. Losh and Wernimont 2014).

Chart Design and Implementation

The reimagined mural chart consists of three layers: a fabric layer that approximates Peabody's original canvas (figure 2, left); a grid of 900 individually-addressable LEDs (figure 2, right); and a soft-button touch interface for toggling each LED off and on (figure 3). The result is an illuminated touch interface that conveys the abstraction of the original grid and the embodied nature of the learning experience, enhanced by contemporary technologies.

Strips of conductive copper tape, arranged in a 30 x 30 matrix and positioned on soft neoprene, are used to register the location of each button press. Two Arudino Megas, daisy-chained together, determine the column and row of the touch. A third Mega, also daisy-chained, takes the location of the button press and illuminates the corresponding LED.

Left: Fabric layer before assembly. Right: LED layer before assembly.



Left: The conductive layers of the touch interface. Right: The assembled touch interface.

Next Steps

Currently, the chart allows the user to touch any square to turn on the corresponding LED. The next steps are to design and implement the interaction that will allow the user to create and input their own events; and to design and implement a color picker, perhaps employing a digital interface. The goal for this phase of the project is to complete a start-to-finish interaction from selecting a historical event, choosing its color and position, and then visualizing it on the mural chart.

References

Klein, L., Foster, C., Hayward, A., Pramer, E., and Negi, S. (2017). The Shape of History: Elizabeth Palmer Peabody's Feminist Visualization Work. *Feminist Media Histories* 3 (3): 149-153.

Ronda, B. (1999). *Elizabeth Palmer Peabody: A Reformer on Her Own Terms*. Cambridge: Harvard University Press.

Sayers, J. (2015). Prototyping the Past. *Visible Language* 49 (3): 156-177.

Wernimont, J. and Losh, E. (2014). Feminist Digital Humanities: Theoretical, Social, and Material Engagements around Making and Breaking Computational Media. https://jwernimont.com/2014/06/02/feminist-digital-humanities-theoretical-social-and-material-engagements-around-making-and-breaking-computational-media/ (accessed 24 April 2018).

TOME: A Topic Modeling Tool for Document Discovery and Exploration

Adam Hayward
adam.hayward@gatech.edu
Georgia Institute of Technology, United States of America

Nikita Bawa
nbawa3@gatech.edu
Georgia Institute of Technology, United States of America

Morgan Orangi
moorangi@gatech.edu
Georgia Institute of Technology, United States of America

Caroline Foster
cfoster2@gatech.edu
Georgia Institute of Technology, United States of America

Lauren F. Klein
lauren.klein@lmc.gatech.edu
Georgia Institute of Technology, United States of America

Introduction and Overview

In the past several years, the utility of topic modeling for the humanities has been clearly established. Scholars can now point to projects that convincingly employ topic modeling to explore the figurative language employed in ekphrastic poetry (Rhody 2012), to trace the "quiet transformations" of literary studies (Goldstone and Underwood

2014), and to distill the epistemic dimensions of novels (Erlin 2017), among others. And yet, broader applications of the technique remain limited by the computational and statistical expertise required to implement a topic model and interpret its results. While there has been some work to develop topic model "browsers" (e.g. Goldstone 2014, Murdock and Allen 2015), these projects are designed to facilitate the exploration of the model itself, rather than to leverage the affordances of topic modeling for humanities scholars. By contrast, our interface was conceived so that non-technical humanities scholars can employ a topic model of their corpus in order to discover the documents most salient to their research (Klein et al. 2015).[1]

Corpus, Model, and Database

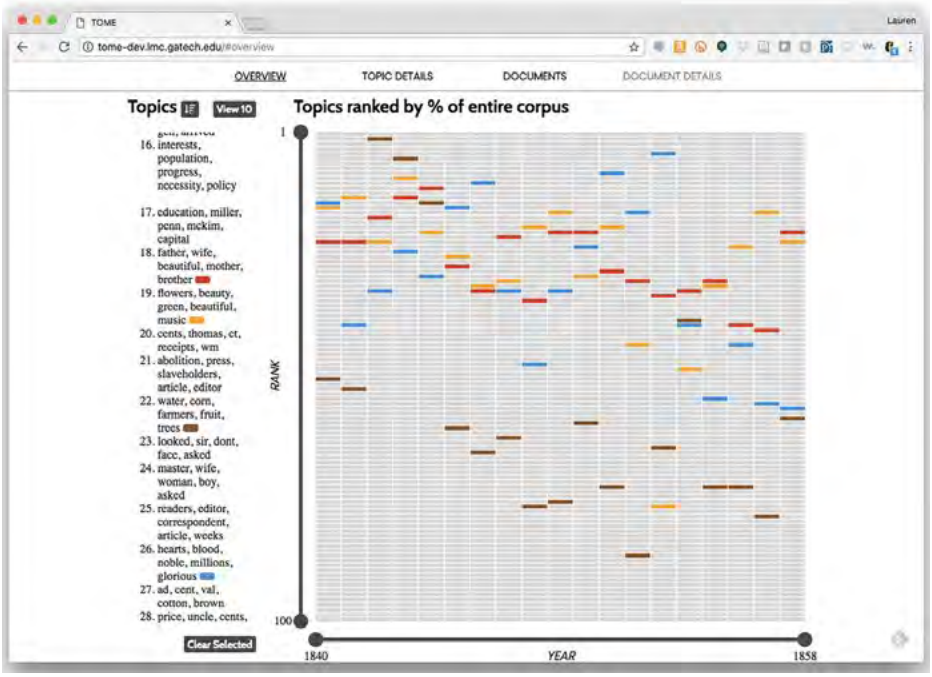
Our corpus consists of nearly 300,000 documents drawn from a collection of nineteenth-century abolitionist newspapers. The documents were scraped from the Accessible Archives website, as per an agreement with Accessible. Additional cleaning of the data, as well as metadata creation, was performed through custom Python scripts.

The topic model of our corpus was created using gensim, the vector space and topic modeling library (Rehurek and Sojka 2010). We employed gensim's wrapper for Latent Dirichlet Allocation (LDA) from MALLET (McCallum 2002). We generated 100 topics after 100 iterations, filtering the 100 most common words. We printed the topics and topical composition of each document to CSV files. We then ingested the data into a MySQL database using Django's ORM framework.[2]

Interface and Sample Interaction

Our interface is the result of a several-month design process during which we considered a variety of user scenarios. Our goal was to scaffold the process of document discovery so that the user could draw new insights as they moved through each section of the interface: Topic Overview, Topic Details, Document Overview, and Document Details.[3]

The user begins with the Topic Overview section (Figure 1), which employs a custom visualization in order to display each of the 100 topics according to its change in rank over time. The user can also filter the topics by keyword or sort according to overall prevalence.



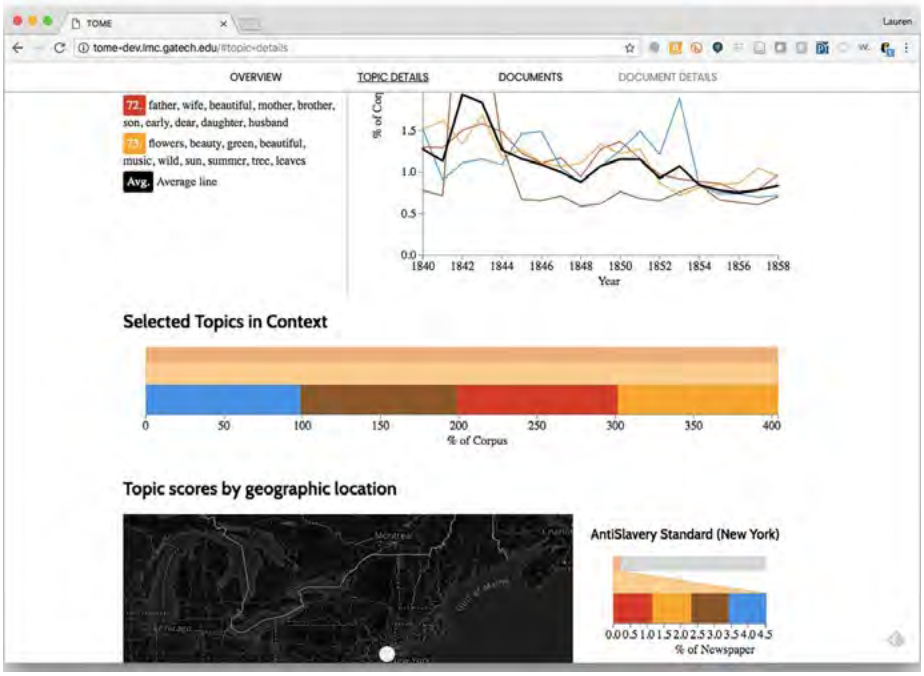
Topic Overview

When the user has selected their topics of interest, they scroll to see details about those topics: change in percentage of the corpus over time; distribution in each newspaper over time; and geographic distribution (Figure 2). These visualizations work together to show which topics were most prevalent at which times; which sources were reporting on which topics at particular times; and where each topic was being reported on. From there, the user can either return to the Topic Overview to further refine the topic set (Gelman 2004), or scroll down to the Document Overview section.

1 The first round of research on TOME was conducted between 2013 and 2015 in collaboration with Jacob Eisenstein, Assistant Professor of Interactive Computing at Georgia Tech, funded by NEH Office of Digital Humanities Startup Grant HD-51705-13. See Klein et al. 2015.

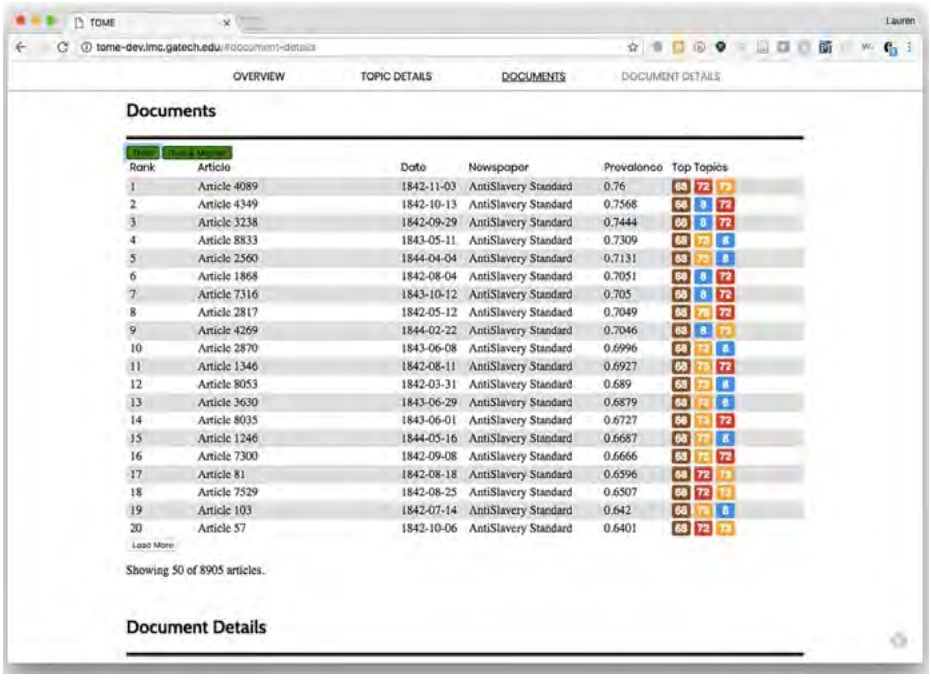
2 The topic model and related processing scripts can be found at: https://github.com/GeorgiaTechDHLab/TOME/.

3 A live version of this interface can be found at: http://tome.lmc.gatech.edu/.



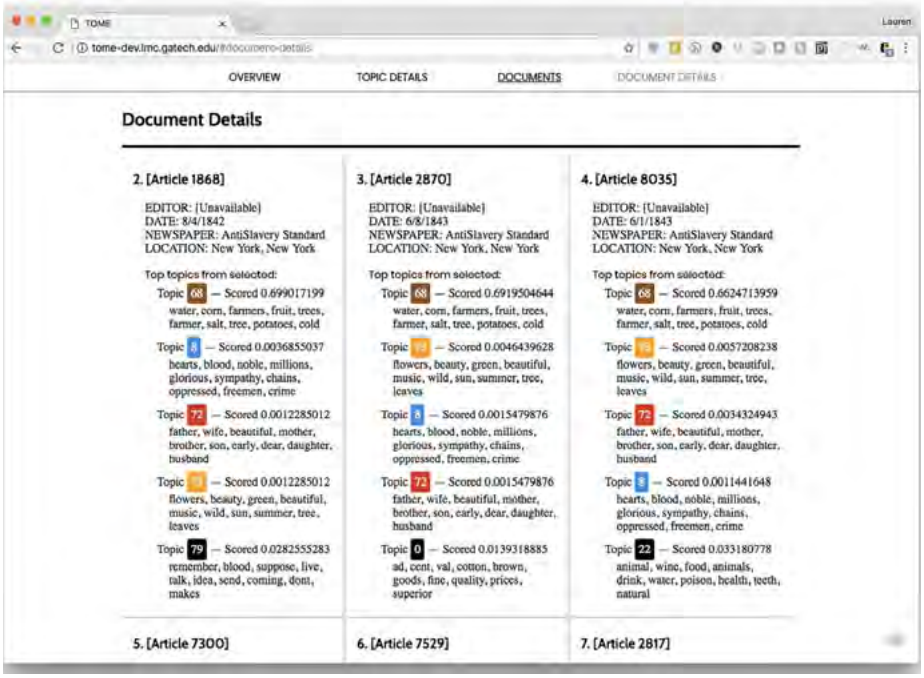
Topic Details

The Document Overview (figure 3) section allows the user to further refine the set of documents they will eventually read. They can toggle between a standard list view of all the documents, ranked in terms of what percentage of the selected topics they contain, and a dust-and-magnets view (Yi et al. 2005).



Document Overview

From there, they move to Document Details (figure 4), which displays the metadata associated with each article in the corpus, ordered according to the percentage of the selected topics they contain. This allows the user to click through to the articles themselves, having narrowed down a set of articles relevant to their research.



Document Details

The interface is implemented using HTML and JavaScript, including D3.js, the JavaScript-based visualization library, and AJAX for client-side data retrieval.

Initial research on TOME was conducted from 2013 to 2015 in collaboration with Jacob Eisenstein, School of Interactive Computing, Georgia Institute of Technology, funded by NEH Office of Digital Humanities Startup Grant HD-51705-13.

References

Erlin, M. (2017). Topic Modeling, Epistemology, and the English and German Novel. *Cultural Analytics*.

Gelman, A. (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics* 13 (4): 755–779.

Goldstone, A., and Underwood T. (2014). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History* 45 (3): 359-384.

Goldstone, A. (n.d.). DfR Browser. https://agoldst.github.io/dfr-browser/ (accessed 25 April 2018).

Klein, L., Eisenstein, J., and Sun, I. (2015). Exploratory Thematic Analysis for Digitized Archival Collections. *Digital Scholarship in the Humanities* 30 (Supp. 1): i130-i141.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (accessed 25 April 2018).

Murdock, J. and Allen, C. (2015). Visualization Techniques for Topic Model Checking. *AAAI Conference on Artificial Intelligence*, Austin, TX, January 2015.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valetta, Malta, May 2010.

Rhody, L. M. (2012). "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2 (1).

Yi, J.S. (2005). Dust & Magnet: Multivariate Information Visualization Using a Magnet Metaphor. *Information Visualization* 4 (4): 239-256.

Bridging Digital Humanities Internal and Open Source Software Projects through Reusable Building Blocks

Rebecca Sutton Koeser
rebecca.s.koeser@princeton.edu
Center for Digital Humanities,
Princeton University, United States of America

Benjamin W Hicks
bhicks@princeton.edu
Center for Digital Humanities,
Princeton University, United States of America

Software development is often an integral aspect of Digital Humanities projects. By working to generalize and build small modules or utilities targeting specific needs rather than large-scale systems, DH software developers have the capacity to generate tools with greater potential for scholarly reuse, which should enable more rapid development on future projects, and allow developers to focus on innovative work. This poster demonstrates a case study of modular software developed as part of ongoing DH projects.

There is a tendency among some institutions, particularly libraries, to adopt existing large-scale Open Source Software solutions and adapt them for local needs; but as Hector Correa points out, this approach results in skipping the work of thinking carefully about users and local needs (Correa, 2017). If large-scale software solutions developed by coalitions of libraries are problematic (Princeton University Library Systems, 2017) where needs are at least similar, even where content structures or workflows differ, this problem is redoubled for research software, which is much more likely bespoke to a particular problem. As Correa argues, single-purpose software is less complex and easier to understand and manage; and understanding the logic of code is crucial for research that is based on or otherwise makes use of software (Koeser, 2015).

Applying best practices from software development such as modular design can mitigate these problems through an emphasis on delivering working components of software and focusing on simplicity of purpose—a single, well-honed and balanced knife rather than a multi-tool with every imaginable attachment. This approach is consistent with the design philosophy from one of the greatest success stories of modern open-source software, UNIX and its derivatives (Raymond, 2003).

There are certainly possible drawbacks and concerns about this approach. It may require more effort, and perhaps different skills, to create, release, and manage independent software packages or modules. According to Glass' *Facts and Fallacies of Software Engineering*, it is "three times as difficult to build reusable components as single use components" (Glass, 2003: 49). In our case, when new software modules were being developed and extended in tandem with an existing software project, finalizing a new release of that project involved releasing and publishing multiple software modules. There is also a danger of generalizing too soon; another familiar rule of thumb in software is that you have to do something three times before you know how to generalize it properly (Glass, 2003).

As a case study, our poster will present an overview of the software written for two annotation projects that were developed at the same time. "Derrida's Margins" analyzes the work of Jacques Derrida through references in *De la grammatologie* and corresponding annotations in the books he cited. "The Winthrop Family on the Page" examines a community of readers connected through books over time via annotations. This software ecosystem includes two project codebases (Koeser et al., 2018; Koeser and Hicks, 2018a) that make use of four new reusable components (Koeser and Hicks, 2018b; Koeser, 2018b), two of which (Koeser, 2018a; Koeser and Hicks, 2018c) were adapted from the "Readux" codebase (Koeser et al., 2017), which was previous developed at Emory University. In the process, we also used and made minor updates to a related, pre-existing module (Koeser, 2018c).

For each of these tools, a use case emerged in one project which could be generalized for other projects, with potential for broader reuse. As an example, "viapy"—a Python module for searching and providing VIAF data to a web framework—was adapted from previous work, and first existed as code for one of the annotation projects, but it proved generalizable. In fact, it proved easier to extract as a reusable component rather than duplicate; one project team discovered a bug that had previously gone undetected, and creating a reusable package allowed us to correct the problem once for both projects. Likewise, code for storing and displaying annotations from the Readux project was ripe for repackaging as a general module because of its relatively direct purpose despite the different intellectual aims of these projects. However, these codebases also contain similar, potentially reusable functionality that is not yet ready for generalization.

These projects provide a view into the ongoing process of balancing customized solutions to DH projects with generalizing focused portions of functionality. Modular design aimed at 'doing one thing and doing it well' offers the possibility of creating an ecosystem of reusable packages that are widely useful and applicable, and can participate in a larger community of open source and other DH software research.

References

Correa, H. (2017). Build your own software *Hector Correa* http://hectorcorrea.com/blog/build-your-own-software/70 (accessed 28 November 2017).

Glass, R. L. (2003). *Facts and Fallacies of Software Engineering*. Addison-Wesley Professional.

Koeser, R. S. (2015). Trusting Others to 'Do the Math'. *Interdisciplinary Science Reviews*, 40(4): 376–92 doi:10.1080/03080188.2016.1165454. http://dx.doi.org/10.1080/03080188.2016.1165454 (accessed 29 June 2016).

Koeser, R. S. (2018a). *Django-Annotator-Store: Django Application to Act as an Annotator.js 2.x Annotator-Store Backend*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/django-annotator-store.

Koeser, R. S. (2018b). *Viapy: VIAF via Python*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/viapy.

Koeser, R. S., Glover, K., Li, Y., Varner, J. and Thomas, A. (2017). *Readux: Django Web Application to Display, Annotate, and Export Digitized Books in a Fedora Commons Repository*. JavaScript Emory Center for Digital Scholarship https://github.com/ecds/readux.

Koeser, R. S. and Hicks, B. W. (2018a). *Winthrop-Django: Django Web Application for the Winthrop Family on the Page Project*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/winthrop-django.

Koeser, R. S. and Hicks, B. W. (2018b). *Django-Pucas: Django App to Streamline CAS Auth and Populate User Attributes from LDAP*. Python Center for Digital

Humanities at Princeton https://github.com/Princeton-CDH/django-pucas.
Koeser, R. S. and Hicks, B. W. (2018c). *Djiffy: Django Application to Index and Display IIIF Manifests for Books*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/djiffy.
Koeser, R. S., Hicks, B. W., Glover, K. and Budak, N. (2018). *Derrida-Django: Django Web Application for Derrida's Margins*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/derrida-django.
Koeser, R. S. (2018). *Piffle: Python Library for Generating and Parsing IIIF Image API URLs*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/piffle.
Princeton University Library Systems (2017). Valkyrie *Princeton University Library Systems by Pulibrary* https://pulibrary.github.io/2017-07-06-valkyrie (accessed 28 November 2017).
Raymond, E. S. (2003). *Art of Unix Programming, The*. Addison-Wesley Professional http://proquest.safaribooksonline.com/book/operating-systems-and-server-administration/unix/0131429019.

Building Bridges Across Heritage Silos

Kalliopi Kontiza
kalliopi.kontiza@ng-london.org.uk
The National Gallery, United Kingdom

Catherine Jones
catherine.jones@uni.lu
University of Luxembourg, Luxembourg

Joseph Padfield
joseph.padfield@ng-london.org.uk
The National Gallery, United Kingdom

Ioanna Lykourentzou
ioanna.lykourentzou@list.lu
Luxembourg Institute of Science and Technology, Luxembourg

Building Bridges aCROSS CULTural Heritage Silos

This research considers how best to cross the divides that exist between: (1) disparate practices between research fields (2) disparate interpretations of shared cultural heritage by the public and (3) disparate cultural heritage objects.

Figure 1The CrossCult Consortium and Partners

Building bridges across disciplines

Within the field of heritage research there still remain, to this day, many silos between researchers in sciences or the humanities, professionals, practitioners and information technologists. In this poster we consider how best to bridge these gaps between the disciplines. We present, as a demonstrator, an H2020 project named CrossCult (http://www.crosscult.eu). The project brings together inter-disciplinary researchers including: Social scientists, Data and Information scientists, Heritage and Digital Heritage Scientists, Engineering, Humanities and Digital Humanities (Archaeologists and Digital Archaeologists, Linguists, Museum Professionals), Practitioners (Conservators, Curators), and Information Technologies (Backend and Front end and app Developers, Programmers,

Semantic Web specialists, Gamers). We achieve collaboration and discussion through shared common goals and research objectives, and we support dialogue through tools. When possible, we use open source technology to support us for Communication, Programming, Data Structuring/Editing, Visualisation, Conceptual Mapping. We follow standards to be compatible with other people's work, produce reusable research outputs and collaborate with other European projects towards the same goal.

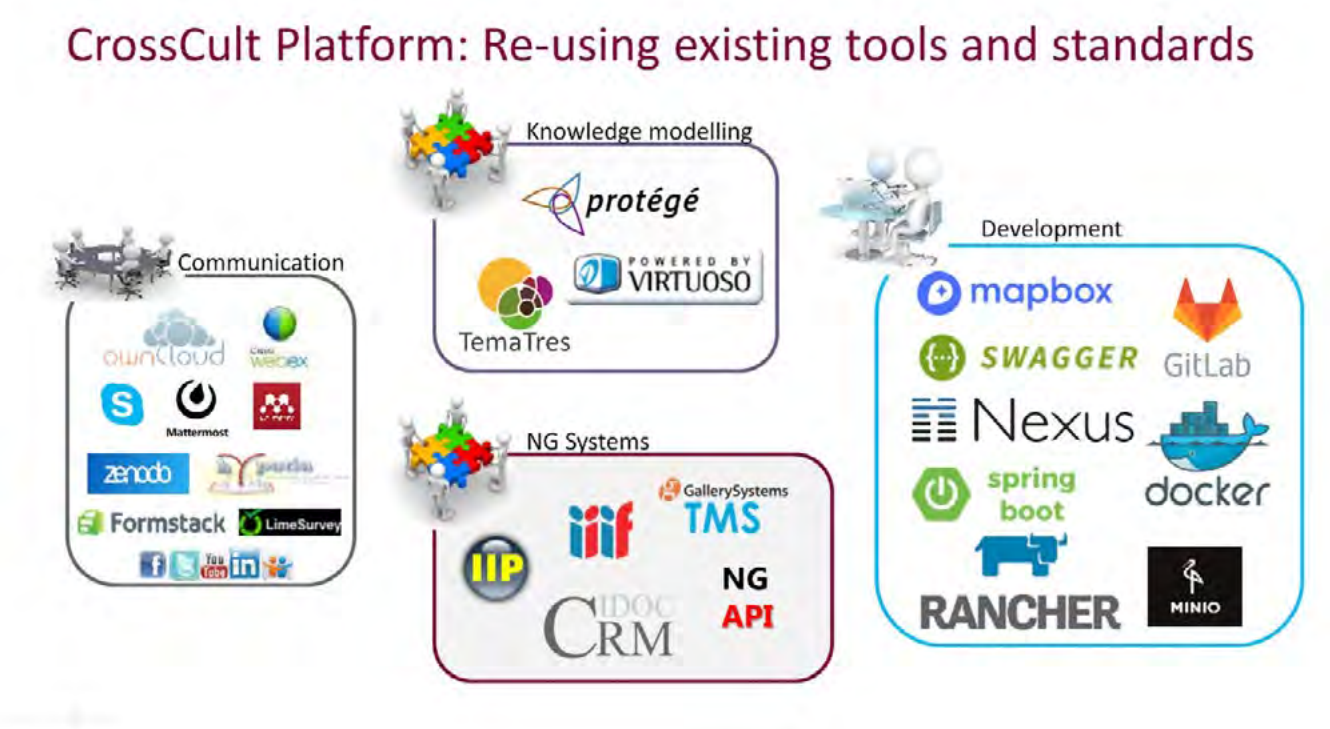


Figure 2 Reusing existing tools and Standards

Building bridges across members of the public

The challenge also extends beyond researchers and continues into the lived experience of our shared Heritage. It raises the challenge of how CROSSCultures can challenge siloed opinions and interpretations of Cultural Heritage (Lykourentzou et al., 2016). At the heart of this project is the desire to build bridges between disciplines to explore innovative practices that can present historic knowledge to non-specialist audiences in an engaging way. European history is an exciting mesh of interrelated facts and events, interpretations and narratives that cross countries and cultures. However, public history is a challenging practice that must be mindful of the audiences, their interests and goals; in this research we are concerned with the museum or the city visitor (Vasilakis et al., 2016).

Building bridges across cultural objects

The final challenge we explore is how to build bridges between disparate objects of our common Heritage. We use heritage objects and historical resources to trigger reflection, individually and collectively, on European history and to showcase the importance to bridge the past and its connection to the present (ERCIM News, 2017).

Using the CROSSCULT project we demonstrate how we are can address the three challenges by developing around four use cases: from large museums to small ones, and from indoors to outdoors. In this presentation, we discuss two of the project's four pilots (Pilot 1 and Pilot 4), which highlight the comparison between the *Indoor and Outdoor Exhibition*; in the first case with the museum/gallery and its paintings and in the second case with the city and its geo-located Places of Interest (POIs). The exhibits (both POIs and paintings) are represented as semantically structured data, linked through our Knowledge Base (Vlachidis et al., 2017). They are our stepping stones to create stories that connect one item to the other, and invite the user (gallery visitor or city traveller) to discover them. The POIs are either discovered outside (in Pilot 4) and can lead to the museum/gallery or vice versa (Pilot 1), eventually bridging the outdoors with the indoors and creating a seamless cultural discovery experience.

Figure 3 CrossCult H2020 project – Overview of the four pilots and their supporting technologies

Pilot 1: Large multi-thematic venue - The National Gallery London- Building narratives through personalisation

We use the gallery's large collection to offer the visitors personalised stories that highlight the connections among people, places and events across European history, through art. Semantic reasoning, recommender systems and path routing optimisation are employed to ensure that each visitor will be navigated through the conceptually linked exhibits that interest them the most, while avoiding congested spaces as much as possible. The experience combines technologies, balancing in a unique way individual visitor needs with museum-wide objectives, can be extended and customised to serve the needs of various other large venues across Europe.

Pilot 4: Multiple cities - City of Valletta in Malta and City of Luxembourg. Building narratives through location based gaming and crowdsourcing.

Pilot 4 takes place outdoors in the two cities to trigger reflection through urban discovery. Focusing on the topic of migration, past for Malta and present for Luxembourg, and using the technologies of location-based services, urban informatics and crowdsourcing, it invites people to walk the two cities, discover and share stories. Visitors and residents engage in comparative reflection that challenges their perception on topics touched by migration such as identity, quality of life, traditions, integration and sense of belonging (Jones et al., 2017).

Acknowledgements:

The work described in this presentation has received funding support from European Union's Horizon 2020 research and innovation programme under grant agreement no 693150.

References

ERCIM News. (2017, October) Reinterpreting European History Through Technology: The CrossCult Project. Retrieved from https://ercim-news.ercim.eu/en111/special/reinterpreting-european-history-through-technology-the-crosscult-project (accessed 02 May 2018)

Jones, C. E., Liapis, A., Lykourentzou, I., Guido, D. (2017). Board Game Prototyping to Co-Design a Better Location-Based Digital Game. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM Press, New York, USA, pp. 1055-64. Available from: https://doi.org/10.1145/3027063.3053348

Lykourentzou, I., Naudet Y., Vandenabeele, L. (2016). Reflecting on European History with the Help of Technology: The CrossCult Project. *Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage*. The Eurographics Association, pp. 67-70. Available from: https://doi.org/10.2312/gch.20161384

Vassilakis, C., Antoniou, A., Lepouras, G., Wallace, M., Lykourentzou, I., Naudet, Y., 2016. Interconnecting Objects, Visitors, Sites and (Hi)Stories Across Cultural and Historical Concepts: The CrossCult Project, in: Ioannides, M., Fink, E., Moropoulou, A., Hagedorn-Saupe, M., Fresa, A., Liestøl, G., Rajcic, V., Grussenmeyer, P. (eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*. Springer International Publishing, Cham, pp. 501–10. Available from: https://doi.org/10.1007/978-3-319-48496-9_39

Vlachidis, A., Bikakis, A., Kyriaki-Manessi, D., Triantafyllou, I., Padfield, J., Kontiza, K. (2017). Semantic Representation and Enrichment of Cultural Heritage Information for Fostering Reinterpretation and Reflection on the European History. *Paper presented to the final Conference of the Marie Skodowska-Curie Initial Training Network for Digital Cultural Heritage, ITN-DCH 2017*. Olimje, Slovenia, 23–25 May.

Voces y Caras: Hispanic Communities of North Florida

Constanza M. López Baquero
constanza.lopez@unf.edu
University of North Florida, United States of America

Voces y Caras: Hispanic Communities of North Florida is an ongoing project that explores the power of digital *testimonio* (Benmayor, 2012) to make visible hidden communities and enable processes of self-discovery by students of Latinx origin in the U.S. The project engages heritage speakers of Spanish in the process of developing questions and recording interviews with members of the Hispanic/Latinx community in North Florida, a population

that has been, according to many, deliberately made invisible.

Since the inception of the project in 2012, 109 interviews have been conducted, recorded, transcribed and archived. The project serves at least four purposes: (1) It recognizes immigrants as an indispensable part of our society in a political environment increasingly hostile to them, (2) it puts students who are heritage speakers of Spanish in contact with their cultural and historical backgrounds, (3) it gives these students the opportunity to recognize themselves in the stories of others, and (4) it serves as a pedagogical tool that creates communities in and outside of the classroom.

Digital *testimonio* provides an important tool for teaching bicultural students who are searching for their own identities, particularly those who live in an area, like North Florida, where they feel pressure to assimilate or avoid the stereotypes that surround being Latinx in the United States. In many cases, these students are largely disconnected from their own histories, as the Hispanic roots of much of the United States, as well as the history of Latin America, are barely present, if represented at all, in mainstream curriculum. As the Latinx community in the U.S. gains visibility, in part through the negative ramifications of the current political climate, these students are increasingly interested in understanding how they fit into a larger Latinx identity, as well as in vindicating the misperceptions or distortions of Latinx people that they witness in the media.

Since our students live in a large geographical area without a center for immigrants, or a specifically Latinx neighborhood like you would find in Orlando or Miami, many feel lost because they are not fully accepted into the mainstream culture. Furthermore, Latinx make up a small percentage of the university's population and this furthers their feeling of alienation. When they come to my class, they learn about the value of community and history. Voces y Caras is a collection of stories that are testimonial and as Rina Benmayor has stated, "*Testimonio*, thus, expresses the central values of situated knowledge production, embodied theorizing, and community engagement, and thus can be considered a signature pedagogy," which can be "grounded in liberatory values and methods." By learning about other Latinx and what they are doing to influence our city, students discover their own stories. The sacrifices and traumas of other immigrants help them shape their own identities and claim their rights to belong to the U.S., and also to the culture where they, or their parents came from. Benmayor highlights the benefit of this type of projects because it "engages students first hand in reproducing the processes of (1) situated knowledge production, (2) embodied theorizing, and (3) collective practice that are foundational to the field. These processes constitute core epistemologies for Latin[x] Studies, ones that we hope all of our students learn to perform in their lives as well as in their professional futures" (2012: 509).

As I ponder upon my project, I believe that its value resides largely with the opportunities it offers for engagement with local communities. As Will Fenton argued in a recent opinion piece in the *Chronicle of Higher Education*, such use of scholarship to connect with the public is sorely lacking in the Digital Humanities today. I believe, furthermore, that this project demonstrates how digital approaches can be deployed in ways that are truly transformational for students from a variety of disciplinary backgrounds.

There is an organic connection between oral history projects and digital humanities. Listening to the stories of others make us more empathetic. These stories arouse feelings of love and compassion because we can recognize our stories in others. In this line, Voces y Caras highlights the achievements of the community. This is particularly relevant in our present political environment where immigrants have been perceived as a problem rather than what they are; an indispensable part of our society that contributes greatly to its growth. The recordings, excerpts of the interviews, and pictures of the interviewees are available online at vocesycaras.weebly.com

References

Benmayor R. (2012). Digital Testimonio as a Signature Pedagogy for Latin@ Studies. *Equity and Excellence in Education.* 45, 507-524.

Benmayor, R. (2008). Digital Storytelling as a Signature Pedagogy for the New Humanities. *Arts and Humanities in Higher Education.* 7, 188-204.

Fenton, W. (2018). Literary scholars should use digital humanities to reach the oft-ignored 'public' (opinion). *Technology and Learning Blog: Inside Higher Ed.* 2018-01.

Empatía Digital: en los pixeles del otro

Carolina Laverde
ca.la1412@gmail.com
Biblioteca Nacional de Colombia, Colombia

Vivimos en la sociedad de la imagen y la información (Manuel Castells, 1996) que se caracteriza por la hiperproducción de conocimiento. Esto representa un verdadero desafío estructural a la hora de formular proyectos relevantes en los que el desarrollo y la investigación no sean los únicos enfoques de un humanista digital: se requiere aplicar la empatía digital como puente que equilibre el desarrollo de productos digitales.

"La empatía digital es un proceso en el cual una persona puede analizar > reflexionar > proyectar > predecir > sentir mediante la comunicación con lo digital" (Friesem, 2105). La empatía es un proceso subvalorado como herramienta en las primeras etapas de la creación

de un proyecto, cuando realmente la empatía aplicada a los contextos digitales es crucial para poder formular y formar productos responsables, sostenibles y cohesivos desde un contexto de creación multidisciplinar.

Asimismo, para desarrollar proyectos en Humanidades Digitales resulta necesario estructurarlos a partir de tres preguntas: ¿qué se quiere generar?, ¿cómo se quiere construir? y ¿cómo se va a presentar?, y de esta manera encontrar **insights** que generen empatía con el usuario, que den cuenta de sus motivaciones y gustos para lograr proyectar un tono de comunicación, línea de pensamiento e interacción, entendiendo desde un plano mucho más profundo las necesidades de del usuario para poder crear un resultado y producto más efectivo (McDonag y Tomas J, 2010) al establecer una conexión emocional que se convierte en una oportunidad creativa.

> Es la habilidad cognitiva y emocional de ser reflexivo y socialmente responsable mientras se utilizan estratégicamente medios digitales (Friesem, 2015).

En otras palabras, enviar el mensaje en el formato adecuado apelando a la sensibilidad del público objetivo y a direccionar una estrategia de valor a través de la emoción lleva a "humanizar los productos digitales", al mismo tiempo que permite observar y analizar más allá de la superficialidad comercial de algunas herramientas como *focus group* o encuestas, al identificarse con estados emocionales, cognitivos y con actitudes de otras personas por medio de la experiencia indirecta, es decir, "ponerse en los zapatos del otro".

En este orden de ideas, mi propuesta es un poster que permita a los participantes del congreso encontrar **insights** de una manera sencilla a través de una caja de herramientas que funcione como base de un proyecto acertado y sostenible. Por lo tanto, este poster permitirá al usuario llevarse algo práctico de él con claves rápidas y pasos simples para empezar a fortalecer la habilidad der ser empático y así utilizarlo como una herramienta para conectar a un nivel emocional como valor agregado a los proyectos digitales.

Sobre el contenido del poster se plantean 3 formatos con ejercicios y técnicas básicas como primer acercamiento al concepto de empatía digital divido en tres secciones a partir de tres preguntas básicas que son ¿qué? ¿cómo? ¿por qué? Con la finalidad de **Sentir + compartir+ reaccionar = experiencia de usuario.**

Finalmente se quiere evidenciar los procesos creativos de la Biblioteca Nacional de Colombia en el área de Humanidades y Desarrollos Digitales y como han se han transformado utilizando este puente como herramienta que serán aplicados en la creación de este poster ya que después de un proceso de conceptualización por medio de metodologías como design thinking entre otras herramientas y por su puesto desde la empatía digital se utilizarán unos colores específicos por conceptos que se van a comunicar basados en la teoría del color para aplicarlos mediante técnicas como: ilustración digital y tipografías que hacen alusión a la estética del mundo digital como los pixeles o el código.

La interacción será análoga en la medida que el usuario pueda entender los insights para ponerlos en práctica en sus procesos al poder revelar el contenido del poster con ayuda de un elemento externo para poder filtrar el contenido de cada color, esto es posible al hacer uso de un recurso visual como la adición de colores primarios RGB. En ese orden de ideas al tener todos los contenidos impresos al mismo tiempo cada tipo de contenido en una tinta (verde, rojo o azul), se genera una recarga o confusión visual resultando complicado para el usuario entender la información en la primera impresión. Al ayudarse con los elementos de filtrado de color pueden obtener la información por medio de filtrado por lo tanto se propone es que haya un cambio de visión y perspectiva como lo requiere la habilidad de ser empático para su posterior aplicación a proyectos de humanidades digitales.

References

Dave M Berry. (2010). *The Computational Turn: Thinking About the Digital Humanities*.

Yonty Friesem. (2016). *Chapter 2 - Empathy for the Digital Age: Using Video Production to Enhance Social, Emotional, and Cognitive Skills*.

IDEO. (2016). *What is Human-Centered Design?*

Jon Kolko. (2010). Abductive Thinking and Sense making: *The Drivers of Design Synthesis*. Vol. 26, No. 1 (Winter, 2010), pp. 15-28.

Jon Kolko. (2015). *Design Thinking Comes of Age*.

Hasso Plattner (2013). *Empathy field guide. Institute of Design at Stanford.*

Mark Considine (2012). *Thinking Outside the Box? Applying Design Theory to Public Policy: Applying Design Theory to Public Policy.*

McDonagh y Deana. (2010). *Rethinking Design Thinking: Empathy Supporting Innovation.*

Sanhueza y Camila Holven. (2012). *Design Empathy in Service Design Methodology*.

Elena González García (2016). *Un nuevo camino hacia las humanidades digitales: el laboratorio de innovación en humanidades digitales de la uned (linhd)*.

Atlas de la narrativa mexicana del siglo XX y la representación visualizada de México en su literatura. Avance de proyecto

Nora Marisa León-Real Méndez
nora.marisa@itesm.mx
Tecnológico de Monterrey, Mexico

En esta presentación se busca mostrar el avance obtenido en un año de trabajo en el proyecto de creación de un mapa literario de México en el que se representen gráficamente las obras y los espacios en las que éstas se desarrollan. El *Atlas de la Narrativa Mexicana del siglo XX* compila y presenta visualmente información geográfica proveniente de las obras más representativas de la literatura mexicana contemporánea con un propósito educativo. Además, el proyecto busca servir como base para la realización de conexiones sociohistóricas que los estudiantes pueden realizar, pues la visualización de las distintas versiones de México presentes en la literatura es un paso importante para la evolución de la identidad cultural del país, así como una manera innovadora de reconocer nuestra realidad dentro de los textos. Por otra parte, este proyecto requiere analizar la narrativa sobre México utilizando herramientas de análisis literario, historia y geografía, a través de medios digitales. Esta naturaleza interdisciplinaria vuelve al proyecto pertinente dentro del marco de las Humanidades Digitales y arroja ya resultados que contribuyen a la metodología de su aplicación en clase.

El primer paso del proyecto (aprobado por la Convocatoria de Experimentación en Innovación Educativa NOVUS en agosto de 2017) ha sido recopilar la información necesaria, creando un corpus de las novelas más representativas de la literatura mexicana del siglo XX (de inicio, por medio de un compilado de Novelas de la Revolución Mexicana: *Los de abajo*, de Mariano Azuela; *El águila y la serpiente*, de Martín Luis Guzmán; *Cartucho*, de Nellie Campobello; y *Los relámpagos de agosto*, de Jorge Ibargüengoitia), considerando la representación narrativa que hacen del espacio mexicano. Luego, se asignó la lectura de los primeros textos a los participantes del proyecto para realizar las anotaciones y capturar los datos. Con esta información se crearon categorías espaciales que puedan ser marcadas en un mapa de México, de acuerdo con el estado, región o población mencionados en las obras. Por otra parte, estos espacios han sido también clasificados en dos categorías narrativas: aquellos en los que sucede la acción de la novela y los que son mencionados como referentes de eventos fuera de la trama. Esta información se ha vertido en un primer borrador del *Atlas*, un mapa digital realizado con herramientas de acceso abierto propias de las HD, en el que se proyecta visualmente la información de forma que se pueda interactuar con ella: conocer qué porciones del territorio mexicano aparecen con mayor frecuencia en las obras, u observar la predominancia de los espacios rurales o urbanos, por ejemplo. Esta información cartográfica nos permite sacar ya algunas conclusiones con respecto a la representación de la Revolución Mexicana en la literatura, considerando los espacios de acción de las obras en su proporción con la extensión geográfica del país y de los hechos sucedidos en la historia de México. Pero, sobre todo, este proceso ha servido como práctica para proponer el método de creación del *Atlas* así como las áreas en las que hay oportunidad de mejora.

Eventualmente, se busca que el *Atlas* pueda ser utilizado como herramienta de enseñanza de la literatura mexicana en cursos de preparatoria y profesional, permitiendo a los estudiantes contribuir en su crecimiento, aportando nuevos datos según sus lecturas. La información recopilada de manera gráfica permitirá continuar encontrando conexiones entre distintas obras y movimientos literarios, que luego podrían ser analizados por estudiantes e investigadores de la literatura mexicana contemporánea.

Al final del proyecto, se espera contar con un producto demostrable y perfectible (el *Atlas de la Narrativa Mexicana del siglo XX*), así como con grupos de estudiantes que han pasado por el proceso de contribuir a su creación y que, a través de ello, han aumentado su interés y desempeño en las clases de literatura mexicana. De manera tangible, los alumnos serán capaces de mostrar en un mapa de México los espacios detectados dentro de las obras literarias leídas, así como de explicar distintas relaciones entre el espacio y la obra.

HuViz: From _Orlando_ to CWRC… And Beyond!

Kim Martin
kmarti20@uoguelph.ca
University of Guelph, Canada

Abi Lemak
alemak@uoguelph.ca
University of Guelph, Canada

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Chelsea Miya
cmiya@ualberta.ca
University of Guelph, Canada

Jana Smith-Elford
smithelf@ualberta.ca
University of Guelph, Canada

The Orlando Visualizer (OViz) was originally conceived in 2010 as a tool that would display extracts from The Orlando Project's textbase as a series of interconnected nodes in a graph. Since then, the project has grown to address digital humanities research more generically. Now called HuViz (fig. 1), the Humanities Visualizer is a browser-based, interactive interface that allows for the exploration of semantic relationships and ontologies represented using Linked Open Data (LOD). LOD is a practice of creating, sha-

ring, and interlinking bits of information on the Semantic Web (linkeddata.org). At its core, LOD is a way of structurally representing data as connected. More broadly, it challenges how information networks are built within digital environments and calls attention to the importance of making these networks and the data they house open and accessible. In the spirit of LOD, HuViz came together as a tool designed to make available the contents, along with the contexts, of portions of the Semantic Web to experts and lay-users alike in ways that are open, editable, and transferable. This poster will provide an overview of HuViz's development, shaped by the results of user-testing, the demands of Orlando's complex data, as well as the growing ontology of the Canadian Writing Research Collaboratory (CWRC), which is building out from the Orlando data. Future possibilities include use by other projects housed by CWRC's infrastructure (see beta.cwrc.ca).

The CWRC ontology team has been using HuViz to visualize the Orlando datasets, translating the textbase's XML-encoded entries on women writers into RDF assertions (also referred to as triples) (Simpson and Brown, 2013). The test extractions made from the Orlando textbase range from Virginia Woolf to Margaret Atwood, encompassing everything from the schools they attended, to the places they lived, the writers they influenced, and the overlapping and often contradictory cultural forms that contribute to an author's social identity. Given the scope of the Orlando data, which contains millions of connections, as well as the immeasurably bigger Semantic Web itself, the task of visualizing massive hoards of data in meaningful ways remains a central question in developing HuViz. This problem of visualizing large-scale datasets is by no means new for digital humanities scholars (Duke, 2005; Sherratt, 2011). With increased attention paid to the value of LOD for humanities scholarship in the past decade, the question of how to make these graphs both interactive and legible has arisen as a major concern (Katifori, 2007; Ghorbel et al., 2016). Beyond interface design, questions of tool mediation and the "avenues of interpretation" (Warwick, 2012) made available to the user are central to discussions surrounding the false neutrality of technology (McPherson, 2012; Nakamura, 2002; Chun, 2005). With these concerns in mind, the design of HuViz incorporates some aspects of D'Ignazio and Klein's "feminist data visualization" principles (2016). The ability to visualize data along with the structure of the ontology that governs it, for instance, aims to enable interrogation, such as Jacqueline Wernimont's, of "how and where we might locate feminist ideology and politics within digital archives" (2013).

In the latest iteration of HuViz, these concerns have materialized in features supporting:

Context awareness (provision of source snippets; ability to visualize ontologies as well as data; support for web annotation data model)

1. Collaboration (HuViz code available on Github; forthcoming edit button)
2. Transparency (users may import their own data and ontology; CWRC ontology extensively documented and published in HTML)

This poster and tool demonstration will show the growth of this tool over the past several years. The poster will provide an overview of feature development and indicate how a growing body of user tests have shaped that process, highlighting a number of enhanced features. These include:

- Enhanced control over shape, colour, size and weight of edges, nodes, and background both for user preference and to aid accessibility
- Visualization of LOD ontologies
- Loading a dataset or an ontology from a URL (eg. GitHub)
- And perhaps most excitingly, the chance for users to upload and explore their own datasets.

The tool demonstration will introduce attendees to basics of HuViz and invite them to play with it. We will have multiple datasets and ontologies for users to explore, and will provide a link to detailed instructions on how to upload their own datasets or ontologies.

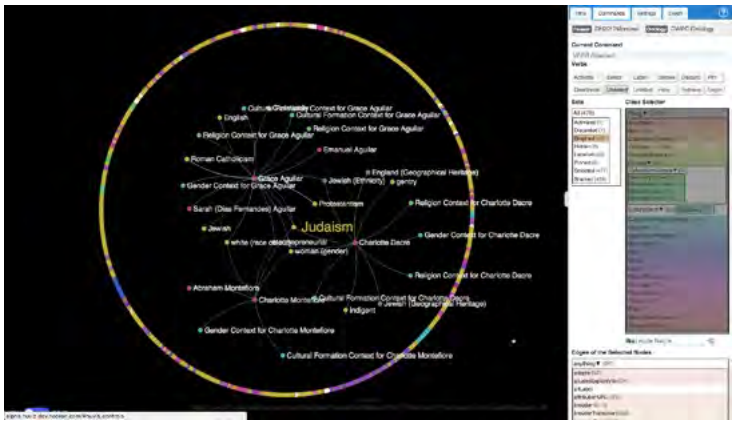


Figure 1. HuViz visualizing Orlando data via CWRC Ontology.

References

Chun, Wendy Hui Kyong. (2005). "On software, or the persistence of visual knowledge." *Grey Room* 18: 26-51.

Canadian Writing Research Collaboratory. beta.cwrc.ca (accessed 25 Nov. 2017).

D'Ignazio, Catherine, and Lauren F. Klein. (2016). "Feminist data visualization." Paper presented at the 2016 IEEE VIS Conference, Baltimore, October 23–28.

Duke, David J., Ken W. Brodie, David. A. Duce, and Ivan Herman. (2005). "Do you see what I mean? [Data visualization]." *IEEE Computer Graphics and Applications* 25.3: 6-9.

Ghorbel, Fatma, Nebrasse Ellouze, Elisabeth Métais, Fayçal Hamdi, Faiez Gargouri, and Noura Herradi. (2016). "MEMO GRAPH: An ontology visualization tool for everyone." *Procedia Computer Science* 96: 265-274.

Humanities Visualizer. http://alpha.huviz.dev.nooron.com/ (accessed 25 Apr. 2018).

Katifori, Akrivi, Constantin Halatsis, George Lepouras, Costas Vassilakis, and Eugenia Giannopoulou. (2007). "Ontology visualization methods—a survey." *ACM Computing Surveys (CSUR)* 39.4: 10.

Linked Data. http://linkeddata.org/ (accessed 25 Nov. 2017).

McPherson, Tara. (2012). "Why are the Digital Humanities so white? Or thinking the histories of race and computation." In M. Gold (ed). *Debates in the Digital Humanities*. Minnesota: University of Minnesota Press, pp. 139-160.

Nakamura, Lisa. (2002). *Cybertypes: Race, Ethnicity, and Identity On the Internet*. London: Routledge.

Sherratt, Tim. (2011). "It's all about stuff: Collections interfaces, power and people." *Journal of Digital Humanities* 1: 1-1.

Simpson, John, and Susan Brown. (2013). "From XML to RDF in The Orlando Project." *Culture and Computing (Culture Computing), 2013 International Conference*. IEEE.

Warwick, Claire. (2012). "Studying users in digital humanities." In *Digital Humanities in Practice*, edited by Claire Warwick, Melissa Terras, and Julianne Nyhan, 1–21. London: Facet Publishing.

Wernimont, Jacqueline. (2013). "Whence Feminism? Assessing feminist interventions in digital literary archives." *DHQ: Digital Humanities Quarterly* 7.1.

Endangered Data Week: Digital Humanities and Civic Data Literacy

Brandon T. Locke
blocke@msu.edu
Michigan State University, United States of America

Endangered Data Week (http://endangereddataweek.org) emerged in the early months of 2017 as an effort to encourage conversations about government-produced, open data and the ways in which it may become endangered due to political, technical, and social factors.

The 2016 US election set off a wave of activism surrounding government data, particularly in the collection and mirroring of environment and climate change data. While much of this attention has been focused on the United States, similar conditions have affected and continue to threaten governments around the world. Endangered Data Week presented an opportunity to funnel even more attention to the issue of potential federal data loss, while also providing opportunities to include lessons on data literacy, civic issues and policy advocacy, data management and curation, technical skills for data capture, and open access and open data in scholarship.

The inaugural Endangered Data Week (April 17-21, 2017) was comprised of 57 formally registered events from 30 institutions and organizations, including virtual participation from hundreds of participants from around the world. The second annual Endangered Data Week will be February 26 - March 2, 2018.

One particularly interesting strain of events in Endangered Data Week is civic data literacy. While so many other projects, including DataRescue, the Preservation of Electronic Government Information (PEGI) project and Environmental Data and Governance Initiative (EDGI) are focused on capturing and preserving government data, Endangered Data Week data literacy events focus on the capacity of the user communities. They seek to enable broader communities to use, interpret, and analyze open data.

The required knowledge and tools for working with civic data overlap significantly with much of the work digital humanists do with data. The creation of datasets often requires scraping information off of the web in flat HTML or confusing databases. Data in both contexts is often irregularly formatted or melded together from multiple sources, requiring the cleaning and reorganization. Meaningful research often requires an iterative process of researching the contexts in which the data was created and the data itself to resolve undocumented meaning in the data. Both contexts also require interpretation for both specialized and non-specialized audiences.

This poster will include a brief overview of Endangered Data Week and will focus on the existing efforts to teach civic data literacy, including an exploratory framework for the most essential skills, knowledges, and tools that are required for diverse communities to use civic data, and the relationship between these events and the broader role of digital humanities faculty, librarians, and staff within our institutions and the communities in which we live.

Herramienta web para la identificación de la técnica de manufactura en fotografías históricas

Gustavo Lozano San Juan
gustavolsj@gmail.com
Instituto de Investigaciones Estéticas
Universidad Nacional Autónoma de México, Mexico

Introducción

Este proyecto consiste de una metodología para identificar el proceso fotográfico en fotografías históricas, está inspirado en el concepto de árbol de decisiones utilizado en las ciencias de datos para clasificar entidades con base en sus diferentes atributos y valores, la implementación ha sido realizada por medio de una herramienta web en idioma español.

Este recurso está dirigido a archivistas historiadores y otros profesionales de archivos históricos en Latinoamérica y les permite identificar el proceso fotográfico entre una gama de 29 alternativas utilizadas a los largo de los siglos XIX y XX, para lo cual los usuarios son guiados paso a paso a través de la metodología por medio de preguntas sobre las características de la fotografía que buscan catalogar.

Figura 1. Diferentes procesos fotográficos históricos

Antecedentes

La identificación del proceso fotográfico es una de las tareas fundamentales que realizan los archivos históricos en el ámbito de la catalogación de fotografías ya que brinda a los investigadores información sobre su temporalidad y características físicas, como el color, el tipo de soporte, el formato, entre otras.

Comúnmente la identificación del proceso fotográfico es una habilidad visual especializada que se transmite de persona a persona de manera empírica mediante la observación detallada de cientos de fotografías y el estudio de su evolución tecnológica. Esta forma de aprendizaje limita la diseminación de este conocimiento entre los profesionales de los archivos y como resultado de ello un gran número de fotografías se encuentran incorrectamente clasificadas dentro de los catálogos.

En la bibliografía sobre conservación de fotografías se han propuesto varios esquemas de clasificación, Lavedrine propone la división inicial de las fotografías por polaridad, posteriormente por soporte y finalmente por tono, aunque este es un modelo útil pone el énfasis en la conservación y no en la identificación (Lavedrine, 2009: 15). Reilly abordan específicamente el tema de la identificación, aunque enfocado únicamente a impresiones del siglo XIX, y no contempla negativos o impresiones a color del siglo XX (Reilly, 1986: 40). El Graphics Atlas (IPI, 2017) es una página que brinda una vasta información que ilustra y describe las características físicas de las fotografías y ayuda al usuario a identificarlas, sin embargo, al igual que las fuentes anteriores se ocupa principalmente de los procesos fotográficos más comunes en los archivos de Estados Unidos y en Europa y su contenido se encuentran en idioma inglés, lo que limita su utilidad y aplicación en archivos de Latinoamérica.

Desarrollo

Una revisión bibliográfica de la literatura en español permitió definir la terminología y los conceptos más adecuados para nombrar cada uno de los procesos fotográficos las características físicas y sus valores (Barra y Gutiérrez, 2000: 19; Boadas et al. 2001: 211; SE, 2016: 20), con esta información posteriormente se elaboró una tabla de datos común para todos los procesos y sus atributos.

Clasificaciones		Atributos comunes				Atributos particulares								
Proceso	Tipología	Soporte primario	Iluminación	Polaridad	Tono	Fechas	Estratigrafía	Magnificación	Tonalidad	Brillo	Superficie	Particularidades Objeto	Texto	Deterioro
Daguerrotipo	Imagen de cámara	Metal	Reflexión	Positivo	Monocromático	1839 - 1860			Neutro	Muy brillante		Positivo-negativo		Delaminación de plata, corrosión de cobre
Ambrotipo	Imagen de cámara	Vidrio	Reflexión	Positivo	Monocromático	1851 - 1865			Café	Brillante		Luces fechosas		
Ferrotipo	Imagen de cámara	Metal	Reflexión	Positivo	Monocromático	1855 - 1890			Café	Semi mate		Magnetico		Corrosión, faltantes, craqueladuras
Cianotipo	Impresión	Papel	Reflexión	Positivo	Monocromático	1840 - 1920	Una	Fibras visibles	Cian	Mate				
Albúmina	Impresión	Papel	Reflexión	Positivo	Monocromático	1851 - 1890	Dos	Fibras visibles	Amarillo, Café rojizo	Semi mat	Textura del pap	Soporte primario delgado, Soporte secundario grueso		Craqueladuras amarillamiento, perdida de densidad en las luces y sombras
Carbón	Impresión	Papel	Reflexión	Positivo	Monocromático	1860 - 1940	Dos		Otro		Relieve en sombras			
Colodión de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1885 - 1910	Tres	No se ven fibras	Purpura_Rojo	Brillante	Textura lisa	Soporte primario grueso, Soporte secundario grueso, Iridiscencia		Abrasiones, perdida de densidad en las luces
Plata gelatina de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1885 - 1910	Tres	No se ven fibras	Amarillo	Brillante	Textura lisa	Soporte primario grueso, Soporte secundario grueso, Iridiscencia		perdida de densidad en las luces
Goma bicromatada	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 1930	Dos		Otro					
Platinotipo	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 1930	Una	Fibras visibles	Neutro					Ghosting
Plata gelatina	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 2018	Tres	No se ven fibras	Neutro					
Colodión mate de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1895 - 1910	Tres	No se ven fibras	Neutro, Café (platino), purpura (oro)	Semi mate	Texturizado	Soporte primario grueso, Soporte secundario grueso y de color,		Ghosting, No hay perdida de densidad
Difusión de plata	Impresión	Papel	Reflexión	Positivo	Monocromático	1942 - 2018	Tres		Neutro		Lisa	Pestañas, borde irregular o perforado, superficie con restos de adhesivo,	Polaroid	Revelado irregular, amarillamiento y desvanecimiento por recubrimiento irregular Craqueladuras
Cromógeno	Impresión	Papel	Reflexión	Positivo	Policromático	1940 - 2018	Tres			Brillante	Lisa o texturizada	Papel de fibra, RC, acetato pigmentado		<1980 amarillamiento, perdida de balance de color y desvanecimiento de colorantes
Difusión de colorantes por transferencia	Impresión	Papel	Reflexión	Positivo	Policromático	1963 - 2018	Tres			Brillante	Lisa	Pestañas, borde irregular o perforado, superficie con restos de adhesivo,	Polaroid	Revelado irregular.
Blanqueo de colorantes	Impresión	Papel	Reflexión	Positivo	Policromático	1963 - 2018	Tres							
Difusión de colorantes por transferencia inferior	Impresión	Papel	Reflexión	Positivo	Policromático	1972 - 2018	Tres			Brillante	Lisa,	Marco blanco con contenedor de químicos de procesado,	Polaroid	Revelado irregular, craqueladuras, migración de colorantes en areas blancas
Colodión húmedo	Negativo	Vidrio	Transmisión	Negativo	Monocromático	1851 - 1885			Café		Aglutinante irregular, Barniz	Vidrio grueso e irregular		Abrasión
Gelatina seca	Negativo	Vidrio	Transmisión	Negativo	Monocromático	1880 - 1925			Neutro		Aglutinante irregular	Vidrio delgado y recto		Espejo de plata
Plata gelatina Nitrato de celulosa	Negativo	Plástico	Transmisión	Negativo	Monocromático	1889 - 1950						Muescas "V"	Nitrate	Amarillamiento del soporte, deformación,
Plata gelatina Acetato de celulosa	Negativo	Plástico	Transmisión	Negativo	Monocromático	1935 - 2018						Muescas "U"	Safety	Canales, burbujas, deformación, olor a vinagre
Plata gelatina Poliester	Negativo	Plástico	Transmisión	Negativo	Monocromático	1955 - 2018						Birrefringencia	Safety	
Cromógeno Acetato de celulosa	Negativo	Plástico	Transmisión	Negativo	Policromático	1942 - 2018							Safety	
Cromógeno Poliester	Negativo	Plástico	Transmisión	Negativo	Policromático	1955 - 2018						Birrefringencia	Safety	
Plata gelatina Vidrio	Transparencia	Vidrio	Transmisión	Positivo	Monocromático	1880 - 1940								
Plata gelatina Acetato de celulosa	Transparencia	Plástico	Transmisión	Positivo	Monocromático	1935 - 2018							Safety	
Plata gelatina poliester	Transparencia	Plástico	Transmisión	Positivo	Monocromático	1955 - 2018						Birrefringencia	Safety	
Procesos aditivos	Transparencia	Vidrio	Transmisión	Positivo	Policromático	1907 - 1935		Retícula			Lineas rectas paralelas	Bajo brillo y saturación		Delaminación, puntos verdes
Cromógeno Acetato de celulosa	Transparencia	Plástico	Transmisión	Positivo	Policromático	1935 - 2018							Safety	
Cromógeno Poliester	Transparencia	Plástico	Transmisión	Positivo	Policromática	1955 - 2018						Birrefringencia	Safety	

Figura 2. Tabla de datos de procesos, características físicas y valores.

Posteriormente la tabla de datos se tradujo en un árbol de decisiones, las características físicas se convirtieron en nodos de decisión, sus valores en ramas y los procesos fotográficos en hojas. Por un lado, este esquema proporciona la estructura de navegación a la página web y por otro le permite al usuario visualizar el panorama del universo posible y comprender las distintas combinaciones de atributos que caracterizan a los procesos fotográficos.

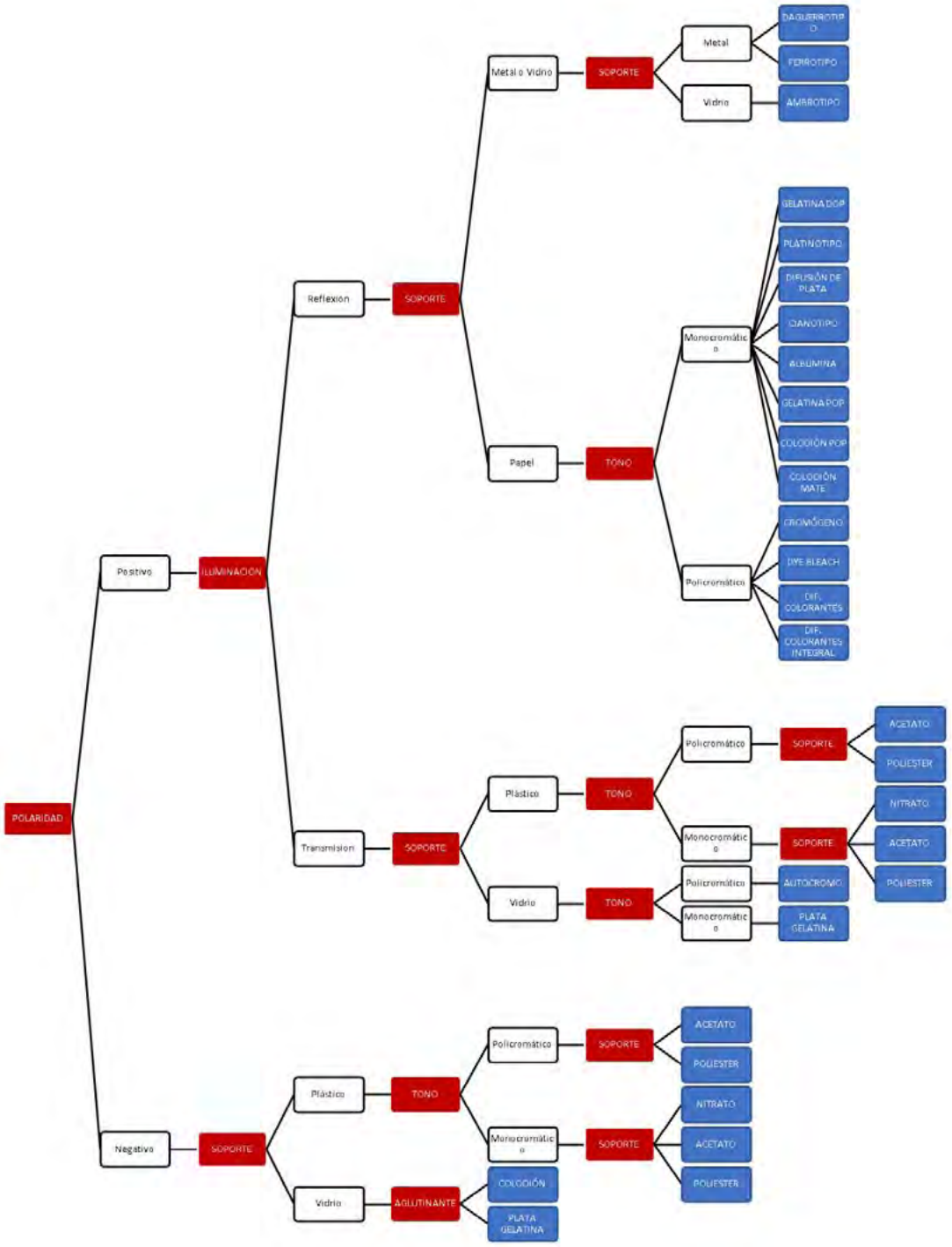


Figura 3. Árbol de decisión para la identificación de procesos fotográficos

Mediante el diseño y programación de la página web se recreó la estructura del árbol de decisión y utilizando preguntas con un lenguaje claro y sencillo se guía al usuario a través de la metodología, las respuestas se ilustran con galerías de imágenes que permiten comparar la fotografía que se busca identificar y encontrar similitudes. El objetivo principal de esta fase del proyecto fue hacer accesibles conceptos que son difíciles de comprender sólo verbalmente pero que son fáciles de reconocer de manera visual.

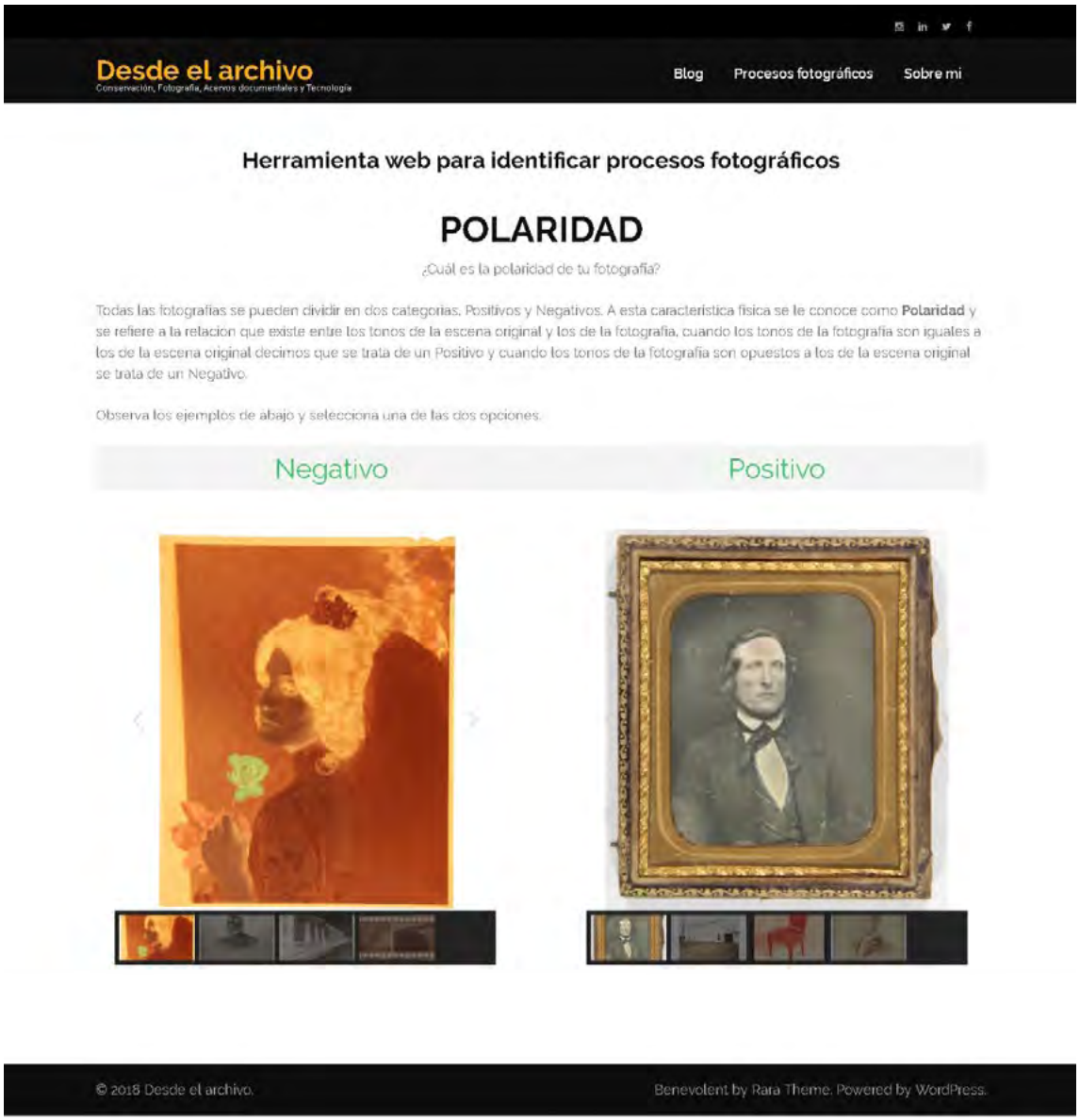


Figura 4. Interfaz de usuario de la herramienta web

Conclusiones

Gracias a las posibilidades comunicativas de la tecnología web, herramientas como esta pueden contribuir a diseminar conocimientos especializados que son poco accesibles, lo cual a su vez permite a un mayor número de personas comprender y valorar la materialidad de las fotografías analógicas resguardadas en los archivos históricos.

Dirección web. http://www.desdeelarchivo.com/procesos-fotograficos/

References

Barra, P., y Gutiérrez, I. (2000). *Normas Catalográficas del Sistema Nacional de Fototecas del INAH*, México: INAH/CONACULTA.

Boadas, J., Casellas, L., y Suqyet, M. (2001). *Manual para la Gestión de Fondos y Colecciones Fotográficas*. Girona: CCG ediciones.

IPI. Image Permanence Institute. (2017). *Graphic Atlas*. http://www.graphicsatlas.org (recuperado el 16 de noviembre de 2017).

Lavedrine, B. (2009). *Photographs of the Past: Process and Preservation*. Los Ángeles: Getty Conservation Institute.

Reilly, J. (2009). *Care and Identification of 19th-century Photographic Prints*. Rochester: Eastman Kodak Co.

SE. Secretaría de Economía. (2016). *Norma Mexicana NMX-R-069-SCFI-2016. Documentos fotográficos. Lineamientos para su Catalogación*. México: Secretaría de Economía.

Propuesta interdisciplinaria de un juego serio para la divulgación de conocimiento histórico. Caso de estudio: la divulgación del saber histórico sobre la vida conventual de los carmelitas descalzos del ex-Convento del Desierto de los Leones

Leticia Luna Tlatelpa
letyludigital@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Fabián Gutiérrez Gómez
fabian.gutierrez.gomez@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Edné Balmori
ednebalmori@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Feliciano García García
felicianogarcia.9@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Luis Rodriguez Morales
luis.rodriguezm12@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Resumen

El surgimiento de las narrativas hipermedia y transmedia así como de los productos culturales propios de la era digital como los *Juegos Serios*, expanden el espectro de los medios de comunicación tradicionalmente usados para divulgar la historia. Con estos nuevos medios es factible realizar divulgación bajo el modelo contextual, el cual, al contrario del modelo de déficit, considera las respuestas del público y la complejidad del fenómeno comunicativo de la divulgación (Leewenstein, 2003). En este sentido, se propone un *Juego Serio* sobre la vida conventual de los Carmelitas Descalzos que habitaron el Desierto de los Leones que estimule una experiencia memorable en jóvenes.

Divulgación de la historia

Se retoma el concepto de comunicación de la ciencia propuesto por Burns *et al.* (2003: 1991) para aplicarlo a la divulgación de la historia. Ellos plantean que la divulgación debe provocar alguna de las siguientes respuestas en la audiencia con relación a la ciencia: sensibilización, disfrute, interés, construción de opinión y/o análisis de contenido.

Se siguió la metodología de interpretación temática propuesta por Ham (2013) para responder a la pregunta de cómo comunicar el conocimiento histórico acerca de los carmelitas a un público joven, este último definido según Feixa (2003), a través de un videojuego y generar lazos de identidad fincados en valores culturales.

Esta metodología resalta la importancia de conocer la audiencia y sus intereses para provocar en ella reflexiones, cuestionamientos o la generación de nuevos significados, de modo que recuerde la experiencia divulgativa. Gándara añade, además, la importancia que tiene la narrativa para crear divulgación significativa (Gándara citado por Sánchez, 2016). Es de destacar también, que las emociones son relevantes en el divulgación, ya que pueden lograr que esta sea más recordada (Bonfil, 2003).

Si bien hay antecedentes del uso de videojuegos para tratar temas históricos, (Caldera Estudios, 2010), (Mulaka, 2018), (Nomdedeu, 2015), (Rodríguez et al, 2017), (Salinas et al, 2017), no se encontró uno cuyo diseño emplee la interpretación temática.

Los Carmelitas Descalzos

Esta orden fue una de las últimas que llegaron a la Nueva España. Aún cuando el cometido de estos frailes fue evangelizar a la población indígena, decidieron que su misión sería continuar su vida contemplativa que, según su cosmovisión, los acercaba a Dios (Ramírez, 2015). Formaron la Provincia de San Alberto con conventos en las principales ciudades de la Nueva España. Así, la Corona les permitió edificar el Desierto en el Monte de Santa Fe. Este yermo fue un espacio donde renunciaban a todo placer sensorial, como el disfrute de la comida, la prohibición de mirar a otra persona a los ojos, y el cumplimiento de la regla del silencio y del claustro para dedicarse a la contemplación (Báez, 1981).

Videojuegos

Se escogió el juego serio ya que "buscan cumplir un propósito más allá del propósito autocontenido de los juegos de entretenimiento" (Mitgush y Alvarado, 2012: 121, citado por González, 2017). Las emociones son observadas desde el enfoque de los videojuegos, como lo propone Lazzaro (2004). Finalmente, para el diseño del videojuego se siguieron los lineamientos propuestos por Shell (2005) y la metodología de interpretación temática. En esta, es primordial escoger un Tema (Ham, 2003) que guíe el contenido del objeto comunicativo.

Después de analizar los documentos históricos y desde la perspectiva de una audiencia jóven, se definió el tema como **los espacios donde se ejerce control sobre las personas anulan la individualidad y fomentan el desarraigo**. A partir de allí, se modeló la vida de los frailes en términos de tentaciones, recuerdos de la vida pasada, castigos, recompensas, el diablo, reglas, obediencia. El tema también determinó la estética gráfica, la animaciones y la música.

El videojuego se titula *Tentación en el Desierto* y es de tipo *Click and Point*. El jugador ayuda a un fraile Carmelita recién llegado al Yermo de Santa Fe a luchar contra tentaciones para que lo acepten como ermitaño mientras explora elementos de diferentes espacios del convento. Hay cinemáticas que muestran los recuerdos del personaje antes de convertirse en fraile; las tentaciones, castigos y retos del diablo son representados como minijuegos que el jugador debe resolver; la obediencia a las reglas es la energía del fraile; hay un diario escrito en primera persona el cual brinda información acerca del contexto histórico.

Conclusiones

Las pruebas con el prototipo analógico del juego mostraron que los jugadores problematizaron la vida carmelita y manifestaron emociones e interés. Se recomienda evaluar con la versión digital para determinar si la interpretación temática aplicada al diseño de un juego serio genera experiencia memorable al divulgar conocimiento histórico.

References

Báez M., E. (1981). *El Santo Desierto. Jardín de Contemplación de los Carmelitas Descalzos en la Nueva España*, México: Universidad Nacional Autónoma de México.

Bonfil, O., M. (2003). Una estrategia de guerrilla para la divulgación: Difusión cultural de la ciencia. *Congreso Latinoamericano Ciencia, comunicación y sociedad, Costa Rica.*

Burns, T.W., O'Connor, D. J., y Stockmayer, S. M. (2003). *Science communication: a contemporary definition. Public understanding of science*, 12(2), 183-202.

Caldera Estudios. (2010). Peluconas. [en linea] disponible en: http://caldera-estudio.com/proyectos/asi-se-veia-mexico-hace-250-anos/ [consultado 20 abril 2018].

Feixa, C. (2003). Del reloj de arena al reloj digital. Sobre las temporalidades juveniles. *Jóvenes, Revista de Estudios sobre la Juventud.* 7 (19), 6-27.

González, A. (2017). *Comunicación de la ciencia en videojuegos: evolución y juegos serios.* Tesis de Maestría. Universidad Nacional Autónoma de México.

Ham, S. H. (2013). *Interpretation: making a difference on purpose*. Fulcrum publishing

Lewenstein, B. (2003). Models of public communication of science and technology. *http://communityrisks. cornell. edu/Background-Materials/Lewenstein2003. pdf*.

Lazzaro, N. (2004). *Why we play games: Four keys to more emotion without story.*

Mulaka. (2018). Mulaka. [en linea] disponible en: https://www.lienzo.mx/mulaka/?lang=es [consultado 20 abril 2018].

Nomdedeu, L. (2015). *RAÍCES, un juego serio social para revalorizar las culturas originarias.* Tesina de Licenciatura. Universidad Nacional de la Plata.

Ramirez, J. (2015). *Los Carmelitas Descalzos en la Nueva España. Del activismo misional al apostolado urbano. 1585 - 1614.*, México, INAH.

Rodríguez, F. C., Palacios D., E., Marín G., P., Ortiz M., B. y Romero Q., G. (2017).Wirikuta. [en linea] disponible en: https://leiva2017.wordpress.com/proyectos/wirikuta/ [consultado 20 abril 2018].

Salinas, I., Hernández, E., Rodríguez, S. (2015). El desarrollo social a través de la valoración del sistema estético-comunicativo de los pueblos nativos de Baja California. [en linea] disponible en: http://www.redalyc.org/pdf/4981/498150319057.pdf [consultado 20 abril 2018].

Sánchez, M. (2016). *La Museología como herramienta de vinculación entre el profesor y el patrimonio. Propuesta de un curso de capacitación a profesores que imparten la asignatura estatal Patrimonio Cultural y Natural del Distrito Federal.* Tesis de Maestría. Escuela Nacional de Conservación, Restauración y Museografía.

Schell, J. (2015). *The art of game design: a book of lenses*. CRC Press.

Digital 3D modelling in the humanities

Sander Münster
sander.muenster@tu-dresden.de
TU Dresden, Germany

For more than 30 years, digital 3D modelling and in particular reconstruction methods have been widely used to support research and education in the digital humanities, especially but not exclusively on historical architecture. While technological backgrounds, project opportunities, and methodological considerations for the application of digital 3D modeling techniques are widely discussed in literature (e.g. Arnold and Geser, 2008, European Commission, 2011, Frischer, 2008, Bentkowska-Kafel et al., 2012, Bentkowska-Kafel, 2013, Kohle, 2013), my interest is to investigate digital 3D modeling in the humanities as a scholarly area and to derive implications for further organizational and methodical development. Against this background, my research investigates the following research questions:

-
- What marks a scholarly culture of 3D modelling in the humanities?
- What are technical and designal implications and workflows for model creation and presentation?
- How can digital 3D modelling techniques be learned and taught?

The research presented is part of an ongoing post doc thesis work dedicated to draw a "big picture" on digital 3D modelling techniques as research tools in the humanities. Against this background, my own and my department's activities include to investigate (1) a scholarly community, (2) usage practices occurring within single projects and to gain implications for further methodical development. We develop (4) technologies and workflows to enhance both, the creation of 3D models and user-centered interfaces and investigate how 3D models are (5) perceived and how 3D modeling techniques can be used in (6) education (Table 1). Research has been carried out since 2010 in 12 projects on local, national and EU level so far.

Area	Research Interest	Investigation
Scholarly community	Who are main authors? What are academic structures?	[A] Social network (c.f. Wellman, 1988) and bibliometric analysis (c.f. De Solla Price, 1963) of publications from major conferences in the field of digital cultural heritage 1990-2015 (n=3917)
	What are topics?	[B] Automated topic mining of 3917 articles, manual classification of 452 articles plus 26 project reports via qualitative content analysis (c.f. Mayring, 2000)
	Who funds projects?	[C] Qualitative content analysis of 518 project activities in the field of digital cultural heritage including
	What marks a disciplinary culture?	[D] Three stage investigation including a questionnaire-based survey during three workshops with 44 participants to gain a general overview; 15 guideline based interviews with researchers to investigate research culture in depth (Mieg and Näf, 2005, Gläser and Laudel, 2009); online survey with 988 participant to quantify findings
Usage practices	What are phenomena and strategies for cooperation?	[E] 4 case studies: Data collection via expert interviews (c.f. Gläser and Laudel, 2009) and observation (c.f. Lamnek, 2005). Data analysis via heuristic frameworks (c.f. Kubicek, 1977) and grounded theory (c.f. Bryant and Charmaz, 2010)
	How to support cooperation in 3D modelling projects?	[F] Employment and evaluation of SCRUM as agile project management approach (Schwaber, 2004) in 2 educational project seminars so far with 13 student teams
Methodological development	What are current challenges?	[G] Three group discussions (c.f. Lamnek, 2005) on workshops at national/ international conferences (~60 participants) to examine; online survey with 650 participants
	How to systematize?	[H] Classification scheme developed and applied for 8 projects yet
Technologies	How to create 3D models?	[I] Development of workflows and toolsets to automatically create 3D models from historic photos [removed for reviewing], and semi-automatic creation from GIS data [removed for reviewing]
	How to improve user interaction with 3D models?	[J] Development and testing of 4D geo browsers; browser-based augmented and virtual reality interfaces for mobile devices
Perception	What factors are influencing perception of 3D models?	[K] 2 expert workshops and literature survey yet to identify influencing factors
	How are virtually represented structures perceived?	[L] Two studies to investigate how virtually represented structures and proportions are perceived, involving 21 persons and using usability testing (c.f. Nielsen, 1993)
Education	How to use 3D modelling techniques in education?	[M] 3 student seminars to develop and test team project-based learning approaches via formative & summative assessment (c.f. Dumit, 2012)

Table 1 - Investigational parts

Some results at a glance

What are some results at a glance? Considering a scholarly community on digital 3D reconstruction and modeling, discourses on major conferences during the last 25 years were mainly led by institutions from European Mediterranean countries, covering primarily technological topics. Especially statues and buildings in Mediterranean countries dating from all periods Anno Domini deliver rich content for such reconstruction. Due to the high complexity and team-based workflows, aspects and usage practices for communication, cooperation, and quality management are of high relevance within 3D reconstruction projects. Especially if people with different disciplinary backgrounds are involved, visual media are intensively used to foster communication and quality negotiations, for example by comparing source images and renderings of the created virtual reconstruction. Furthermore, several projects successfully adopted highly standardized conventions from architectural drawings for interdisciplinary exchange. To support a methodological development I ran five workshops to identify prospects and demands for further development, involving around 60 researchers and an online survey to verify findings from these workshops. Costs and training were named most frequently as currently pressuring issues. With regards to technologies, a big hurdle to overcome in order to use augmented and virtual reality is the current need to download and install additional software. Since current browser generations allow the visualization of 3D content natively, our focus is on user-friendly interaction concepts to access both, visualizations and underlaying informations. Regarding the perception of virtual 3D models relatively little visual information is needed to allow observers to distinguish buildings from each other or to identify a single building and to gain information about its spatial relation and shape [removed for reviewing]. Moreover, we adopted and evaluated team project-based learning approaches to support student education in digital 3D reconstruction. As observed in two courses so far, a development of procedures and strategies for cooperation within student project teams for creating virtual representations evolves slowly, and mostly as reaction of upcoming problems and demands. Related competencies are based highly on implicit knowledge and experience. As consequence, a teaching of best practices prior to a project work is less effective than coaching during the project work.

Next steps

What are next steps? Since 3D models in the humanities are primarily accessed via visualizations, a toolset for assessing visualization and interactivity of 3D models and presentations is currently missing and will be in focus of a next research stage. Many of the already completed investigations are of qualitative nature or focus on particular aspects. Consequently, a further validation for adjacent aspects as well as a verification of findings are alltime tasks. To proceed, further investigations on the scholarly use of 3D models and historical photographs or the design of interfaces for virtual museums are under development as well as a survey to further investigate challenges and perspectives of 3D modeling. Since the research is intended to enhance the validation and dissemination of 3D modeling technologies in the humanities both, education and organizational development are key issues. Beside the further development and establishment of teaching concepts and university courses, especially strategies for self-driven and scalable learning as MOOCs or open educational resources seems promising. Finally, beneficial and methodologically grounded best practice examples, an institutionalization of chairs and institutes as well as an increased awareness seem to be crucial for a further organizational establishment.

References

ARNOLD, D. & GESER, G. 2008. *EPOCH Research Agenda – Final Report,* Brighton.

BENTKOWSKA-KAFEL, A. 2013. Mapping Digital Art History. Available: https://bentkowska.files.wordpress.com/2013/05/annabentkowska-kafel__gettyd-ah-lab_2013.pdf.

BENTKOWSKA-KAFEL, A., DENARD, H. & BAKER, D. 2012. *Paradata and Transparency in Virtual Heritage,* Burlington, Ashgate.

BRYANT, A. & CHARMAZ, K. 2010. *The SAGE Handbook of Grounded Theory,* Thousand Oaks, SAGE.

DE SOLLA PRICE, D. 1963. *Little Science - Big Science,* New York, Columbia Univ. Press.

DUMIT, N. Y. 2012. *Diagnostic/Formative/Summative Assessment,* n.n.

EUROPEAN COMMISSION 2011. *Survey and outcomes of cultural heritage research projects supported in the context of EU environmental research programmes. From 5th to 7th Framework Programme,* Brussels, European Commision.

FRISCHER, B. 2008. *Beyond illustration : 2D and 3D digital technologies as tools for discovery in archaeology,* Oxford, Tempus Reparatum.

GLÄSER, J. & LAUDEL, G. 2009. *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen,* Wiesbaden, VS Verlag für Sozialwissenschaften.

KOHLE, H. 2013. *Digitale Bildwissenschaft,* Glückstadt.

KUBICEK, H. 1977. Heuristische Bezugsrahmen und heuristisch angelegte Forschungsdesigns als Element einer Konstruktionsstrategie empirischer Forschung. *In:* KÖHLER, R. (ed.) *Empirische und handlungstheoretische Forschungskonzeptionen in der Betriebswirtschaftslehre.* Stuttgart.

LAMNEK, S. 2005. *Qualitative Sozialforschung. Lehrbuch,* Weinheim.

MAYRING, P. 2000. Qualitative Content Analysis. *Forum Qualitative Sozialforschung,* 1**,** Art. 20.
MIEG, H. A. & NÄF, M. 2005. *Experteninterviews,* Zürich.
NIELSEN, J. 1993. *Usability Engineering,* Salt Lake City, Academic Press.
SCHWABER, K. 2004. *Agile Project Management with Scrum,* Redmond.
WELLMAN, B. 1988. Structural Analysis. From Method and Metaphor to Theory and Substance. *In:* WELLMAN, B. & BERKOWITZ, S. D. (eds.) *Social Structures: A Network Approach.* Princeton: Princeton University Press.

Question, Create, Reflect: A Holistic and Critical Approach to Teaching Digital Humanities

Kristen Mapes
kmapes@msu.edu
Michigan State University, United States of America

Matthew Handelman
handelm@msu.edu
Michigan State University, United States of America

Teaching digital humanities at the undergraduate level is as much about issues of critical theory, inclusion, and diversity as it is about teaching digital tools and methods. Examining DH methods such as topic modeling introduces students to the concept of algorithmic bias, pointing to the algorithms that shape our daily lives. Working with DH tools such as Palladio enables students to confront and reveal the layers of representation (and inequality) that structure the virtual and physical spaces that we inhabit. And creating digital archives with platforms such as Omeka challenges students to question the purpose and limits of digital tools, offering opportunities to reflect on the ethics of (digital) representation. The dialectics of teaching new DH tools and questions of critique, the archive, and representation central to the humanities form the basis of the undergraduate Digital Humanities Minor at our institution, in which students take two sequential, required courses: "Introduction to Digital Humanities" and the "Seminar in the Digital Humanities". Our talk will explore how we weave together these courses to create a holistic and critical approach to the foundations of digital humanities at the undergraduatelevel.

In "Introduction to Digital Humanities," students examine a range of DH methods and activities and create a final project of their own choosing. We explore DH approaches to humanities questions by evaluating digital projects that engage with the Harlem Renaissance and its context. By centering students' exposure to DH on one broad but unifying topic, we can avoid the trap of the carousel of tools into which an Intro DH class could fall. The Harlem Renaissance centers the course because it touches on cultural areas of critical interest spanning disciplines – art, music, literature, economic history, social history, political history, and urban planning – and has several DH projects either directly on the Harlem Renaissance or on related topics. By rooting the course in a historical cultural period, students are introduced to structural trends and issues that reverberate today.

In analyzing digital projects as a class, we critique the data behind the project, its presentation - in terms of style, effectiveness, and accessibility - and the structures in which it was made. We discuss what role grant funding plays in promoting certain types of projects, how crowdsourcing relates to labor ethics and the digital, who the project's users may be, and what its long term preservation prospects are. We then apply this critical framework to projects ranging from a digital edition (such as Claude McKay's Early Poetry) to large scale image analysis (such as On Broadway) to linked data and network analysis (such as Linked Jazz). We also talk to project leaders (from Virtual Harlem, Umbra Search, and the Mapping the Second Ku Klux Klan projects) to get a behind-the-scenes perspective on project management, origins, and goals. The bulk of the second half of the semester is spent on student projects. Students choose any topic they like and develop a critical research question. It is then is up to each student to choose a DH method and to find, gather, and clean their data. Class time is built in for one-on-one assistance from the professor and the embedded librarian to guide the students through the frustration and joy of the iterative DH project. By the end of the semester, the same digital project evaluation framework is used to analyze the students' projects.

The second semester in this year-long sequence, "Seminar in Digital Humanities," deepens students' skills with DH tools and methods, applies these skills in a semester-long DH project, and combines students' DH knowledge with the reflective practices of critical theory. As both "Text, Technology, and the Body" (spring 2016) and "Digital Humanities and Critical Theory" (spring 2017), students participate in a collaborative DH project, in which they design and build an online collection using archival materials from our institution's Special Collections as well as analyze and reflect on their digital work and the content of our archive. Whether it is digitizing criminology broadsheets from 17th Century Europe or early-twentieth century comics, this course frames DH as a continuation of - instead of a break with - critical debates over media, technology, and culture - from classics such as Walter Benjamin to current critical voices in DH such as Alan Liu and Laura Klein. The goal of these projects is not only to enable students to conceptualize and execute a student-led DH project, but also to develop their ability to read and critique digital tools and recognize their affordances, limitations, and political implications.

Exploring and employing a variety of digital techniques, "Seminar in the Digital Humanities" adapts and expands on the "read, play, build" approach to teaching DH proposed by Joanna Swafford at DH 2016. The semester is divided into seven units, the first two of which position DH within contemporary (and critical) debates in the humanities and introduce students to the historical and disciplinary context pertaining to our subject matter. For each unit, students read theoretical texts and articles that contextualize the tool under consideration as part of a larger historical-critical discourse within media studies, critical theory, and the history of DH. These readings provide the background in which students then learn how to implement these tools and explor examples aided by guest DH specialists from around our institution. The final phase of each unit provides a collaborative space for class members to create a working plan to apply this technique to our project - in order, for example, to clean our metadata, digitize our selected archival materials, and set up the Omeka site. Finally, students execute this plan as their individual project and compose a reflective essay that positions their work in the critical debates and comments on the technological, epistemological, and ethical choices that went into their digital work. These individual projects and critical reflections provide a self-reflective context for our digital collection, while allowing the students to cultivate their identities as critical thinkers and digital humanists.

Taken together, these two undergraduate courses expose students to a range of digital tools and methods for humanistic inquiry, providing them with experience overseeing their own DH project from conception to completion as well as participating in a semester-long team project. In different ways, the courses introduce students to critical frameworks for asking humanities questions of the digital and for using the digital to ask humanities questions. Teaching DH and critical thought as two sides of the same coin, this DH sequence provides students with tools to not only understand, but also intervene in a world increasingly mediated by digital processes.

"Smog poem". Example of data dramatization

Piotr Marecki
piotr.marecki@ha.art.pl
Jagiellonian University, Poland

Leszek Onak
leszek.onak@gmail.com
Jagiellonian University, Poland

The proposed poster is a visual presentation of the literary experiment "Smog Poem" (2018) by a Polish poet Leszek Onak developed in the UBU lab at the Jagiellonian University run by dr Piotr Marecki. Drawing on the terminology of expressive processing developed by Noah Wardrip-Fruin, and platform studies by Nick Montfort and Ian Bogost the authors of the poster present the process of the creation of the work.

According to World Health Organization ambient particle pollution kills about 6.5 million people annually affecting all regions of the world. In Poland, it amounts to nearly 50 thousand deaths each year. Krakow, the former capitol of Poland, is ranked third among the European cities with the highest levels of particulate matter (PM 10).

"Smog Poem" is a text and graphics generator that uses the data on the environmental pollution to change the tissue of the text, its graphic elements, and other components depending on the pollution's intensity. The algorhythm has a form of an internet browser plugin; after its installation, the users browsing through the internet will experience the air pollution in front of their own eyes through the glitches appearing on the websites they use, the replacement of the photos and text modification. Some articles will be replaced by a separate generated text based on the syntactic mechanisms and by using the rules of the "Game of Life" by John Conway.

The piece consists of two main engines. One mechanism is pulling data on the actual air pollution with Particulate Matter (PM 10 and PM 2.5), Nitrogen Dioxine NO2, Sulfur Dioxide SO2 and Carbon Monoxide CO. Each of those indicators is responsible for a different element distorting the content. The second mechanism is responsible for the upload of data from the websites and its modification depending on the particle pollution. If the air quality does not exceed the norms, the content of websites remains unchanged.

The algorhythm is representative of the growing trend of digital art based on resources and presenting them in a way to influence the user's consciousness. It refers to the concept of 'data dramatization' by Liam Young, who once said: 'Data Dramatization, as opposed to data visualization presents a data set with not only legibility or clarity but in such a way as to provoke an empathetic or emotive response in its audience.'

„Smog poem" is one of the of the digital works developed in the UBU lab at the Jagiellonian University. The lab primarily produces digital works that can function in a few fields of the demoscene, electronic literature, video games and media art. The research conducted in the lab focuses on, among other things, local phenomena in the digital media field, e.g. strategies for cloning platforms in Central and Eastern Europe, as well as the digital genres and their specific features in Central and Eastern Europe. The artists, programmers and scholars affiliated with the lab develop new genres and communication practices (technical reports, open notebook science) to describe the creative process in its widest definition in the era of digital textuality. The project has been made possible through the support of the Polish Ministry of Science and Higher Education "National Programme for the Development of Humanities".

ANJA, ¿dónde están los encabalgamientos?

Clara Martinez-Canton
cimartinez@flog.uned.es
LINHD, UNED, Spain

Pablo Ruiz-Fabo
pablo.ruiz@linhd.uned.es
LINHD, UNED, Spains

Elena González-Blanco
egonzalezblanco@flog.uned.es
LINHD, UNED, Spain

Introducción

Encabalgamiento es el desajuste entre la pausa métrica y la sintáctica (Domínguez Caparrós, 2000: 103) que ocurre cuando una unidad de sentido se rompe entre dos versos. Este fenómeno, desde siempre utilizado con distintos fines expresivos (énfasis, ambigüedad, etc.) es difícil de delimitar formalmente.

El estudio más sistemático realizado para su caracterización en español sigue siendo el realizado en su tesis por Quilis (1964). El estudioso experimentó con lecturas de prosa, buscando demostrar qué unidades sintácticas no permiten pausa de sentido en su interior. Basándose en los resultados definió una serie de categorías gramaticales y sintácticas cuya separación en versos distintos produce encabalgamiento. La tipología allí establecida se considera ya clásica. El estudio de Quilis proporciona una definición formal y empírica del fenómeno. Con base en sus reglas se ha creado una herramienta capaz de detectar el encabalgamiento y sus tipos.

Este póster presenta la interfaz ANJA para el análisis automático del encabalgamiento desde una sencilla aplicación web: http://prf1.org/anja/index/, desarrollada dentro del proyecto ERC POSTDATA GA- 679528[1].

Estado del arte

La naturaleza formal del análisis métrico lo hace un campo propicio para su tratamiento computacional (Birnbaum and Thorsen, 2015; Delente and Renault, 2015). El procesamiento del lenguaje natural (PLN) ofrece muchas posibilidades para la métrica, pues las reglas de definición lingüística permiten llevar a cabo análisis y extracción automática de grandes cantidades de información de corpus textuales.

Para la automatización del análisis métrico en español destacamos los estudios de escansión silábica y acentual de Navarro-Colorado (2017), Agirrezabal (2017) y Gervás (2000). También los trabajos de generación automática de poesía con patrones métricos (Gervás, 2000b) y (Gervás, 2015).

En el campo de las interfaces cabe distinguir entre aquellas que exploran datos de textos ya analizados, recogidos en una base de datos, y aquellas que permiten la entrada y análisis de cualquier poema. Del primer tipo destacamos For Better For Verse[2] (Tucker, 2011) y Database of Czech Metre[3] (Plecháč and Kolár, 2015). Entre las que permiten introducir textos destacamos, en español, la ligada a la herramienta de Navarro-Colorado[4], que analiza versos endecasílabos. Otros sitios con interfaz de entrada para análisis métrico son Separarensílabas[5] o Lexiquetos[6]. En otras lenguas destacamos Metricalizer[7] (Bobenhausen and Hammerich, 2015) para alemán, Aoidos[8] (Mittmann, 2016) para portugués y español, y RhymeDesign[9] (McCurdy et al., 2015) especializado en rima en inglés.

Una interfaz para el análisis del encabalgamiento representa, sin embargo, una novedad en el campo.

Herramienta y resultados

El programa de detección del encabalgamiento en español, basado en PLN, se desarrolló en 2016-2017 y fue evaluado sobre dos corpus de test de distintos periodos (Ruiz et al., 2017). ANJA proporciona una interfaz web simple para este programa. El sistema consta de tres componentes: módulo de preprocesado para uniformar el formato de los poemas, pipeline de PLN (basada en IXA Pipes (Agerri et al., 2014) para POS-tagging, constituyentes y dependencias sintácticas) y módulo de detección de encabalgamiento (basado en reglas y diccionarios) y ampliamente documentado en el sitio web[10]. Se ha utilizado esta herramienta para etiquetar un corpus de más de 4000 sonetos alojado y documentado en https://github.com/postdataproject/disco.

El código de la herramienta de detección de encabalgamientos está disponible en https://bitbucket.org/pruizf/anja_public/.

Interfaz gráfica de usuario

ANJA es una interfaz pública y gratuita, alojada en: http://prf1.org/anja/index/. Permite cargar los poemas que el

1 Este trabajo se enmarca dentro del proyecto de investigación Starting Grant Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528), financiado por el European Research Council (ERC) bajo el programa: European Union´s Horizon 2020 research and innovation programme, dirigido como Investigador Principal por la profesora Elena González-Blanco, LINHD UNED (http://postdata.linhd.es/).

2 http://prosody.lib.virginia.edu/
3 http://versologie.cz/v2/web_content/
4 http://adso.gplsi.es/index.php/es/demostracion/
5 http://www.separarensilabas.com/index.php
6 http://lexiquetos.org/silio/
7 https://metricalizer.de/en/metrikanalyse/poem
8 http://aoidos.ufsc.br/
9 http://www.sci.utah.edu/~nmccurdy/rhymeDesign/
10 https://sites.google.com/site/spanishenjambment/

usuario decida y analizarlos en el momento. También ofrece la carga de archivos ZIP que contengan archivos en texto plano.

La interfaz de usuario está construida con el framework Django (Python), con las plantillas de Bootstrap 3. Las vistas de Django se llaman con AJAX para poblar los elementos de la UI. Para el análisis de PLN, Django accede a servicios web Java (IXA Pipes) implementados en nuestro servidor.

ANJA presenta dos ventanas de navegación (Fig. 1), la principal, para introducir poemas, a la derecha y, a la izquierda, una mínima guía de uso que explica su funcionamiento y enlaza a la web del proyecto.

Captura de ANJA

Los resultados se ofrecen dos formatos: *Standoff* (tipo de encabalgamiento y línea), e *Inline* (etiquetado gramatical y tipo de encabalgamiento por línea, ver Fig. 2 para *Inline*). Las anotaciones PLN en que se basa en sistema se ofrecen en las pestañas *PosTags* (etiquetas gramaticales) y *FullNLP* (pipeline completa).

El enlace **legend**[11] da acceso a la leyenda que explica los tipos de encabalgamiento, las etiquetas gramaticales y otras convenciones de representación:

Standoff | Inline | PosTags | FullNLP [legend]

#	Text	Position	Enjambment Type
1	{Ya A} {se Q} {doran V} {las D} {hojas N} {, O} {ya A} {las D} {tardes N}	B	ex_subj_verb
2	{Son V} {lo D} {mejor G} {del P} {día N}	I	ex_subj_verb
3	{Bajo P} {su D} {luz N} {madura G} {de P} {sosiego N} {. O}	O	
4	{Ya A} {comienza V} {el D} {curso N} {. O}	O	
5	{Aprender V} {, O} {enseñar V} {, O} {lecciones N} {, O} {aulas N} {. O}	O	
6	{Siempre A} {en P} {octubre O} {comenzaba V} {el D} {año N} {. O}	O	
7	{¡ O} {Y C} {cuántas Q} {veces N} {esa D} {luz N} {de P} {otoño N}	O	

Anotaciones de encabalgamiento en formato *Inline*

La existencia de una aplicación web simple para la utilización esta herramienta la hace accesible para una gama mucho más amplia de usuarios.

References

Agerri, R., Bermudez, J. and Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. *Proceedings of LREC 2014, the 9th International Language Resources and Evaluation Conference*, vol. 2014. Reykjavik,Iceland, pp. 3823–3828

11 https://sites.google.com/site/spanishenjambment/legend

http://www.lrec-conf.org/proceedings/lrec2014/pdf/775_Paper.pdf (accessed 20 April 2017).
Agirrezabal, M. (2017). Automatic Scansion of Poetry San Sebastián/Donosti: Universidad del País Vasco.
Birnbaum, D. J. and Thorsen, E. (2015). Markup and meter: Using XML tools to teach a computer to think about versification. *Balisage: The Markup Conference* http://www.balisage.net/Proceedings/vol15/print/Birnbaum01/BalisageVol15-Birnbaum01.html (accessed 22 April 2017).
Bobenhausen, K. and Hammerich, B. (2015). Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer2. *Langages*(3): 67–88.
Delente, É. and Renault, R. (2015). Outils et métrique: un tour d'horizon. *Langages*(3): 5–22.
Domínguez Caparrós, J. (2000). *Métrica Española*. Madrid: Síntesis.
Gervás, P. (2000a). A Logic Programming Application for the Analysis of Spanish Verse. *Computational Logic—CL 2000*. Berlin: Springer Berlin Heidelberg, pp. 1330–44.
Gervás, P. (2000b). Wasp: Evaluation of different strategies for the automatic generation of spanish verse. *Proceedings of the AISB-00 Symposium on Creative & Cultural Aspects of AI*. pp. 93–100 https://www.researchgate.net/profile/Pablo_Gervas/publication/228609235_Wasp_Evaluation_of_different_strategies_for_the_automatic_generation_of_spanish_verse/links/00b4952aada6407047000000.pdf (accessed 22 April 2017).
Gervás, P. (2015). Tightening the Constraints on Form and Content for an Existing Computer Poet. *AISB Convention 2015* http://eprints.sim.ucm.es/37000/ (accessed 22 April 2017).
McCurdy, N., Srikumar, V. and Meyer, M. (2015). Rhymedesign: A tool for analyzing sonic devices in poetry. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. pp. 12–22.
Mittmann, A. (2016). Escansão automática de versos em português. https://repositorio.ufsc.br/handle/123456789/175819.
Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities* doi:10.1093/llc/fqx009. https://academic.oup.com/dsh/article-abstract/doi/10.1093/llc/fqx009/3064339/A-metrical-scansion-system-for-fixed-metre-Spanish (accessed 19 April 2017).
Plecháč, P. and Kolár, R. (2015). The Corpus of Czech Verse. *Studia Metrica et Poetica*, 2(1): 107–118.
Quilis, A. (1964). *Estructura Del Encabalgamiento En La Métrica Española*. Consejo Superior de Investigaciones Científicas, patronato' Menéndez y Pelayo,' Instituto' Miguel de Cervantes,'.
Ruiz Fabo, P., Bermúdez Sabel, H., Martínez Cantón, C. I., González-Blanco, E. and Navarro-Colorado, B. (2018). The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings. Humanidades Digitales 2018 (DH 2018). Ciudad de México, México.
Ruiz Fabo, P., Bermúdez-Sabel, H., Martínez Cantón, C. I. and Calvo Tello, J. (2017). *Diachronic Spanish Sonnet Corpus (DISCO). Madrid: UNED.* Madrid https://doi.org/10.5281/zenodo.1012567.
Ruiz, P., Martínez Cantón, C., Poibeau, T. and González-Blanco, E. (2017). Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, pp. 27–32.
Tucker, H. F. (2011). Poetic data and the news from poems: A for better for verse memoir. *Victorian Poetry*, 49(2): 267–281.

Combining String Matching and Cost Minimization Algorithms for Automatically Geocoding Tabular Itineraries

Rui Santos
ruipds@gmail.com
IST and INESC-ID,
University of Lisbon, Portugal

Bruno Emanuel Martins
bruno.g.martins@ist.utl.pt
IST and INESC-ID,
University of Lisbon, Portugal

Patricia Murrieta-Flores
p.murrietaflores@chester.ac.uk
Digital Humanities Research Center,
University of Chester, United Kingdom

Historical itineraries, often accessible as tables or as sequential lists of names for the places visited in the context of a particular journey, are abundant resources and also important objects of study for Humanities scholars, providing 'snapshots' of particular socio-cultural events, insights into the development of human mobility, and invaluable information related to the establishment of road networks. Well-known examples include the 3rd century *Itinerarium Antonini Augusti* or the *Itinerarium Burdigalense,* written between the 8th and 10th centuries, among others. Many historical manuscripts and/or transcriptions containing information on itineraries, dating from the medieval period to the 20th century, are nowadays available in digital formats, through initiatives such as Europeana or the Internet Archive, or in the context of Digital Humanities projects like Pelagios.

Few historical tabular itineraries are nonetheless directly associated with map-based representations and, in many cases, there is little information on the actual routes

taken in between locales. As such, there are many interesting questions related to early traveling routes, in need of further study. We believe that the analysis of historical itineraries (e.g., for consistency checking, or enabling new inquires/inferences about the routes) can be facilitated through the analytical tools of Geographical Information Systems (GIS) and/or through map-based representations. The research reported in this poster concerns with automatically geocoding historical itineraries, leveraging innovative methods that explore the idea that travelers tend to choose the most efficient routes (e.g., itineraries will likely minimize the distance between locations).

In brief, we propose an automated method for geocoding tabular itineraries based on a sequence of four stages, combining string similarity search and well-known optimization procedures (Santos et al., 2017b). On the first stage, we use string similarity to look for candidate disambiguations in a large-coverage gazetteer. State-of-the-art string matching methods (Santos et al., 2017a, 2018), leveraging supervised learning, can then optionally be used to further filter/restrict the set of disambiguation candidates. A least-cost path between pairs of candidates, visited in sequence over the itinerary, is afterwards estimated on the third stage. We tested geodesic paths over the Earth's surface, or least-cost path calculations (Douglas, 1994) leveraging terrain slope and land-coverage for estimating movement costs. Finally, Step 4 leverages the distance associated to each of the paths between candidate pairs, computed in Stage 3, to find an overall best path for the entire itinerary, also disambiguating each of the toponyms to the most likely candidate. A dynamic programming algorithm similar to Viterbi decoding (Forney, 1973) is used at this stage to efficiently compute the global path that minimizes the traveled distance.

The proposed method was tested with manually geocoded itineraries (e.g., measuring the distance between the estimated disambiguation and ground-truth geo-spatial coordinates for the places in each itinerary). We relied on a dataset of well-known European historical itineraries (see http://www.peterrobins.co.uk/itineraries/list.html), containing 24 instances corresponding to sequences of varied lengths. We also used the GeoNames gazetteer for supporting the disambiguation of toponyms into geo-spatial coordinates, i.e. a resource which focuses on the modern administrative geography that nonetheless lists many historical variants as alternative place names. Our experiments showed that while approximate string matching can already achieve very low median errors (e.g., many of the itinerary toponyms match exactly with entries in GeoNames, and thus the median distance towards the correct disambiguations is quite low), the combination with cost optimization can significantly improve results in terms of the average distance. Moreover, using Least-Cost Paths (LCPs) for reconstructing the most likely routes can enable new inquiries and inferences. Although LCP analysis is commonly used within computational archaeology (Murrieta-Flores, 2012), the application that is reported through this poster is particularly innovative.

Our work shows that methods leveraging the intuition that travelers tend to choose the least-costly routes, in combination with approximate string matching for finding gazetteer entries that corresponding to historical toponyms, are indeed effective for automatic geocoding. We focused on the validation of the automated method but we believe that, if implemented within plugins for popular GIS environments, the proposed ideas can effectively help Humanities scholars in the analysis of data pertaining to historical itineraries.

Figure 1 - Ground-truth trajectory for the pilgrimage of Jehan de Tournay from Valenciennes to Venice (left), compared to the estimated trajectory for the same itinerary (right).

Acknowledgements

This research was supported by the Trans-Atlantic Platform for the Social Sciences and Humanities, through the Digging into Data project with reference HJ-253525. The researchers from INESC-ID also had financial support from Fundação para a Ciência e Tecnologia (FCT), through the INESC-ID multi-annual funding from the PIDDAC program, (UID/CEC/50021/2013)

References

Douglas, D. H. (1994). Least-cost Path in GIS Using an Accumulated Cost Surface and Slopelines. Cartographica: *The International Journal for Geographic Information and Geovisualization*, 31(3).

Forney, G. D. (1973). *The Viterbi Algorithm. Proceedings of the IEEE*, 61(3).

Murrieta-Flores, P. (2012). *Traveling through past landscapes - Analyzing the dynamics of movement during Late Prehistory in Southern Iberia with spatial technologies*. Ph.D. Dissertation, University of Southampton.

Santos, R., Murrieta-Flores, P. and Martins, B. (2017a). Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth*.

Santos, R., Murrieta-Flores, P. and Martins, B. (2017b). *An Automated Approach for Geocoding Tabular Itineraries. Proceedings of the ACM Workshop on Geographic Information Retrieval*. New York: ACM Press.

Santos, R., Murrieta-Flores, P., Calado, P. and Martins, M. (2018). Toponym Matching Through Deep Neural Networks. *International Journal of Geographical Information Science*, 32(2)

How We Became Digital? Recent History of Digital Humanities in Poland

Maciej Maryl
maciej.maryl@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences, Poland

Digital humanities suddenly erupted in Poland in the second decade of the 21st Century: first digital humanities centres were established (2013-2015); Poland joined important European networks and consortia like CLARIN (2013), NeDiMAH (2014), DARIAH (2015), or OPERAS (2017) while establishing national consortia CLARIN-PL (2013), and DARIAH-PL (2015); finally, it hosted important international conferences: CLARIN 2015 in Wrocław and ADHO's Digital Humanities 2016 in Kraków. Yet, this sudden eruption by no means marks the beginning of DH in Poland. The first digital projects in the humanities could be traced backed to early 2000s as the data collected in the survey by Werla & Maryl (2014) suggest. Those events should then be understood as landmarks in the process of the institutionalization of digital humanities in Polish scholarship.

This paper explores the specificity of digital humanities in Poland through the analysis of the events and projects which lead to this institutionalization. As O'Sullivan et al. 2015 point out "Tracing the emergence of academic disciplines in a national context is a useful undertaking, as it goes beyond the definition of a field to an assessment of its evolution within a more specific cultural context." They also claim that the emergence of the field is closely connected to the social as well as economic trends. It is true for Poland, where humanities computing evolved slowly due to technological deficiencies and budgeting problems. Moreover, Polish humanities in the 1990s (especially in the field of literature, culture and history) were also preoccupied with removing the "white spots", i.e. conducting research on topics that could not have been accounted for before 1989 for political reasons. On the other hand, when discussing the development of DH in a country which was hardly a forerunner of digital methods, but rather its late adopter, heavily influenced by the experiences of foreign institutions, it is extremely difficult to pinpoint the regional specificity of digital research practices (cf. Schreibman 2012). Is there any local flavour of the practices, materials, or tools selected? Does it go beyond mere linguistic differences? Are region-specific research questions being asked?

The discussion will be based on selected projects (Werla & Maryl 2014), conferences, as well as on the observations of the forming phase of DARIAH-PL consortium (2013-2015), which would serve as a case-study. The issue of national specificity of DH in Poland in comparison to other European countries will be addressed in the light of the results of DARIAH VCC2 survey on digital methods (Dallas et al. 2017), conducted in 2014-15 by the Digital Methods and Practices Observatory (DiMPO) Working Group of DARIAH-EU. The discussion will be informed by Roopika Risam's concept of "DH accent" which allows to account for "both local specificity and global coherence in DH" (2017:378).

Although the authors of *Digital_Humanities* claim that "The mere use of digital tools for the purpose of humanistic research and communication does not qualify as Digital Humanities" (ibid.) The results of DARIAH VCC2 survey on digital methods and tools in the humanities show that the application of digital methods in the humanities is gradual. The tools like word processors, web search engines and various online resources (digital libraries, archives, journals) are widely adopted. Yet, a bit more advanced tools (e.g. bibliography managers or specialized note-taking applications) are relatively less popular. And there are still some types of applications (e.g. databases,

Content-Management-Systems, or use of social media in scholarly practices) which are used only by a small group of scholars. Therefore being a digital humanist means placing oneself on the scale ranging from the basic tools nearly all of us use to the most advanced stage on which new methods and software capacities enable us to pose completely new research questions (or to answer the old ones in a fundamentally different manner).

This process of *becoming* digital, i.e. adopting digital methods and practices by scholars in the humanities, will be analysed through the conceptual framework of "three waves" of digital humanities: (1) early remediation of traditional methods of scholarly inquiry (cf. Svensson 2009); (2) taking the advantage of the new medium in creating new methods and genres (Pressner 2011; Davidson 2008; Svensson 2010) (3) critical scrutiny of the epistemic constraints of the medium (Berry 2011, Rogers 2015). Those waves, although sometimes understood chronologically, are here considered as co-occuring in a DH community.

Polish sample of the DARIAH survey does not differ greatly in terms of the digital tools applied by scholars in comparison to the European sample. They use less often bibliography managers or personal databases, but Polish results seem to be rather consistent with European sample, what – in turn – shows that Polish DH, although developed beyond the existing networks, show similar patterns of growth. There are however important differences in terms of disciplinary background, career status and perceived needs of the Polish scholars, who were more interested in enhancing their existing research practices (improved access to the sources or software, networking), and are less open to new methods and tools (advice, courses, support options).

By means of such comparative perspective this paper engages with the conference topic, discussing how digital approaches may be instrumental in building 'bridges' between various research communities, which in turn may contribute to levelling the differences with regards to centres and peripheries of contemporary DH. Understanding the tension between local and transnational initiatives is important to capture the specificity of Polish DH, which could be viewed also as a heavily institution-related. Poland participates in CLARIN and DARIAH, yet Polish scholars are not that active in ADHO (there is no Polish Association of DH). Given the emerging national and international DH initiatives in Eastern Europe, as well as the plans to establish DARIAH Hub for the region, it may be a good moment to reflect on the interplay of regional and external factors of this process. A better understanding of how we have become digital humanists, offered here on the example of Poland, may inform those initiatives.

References

Berry, D.M. (2011). The Computational Turn: Thinking About the Digital Humanities. *Culture Machine,* vol. 12 , https://www.culturemachine.net/index.php/cm/article/view/440/470.

Dallas, C., Chatzidiakou, N., Benardou, A., Bender, M., Berra, A., Clivaz, C., Cunningham, J., et al. (2017). *European Survey on Scholarly Practices and Digital Needs in the Arts and Humanities - Highlights Report*. Zenodo. doi:10.5281/zenodo.260101.

Davidson, C. N. (2008) "Humanities 2.0: Promise, Perils, Predictions". *Publications of the Modern Language Association of America (PMLA)* 123(3), 707-717.

O'Sullivan, J., Murphy, O. and Day, S. (2015). The Emergence of the Digital Humanities in Ireland. *Breac: A Digital Journal of Irish Studies*, https://breac.nd.edu/articles/the-emergence-of-the-digital-humanities-in-ireland/

Presner, T. (2010). Digital Humanities 2.0: A Report on Knowledge. *OpenStax CNX. 8* http://cnx.org/contents/2742bb37-7c47-4bee-bb34-0f35bda760f3@6

Risam, R. (2017). Other worlds, other DHs: Notes towards a DH accent. *Digital Scholarship in the Humanities*, 32(2), 377-384.

Rogers, R. (2015). *Digital methods*. Cambridge: MIT press.

Schreibman, S. (2012). Controversies around the Digital Humanities. *Historical Social Research / Historische Sozialforschung,* 37(3):141, 46-58.

Svensson, P. (2009). Humanities Computing as Digital Humanities. *Digital Humanities Quarterly* 3: 3.

Svensson, P. (2010). "The Landscape of Digital Humanities" *Digital Humanities Quarterly* 4:1.

Werla, M., and Maryl, M. (2014). *Humanistyczne projekty cyfrowe w Polsce*. Poznań-Warszawa, http://lib.psnc.pl/publication/831.

Hacia la traducción automática de las lenguas indígenas de México

Jesús Manuel Mager Hois
mmager@turing.iimas.unam.mx
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Mexico

Ivan Vladimir Meza Ruiz
ivanvladimir@turing.iimas.unam.mx
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Mexico

Introducción

En México existen 68 lenguas indígenas oficialmente reconocidas (Diario oficial, 2013). Esta riqueza lingüística forma parte del mosaico multicultural que define la identidad de nuestro país. Sin embargo, la predominancia cultural del español y el rezago generalizado del acceso a las tecnologías de información (Sandoval-Forero, 2013) por parte de los hablantes de estas lenguas crea barreras

culturales que dificultan la transferencia del conocimiento entre los pueblos indígenas.

En los últimos años se ha consolidado el campo de traducción automática. Parte de la consolidación de la traducción automática se debe a la traducción estadística (SMT) (Koehn, 2009; Lopez, 2008). Ésta metodología usa ejemplos de oraciones en ambas lenguas (corpus paralelos) para determinar los parámetros de un modelo estadístico que permite tal traducción. Adicionalmente, en los últimos años se han abierto paso a los modelos de traducción automática basados en redes neuronales (NMT) (LeCun *et al.*, 2015), los cuales permiten traducción multilingüe, en donde se crea un modelo de traducción común entre múltiple lenguas, el cual se utiliza posteriormente para mejorar la traducción entre pares de lenguas (Cho *et al., 2014*).

Metodología y resultados

En este proyecto presentamos nuestros avances en la creación de traductores automáticos para cinco lenguas indígenas al español: wixarika, náhuatl, yorem nokki, purépecha y mexicanero. Para obtener una visión completa sobre el campo decidimos hacer una comparación entre SMT y NMT. En ambos casos entrenamos los modelos usando segmentación morfológica que ha mostrado mejores resultados para lenguas polisintéticas (Mager, et. Al, 2016).

Para SMT fue utilizado el traductor por frases MOSES (Kohen, *et. al*., 2007) junto con el alineador GIZA++ (Och y Ney, 2003). Para los experimentos de NMT fue utilizado el modelo neuronal Codificador-Decodificador (Seq2Seq) con Redes Neuronales Recurrentes Bidireccionales (BiRNN) y con celdas de Unidades Recurrentes con Compuertas (GRU) (Cho., *et. al.*, 2014). Las pruebas fueron llevadas a cabo con OpenNMT (Klein, *et. al.*, 2017) con un corpus que consta de 985 frases traducidas a los 5 idiomas y que incluyen notación morfológica (Gómez y López, 1999; Chamoreau, 2003; Freeze, 1989; Lastra, 1980). Cada modelo ha sido evaluado de manera automática usando Bilingual Evaluation Understudy (BLEU) (Papineni, *et. al.*, 2002), y su salida fue valorada de manera manual, de tal manera que ha sido posible identificar los retos y limitaciones de las propuestas.

	NMT	SMT
Mexicanero-Español	2.95	23.47
Náhuatl-Español	3.04	10.14
Purépecha-Español	0	5.38
Wixarika-Español	0	0
Yorem Nokki-Español	0	2.44

Tabla 1: BLEU de los resultados experimentales de traducción de los cinco pares de idiomas con NMT y SMT

Como podemos ver en la tabla 1, los resultados de SMT superan los de NMT debido al corpus tan reducido con que se entrenaron. Mexicanero y náhuatl tuvieron un mejor desempeño que el wixarika, dado que el wixarika es una lengua con morfología con mayor cantidad de morfemas por palabra que el náhuatl (Kann, et. al. 2018).

Discusión

Si bien, se lograron mejorar las traducciones de manera importante, estos no son suficientes para ser usadas en la práctica cotidiana de manera autónoma o para asistencia humana. A través del desarrollo de estos traductores que hemos identificado los siguientes retos:

- **Escasez de los recursos.** Para poder generar un traductor automático es necesario contar con cientos de miles de pares de oraciones entre las dos lenguas; sin embargo, el poco uso de tecnologías de las comunidades nativo hablantes hace difícil la construcción de este corpus.
- **Complejidad morfológica**. Dada la naturaleza polisintética de estas lenguas, se necesita mejorar la segmentación morfológica automática para evitar la dispersión de datos (Kann, et. al. 2018).
- **El español es una lengua distante a los idiomas indígenas** que, en su gran mayoría tienen una topología morfológica polisintética, a diferencia del español que es fusionante y con orden Sujeto-Verbo-Objeto.
- La **falta de estandarización de la ortografía de las lenguas y el amplio espectro dialectal interno en las lenguas**.

Conclusiones

El presente trabajo expone primeros avances en traducción automática de cinco lenguas indígenas al español con SMT y NMT, identificando retos y limitaciones. Para trabajos futuros planteamos; mejorar el análisis y la segmentación morfológica de las lenguas indígenas, dada la fuerte correlación entre traducción y calidad de segmentación; la generación de corpus paralelos sintéticos a partir de modelos de aumento de datos; y la recopilación de más datos paralelos escritos para todos los idiomas indígenas trabajados, además de incorporar más idiomas.

References

Bahdanau, D., Cho, K., y Bengio, Y. (2014). 'Neural machine translation by jointly learning to align and translate'. *arXiv preprint arXiv:1409.0473*.

Canger, U. (2001). *Mexicanero de la sierra madre occidental*. El Colegio de México.

Chamoreau, C. (2003). *Purépecha de Jarácuaro* (p. 162). El Colegio de México.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., y Bengio, Y. (2014). Learning

Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724-1734.
Diario Oficial. (2014) Programa Especial de los Pueblos Indígenas 2014-2018, *Diario Oficial de la Federación*, México, Distrito Federal, 20 de abril.
Freeze, R. A. (1989). *Mayo de Los Capomos, Sinaloa (Mayo of Los Capomos, Sinaloa)*.
Gómez, P., & López, P. G. (1999). *Huichol de San Andrés Cohamiata, Jalisco* (Vol. 22). El Colegio de México.
Kann, K., Mager, M., Meza, I. Schütze, H. (2018) Fortification of Neural Morphological Segmentation Models for Polysynthetic Minimal-Resource Languages *16th Annual Conference of NAACL-HLT 2018*, New Orleans, Louisiana, US.
Klein, G., Kim, Y., Deng, Y., Senellart, J., y Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation.*En Proceedings of ACL 2017, System Demonstrations*, pp. 67-72.
Koehn, P. (2009). *Statistical machine translation*. Cambridge: Cambridge University Press.
Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ..., y Dyer, C. (2007) Moses: Open source toolkit for statistical machine translation. *En Proceedings of the 45th annual meeting of the ACL*. Association for Computational Linguistics, pp. 177-180.
Lastra de Suárez, Y. (1980). Náhuatl de Acaxochitlán (Hidalgo). *Archivos de lenguas indígenas de México. DF: Colegio de México*.
LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, 521(7553): 436.
Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3): 8.
Mager Hois, J. M., Barrón Romero, C., y Meza Ruiz, I. V. (2016). Traductor estadístico wixarika-español usando descomposición morfológica. *Memorias de COMTEL*. Lima, Perú,
Och, F. J., y Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1): 19-51.
Papineni, K., Roukos, S., Ward, T., y Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311-318.
Sandoval-Forero, E. A. (2013). Los indígenas en el ciberespacio. *Agricultura, sociedad y desarrollo*, 10(2): 235-256.

Towards a Digital History of the Spanish Invasion of Indigenous Peru

Jeremy M. Mikecz
mikecz@usc.edu
University of Southern California, United States of America

What role did indigenous activity play in shaping the events of 'conquest'? How can digital tools aid in the reconstruction of this activity? These are the core questions driving my research on the Spanish invasion of Peru.

My research experiments with the use of digital methods to assist in the rewriting of indigenous history during the early period of European invasion. In this poster, I will introduce some of these digital methods – particularly the use of data and geo-visualizations to a) identify gaps and silences in colonial sources, and b) to fill in some of those gaps with information recovered from indigenous sources.

These methods draw on diverse inspirations. In reconstructing the hidden geography or spatiality of historical texts, it follows literary geographers' recent innovative mapping of fictional sources.(Cooper et al., 2016; Cooper and Gregory, 2011; Moretti, 1998) In visualizing and recognizing patterns within these sources through the creation of a wide variety of data visualizations, it draws on work ranging from nineteenth-century information graphics to twenty-first-century data science. Finally, it also finds inspiration in pioneering work in Historical GIS, spatial history, and qualitative and even indigenous cartography. (Eltis and Richardson, 2015; Knowles et al., 2014; For indigenous cartography, see the work of Margaret Wickens Pearce, including: Pearce and Hermann, 2010)

The role of geography and indigenous activity in European invasions of the indigenous Americas – first elided or erased by colonial authors – has remained largely overlooked by modern scholars. In Inka Peru, Spanish conquistadors encountered a complex imperial infrastructure and labor system that mitigated much of the geographic challenges of an invasion of the Andes. Native guides showed them the way, native informants advised them on potential obstacles to their journey, native allies offered military and political support, native messengers relayed information between the Spanish and their allies, native porters carried their supplies, and native villagers provided them with lodging and support.

While recent work – especially increased use of indigenous sources – has begun to reconstruct some of this activity(Matthew and Oudijk, 2007), I propose a new methodology to more fully reconstruct indigenous geography and activity and to present an alternate vision of the invasion of Peru. This is accomplished in two steps. First, I use digital text analysis methods to examine how colonial sources hid or erased indigenous activity. Second, I use geovisualizations to reconstruct indigenous activity in conquest-era events as it played out across space and time. This reconstruction of indigenous activity draws on a diverse range of indigenous sources. These include: 1) indigenous polities' petitions to the Crown documenting the service they provided the conquistadors during the invasion, 2) *cacicazgo* cases which document an indigenous group's history (for disputes over hereditary succession to leadership positions) and often include some references to the conquest era, and 3) the trial testimony of indigenous witnesses describing their experiences during the period.

This reconstruction and mapping of indigenous activity will be the focus of this poster. I will provide examples of four types of data and geo-visualizations I use to reconstruct this indigenous activity. These include:

1. **Geographic Knowledge Maps:** Mapping the geographic extents (and limits) of European knowledge of the Americas– places known and unknown – makes clear just how limited their knowledge and, by extension, their power was.
2. **Mood Maps:** First created by literary geographers, mood maps allow the mapping of an author's subjective experiences of a landscape.
3. **Density Plot of Events:** Graphing the density and range of events described in historical literature allows the comparison and contrast of how the story of the conquest of Peru has changed over time.
4. **Indigenous Activity Maps,** which trace the often hidden role of indigenous actors in conquest events.

References

Cooper, D., Donaldson, C., Murrieta-Flores, P. (Eds.), 2016. *Literary Mapping in the Digital Age, New edition edition*. ed. Routledge, Farnham, Surrey, England ; Burlington, VT.

Cooper, D., Gregory, I.N., 2011. Mapping the English Lake district: A literary GIS. *Trans. Inst. Br. Geogr.* 36, 89–108.

Eltis, D., Richardson, D., 2015. *Atlas of the transatlantic slave trade*. Yale University Press, New Haven, CT.

Knowles, A.K., Cole, T., Giordano, A., 2014. *Geographies of the Holocaust*. Indiana University Press.

Matthew, L.E., Oudijk, M.R., 2007. *Indian conquistadors: indigenous allies in the conquest of Mesoamerica*. University of Oklahoma Press, Norman.

Moretti, F., 1998. *Atlas of the European novel, 1800-1900*. Verso, London; New York.

Pearce, M.W., Hermann, M.J., 2010. Mapping Champlain's Travels: Restorative Techniques for Historical Cartography. *Cartogr. Int. J. Geogr. Inf. Geovisualization*. https://doi.org/10.3138/carto.45.1.32

Style Revolution: Journal des Dames et des Modes

Jodi Ann Mikesell
jm4470@tc.columbia.edu
Columbia University, United States of America

Avery Schroeder
abschroeder4@gmail.com
City University of New York, The Bard Graduate Center, United States of America

Anne Higonnet
ahigonne@barnard.edu
Columbia University, United States of America

Alex Gil
agil@columbia.edu
Columbia University, United States of America

AnaKaren Aguero
agueroak@gmail.com
Columbia University, United States of America

Sarah Bigler
scb2180@columbia.edu
Columbia University, United States of America

Meghan Collins
mmc2267@columbia.edu
City University of New York, The Bard Graduate Center, United States of America

Emily Cormack
emily.cormack@bgc.bard.edu
Columbia University, United States of America

Zoë Dostal
azd2103@columbia.edu
Columbia University, United States of America

Barthelemy Glama
bg2601@columbia.edu
Columbia University, United States of America

Brontë Hebdon
bah416@nyu.edu
New York University, Institute of Fine Arts, United States of America

Recently rediscovered at The Morgan Library, fashion plates from the *Journal Des Dames et Des Modes*, taught all Europeans how to look, read, and entertain themselves as modern individuals. Dating from 1797-1804, they represent the most radical changes in all of clothing history. This revolution in consumer culture signals the birth of fashion as we know it and transformed conceptions of identity, gender, and power. Their revolutionary representations of fashion generates cult followings within both academic and hobbyist circles; among whom are art historians, antiquarian bibliophiles, and historical fashionistas. However, the plates lack circulation and few digital sources present research that is both academically rigorous and accessible to learners of all levels. Our work seeks to remedy this issue and bridge the accessibility gap by creating a digital exhibit of the most rare and stylistically revolutionary plates. In doing so, we have produced our exhibit using minimal computing approaches developed at Columbia University Library and the Group for Experimental Methods in the Humanities.

Our website serves as a resource for viewing the *Journal Des Dames et Des Modes* color plates themselves, but also includes resources which contribute to furthering the observer's contextual understanding. We've done this by providing concise and easily digestible academic essays,

translation glossaries for both terms and color, a historical timeline, and an interactive map which visually situates the fashion plate figures within 18th century Paris. Our conference poster reflects the importance of our topic's historically democratic roots, describes our use of Wax (a suite of tools for minimal exhibitions), and collaboration structures; and directly links our undertaking to the democratic production and dissemination of knowledge through the aesthetics of minimal computing. By creating an accessible public-facing entry into a collection of art historical objects we create a channel to information without which *Journal Des Dames et Des Modes* scholarship would remain siloed in an institution's basement.

Ten graduate students—whose diverse institutional affiliations range from Columbia University, NYU, and The Bard Graduate Center—collaborated under the direction of Professors Anne Higonnet and Alex Gil to accomplish an unprecedented digital archive and scholarly online resource. The course, "Style Revolution," was a hybrid between traditional Art History seminar and an innovative Digital Humanities seminar. Students enrolled in the course had had no prior knowledge of coding in any of the languages used (HTML, CSS, Markdown, Bash, YAML, etc) nor familiarity with any of the additional software tools that were employed to create our final site. A wide range of literacies were taught, practiced, shared and acquired, from multiple lenses and disciplines, through multi-directional pedagogy, where all became teachers for one another at some point.

The site's main functionality was built using an early version of Jekyll Wax, which creates iiif compliant tiles and manifests, and generates pages with complete sets of YAML metadata converted from a spreadsheet. The iiif in turn allows our use of Open Sea Dragon for interacting with high resolution images, without burdening the browser with front-loaded data. The spreadsheet made it possible for all graduate students, regardless of technical inclination, to contribute metadata to each plate in the archive without the need for a database or forms. Additionally, because the resulting data is in CSV format and the complete site lives on GitHub, we share all data with the public directly. Leveraging the power of markdown and Jekyll, each student was able to contribute unique multimodal 'essays' to the project, from mapping exercises to digital art, based on original research.

By providing an online resource for the *Journal Des Dames et Des Modes* we are engaging the public in creating a greater understanding of current fashion phenomena, but one for which we lack a historical framework. The *Journal Des Dames et Des Modes* helps to create this framework and guides the viewer to a deeper, more meaningful understanding of how a seemingly inconsequential fad within fashion can create a paradigm shift in societal conceptions of consumer culture and its importance in material representations of our modern day identity. Simultaneously, we are modeling how collaborative work in the beginner digital humanities classroom can achieve almost complete control of an online exhibit of public import. This work will act as the foundation for an ongoing, larger project— and has already begun to be added upon. We look forward to the constant evolution of new projects, as we believe the increased attention our site provides will generate a response of scholarship, with which, we will continue to expand our project.

The Two Moby Dicks: The Split Signatures of Melville's Novel

Chelsea Miya
cmiya@ualberta.ca
University of Alberta, Canada

There has been a longstanding debate over the cetology sections in Herman Melville's Moby Dick. These chapters, which are interwoven into the mid-section of the novel, are curiously devoid of characters or plot development and instead describe whaling biology and behavior. Some Melville scholars, including Charles Olson and Lawrence Buell, have suggested that the novel might have been written as two separate texts that were spliced together in the final stages. As the original manuscripts have been lost, this has never been confirmed. However, I hope to show that the way in which the chapters cluster together reveals that the novel does indeed have two unique stylistic signatures. This is perhaps compelling evidence in favor of the "two Moby Dicks," a phenomenon that has been much speculated upon but never proven.

References

Bastian M., Heymann S., Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

Eder, Maciej. Kestemont, Mike and Rybicki, Jan. (2015). *'Stylo': a package for stylometric analyses*.

devochdelia: el Diccionario Etimolójico de las Voces Chilenas Derivadas de Lenguas Indíjenas Americanas de Rodolfo Lenz en versión digital

Francisco Mondaca
f.mondaca@uni-koeln.de
Universität zu Köln, Germany

devochdelia es la versión digital y en línea[1] del *Diccionario Etimolójico de las Voces Chilenas Derivadas de Lenguas*

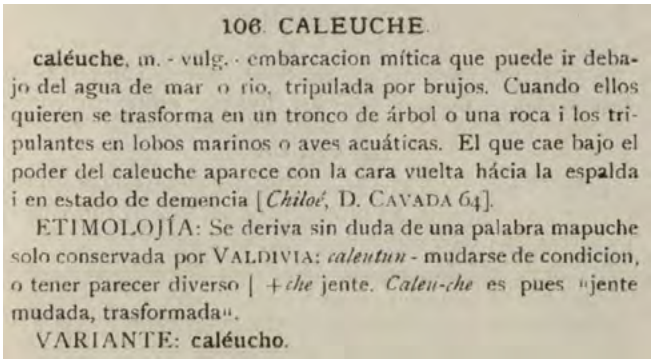
1 http://devochdelia.cl

Indíjenas Americanas (1905-1910) (Diccionario) compilado por el lingüista alemán-chileno Rodolfo Lenz. Esta obra ha sido fundamental en el desarrollo de la lexicografía chilena e hispanoamericana por su innovador y minucioso método de compilación. La digitalización de textos antiguos y valiosos como lo es el Diccionario presenta problemas engorrosos que dificultan el proceso en sí y el acceso a los datos obtenidos. En este proyecto se pueden apreciar soluciones accesibles a este tipo de dificultades facilitando tanto la digitalización de diccionarios impresos como su consulta en línea.

Acerca del diccionario impreso

La relevancia del Diccionario para lexicografía chilena radica en su enfoque descriptivo[2], que lo distingue de los diccionarios publicados en Chile hasta ese entonces. Si bien ya se habían publicado obras de americanismos con esta perspectiva, tanto en España (De Alcedo 1789) como en Cuba (Pichardo 1836), el Diccionario presenta innovaciones que lo destacan a nivel mundial. Entre ellas cabe mencionar la clara y detallada descripción del método de compilación empleado y de la teoría subyacente; la coherencia en la estructura y tipografía de los artículos, así como en la clasificación geográfica del área de empleo de los vocablos (Lenz 1905-1910[1980]:16).

Como nunca antes en la lexicografía chilena, un autor realiza un trabajo tan exhaustivo al comparar la información recabada con diccionarios publicados en Chile e Hispanoamérica. Pero no se limita a eso, también organiza conferencias con colegas, estudiantes e interesados en el tema para verificar la información reunida y añadir a su manuscrito nuevas palabras de origen indígena (Lenz 1905-1910:22ff).



106. CALEUCHE.

caléuche, m. - vulg. - embarcacion mítica que puede ir debajo del agua de mar o rio, tripulada por brujos. Cuando ellos quieren se trasforma en un tronco de árbol o una roca i los tripulantes en lobos marinos o aves acuáticas. El que cae bajo el poder del caleuche aparece con la cara vuelta hácia la espalda i en estado de demencia [*Chiloé*, D. CAVADA 64].

ETIMOLOJÍA: Se deriva sin duda de una palabra mapuche solo conservada por VALDIVIA: *caleutun* - mudarse de condicion, o tener parecer diverso | + *che* jente. *Caleu-che* es pues "jente mudada, trasformada".

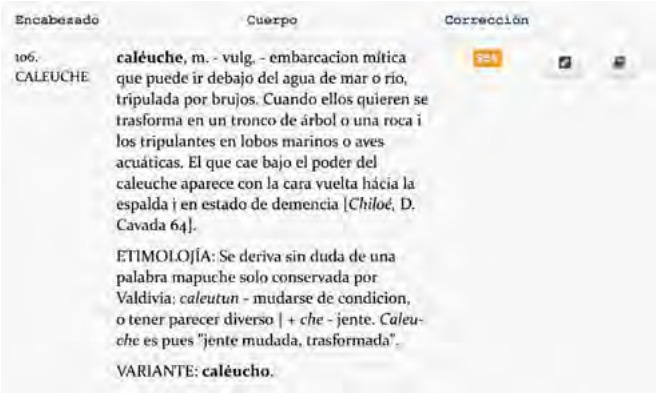
VARIANTE: **caléucho**.

La entrada ‚caleuche' en el Diccionario (Lenz 1905-1910:163)

El Diccionario cuenta con 1665 entradas que se dividen en encabezado y cuerpo. En el primero se aprecia la voz indígena propiamente tal y en el segundo se tratan las palabras chilenas derivadas de ella. Como suele ser tradición en los diccionarios semasiológicos, luego del lema se aprecia la categoría gramatical y el significado. Siempre se encuentra la „etimología", pudiendo no estar presentes secciones como „variantes" o „derivados".

Acerca del diccionario digital[3]

Un diccionario es un objeto cultural cuya función es aclarar dudas de carácter lingüístico. Por otra parte, el proceso de extracción de texto desde imágenes (OCR), es propenso a generar errores, lo que no se espera encontrar en ningún texto, menos en diccionarios. Las decisiones técnicas en este proyecto se tomaron bajo la premisa de poner en línea una versión digital del Diccionario con la menor cantidad posible de errores y, al mismo tiempo, acceder a todas las entradas del mismo. El formato elegido para la generación de texto en OCR fue Hypertext Markup Language (HTML), porque permite mantener cursivas y negritas, además de presentarse en un navegador de Internet sin problemas. Corregir todos los encabezados de las entradas, permitió la extracción de las 1665 entradas dentro de sus límites, e hizo posible buscar y encontrar las entradas mediante el número que Lenz les asignó o por el texto del encabezado. De los 1665 cuerpos, 1000 han sido corregidos.



La entrada ‚caleuche' en *devochdelia*

Una vez extraídas las entradas, se creó una aplicación web donde se pueden buscar y corregir las entradas, la cual está hecha con el *framework* Maalr (Neuefeind y Schwiebert 2013). En su versión básica, Maalr permite trabajar con entradas de diccionario en formato de texto simple. Como el fin de *devochdelia* es permitir que los usuarios ayuden a corregir las entradas, hubo que hacer dos modificaciones a Maalr:

a) que se pueda mostrar y editar texto en formato HTML,
b) que se puedan mostrar las imágenes correspondientes a cada entrada para que los usuarios vean la fuente impresa, y también editar las entradas de manera adecuada.

Cada entrada puede ser corregida y estas modificaciones ser vistas sin la necesidad de registrarse o iniciar sesión. Asimismo, cada corrección tiene que ser autoeva-

2 „I la ciencia exije que no escluyamos nada, que no dejemos de apuntar ninguna palabra" (Lenz 1905-1910:20)

3 Para más detalles, ver: http://www.devochdelia.cl/about

luada por el corrector, comunicando el nivel de la corrección a otros usuarios y a los editores.

Este proyecto muestra que, con pocos recursos, es posible digitalizar obras lexicográficas complejas haciendo partícipes en el proceso a quienes se interesan por ellas. Asimismo sirve de base para digitalizar diccionarios a otra escala.

References

De Alcedo, A. (1789). *Diccionario geográfico-histórico de las Indias Occidentales ó América*. Tomo V. Madrid: Imprenta de Manuel González.

Lenz, R. (1905-1910). *Diccionario Etimolójico de las Voces Chilenas Derivadas de Lenguas Indíjenas Americanas*. Santiago: Imprenta Cervantes.

Lenz, R. ([1905-1910] 1980). *Diccionario Etimológico de las Voces Chilenas Derivadas de Lenguas Indígenas Americanas.* Edición dirigida por Mario Ferreccio Podestá. Santiago: Universidad de Chile.

Neuefeind, C. y Schwiebert S. (2013). Introducing Maalr: A Modern Approach to Aggregate Lexical Resources. *Language Processing and Knowledge in the Web, the proceedings of the 25th Conference of the German Society for Computational Linguistics (GSCL 2013)*, Darmstadt, Alemania, 25-27 febrero 2013. https://gscl2013.ukp.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/conferences/gscl2013/demo_maalr-gscl2013.pdf (consultado el 25 de abril de 2018)

Pichardo, E. (1836). *Diccionario Provincial de Voces Cubanas*. Matanzas: Imprenta de la Real Marina.

Unsustainable Digital Cultural Collections

Jo Ana Morfin
jo.morfin@gmail.com
Universidad Nacional Autónoma de México, Mexico

This paper analyzes how in the context of Mexican museums, the lack of policies, frameworks and strategic planning has led to the creation of unsustainable cultural digital collections. It focuses on the challenges in rescuing the digital collection "Bienal Internacional de Poesía Visual y Experimental" [Biennale of Visual and Experimental Poetry], held at the Mediateque of the Museo Universitario del Chopo.

The Mexican artists Araceli Zúñiga y César Espinosa organized the International Biennales of Visual and Experimental Poetry between 1985 and 2009. These events brought together practitioners from all over the world whose work is placed at the intersections of the fields of contemporary visual writing, copy art, concrete music, mail art and performance.

Throughout the years, Zúñiga and Espinosa became interested in creating a "memory" of these events. Therefore, they started to gather videos, mail art works, photography, artistic electrography, from each event. The collection was stored at their house and classified and organized by the artists themselves. Through the years, the collection became a key source for researching and tracing the development of alternative and experimental art practices in Mexico.

Given the significance of this collection and with the aim of preserving and providing greater access to its contents, Zúñiga and Espinosa agreed with the Museo Universitario del Chopo in digitizing the materials and donating a digital version to be included in the collection of the museum. Over 2,000 artworks were digitized. In 2015 the museum received a grant to put these contents online. However, during the development of the project we realized that most of the digital objects were unstable. Given this situation, the project focused on rescuing this digital collection from the oblivion.

The project brought to light several concerns, such as the lack of a digital preservation planning, the deficient use of metadata standards, the shortage of expertise, and more importantly, the lack of institutional policies to create sustainable digital collections. The museum´s team did not follow clear guidelines, standards and best practices for the creation of digital objects and their subsequence management. Thus affecting the ability to read, access and understand the digital materials.

This poster describes the rescuing process and the guidelines we create in order to prevent the creation of unsustainable digital collections within cultural memory institutions.

La automatización y "digitalización" del Centro de Documentación Histórica "Lic. Rafael Montejano y Aguiñaga" de la Universidad Autónoma de San Luis Potosí, mediante la autogestión y software libre

José Antonio Motilla
jamotilla@gmail.com
Universidad Autónoma de San Luis Potosí, Mexico

Ismael Huerta
ismaelhuerta.ten@gmail.com
Universidad Autónoma de San Luis Potosí, Mexico

La presente ponencia tiene como objetivo presentar el estudio de caso del proceso de modernización del Centro de Documentación Histórica "Lic. Rafael Montejano y Aguiñaga" de la Universidad Autónoma de San Luis Potosí (CDHRMA-UASLP), México, constituido por una colección de aproximadamente 100 mil volúmenes bibliográ-

ficos, una amplia sección de manuscritos, publicaciones periódicas, e impresos , y un gran acervo documental que incluye el Archivo Histórico de la UASLP.

Hacia el año 2014, el CDHRMA-UASLP trabajaba con un sistema fundamentalmente análogo, al no contar con un catálogo electrónico ni de herramientas tecnológicas que le permitieran preservar y difundir sus materiales. En ese contexto, se emprendió un profundo diagnóstico del Centro, que buscaba detectar sus carencias con el fin de hacer más eficientes sus procesos. El análisis arrojó como resultado la necesidad de fortalecer cuatro áreas fundamentales: la implementación de un Sistema Integral de Automatización de Bibliotecas (SIAB); el manejo de los inventarios y catálogos mediante bases de datos eficientes; la digitalización de los materiales de alta demanda para garantizar su conservación; y la investigación académica de sus fondos y colecciones.

Ante la falta de presupuesto institucional, el equipo encargado del desarrollo del proyecto tomó la decisión de desarrollar el proyecto mediante el empleo de Software Libre y desarrollar estrategias para reducir al máximo los costos; así, para el desarrollo del SIAB se recurrió a la plataforma de acceso libre Koha; se migraron y sistematizaron las bases de datos en la plataforma File Maker (único software de paga que fue utilizado); para la digitalización de materiales se adquirió una cámara de alta resolución y se creó un soporte con iluminación no profesional para digitalizar documentos; y se implementó un equipo de investigación, coordinado por el departamento de investigación del Centro, con el apoyo de becarios, para crear bases de datos y analizarlas bajo el paradigma de las humanidades digitales.

Como resultado, al día de hoy se cuenta con un catálogo electrónico con más de 7 mil registros, dos periódicos del siglo XIX completamente digitalizados y en consulta, un inventario general de la biblioteca, la descripción detallada y digitalización de algunos fondos del archivo histórico, y herramientas de investigación realizadas mediante minería de texto.

La experiencia y reflexión planteada en ésta ponencia, busca poner sobre la mesa la importancia que herramientas como el software libre, y el desarrollo de aplicaciones tecnológicas e informáticas, puede impactar de manera favorable en la conservación y difusión de acervos bibliográficos y documentales de alto valor patrimonial, y garantizar el acceso a ellas por parte de los investigadores y público interesado tanto del presente como de generaciones futuras.

A Comprehensive Image-Based Digital Edition Using CEX: A fragment of the Gospel of Matthew

Janey Capers Newland
janeycapers.newland@furman.edu
Furman University, United States of America

Emmett Baumgarten
emmett.baumgarten@furman.edu
Furman University, United States of America

De'sean Markley
desean.markley@furman.edu
Furman University, United States of America

Jeffrey Rein
jeffrey.rein@furman.edu
Furman University, United States of America

Brienna Dipietro
brienna.dipietro@furman.edu
Furman University, United States of America

Anna Sylvester
anna.sylvester@furman.edu
Furman University, United States of America

Brandon Elmy
brandon.elmy@furman.edu
Furman University, United States of America

Summey Hedden
summey.hedden@furman.edu
Furman University, United States of America

This poster (with accompanying downloadable dataset and application) will demonstrate as a proof-of-concept a comprehensive image-based publication and analysis of a text bearing artifact, [catalog number redacted for anonymous review], a hitherto unpublished 10th Century palimpsest fragment of the Gospel According to Matthew. The fragment contains most of the "Parable of the Sower".

In editing this text, we sought to be as comprehensive as possible, capturing:

- Natural light and UV images, both overview images and details
- A diplomatic transcription of the overwritten text of Matthew and any legible characters from the undertext
- A word tokenization of the diplomatic transcription, mapping to each token:
  - a normalization
  - editorial status
  - lexical status
  - morphology, part of speech, and syntactic relations
  - alignment to the image data
- A character tokenization, aligned to the image data
- An edition of the whole Gospel according to Matthew from a critical edition, for comparison and context
- Translations aligned to the text
- Editors' comments

In publishing it, we sought simplicity, longevity, and clarity. While we use TEI XML as a format for capturing an initial transcription, the overlapping analytical categories, many-to-many alignments of text and image, and open ended possibilities for commentary precluded implementing a coherent data model fully in XML. At the same time, we wanted a concise and integrated publication.

By using the CEX format[1], a plain text, self-documenting format based on the CITE/CTS architecture, we are able to bring together these many levels of analysis in a form that is at once disaggregated, with each scholarly primitive explicitly and unambiguously citable, while still united in a single file. CEX allows us to distribute a fully integrated dataset in the form of a single plain text file and a single directory of images.

Our publication, a CEX file and a directory of images, is technology-agnostic readable by humans, but also able to serve as the data for an end-user application. We will describe, and have available for download and on USB thumbdrives, a lightweight, zero-configuration single page web application (SPA), fully usable offline, that integrates the data and images for this publication for end-users.[2]

Finally, we will outline the low cost, low technology, collaborative work behind the digitization and editing of this manuscript fragment: off-the-shelf cameras, simple handheld UV lighting, readily available FOSS software.

We believe that this work will be of interest to the international Digital Humanities research community both as a new publication of a Byzantine Greek text and as a demonstration of a replicable and sustainable combination of technology and workflow. We think this approach provides a compelling alternative to XML or RDF editions and complex database-driven end user applications, offering advantages both on the back end (a flexible, scalable, and self-documenting format for implementing diverse data models), and on the front end (lightweight and portable presentation for readers). At the same time the data we present as CEX and images is easily transferrable to other standard formats.[3]

All project data is under version control in a public GitHub repository, and licensed under a CC-BY license. All source code is under either a GPL or MIT public license.

1 CEX (CITE Exchange Format) is a plain text format for capturing data about texts and collections, based on the CITE/CTS architecture and developed by C. Blackwell (*Homer Multitext*), T. Köntges (University of Leipzig), and N. Smith (*Homer Multitext*). For implementations and projects using CEX, see: T. Köentges, (Meletē)ToPān (topic modelling environment): https://thomask81.github.io/ToPan/; C. Blackwell, N. Smith, CEX Library (Scala): https://github.com/cite-architecture/cex; C. Blackwell, N. Smith, CEX Dataset Repository: https://github.com/cite-architecture/citedx

2 This application is based on the ScalaJS implementation of "CITE App" by C. Blackwell and N. Smith: https://github.com/cite-architecture/CITE-App

3 Existing code libraries for working with CEX include a microservice framework that delivers textual and other data from CEX files as JSON objects, via HTTP requests (see https://github.com/cite-architecture/scs-akka) and libraries that export CEX data into other formats, such as 2-column tabular data or 82XF (see https://github.com/cite-architecture/scm).

Using Zenodo as a Discovery and Publishing Platform

Daniel Paul O'Donnell
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

Natalia Manola
natalia@di.uoa.gr
OpenAIRE

Paolo Manghi
paolo.manghi@isti.cnr.it
Zenodo, Switzerland; CNR, Italy

Dot Porter
dot.porter@gmail.com
University of Pennsylvania

Paul Esau
paul.esau@gmail.com
University of Lethbridge, Canada

Carey Viejou
c.viejou@uleth.ca
University of Lethbridge, Canada

Roberto Rosselli Del Turco
robertorossellidelturco@gmail.com
University of Pisa, Italy; University of Turin, Italy

We are 25 years into the World Wide Web revolution. While Humanities researchers have been at the forefront of many uses of networked communication to disseminate their research, they have lagged other disciplines in their adoption of formal discovery and organisational tools (Spiro, 2016; Borgman, 2009; Anderson et al., 2012). Some of the core tools that characterise current best practice in other disciplines—ORCID, DOIs, discipline-wide repositories, mega and overlay journals—have seen slow or limited adoption in the case of Humanities researchers. Data Management and Citation practices tend to be less well-developed and widely practised in the Humanities than in other areas. Humanities publishing, too, especially scholar-led publishing, still commonly involves less than optimal practice—custom, project-held URLs, storage on private/commercial data servers, a lack of formal attention to versioning, backups, and long-term preservation (Copland et al., 2016).

This poster shows how two projects at the University of Lethbridge are addressing these long-standing problems through the use of OpenAIRE/Zenodo (the final form of the poster is O'Donnell et al., 2018). In one case, the project is looking for an open and FAIR (Findable, Accessible, Interoperable, and Reusable) method of publishing project data—a small (by cross-disciplinary standards) set of 2D and 3D images and point clouds, annotations, and textual transcriptions involving medieval

cultural and textual heritage. The goal here is to establish an expansible repository that will allow for non-negotiated additions and reuse by external projects and survive and remain citable long after the originating project has concluded and funding has run out.

The second is the publication platform for a graduate-student run journal. In this case, the students needed a platform that would provide their early career authors with some guarantee of permanent archiving and discoverability while recognising and accommodating the inherently unstable nature of a graduate-student run editorial board: while this year's board is enthusiastic about the project, we have no way of guaranteeing that this will be true of future generations of graduate students.

Although other options exist to solve both these problems, our poster demonstrates the degree to which OpenAIRE/Zenodo provides an extremely simple and durable platform for ensuring the long-term discoverability and preservation of Humanities research in these common use cases.

References

Anderson, D. E., Dwyer, G. and Leahy, S. (2012). Fine-Tuning the Institutional Repository: Evaluating the Self-Archiving Behavior of Researchers in Music. *The Serials Librarian*, 63(3-4). Routledge: 277–87 doi:10.1080/0361526X.2012.722594. https://doi.org/10.1080/0361526X.2012.722594.

Borgman, C. L. (2009). DHQ: Digital Humanities Quarterly: The Digital Future is Now: A Call to Action for the Humanities http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html (accessed 20 June 2017).

Copland, C., Carrell, S., Davidson, G., Grandfield, V. and O'Donnell, D. P. (2016). Kiernan, Kevin S. 2015. Electronic Beowulf - Fourth Edition. *Digital Medievalist*, 10 http://www.digitalmedievalist.org/journal/10/copland/ (accessed 25 April 2017).

O'Donnell, D. P., Manola, N., Manghi, P., Porter, D., Esau, P., Viejou, C., Del Tuco, R. R. and Singh, G. (2018). Using Zenodo as a Discovery and Publishing Platform Paper presented at the DH 2018, Mexico doi:10.5281/zenodo.1234474. https://zenodo.org/record/1234474.

Spiro, L. (2016). Studying how digital humanists use GitHub *Digital Scholarship in the Humanities* https://digitalscholarship.wordpress.com/category/open-access/ (accessed 27 November 2017).

SpatioScholar: Annotating Photogrammetric Models

Burcak Ozludil Altin
bozludil@njit.edu
New Jersey Institute of Technology, United States of America

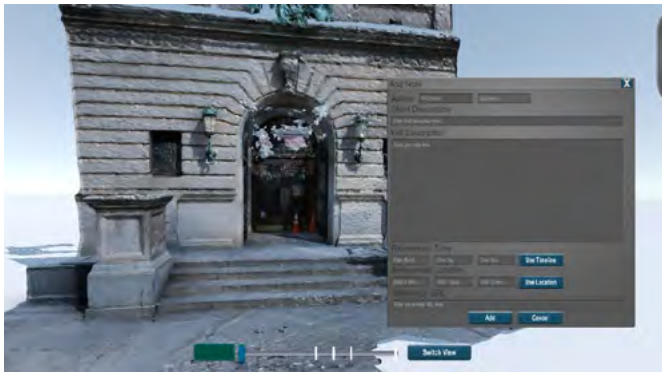
Augustus Wendell
wendell@njit.edu
New Jersey Institute of Technology, United States of America

This poster presents a new phase in the development of *SpatioScholar* which is a platform for studies that require spatial and temporal processing, visualization, and analysis, including art/architectural history and urban research (Wendell et al., 2016). The platform is a scholarly application created in Unity3D synthesizing 3D models with textual information and research documents. (Figure 1) *SpatioScholar* provides a computational close reading system for spatial and temporal data sets by using the following functionalities: (1) through a timeline slider, it demonstrates the phases through which a certain building or location passed; (2) through a simulation, it provides the viewer with the ability to experience the space in first person, and to track any desired aspect of life inside buildings or locations; (3) through a reconnection of the primary materials and the conclusions derived from them, viewers can browse and review the relevant information (photographs, drawings, textual primary documents etc.) that are cross-referenced with the "scene;" and (4) through a "Shared Scholarship" feature, viewers and users can leave notes, comments or browse others' notes.

SpatioScholar interface displaying time slider, mini-map, primary source panel and an enlarged historic photograph that shows the same scene in the 1910s (Photograph source: Sihhat Almanaki, 1933.)

At this phase of the *SpatioScholar* development, we are testing the platform with photogrammetry models. (Figure 2) Photogrammetry is a computational process coordinating optical data recorded in a series of photographic images, solving matched data points for a 3D point cloud, and outputting a 3D model with applied photographic textures. The benefits of photogrammetry to digital art/architectural history and digital heritage in comparison to traditional 3D model building are well debated (Allen et al., 2003; El-Hakim et al., 2007; Webb, 2016). While other scholars have included photogrammetry data in spatial simulations (Ozer and Nagakura, 2016), we are extending this line of work by integrating a shar-

able spatial annotation feature within a single distributed application.



A photogrammetric model imported into SpatioScholar. The "Add Note" interface element shown is used for the sharable spatial annotation

Incorporating photogrammetry in *SpatioScholar* presents multiple advantages: first, it eliminates the need to create a 3D model from scratch for projects that are not previously modelled. Second, the use of simple photographic data allows non-technical or non-traditional users to capture, research, and create accurate 3D representations of space, even with smart phones (6). Adapting *SpatioScholar* to photogrammetry will widen the user base as this technology becomes more readily available and accessible in the field.

SpatioScholar implements a custom developed space based annotation toolset that allows notation of the photogrammetric 3D model through a web accessible database. This feature, combined with a WebGL delivery mode allows a research project to be delivered via the web in the same interface for input, comments, and collaborations without the need to transfer or use another medium. This single interface in *SpatioScholar* combines the **research phase** inherent to scholarly production and its **sharing** with the outside world.

The components that create the *SpatioScholar* functionality within Unity3D are programmed elements that actively manage models, database interaction, user interface, and primary source document coordination. As it stands now, the user imports an FBX format version of photogrammetry model into a Unity3D enabling all the functionalities of the platform by using a previously created "SpatioScholar Unity3D Template Project." By dropping their imported FBX file onto the coordinating *SpatioScholar* component, Unity3D creates temporal, primary document and annotation associations based on existing metadata mapping within the FBX file.

SpatioScholar was conceptualized first and foremost as a platform to facilitate and share research, not as a tool to merely navigate the virtual reconstruction of a building or site. The incorporation of photogrammetry as a fairly accessible technology into the platform paves the path to opening of the platform to a wider user-base that can employ its functionalities to foster research. This poster demonstrates the potentials in bringing spatial data into *SpatioScholar* to create a web-distributable spatial research project, by enlisting temporal awareness, trajectory tracking, primary document coordination, and shared annotation features.

References

Allen, P. K., Troccoli, A., Smith, B., Murray, S., Stamos, I., Leordeanu, M. (2003). New methods for digital modelling of historic sites. In *IEEE Computer Graphics and Applications*, 23(6), 2003, pp. 32–41.

El-Hakim, S, Gonzo, L., Voltolini, F., Girardi, S., Rizzi, A., Remondino, F., Whiting, E. (2007). Detailed 3D Modelling of Castles. In *International Journal of Architectural Computing* 5(2): 200-220.

Fassi, F. (2012). Complex architecture in 3D from survey to web. *International Journal of Heritage in the Digital Era*, 1(3): 379-398.

Osman, M. (1933). *Sihhat Almanaki*, Kader Matbaasi, Istanbul.

Ozer, D.G. and Nagakura, T. (2016). Simplifying architectural heritage visualization – *AUGMENTEDparion.* In Herneoja, A., Österlund T. and Markkanen, P. (eds.), *Complexity & Simplicity - Proceedings of the 34th eCAADe Conference* - Volume 2, University of Oulu, Oulu, Finland, 22-26 August 2016, pp. 521-528.

Webb, N., Buchanan, A. and Peterson, J.R. (2016). Modelling medieval vaults: comparing digital surveying techniques to enhance our understanding of gothic architecture. In Herneoja, A., Österlund T. and Markkanen, P. (eds.), *Complexity & Simplicity - Proceedings of the 34th eCAADe Conference* - Volume 2, University of Oulu, Oulu, Finland, 22-26 August 2016, pp. 493-502.

Wendell, A., Ozludil Altin, B. and Thompson, U. (2016). Prototyping a temporospatial simulation framework:case of an ottoman insane asylum. In Herneoja, A., Österlund T. and Markkanen, P. (eds.), *Complexity & Simplicity - Proceedings of the 34th eCAADe Conference* - Volume 2, University of Oulu, Oulu, Finland, 22-26 August 2016, pp. 485-491.

Decolonising Collections Information – Disrupting Settler Colonial Power In Information Management in response to Canada's Truth & Reconciliation Commission and the United Nations Declaration on the Rights of Indigenous Peoples

Laura Phillips
laura.phillips@queensu.ca
Queens University, Canada

Standard collections information management principles in use by settler colonial cultural institutions derive from the foundation of museums as repositories to showcase the extent of empire and, as with all 'Euro-Western' disciplines, are not capable of objectivity in approach or in reflecting the multiplicity of identities in non-Western world views (Garneau, 2016). As a reflection of contemporary society, cultural institutions must be at the forefront of the decolonisation movement, and not simply initiate projects that perpetuate the museum as the authority to further (consciously or unconsciously) settler colonialist aspirations as one of the "…lasting effects of European colonialism on the multiple stagings and worldings of nations and societies across the globe" (Byrd, 2017: 176). Decolonisation in Canada means critically reflecting on the colonial bias for accepted 'truths' projected by the actions and ethos of cultural institutions, especially museums, to analysis bias in information management from the point of ingestion to management and re-presentation.

Having participated in efforts to build museums based on non-Western world views in both Qatar (Taylor, 2014) and the Cree Nation in Eeyou Istchee (Pashagumskum, 2016), my research continues my progression in deconstructing professional museum practice by exploring these questions:

- How can contemporary museology incorporate Indigenous perspectives to address the power imbalance that perpetuates colonial mythology and the related presumption of ownership rights?
- What practical methods can reframe methods of engagement between Indigenous communities and museums?
- How can critiques of museum practice by Indigenous knowledge keepers be presented to museums to change established procedures?
- How can Indigenous values and traditional knowledge be shared with museums to centre the Indigenous perspective, while respecting unique traditions for each community and safeguarding their intellectual property rights?
- How can museums and curators identify the settler colonial bias in their work?
- Is any of this even possible given that museums are founded on 'scopophilia' (Garneau, 2016)?

My innovative, community-centric research approach will improve the efficacy demonstrated in existing case studies of community based research (Smith, 1999; Tuck, 2009; Tuck and Wang, 2012; Tuck and Wang, 2014). The focused application of Indigenous knowledge to museology, including collections information management, will generate guidance required for imperative revisions in museum policies and procedures to become consistent with Canada's Truth & Reconciliation Commission (Truth and Reconciliation Commission, 2015) and United Nations Declaration on the Rights of Indigenous Peoples (United Nations, 2008).

My poster will present ideas for shifts in information management as perceived during my Ph.D. research in Cultural Studies, including case studies from Indigenous institutions in Canada to demonstrate ways to disrupt the current colonial power structures. The ideas presented will provoke discussion that will ultimately help to create self-empowering principles to engage international, national and provincial cultural institutions to form the basis of new standards of decolonised cultural information management. My poster will include examples of decolonisation efforts taking place in Canada to de-centre the settler colonial hegemony, an overview of theoretical approaches used as the foundation for this shift, and explain how militant research principles (Colectivo Situaciones, 2003; Brown, 2013) can be applied to day to day cultural information management on an individual level to disrupt the current paradigm.

References

Brown, N. (2013). *Militant Research Handbook*. New York: New York University.

Byrd, J. (2017). American Indian Transnationalisms. In Goyal, Y. (ed), *The Cambridge Companion to Transnational American Literature*. Cambridge: Cambridge University Press, pp. 174–89.

Colectivo Situaciones (2003). On the Researcher-Militant *European Institute for Progressive Cultural Policies* http://eipcp.net/transversal/0406/colectivo-situaciones/en.

Garneau, D. (2016). Imaginary Spaces of Conciliation and Reconciliation: Art, Curation, and Healing. In Robinson, D. and Martin, K. (eds), *Arts of Engagement: Taking Aesthetic Action In and Beyond the Truth and Reconciliation Commission of Canada*. Waterloo: Wilfred Laurier University Press, pp. 21–41.

Pashagumskum, S., Menarick, P., Phillips, L., Laurendeau, G. and Scott, K. (2016). Seeing Ourselves: The Path to Self-curation, Cultural Sovereignty and Self-Representation in Eeyou Istchee. In Hele, K. (ed), *Survivance and Reconciliation: 7 Forward / 7 Back: 2015 Canadian Indigenous Native Studies Association Conference Proceedings*. Manitoba: Aboriginal Issues Press, pp. 60–87.

Smith, L. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. 2nd ed. London: Zed Books Ltd.

Taylor, D., Phillips, L., Al Malek, N. and Alathbah, N. (2014). Collective Opportunities: Collections Management in Qatar. In Erskine-Loftus, P. (ed), *Museums and the Material World: Collecting the Arabian Peninsula*. Edinburgh: Museums Etc, pp. 412–52.

Truth and Reconciliation Commision of Canada (2015). Honouring the Truth, Reconciling for the Future: Summary of the Final Report of the Truth and Reconciliation Commission of Canada http://www.trc.

ca/websites/trcinstitution/File/2015/Findings / Exec_Summary_2015_05_31_web_o.pdfhttp://www.trc.ca/websites/trcinstitution/File/2015/Findings/Exec_Summary_2015_05_31_web_o.pdf (accessed 1 June 2017).
Tuck, E. (2009). Re-visioning Action: Participatory Action Research and Indigenous Theories of Change. *Urban Review*, 40(11): 47–65.
Tuck, E. and Ree, C. (2013). A Glossary of Haunting. In Jones, S., Adams, T. and Ellis, C. (eds), *Handbook of Autoethnography*. London: Routledge, pp. 639–58.
Tuck, E. and Wang, K. W. (2012). Decolonization is not a metaphor. *Decolonization: Indigeneity, Education & Society*, 1(2): 1–40.
Tuck, E. and Wang, K. W. (2014). R-Words: Refusing Research. In Paris, D. and Winn, M. (eds), *Humanizing Research: Decolonizing Qualitative Inquiry with Youth and Communities*. Thousand Oakes: Sage Publications, pp. 223–47.
United Nations (2008). United Nations Declaration on the Rights of Indigenous Peoples , http://www.un-.org/esa/socdev/unpfii/documents/DRIPS_en.pdf.
Wilson, J. (2016). Gathered Together: Listening to Musqueam Lived Experiences. *Biography*, 39(3): 469–94.

An Ontological Model for Inferring Psychological Profiles and Narrative Roles of Characters

Mattia Egloff
mattia.egloff@unil.ch
University of Lausanne, Switzerland

Antonio Lieto
lieto@di.unito.it
University of Turin, CAR-CNR, Italy

Davide Picca
davide.picca@unil.ch
University of Lausanne, Switzerland

Introduction

The modelling of the inner world of narrative characters and the ability to capture and formally shape their deep psychological characteristics are at the center of the reflection of a part of literary criticism and remains, today, an open challenge in the Digital Humanities. In this paper, we present an ongoing work of a preliminary version of the Ontology of Literary Characters (OLC), that allows to capture and inference psychological characters' traits starting from their linguistic descriptions as they appear in literary texts.

The ontology of literary characters

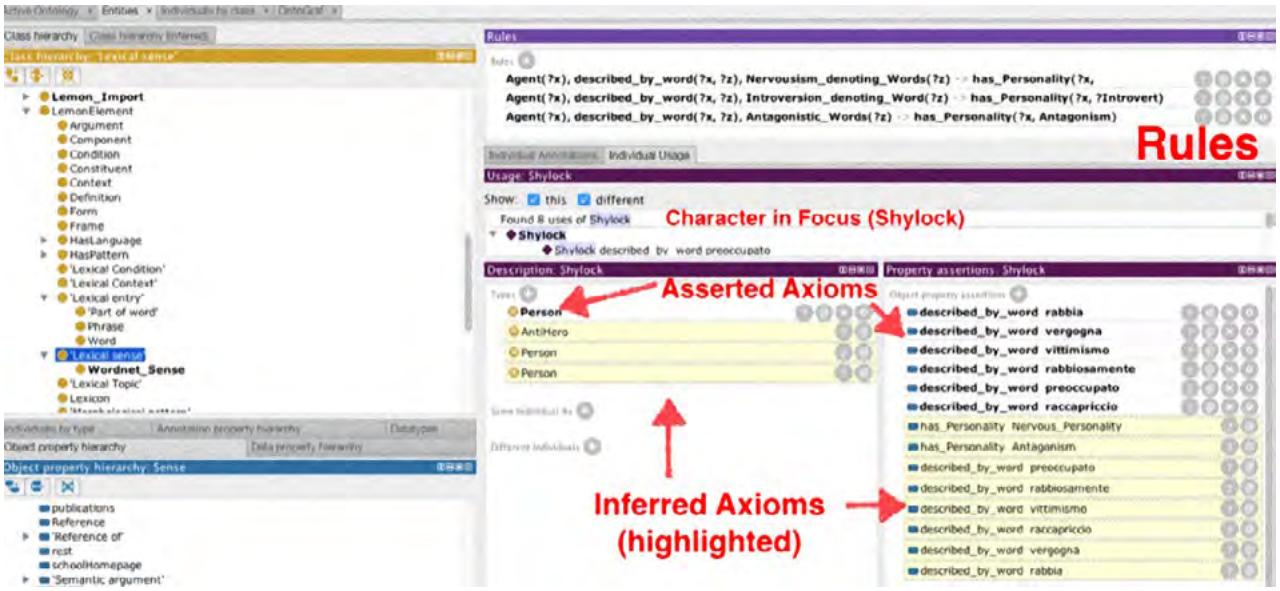
The ontology of literary characters (OLC) integrates different ontological models already available in conceptual models literature. In particular, it integrates the ontology framework LEMON (The Lexicon Model for Ontologies, (McCrae et al., 2011)) and the Ontology of Emotion (OE) (Patti et al., 2015) (encoding affective knowledge in emotional categories based on both Plutchik's (Plutchik, 1997)) and Hourglass's models in (Cambria et al., 2012)) with two additional models:

- a preliminary ontology of narrative roles
- a model of psychological profiles relying on the model of the Big 5 personality traits (Digman, 1990).

In our ontology, the word level is encoded in our model as Lexical Entry in the LEMON module. Lexical Entries are linked to their corresponding Emotion through the property *describes emotion*. The different set of Emotions is represented with the OE model that currently includes 32 emotional concepts. Each of such concept, as specified above, is connected to the word level and, in addition, is connected with specific concepts represented the micro-ontology of the Big Five Personalty Traits. The latter integrated model allows to categorize the psychological profiles of the characters along the axes of Openness to experience Conscientiousness, Extraversion, Agreeableness and Neuroticism. Finally, the concepts of Big Five micro-ontology are connected with those represented in an additional module that allows to represent the narrative roles played by the characters in a given story. Such integrated micro-ontology of narrative roles has been based on the archetypes of HERO, ANTI-HERO and VILLAIN which are commonly used in the narrative realm (Lieto and Damiano, 2014). Regarding the HERO class is represented with the following relevant narrative features: e.g. the fact that it is characterized by his/her fights against the VILLAIN of a story, the fact that his/her actions are necessarily guided by general goals to be achieved in the interest of the collectivity, the fact that they fight against the VILLAIN in a fair way and so on. The ANTI-HERO, on the other hand, is described as characterized by the fact of sharing most of its typical traits with the HERO (e.g. the fact that it is the protagonist of a plot fighting against the VILLAIN of the story); however, his/her moves are not guided by a general spirit of sacrifice for the collectivity but, rather, they are usually based on some personal motivations that, incidentally and/or indirectly, coincide with the needs of the collectivity. Furthermore the ANTI-HERO may also act in a not fair way in order to achieve the desired goal. A classical example of such archetype is Shylock which is described with the words "rabbia"/"anger", „vergogna"/"shame", etc (See Figure 1) . Each of these words is associated with a specific emotion of the OE ontology. In addition, each emotion is linked in the ontology to a

particular Psychological Profile from the Big Five Model. Finally, each Personality of the Big Five Model is semantically connected with a particular narrative role. Finally the VILLAIN is represented as a classic negative role in a plot and is characterized as the main opponent of the protagonist/HERO.

The overall integrated ontological model allowed us to show how a given character (e.g. Shylock in figure 1) described in the text with some particular psychological-denoting words (e.g. described by the words "rabbia"/"anger" …) can be automatically associated to one of the 5 classes of the personality traits of the Big Five and, as as consequence, also to the corresponding narrative role played in a story. Such semantic association is performed by using the ontological connections between the lexical level and the Emotional Concepts and an additional layer of SWRL rules connecting specific types of Words to specific Personality Traits, (See Figure 1).



Example Shylock.

Conclusion

In this paper, we presented an ongoing work on a first version of the Ontology of Literary Characters (OLC),As already observed by (Egloff et al., 2016) this ontology highlights the close relationship between character and language. In particular, where words play a significant role is crafting what we would now call the "personalities" in literature. As a result of these semantic connections it is possible to infer, starting from the natural language description of a given character, which is his/her psychological profile and his/her role played in the plot. In the case of Shylock, the system automatically infer that this character plays the role of ANTI-HERO in the plot. This ontological approach offers a new mean to scholar in order to isolate and analyze these verbal features of character going from natural language description of literary characters to the automatic assignment of their narrative role.

References

Cambria, E., Livingstone, A. and Hussain, A. (2012). The hourglass of emotions. *Cognitive Behavioural Systems*: 144–157.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1): 417–440.

Greenwade, G. D. (1993). The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3): 342–351.

Lieto, A. and Damiano, R. (2014). A hybrid representational proposal for narrative concepts: A case study on character roles. *OASIcs-OpenAccess Series in Informatics*, vol. 41. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Egloff, M., Picca, D. and Curran, K. (2016). How IBM Watson Can Help Us Understand Character in Shakespeare: A Cognitive Computing Approach to the Plays. *In Digital Humanities 2016: Conference Abstracts*. Jagiellonian University and Pedagogical University, Kraków, pp. 488–92.

McCrae, J., Spohr, D. and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *Extended Semantic Web Conference*. Springer, pp. 245–259.

Patti, V., Bertola, F. and Lieto, A. (2015). ArsEmotica for arsmeteo. org: Emotion-Driven Exploration of Online Art Collections. *FLAIRS Conference*. pp. 288–293.

Plutchik, R. (1997). The circumplex as a general model of the structure of emotions and personality.

A Graphical User Interface for LDA Topic Modeling

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Severin Simmler
severin.simmler@stud-mail.uni-wuerzburg.de
University of Würzburg, Germany

Thorsten Vitt
thorsten.vitt@uni-wuerzburg.de
University of Würzburg, Germany

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

Using LDA (Latent Dirichlet Allocation) for analyzing the content structure of digital text collections is a possibility, that aroused the interest of many digital humanists in the recent years. The method allows to generate a so called 'topic model' from a text corpus, each 'topic' in the model being represented by a probability distribution over the words in the corpus. In each of these topics, another group of semantically related words appears with high probability scores. By labeling topics with their most probable words and then calculating the relative contributions of the topics to each text or text segment, researchers can use LDA as an unsupervised method to survey the contents of a text corpus (Blei 2012, Steyvers and Griffiths 2006).

However, to actually use LDA, technical skills lacked by the majority of humanities scholars is necessary. There is a number of accessible implementations of the LDA algorithm, the most popular being in MALLET (McCallum 2002), a Java program that has to be run and controlled from the command line and Gensim (Rehurek und Sojka 2010), a text analysis library for the Python programming language. Basically, most existing implementations of the algorithm require programming skills to be used efficiently, and for most use cases one has to switch between systems, tools and programming languages to complete the entire workflow from preprocessing to the analysis of results.

With the aim of lowering the threshold to use LDA for humanities scholars, we developed a programming library in Python that significantly reduces the complications to control the whole process of topic modeling from preprocessing to the visualization of results with a single Python script. The library, developed with funding from the European infrastructure project DARIAH (https://de.dariah.eu/), allows to choose from three different LDA implementations (MALLET, Gensim, and the 'LDA' package by Allan Riddell; https://pypi.python.org/pypi/lda). It provides a number of interactive, extensively annotated jupyter notebooks (http://jupyter.org/) that can be used as tutorials for beginners and template workflows that can be adjusted to individual needs.

Many potential users are not yet familiar with programming at all, but interested in the method and eager to experiment with it a little before deciding if it is worth learning a new set of skills to use it to its full extent. For them the learning curve of a jupyter notebook is still too steep. That at least was the feedback we received in our workshops which we organized to get feedback from scholars: the wish for a GUI to access at least the basic functionalities was expressed frequently. To meet this demand, we started the development of a 'GUI Demonstrator' that mirrors the working steps and explanations in the notebooks, and allows users to analyse their own texts using LDA with a limited set of options.

The current version, that is implemented in the FLASK microframework (http://flask.pocoo.org/) and runs within a browser window (Fig 1.), includes all steps necessary to get from a number of raw text files (txt and xml file formats are supported) to a visualized output, currently an interactive heat map showing the distribution of topics over texts (Fig. 2). As the quality of results depends on removing frequent words that appear in all texts, users can decide on the number of most frequent words to remove, or provide their own stopword list. They can control the number of topics to be generated, and the number of iterations the algorithm should run. The latter is important, because a large number of iterations will produce more stable results, but the algorithm will take longer to finish the task.

The next working steps include the implementation of standalone graphics in the Qt library (https://www1.qt.io/), and in allowing for flexibility in the choice and use of the results and outputs users are specifically interested in. The possibility to include metadata and evaluation results is another focus for upcoming developments, e.g. to sort text in the output heatmap according to different categories, or topics according their quality indicated by evaluation metrics.

Both the library and the Demonstrator as a standalone executable for Windows and OSX are open source and available on Github (https://github.com/DARIAH-DE/Topics).

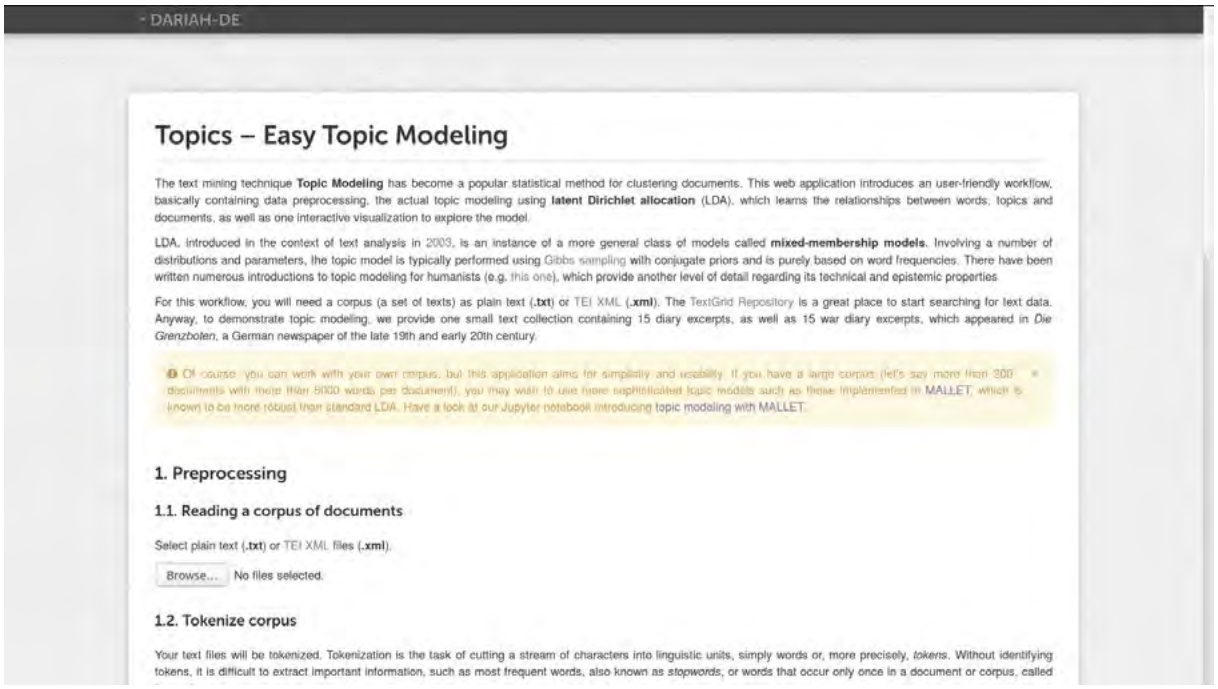


Figure 1: Screenshot of the upper end of the input screen in the current version of the GUI Demonstrator.

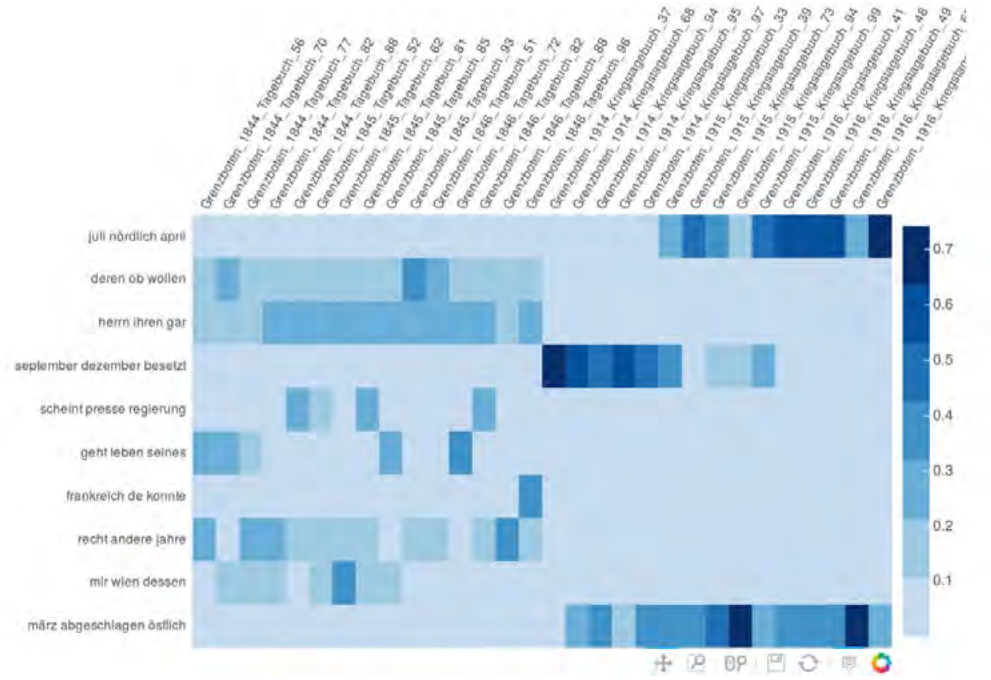


Figure 2: Example for an interactive heatmap output in the current version of the GUI Demonstrator.

References

Blei, David M. (2012): „Probabilistic Topic Models“, in *Communication of the ACM* 55, Nr. 4 (2012): 77–84. doi:10.1145/2133806.2133826.

McCallum, Andrew K. (2002): *MALLET : A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Rehurek, Radim/ Sojka, Petr (2010): "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Steyvers, Mark/ Griffiths, Tom (2006): „Probabilistic Topic Models“, in *Latent Semantic Analysis: A Road to Meaning*, herausgegeben von T. Landauer, D. McNamara, S. Dennis, und W. Kintsch. Laurence Erlbaum.

Eliminar barreras para construir puentes a través de la Web semántica: Isidore, un buscador trilingüe para las Ciencias Humanas y Sociales

Sthephane Pouyllau
stephane.pouyllau@cnrs.fr
CNRS, Huma-Num, France

Laurent Capelli
laurent.capelli@huma-num.fr
CNRS, Huma-Num, France

Adeline Joffres
adeline.joffres@huma-num.fr
CNRS, Huma-Num, France

Desseigne Adrien
adrien.desseigne@huma-num.fr
CNRS, Huma-Num, France

Gautier Hélène
helene.gautier@huma-num.fr
CNRS, Huma-Num, France

"ISIDORE" es un buscador creado por una infraestructura francesa de investigación: la TGIR Huma-Num. No solamente ofrece una plataforma de búsqueda, sino que también normaliza y enriquece los datos y metadatos que cosecha, integrándolos en la Web semántica.

Desde hace dos años, la plataforma "ISIDORE" lanzada en diciembre de 2010 puede enriquecer e indizar metadatos y recursos digitales en Ciencias Humanas y Sociales (CHS) en 3 idiomas: francés, inglés y español. Esta posibilidad es un gran avance para "ISIDORE" y abre perspectivas de colaboración científica en distintos continentes. Conforme a los principios de ciencia abierta y respetuosa de los principios "FAIR", este enfoque permite el intercambio cada vez más estrecho de numerosos datos integrados en la Web de datos.

De hecho, ahora cuenta con más de 5 millones de recursos digitales (libros, revistas científicas, artículos científicos, anuncios y programas de eventos, convocatorias, blogs, mapas, archivos, documentos audiovisuales, etc.) interconectados mediante referenciales, indizados por un motor de búsqueda. Estos datos enriquecidos son accesibles en tres formas : un portal web (http://www.rechercheisidore.fr/), una API (http://www.rechercheisidore.fr/api) y un acceso unificado (http://www.rechercheisidore.fr/sparql) en una óptica de metadatos abiertos según el formalismo RDF. De hecho "Isidore" promueve el uso de estándares interoperables.

Así, "ISIDORE" es capaz de cosechar corpus y bases de datos en español y en inglés, pero ofrece también enriquecimientos multilingües enlazados entre sí mediante las posibilidades ofrecidas por el linked data. Para lograrlo, "ISIDORE" utiliza las alineaciones de los conceptos entre tesauros y vocabularios disponibles en la web semántica como los Registros de Autoridad y Referencia de Materia de la Biblioteca Nacional de España (http://datos.bne.es/temas) para los datos en español, o bien los encabezamientos de materias del referencial de la Biblioteca del Congreso de EEUU (Library of Congress Subject Headings – LCSH, http://id.loc.gov/authorities/subjects.html) para los datos en inglès. De esta manera, los conceptos de estos dos referenciales mayores están alineados en parte con los conceptos del referencial francès Rameau de la BnF (Biblioteca Nacional Francesa).

Junto con tesauros multilingües ya integrados en "ISIDORE" (como Pactols, Lexvo, GeoEthno, GEMET, etc.), y con el sistema de categorización/clasificación también multilingüe (categorías del sistema francés de archivos abiertos HAL-SHS y del sistema "Calenda" de anuncios de eventos científicos y convocatorias del CLEO-CNRS), "ISIDORE" es capaz de proponer un sistema de enriquecimientos/clasificación en 3 idiomas con la posibilidad de cambiar de idioma durante la búsqueda en la interfaz del portal www.rechercheisidore.fr y de la interfaz para tableta/smartphone (http://m.rechercheisidore.fr/?lang=es).

Esta característica permite al investigador no-francófono de tener acceso a nuevos datos con enriquecimientos, enlaces y clasificaciones en tres idiomas, permitiéndole medir, por ejemplo, el interés de fuentes en idioma francés sugeridas por "ISIDORE" en la interfaz (bien sea en inglés o en español).

De momento, casi 220 000 documentos en español se encuentran en "ISIDORE" y la plataforma contempla cosechar aún más en el futuro.

En paralelo, otros desarrollos que hacen de "ISIDORE" una herramienta cada vez más personalizada, han venido completando sus funcionalidades y abriendo perspectivas. Es el caso del widget "IMoCO", ISIDORE Motor Constructor que permite crear sólo en unos clics, una interfaz de consulta personalizada de los recursos disponibles en la plataforma "ISIDORE" (por ejemplo recursos específicos sobre un tema). Así, "IMoCO" está diseñado para los usuarios que deseen incluir en su sitio Web el buscador "ISIDORE" haciendo una simple copia/pega de un código HTML. Simple y neutral, se adapta a la mayoría de los sitios Web. Además, IMoCO puede ser totalmente personalizado con sus estilos CSS. También el widget multilingue WordPress "ISIDORE suggestions" (https://fr.wordpress.

org/plugins/isidore-suggestions/) permite al usuario de blogs WordPress conseguir sugerencias de documentos presentes en "ISIDORE". Estas sugerencias se hacen basadas en palabras claves asociadas al artículo que el usuario esté leyendo. Es posible afinar su búsqueda, subrayando el contenido del artículo consultado o seleccionando una o varias disciplinas.

Con este poster, quisiéramos mostrar todas las posibilidades que ofrece actualmente "ISIDORE" para el mundo hispánico en CHS, con lo que nos permite también contemplar colaboraciones fructuosas que contribuirán sin duda a alimentar esta plataforma y, al final, a enriquecer las búsquedas de investigadores o estudiantes francófonos de "ISIDORE" que tendrían acceso a más recursos en español, así como las investigaciones de usuarios hispanohablantes y angloparlantes. También tener la oportunidad de presentar este póster en el cuadro del congreso DH en México permitiría intercambiar con usuarios potenciales sobre sus necesidades, y alrededor de los futuros desarrollos de la plataforma.

SSK by example. Make your Arts and Humanities research go standard

Marie Puren
marie.puren@inria.fr
INRIA, France

Laurent Romary
laurent.romary@inria.fr
INRIA, France; Centre Marc Bloch, Germany

Lionel Tadjou
lionel.tadonfouet@inria.fr
INRIA, France

Charles Riondet
charles.riondet@inria.fr
INRIA, France

Dorian Seillier
dorian.seillier@inria.fr
INRIA, France

Arts and Humanities research has to address new challenges raised by the increasing amount of digital sources, contents and tools. New digital practices and protocols, new digital methodologies and services, new software and databases, offer a completely renewed framework for research, and encourage the emergence of a next generation of digitally-aware scholars.

Digital infrastructures, such as PARTHENOS, aim at supporting and accompanying the rise of this new generation of scholars by offering innovative solutions to connect digital tools and contents to Arts and Humanities researchers' needs. PARTHENOS has thus acknowledged the growing importance to develop a data-centered strategy for the management of scientific data (European Commission, 2010), and is currently developing the Standardization Survival Kit ("SSK") to help Arts and Humanities scholars understand the crucial role that proper data modelling and standards have to play in making digital contents sustainable, interoperable and reusable.

Accompanied by a live demo of the website[1], the poster will be composed of three parts: introducing the Standardization Survival Kit or "SSK", using the SSK, customizing the SSK.

Even if it is not obvious that the Arts and Humanities would be well-suited to taking up the technological prerequisites of standardization, it is yet essential that standardization takes a crucial role in the management of Arts and Humanities data. In this framework, this poster will present the Standardization Survival Kit, an overlay platform dedicated to promote a wider use of standards within Arts and Humanities. This comprehensive interface aims at providing documentation and resources concerning standards (especially authoritative references for each standard such as sources, Standard Development Organizations), and at covering three types of activities related to the deployment and use of standards in the Arts and Humanities scholarship: documenting existing standards by providing reference materials, supporting the adoption of standards, and communicating with all Arts and Humanities research communities.

The SSK is designed as a comprehensive interface for guiding Arts and Humanities scholars through all available resources (collected within a dedicated Zotero library[2]), on the basis of reference scenarios identified since the beginning of the project (PARTHENOS, 2016). The interface intends to provide a single entry point for both novice and advanced scholars in the domain of digital methods, so that they can have quick access to the information needed for managing digital content, or applying the appropriate method in a scholarly context. Users will be able to explore the platform according to their needs, thanks to precise research criteria: disciplines, standards, research activities and research objects. The poster will show how an Arts and Humanities scholar can navigate the Standardization Survival Kit website, by taking the example of an actual reference scenario. A live demo of the interface will also accompany the presentation, so that those interested in the poster will be able to search the website according to their needs.

To stress the importance of standards for Arts and Humanities scholarly work, let us take the example of a sociologist who is a novice in digital methods, but who wants to disseminate a collection of field survey data online, so that they could be used by other researchers in the long-term. By browsing in the SSK, she or he will find

1 The beta-version of the website can be found here: https://ssk-application.parthenos.d4science.org/ssk/#/scenarios
2 https://www.zotero.org/groups/427927/parthenos-wp4

a standardized scenario that could be perfectly suited to her or his needs: "Encode and modelize field surveys for their online dissemination". The poster will follow this researcher exploring this reference scenario, and going through its nine steps[3] with the associated resources. Let us take some of the scenario's steps as examples:

- the fourth step "Anonymize" offers a curated and up-to-date list of resources to help the researcher respect ethical practices and adopt proven techniques for anonymizing the collected data.
- the second and sixth steps stress on the importance of using tested standard - such as EAD to "Collect and classify" the data, and TEI to "Transcribe the interviews" -, highlight the importance of proper data modelling before disseminating them, and give access to appropriate resources on the subject.

More advanced users will also be able to edit the scenarios themselves, by submitting new resources or adding new steps. They can also create new scenarios. The SSK scenarios and steps can be easily extended, reused and customized, thanks to their flexible data model in TEI[4]. A dedicated interface in the Standardization Survival Kit will enable users to make suggestions, automatically converted in TEI according to the appropriate schema. The poster will present this interface and the associated functionalities. And for those who will be eager to test it, a live demo will be provided.

References

Romary, L., Banski, P., Bowers, J., Degl'Innocenti, E., Ďurčo, M., Giacomi, R., Illmayer, K., et al. (2017). *Report on Standardization (Draft)*. Technical Report Inria https://hal.inria.fr/hal-01560563 (accessed 27 April 2018).

Romary, L., Degl'Innocenti, E., Illmayer, K., Joffres, A., Kraikamp, E., Larrousse, N., Ogrodniczuk, M., Puren, M., Riondet, C. and Seillier, D. (2016). *Standardization Survival Kit (Draft)*. Research Report Inria https://hal.inria.fr/hal-01513531 (accessed 27 April 2018).

(2018). *SSK: Development of the Standardization Survival Kit*. XSLT ParthenosWP4 https://github.com/ParthenosWP4/SSK (accessed 26 April 2018).

Riding the Wave. How Europe can gain from the rising tide of scientific data, *FOSTER FACILITATE OPEN SCIENCE TRAINING FOR EUROPEAN RESEARCH* https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data (accessed 26 April 2018a).

Standard Survival Kit https://ssk-application.parthenos.d4science.org/ssk/#/ (accessed 26 April 2018b).

3 1. Obtain the informed consent of the participants, 2. Collect and Classify, 3. Select and digitize, 4. Anonymize, 5. Convert into sustainable formats, 6. Transcrive the interviews, 7. Add metadata, 8. Contextualize the research, 9. Disseminate and archive.
4 https://github.com/ParthenosWP4/SSK/spec

Monroe Work Today: Unearthing the Geography of US Lynching Violence

RJ Ramey
rj@findauut.com
Auut Studio, United States of America

MonroeWorkToday.org, launched in November 2016, is a digital history project that synthesizes current historical research on the scope of American lynchings. The website was updated again in October 2017 with additional content digitized from Tuskegee University Archives.

Lynchings in the United States were perpetrated as homegrown acts, not orchestrated regionally in any way. This exhibit focuses on people of color murdered over 100 years in this fashion under the pretext of white supremacy. Yet unlike most academic studies, the project does not compartmentalize by region (e.g. the South or West) or by group (e.g. Mexican-Americans). By contrast, *Monroe Work Today* is the first of its kind to use web technologies to visualize the entirety of these documented events, connecting scholarship about African Americans, Native Peoples, Mexicans, Sicilians and Chinese immigrants across the United States (Carrigan and Webb, 2013) (Frazier, 2015) (Pfaelzer, 2007) (Pfeifer, 2013) (and others). Through four years of work, Auut Studio meticulously created a database and directory in the form of a map, compiling all modern academic research with century-old archives of the Tuskegee Institute. This national map carries the names of 4166 victims of lynchings and nearly 600 other victims of racialized mob violence. The project gives clarity to the sheer extent of the murders.

Previous inquiries into the lynching record have relied on tabulations and statistics, enumerating one tally for each state or county – such as 531 lynchings in Georgia vs. 205 in Kentucky, etc. (Tolnay and Beck, 1995) (Guzman, c.1960) (Pfeifer, 2013). This project, however, transforms the public's interaction with **each** lynching using maps and extensive contextual narrative. Its goal is to spawn a public discussion about the logic of white supremacy.

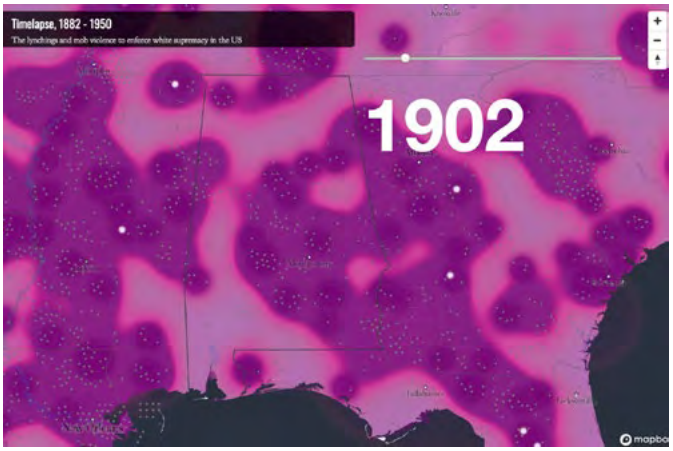
As a second phase to the project, the author now proposes a novel approach to using GIS to understand these murders. Acts of lynching are better examined like other crime data: not as tallies, but rather as incidents with a geographic location. As the commission of overt intimidation over people of color, they were in fact perpetrated with a specific geography in mind. The terrorizing effect was intended to carry over the nearby locale: it enforced the racial order "around **here**." In this context, maps of smaller areas may better recreate the historical truth about lynching, and a geospatial visualization of the regional landscape may better illuminate the original effect of these individual violent acts.

A new computer model created by the author animates regional maps of the USA, weighting the nearby radius

around a murder but also the persistence of its memory over a span of many years. The model makes certain blanket assumptions about the duration of trauma and fear–how long does the grotesque murder of a neighbor dissuade one's actions? These are starting assumptions which the author readily admits may be **wrong**, but they are coded as parameters. This allows different scholars for the first time to test their various interpretations of historical trauma and compare the visual output of competing viewpoints in the model.

This geo-temporal-visual model has the potential to drastically reframe the academic interpretation of lynching by unearthing multiple evolving shapes of the pockets of terror in the historical United States. As a stepping point for future research, this model for the broad reach of real, acute fear could be laid upon a map with other major events in the history of the Jim Crow South and brave acts of popular resistance.

In this poster session, the author will demonstrate the software model to attendees, exchange ideas and suggestions, as well as interrogate on-screen with them several new maps created with the model.



References

Berg, M. (2011). *Popular Justice: A History of Lynching in America*. Chicago: Ivan R. Dee.

Carrigan, W. (2004). *The Making of a Lynching Culture: Violence and Vigilantism in Central Texas 1836-1916*. Urbana: University of Illinois Press.

Carrigan, W. and Webb, C. (2013). *Forgotten Dead: Mob Violence against Mexicans in the United States, 1848-1928*. New York: Oxford University Press.

Frazier, H. (2015). *Lynchings in Kansas, 1850s-1932*. Jefferson, NC: McFarland Publishers.

Frazier, H. (2009). *Lynchings in Missouri, 1803-1981*. Jefferson, NC: McFarland Publishers.

Gonzales-Day, K. (2006). *Lynchings in the West, 1850-1935*. Durham, NC: Duke University Press.

Guzman, J (ed.). (c.1960). Lynching records of Tuskegee Institute as a database typewritten on paper. Tuskegee, AL: Tuskegee University Archives.

Leonard, S. (2002). *Lynching in Colorado, 1859-1919.* Boulder: University Press of Colorado.

Loewen, J. (2005). *Sundown Towns: A Hidden Dimension of American Racism*. New York: New Press.

Newkirk, V. (2009). *Lynching in North Carolina: A History, 1865-1941*. Jefferson, NC: McFarland & Company Inc.

Pfaelzer, J. (2007). *Driven Out: The Forgotten War Against Chinese Americans*. New York: Random House.

Pfeifer, M (ed.). (2013). *Lynching Beyond Dixie: American Mob Violence Outside the South*. University of Illinois Press.

Phillips, P. (2016). *Blood at the Root: A Racial Cleansing in America.* W.W. Norton & Company.

Rushdy, A. (2012). *American Lynching.* New Haven: Yale University Press.

Tolnay, S. and Beck, E.M. (1995). *A Festival of Violence: An Analysis of Southern Lynchings, 1882-1930*. Urbana: University of Illinois Press.

Thompson, V. (2014). *Clinton, Louisiana: Society, Politics, and Race Relations in a Nineteenth-Century Southern Small Town*. Lafayette: University of Louisiana at Lafayette Press.

Educational Bridges: Understanding Conservation Dynamics in the Amazon through The Calha Norte Portal

Hannah Mabel Reardon
hannahmreardon@gmail.com
McGill University, Canada

Calha Norte is the northernmost region of the Brazilian Amazon, and the largest mosaic of protected areas in the world, encompassing nearly 14 million hectares. Given the vastness of this area, government enforcement of parks and conservation zones can be poor, and scarce resources prevent authorities from providing much-needed support to the inhabitants of protected areas. This poster focuses on the Calha Norte Portal, a digital project that constitutes a personal initiative to encourage awareness of conservation efforts in the region. The portal is an educational tool intended to demonstrate the power of digital technologies for fostering greater transparency in conservation management. It also aims to provide a clearer understanding of the social, political, economic and historical dynamics which have shaped the challenges to protecting the Amazon forest today.

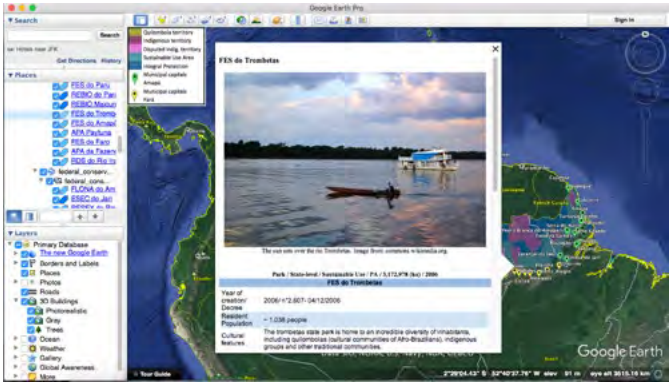
The data for the Calha Norte Portal was gathered during my work with the Social Policy department of the Amazonian Institute for Man and the Environment (Imazon), an environmental NGO based in Belém. In accordance with the department's focus on communities in

Ex. 1&2: A bauxite mine in the Saraca-Taquera national park. Top, a satellite image of the mine in 1986, bottom, the same mine in 2017.

the Calha Norte region, I compiled information from various sources about the region's history, cultural diversity, transportation networks, governing bodies, development indices, demographics, economic activities, protected area implementation, and accessibility. This data was then used to create the Calha Norte Portal, a website and blog with an interactive Google map of the municipal capitals and protected areas in the region. The Google Earth application allows the user to navigate through protected areas, indigenous territories, maroon communities, and municipalities. At a click, each area on the map displays a pop-up window with historical information, demographic statistics, economic and political data, photos, deforestation figures and an implementation index for protected areas. Furthermore, users can look back in time at satellite images from 1960 to the present and visualize patterns of deforestation, and urban sprawl over time.

The project focuses mainly on political, economic, historical, cultural and social data for populations in protected areas and the surrounding municipalities. As an anthropologist, I am particularly interested in dispelling the myth of Amazonia as an uninhabited biological entity, and exposing the important historical dynamics which have shaped the Amazon region as it is today. Understanding the human forces which have pushed the economic development of the region is a crucial first step for conservation policy which can protect both human livelihoods and biodiversity, in line with current sustainable development benchmarks. I also hope to draw attention to the power of digital technologies for overcoming communication barriers between isolated regions and institutional bodies, a major issue in developing informed and tailored conservation policy.

My hope is that, in breaking down the collected data in a visual, interactive format, the uninitiated user will be able to play with the information and learn about the region in any way that suits their interests. The user's guide and tutorials available on the portal offer a guided introduction, but the stand-alone map itself is meant to be played with, manipulated and explored, in ways that dismantle a traditional historical narrative. This poster presentation will elaborate on the features of the Calha Norte Portal and its contribution to greater awareness of regional conservation efforts. The overarching aim is to convey the importance of transparency in the institutionalization of protected areas and to encourage a more thorough understanding of the cultural fabric of the Northern Amazon region, so that research and conservation initiatives might be better tailored to the realities of local communities and their involvement in the protection of the natural resources upon which their livelihoods depend.



Ex.3: Calha Norte in Google Earth. The portal offers users the opportunity to navigate through the online version of the map, or the option to download Google Earth and the Calha Norte KMZ file, for a more complete user experience.

References

Reardon, H. (2018). *Calha Norte Portal.* [Online] Available at: calhanorteportal.com

Building a Community Driven Corpus of Historical Newspapers

Claudia Resch
claudia.resch@oeaw.ac.at
Austrian Academy of Sciences, Austria

Dario Kampkaspar
dario.kampkaspar@oeaw.ac.at
Austrian Academy of Sciences, Austria

Daniela Fasching
daniela.fasching@oeaw.ac.at
Austrian Academy of Sciences, Austria

Vanessa Hannesschläger
vanessa.hannesschlaeger@oeaw.ac.at
Austrian Academy of Sciences, Austria

Daniel Schopper
daniel.schopper@oeaw.ac.at
Austrian Academy of Sciences, Austria

Faced with the challenge of organizing the digital processing and publication of a large collection of historical newspaper data from the 18th century publication known as the *Wien[n]erisches Diarium*, a small project located at the Austrian Centre for Digital Humanities (ACDH) in Vienna has opted for a user-centred, participatory approach and employs methods of community involvement to tackle the specific challenges that arise from the particular qualities of the historical source material.

Founded in 1703, the newspaper under investigation is among the oldest periodical publications still being published today, and was regarded as the most important newspaper of the Habsburg Monarchy for a considerable time span during the 18th century. The value and significance of the newspaper as a source is undeniable, not only due to the density of the information it contains, but also because of the virtually gapless preservation of its run from its foundation in 1703 up until today and the full availability of these original sources. So far, no computer-based processing of this historical data cache has been undertaken. The ACDH project aims at facilitating the use of the source in a digital environment and creating a cornerstone resource, making the *Diarium* freely and easily available to researchers everywhere.

The more than 10.000 issues from the 18th century constitute a mass of text and data. As resources are limited, a number of issues manageable within the project's run had to be selected. For now, the project will thoroughly edit a corpus of approximately 500 issues from all decades of the 18th century. The priority is the quality of the data and the creation of a reliable HTR model that will improve automatic processing and pave the way for expanding or completing the existing corpus at a later point.

As not all queries and research questions that may be posed to the sources can be anticipated, it is the project's primary aim to secure and process the full text of the newspaper in a way that does not disregard or omit any of the relevant information – regardless of the querying researcher's field or discipline. In order to determine which aspects are of particular relevance, where the interests of different disciplinary fields overlap, and how the issues should be prepared and presented to make them useful for the largest number of (academic) users, the digitization project has devised a way to work closely with researchers from various backgrounds.

The project's **community-driven approach** invites and relies on participation on several levels, effectively allowing future users to follow, accompany and shape the project throughout the course of its duration. The following three methods of user involvement were or are being employed in the course of the digitization and annotation process:

1) In spring 2017, a **call for nominations** promoted via digital channels and the print version of the newspaper provided an opportunity for prospective users to nominate specific issues or sets of issues for digitization.
2) While the text recognition process does not involve users, the project team nevertheless upholds the principle of transparency by allowing users to track the progress of the procedure: A **reporting tool** developed for this purpose is accessible via the project website, provides a current list of the issues selected for processing and allows users to track the daily progress in real time.
3) A series of community-driven **annotate-a-thons** allow the project team to survey and adapt to the user community's needs. Consulted as experts and prospective users, (peer) researchers are involved in the annotation process early on and contribute specialised knowledge to the enrichment of the data.

To ensure users' ongoing engagement with the texts even beyond the initial phase and to provide a way to preserve and publicize the results, the platform has been designed with continuous annotation activities in mind. Any user shall be able to make annotations and contribute to the encoding source via the web-app, which will support four basic types of annotations: 1) full text, 2) named entity identification, 3) text or layout corrections, and 4) semantic or structural annotations.

In pioneering a user-centred approach in the development of a digital newspaper resource, the *Diarium* project generates new insights in the potential of community involvement for similar projects. It roadtests methods for motivating both digital and 'traditional' humanities researchers to contribute to a collaborative resource and for creating highly sustainable and re-usable resources

that will meet the needs of diverse user communities, and encourage ongoing engagement.

Expanding Communities of Practice: The Digital Humanities Research Institute Model

Lisa Rhody
lrhody@gc.cuny.edu
CUNY Graduate Center, United States of America

Hannah Aizenmann
haizenmann@gc.cuny.edu
CUNY Graduate Center, United States of America

Kelsey Chatlosh
kchatlosh@gradcenter.cuny.edu
CUNY Graduate Center, United States of America

Kristen Hackett
khackett@gradcenter.cuny.edu
CUNY Graduate Center, United States of America

Jojo Karlin
jojo.karlin@gmail.com
CUNY Graduate Center, United States of America

Javier Otero Peña
javo01@gmail.com
CUNY Graduate Center, United States of America

Rachel Rakov
rrakov@gradcenter.cuny.edu
CUNY Graduate Center, United States of America

Patrick Smyth
patricksmyth001@gmail.com
CUNY Graduate Center, United States of America

Patrick Sweeney
pswee001@gmail.com
CUNY Graduate Center, United States of America

Stephen Zweibel
szweibel@gc.cuny.edu
CUNY Graduate Center, United States of America

In his preface to *Doing Digital Humanities: Practice, Training, Research* (2016), Ray Siemens points out that imagining digital humanities as a community of practice wherein participants come into conversation with one another over shared approaches to craft establishes a "methodological commons" where fields intersect by sharing their work processes. Presenting a taxonomy of approaches to training that span from the informal to the formal within the methodological commons, Siemens suggests that the variety of possible approaches builds an infrastructure for "self-determination" in humanists' approach to learning useful skills. Somewhere between informal consultations and formal degree programs, short courses and "bootcamps" offer professional and research skill development opportunities that scholars can choose from based on their most pressing needs.

Digital humanities skill development cannot be automated; it is resource intensive. It depends upon a limited number of people to deliver highly personalized training to relatively small cohorts of scholars--a model that is difficult to fund and harder to scale. As interest in and demand for training in digital humanities research methods continues to increase, overall capacity to reach the needs and interests of diverse populations of scholars in the wide range of institutional contexts where they do their work has not kept pace.

Committed to building a vibrant community of scholars who deploy a critical use of digital technologies in their teaching and research, the CUNY Graduate Center will run its fourth week-long digital research institute in January 2018. Between 2016 and 2017, GC Digital Initiatives offered a combined 100 hours of instruction on digital research methods to more than 100 students, faculty, staff, and librarians across the CUNY system.[1] Our institute model has focused on reducing the time required to develop new curricula through sharing and versioning, expanding the number of participants per institute through collaborative learning environments, and supporting participants through community-building. The success of our model is demonstrated by continued, growing interest from students, faculty, and staff each year.

As interest in digital humanities at universities, museums, libraries, and archives increases, so too does the demand for faculty, administrative staff, librarians, postdocs and graduate students who are tasked with expanding DH research and teaching capacity with relatively few resources. With funding from the National Endowment for the Humanities, we will be expanding our model to create a sustainable, reproducible model for digital methods training that can be adapted and used in a variety of institutional contexts. Our institute model is designed to integrate feedback so that it can be replicated, modified, and reproduced in new contexts, lowering the barrier to entry for digital humanities scholars by meeting scholars where they are rather than requiring participants to travel to receive training.

In June 2018, 15 individual participants will participate in the first Digital Humanities Research Institute. The DHRI emphasizes foundational technical skills, such as the command line, git, Python, and databases, that provide a flexible technology "stack" and that better enable DH researchers to become more confident autodidacts and mentors in their own right. While participants develop

1 GC Digital Research Institute http://cuny.is/gcdri

familiarity with useful tools, they learn more importantly how to navigate a computer's information architecture, read technical documentation, and reason through simple systems, leading to a greater conceptual vocabulary and increased confidence approaching technology with a critical eye. As participants learn skills to support their individual research goals and professional growth, they will also learn how to lead similar digital humanities institutes in their local communities over the following academic year. Through the process of iterating, refining, and building the institute model, we intend to share the lessons learned to increasingly wider communities of learners and build a network of curricular models and support.

Our poster will feature curricula, pedagogical materials such as datasets, and resources developed for the ten-day residential institute, where participants will explore interdisciplinary digital humanities research and teaching with leading DH scholars, develop core computational research skills through hands-on workshops, and begin developing versions of the DHRI for their own communities. We will share lessons learned and provide information about forthcoming institutes. Short video clips will feature our unique approach to digital humanities pedagogy and interviews with previous institute instructors and participants.

References

Crompton, Constance, Richard J. Lane, and Ray Siemens. *Doing Digital Humanities: Practice, Training, Research*. Routledge, 2016.

Hispanic 18th Connect: una nueva plataforma para la investigación digital en español

Rubria Rocha
rubria@tamu.edu
Texas A&M University, United States of America

Laura Mandell
mandell@tamu.edu
Texas A&M University, United States of America

18thConnect.org es una comunidad en línea de académicos que realizan revisión por pares de materiales digitales obteniendo metadatos de los mismos para colocarlos en nuestro buscador que está disponible de forma gratuita. Los materiales patentados, tales como Early English Books Online (EEBO) y Eighteenth-Century Collections Online (ECCO) también se pueden buscar a través de nuestro asistente de búsqueda. Además, los libros y documentos de las colecciones EEBO y ECCO de la literatura moderna temprana están disponibles en 18thConnect para que los usuarios corrijan sus transcripciones mecánicas, a través de nuestra herramienta TypeWright. Cualquier persona que corrige un documento puede, entonces, tenerlo tanto en formato de texto plano como en XSLT. Exhortamos a los especialistas a corregir textos, crear ediciones digitales en GitHub o enviarlas al TEI Archiving and Publishing Access Service (TAPAS), así como a enviar sus ediciones a 18thConnect para su revisión por pares y para publicarlas en acceso abierto.

Mientras que 18thConnect ha estado en línea desde el 2009, la idea de crear Hispanic 18th Connect, resultó de la necesidad de ayudar en el proyecto *Primeros Libros* de la Texas A & M University, financiado por la NEH, para desarrollar OCR para documentos históricos escritos en español. Dados los resultados positivos en el proyecto *de Primeros Libros* creemos que es momento de extender nuestros recursos para su uso en otras bibliotecas hispanas y para hacer disponibles sus colecciones en nuestro sitio.

Para comenzar este proceso, elegimos traducir y adaptar nuestra interfaz al idioma español y a la cultura hispana. Nuestra justificación es que el español es la segunda lengua materna más hablada en el mundo, así mismo, el 18% de los habitantes en los Estados Unidos habla español y se estima que para el 2060, E.U. sea el segundo país con mayor número de hispanohablantes después de México (Llorente, 2017). Además, la cultura hispana permea en múltiples países, donde también se puede observar una especial motivación por conocer más de esta cultura. Esto último, se ve reflejado en el creciente interés por hacer investigación y desarrollar proyectos relacionados a la lengua y cultura hispanas desde las humanidades digitales (Gutiérrez y Ortega, 2014 y AtlasCS-HD, 2015).

El proyecto de Hispanic 18th Connect consiste en 6 etapas: 1) traducción al español; 2) revisión del funcionamiento de la interfaz; 3) prueba piloto con colegas humanistas cuyos intereses sean en estudios hispánicos con y sin experiencia previa en la interfaz de 18th Connect en inglés; 4) análisis de los resultados de la prueba piloto; 5) adición o modificación de contenidos de acuerdo a las respuestas y comentarios de la prueba piloto; 6) presentación oficial de la interfaz de Hispanic18th Connect.

La presentación del póster de Hispanic 18th Connect tiene varios aspectos a cubrir: por un lado, dar a conocer que la plataforma 18thConnect será más accesible para la comunidad hispanohablante por tener la opción de navegar en su sitio en español; presentar los retos que implicó la traducción de esta plataforma tanto en cuestión de términos, como en relación a los aspectos culturales que creemos pueden impactar (resultados de la etapa 4) y es dónde se pudiera visualizar cómo las características de la comunidad podrían modificar las instrucciones y/o las herramientas con las que cuenta 18thConnect para que pueda ser relevante en el estudio del siglo 18 hispano.

El principal objetivo para este primer momento, es ofrecer esta plataforma traducida al español, y darla

a conocer con el material de las colecciones existentes, y en un segundo momento, que es nuestro objetivo a mediano plazo, incluiremos nuevas colecciones y buscaremos identificar las posibles necesidades de nuevas herramientas que sirvan al estudio del siglo 18 hispano.

Nuestro póster mostrará el trabajo realizado en las 6 etapas, los retos, los hallazgos, los cambios y las diferencias respecto a la interfaz en inglés. Así mismo, contendrá la información más importante de la interfaz y ejemplos que demuestren la manera en que se realizarán las búsquedas de documentos en español (provenientes de EEBO y ECCO), la forma en que se corrigen documentos con TypeWright, y cómo se crean las ediciones digitales para que puedan ser sometidas a revisión por pares en Hispanic 18th Connect.

References

18thConnect Eighteenth-century Scholarship Online (2009). Available at: http://www.18thConnect.org (Accessed 17 November 2017).

AtlasCSHD (2015). Atlas de Ciencias Sociales y Humanidades Digitales. Available at: http://medialab.ugr.es/proyectos/atlas-de-ciencias-sociales-y-humanidades-digitales/ Mapa: http://grinugr.org/mapa/#

Llorente, A. (2017). ¿En qué países se habla español fuera de España y América Latina? *BBC Mundo*. Available at: http://www.bbc.com/mundo/noticias-america-latina-38021392 (Accessed 17 November 2017).

Ortega, É. y Gutiérrez, S. (2014). MapaHD. Una exploración de las Humanidades Digitales en español y portugués. In Romero, E. y Sánchez M. (eds), *Ciencias Sociales y Humanidades Digitales Técnicas, herramientas y experiencias de e-Research e investigación en colaboración*. CAC, Cuadernos Artesanos de Comunicación, pp. 101-128.

TAPAS, TEI Archiving and Publishing Access Service (2014). Available at: http://tapasproject.org/ (Accessed 30 April 2018).

Lorenzetti Digital

Elvis Andrés Rojas Rodríguez
elarojasrod@unal.edu.co
Universidad Nacional de Colombia, Colombia

Jose Nicolas Jaramillo Liévano
jonjaramilloli@unal.edu.co
Universidad Nacional de Colombia, Colombia

Lorenzetti Digital es un proyecto de historia digital e historia pública que pretende mostrar perspectivas de la Edad Media desde Latinoamérica a un público especializado, escolar y no especializado. Esto se hace buscando conexiones en el mundo medieval desde los frescos de Ambrogio Lorenzetti *Le Allegorie del Buono e Cattivo Governo e dei loro Effetti*, pintados en la ciudad de Siena, Italia, en el siglo XIV. A través de los personajes alegóricos del fresco se relacionan fuentes pictóricas y fuentes primarias textuales para reconstruir el contexto histórico de cada personaje.

El proyecto nace como un ejercicio académico estudiantil para las materias **Historia Digital** e **Historia Medieval**, a través de la plataforma wix. En principio solo se proyectaba como una herramienta digital de difusión del conocimiento histórico desde los estudiantes. Sin embargo, ahora las ambiciones del proyecto son más grandes. Lorenzetti Digital se proyecta como una herramienta de comunicación de la historia medieval y, por otro lado, como un repositorio de fuentes primarias medievales. Esto significa que el sitio web resolverá las necesidades de los usuarios brindándoles un primer acercamiento interactivo a la historia medieval, para luego profundizar en diferentes niveles de investigación a través de las fuentes primarias del repositorio. Para esto, implementamos herramientas como el HTML5 en vez de la plataforma wix. Para finales de 2018, se espera que el proyecto tenga un sitio web con dominio propio y que estén consolidados tanto los aspectos didácticos, visuales e investigativos como el repositorio.

Se trata, también, de dar una perspectiva de la historia europea medieval desde Latinoamérica. Consideramos que hay un **mercado** para la historia medieval en Colombia y Latinoamérica en general, propiciado por la industria del entretenimiento y desaprovechado por los historiadores. Por tanto, Lorenzetti Digital es una herramienta para los curiosos, los estudiantes y los investigadores por igual.

Lorenzetti Digital ha participado en las dos últimas conferencias de la International Federation for Public History, donde recibió críticas y comentarios útiles para el proyecto. El vínculo entre las humanidades digitales y la historia digital en este caso radica en la necesidad de aplicar herramientas informáticas como el desarrollo de sitios web para la divulgación del conocimiento histórico. Esto es, enseñar un período de la historia que recibe mucho interés por parte del público. Se trata de un ejercicio de estudiantes de historia para responder a una necesidad de la sociedad, que debe ser tratada de una forma no convencional para el historiador. Es decir, desde formas digitales con contenidos con potencial hipertextual.

Por supuesto, esto presenta muchos desafíos, preguntas y problemas de entrada. Primero, el hecho de que algunos estudiantes y profesores de historia se enfrenten al desarrollo y diseño web es algo para resaltar. Esto radica en un reto de interdisciplinariedad para lograr un desarrollo multimedia equilibrado entre historia, estética y funcionalidad.

Por otro lado, está el problema de la investigación. El ejercicio curatorial y de investigación que hay detrás

es un entramado de conexiones y redes complejas entre el fresco de Lorenzetti, la tradición iconográfica medieval, renacentista y antigua y las fuentes textuales como *La Divina Comedia*, *El Decamerón*, los mitos grecorromanos y los mitos judeocristianos. Lorenzetti Digital también se trata de un ejercicio investigativo nativo digital, donde la mayoría de esas fuentes están disponibles en línea. Sumado a esto nos enfrentamos a un reto de carácter epistemológico y temático. Sitios web sobre historia medieval existen en grandes cantidades con contenidos precisos y de alta rigurosidad, esto implica reconfigurar la estructura y narrativa del sitio web, manteniendo el nivel de asertividad sobre el pasado que se quiere comunicar. Esto a través de una lectura **por capas** o **hipertextual** de los personajes del fresco.

Por último, está el problema de hacer la plataforma más participativa, no solo proveyendo información, datos, interpretaciones y fuentes, sino también recibiendo ideas, comentarios, nuevos trabajos y retroalimentación en general. Así, no intentamos desarrollar un sitio web plano y estático, sino más bien uno dinámico y participativo, digno de la Web 2.0.

Hasta la fecha, no tenemos conocimiento de proyectos similares desarrollados desde Latinoamérica. Sin embargo, sí hay referencias de otras obras de arte que a través de sus personajes narran o explican ciertas ideas. Por eso mismo, el proyecto plantea más retos que soluciones, se trata de explorar el campo de las humanidades digitales, el diálogo con otros saberes fuera de las ciencias sociales y humanas, y de entablar una conversación virtual con el público del proyecto.

References

"An Empirical Framework For Learning (Not a Methodology)". Consultado el 11 de marzo de 2018. http://scrummethodology.com/.

Gentile, Gianni. Luigi Rogna y Anna Rossi. *Multistoria 1. La civilità medievale*. Vicenzo Bona: Editrice La Scuola, 2013

"Exposition Monet 2010 - RMN - Grand Palais - Paris". Consultado el 12 de marzo de 2018. http://www.monet2010.com/.

"Jheronimus Bosch - de Tuin der Lusten". Consultado el 12 de marzo de 2018. https://tuinderlusten-jheronimusbosch.ntr.nl/.

Skinner, Quentin. *El artista y la Filosofía Política.* Madrid: Cambridge University Press: 2009

Traditional Humanities Research and Interactive Mapping: Towards a User-Friendly Story of Two Worlds Collide

Vasileios Routsis
v.routsis@ucl.ac.uk
University College London, United Kingdom

Background

Historians have been using printed maps to illustrate movements of people, trends or any other kind of information for a long time. However, only recently the technological advances made it possible to produce digital interactive environments. Digital Humanities is born out of the need to use computational methods to facilitate humanities research, and data visualisation and digital cartography are two important areas within the Digital Humanities spectrum of research fields.

Objectives

This poster draws on the conclusion of the first phase of the *Mapping the Enlightenment: Intellectual Networks and the Making of Knowledge in the European Periphery* [1] (MtE) project funded by the Research Centre for Humanities in Greece[2]. The major deliverable was the creation of an online interactive mapping tool capable of indexing and visualising data of movements of Greek-speaking scholars during the Enlightenment Era. The first public version of the tool was released in late December 2017.

The project's goal is to enhance users' understanding of the emergence of modern science and technology as the expression of a dynamical geography. Addressing the spatiality of knowledge, it focuses on associating particular cultural traits with specific points on a map, and work on tracking down the various paths and encounters through which such cultural traits and the respective knowledge practices evolved.

By digitising and mapping the original data in a user-friendly way and using the latest modern technology available, the team behind this project hopes to re-emerge existing knowledge out of obscurity and ideally cultivate the ground that can lead to the development of new knowledge around this topic. Two of the major benefits of creating the digital tool include: i) availability/access to information: It is easier to access a website than a printed copy and ii) understanding of information: Interactive visualisation helps users explore and retrieve the information they want easier and in ways that may engage them further.

1 https://mapping-the-enlightenment.org/
2 https://www.rchumanities.gr/en/

The mapping tool

The tool uses a holistic approach to deliver the data with a unified all-in-one interface. Within this framework, there are no separate web pages, and the entirety of the available information is accessible via the tool's dashboard. Communication between the server and the clients is asynchronous. A considerable effort has been put to enrich the user experience by providing flexibility of the interface to improve data comprehension and to accommodate users' diverse navigational preferences and different screen resolutions (see Figure 1: The tool interface with its dashboard sidebar collapsed, and different windows opened at the same time.Figure 1,Figure 2: The tool interface with the sidebar open, the timeline placed at the bottom of the screen and an informational window opened. Figure 2 anFigure 3: In this screenshot, the timeline is contained within the sidebar with various data graphs open at the same time on top of the map.d Figure 3 in Appendix).

The tool is custom-built, and its technical infrastructure supports open-source software. On the server side, Apache, PostgreSQL, PHP, and GeoServer with PostGIS library is used. On the client side, the latest versions of the web standards model HTML5, CSS3, and JavaScript provide a modern and user-friendly user interface. Leaflet.js and D3.js are the main libraries that drive the mapping and visualisation system core. The combination of these technologies gives life to the historical data of the project by combining powerful visualisation components and a data-driven approach to DOM manipulation.

Discussion

Stemming from our own experience developing MtE, the poster intends to discuss and exchange ideas on how modern geohumanities projects can be designed and delivered successfully from their early to final stages. As it is known amongst Digital Humanities scholars, digitisation of information is far from being straight-forward and often involves highly complicated techniques to extract and transform the data to the desired format. Furthermore, as the digital age expands and the underlying technologies change, the problem of digital obsolescence lurks, the situation when a digital resource is no longer supported and readable. There may also be challenges in keeping the necessary balance between offering a simple and user-friendly environment without at the same time compromising the integrity and richness of the original data. In addition, each project may have different needs, peculiarities, and objectives. The discussions that are hoped to be made through this poster aim to lead to an exchange of knowledge from both technical and theoretical perspectives that will help to build better similar digital humanities projects in the future.

Finally, instead of a conclusion, it is worth mentioning that such projects and tools are especially valuable if they contribute in raising the academic and public interest in historical, cultural and societal matters - especially if these engage within a critical discourse. In this context, digital technology is used as a tool and means for these purposes and not as a self-referencing end.

Appendix

Figure 1: The tool interface with its dashboard sidebar collapsed, and different windows opened at the same time.

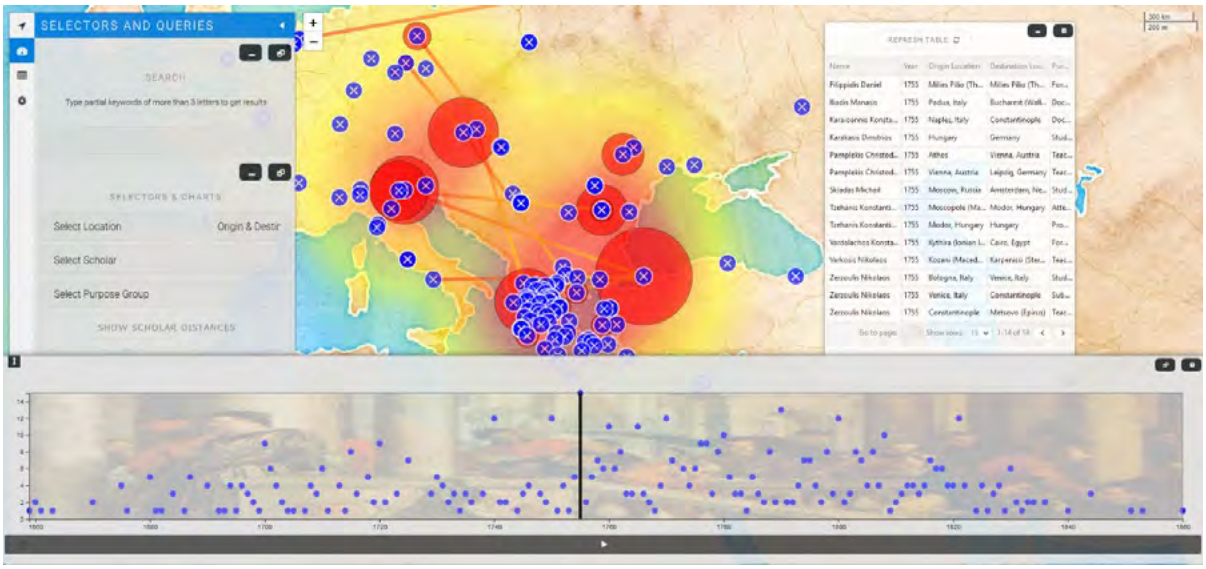


Figure 2: The tool interface with the sidebar open, the timeline placed at the bottom of the screen and an informational window opened.

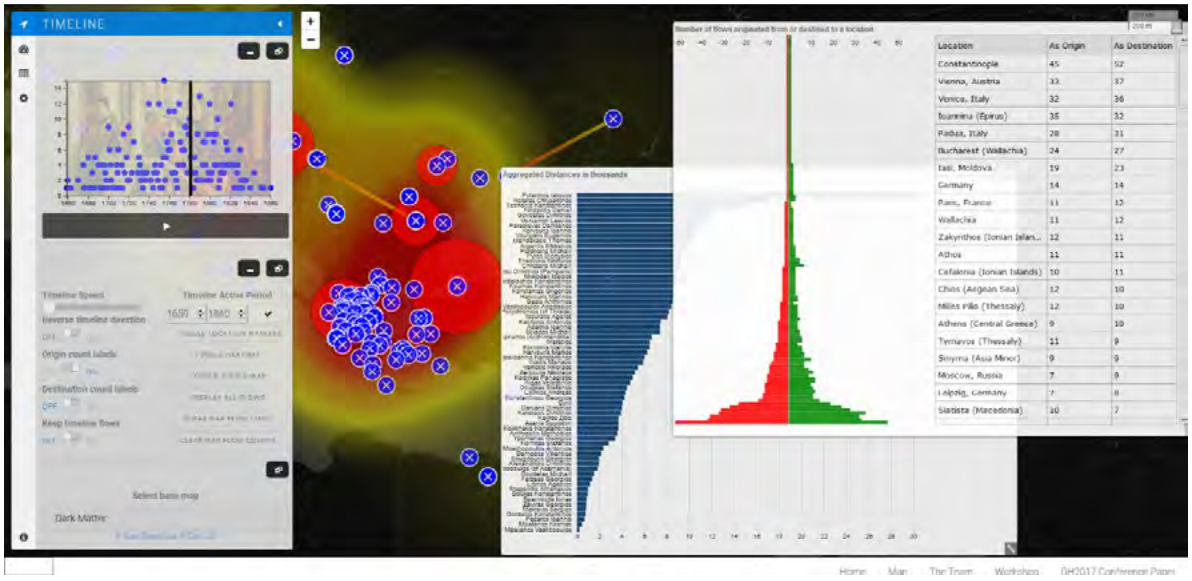


Figure 3: In this screenshot, the timeline is contained within the sidebar with various data graphs open at the same time on top of the map.

Digital Humanities Storytelling Heritage Lab

Mariana Ruiz Gonzalez Renteria
mruizgo1@asu.edu
Arizona State University, United States of America

Angélica Amezcua
aamezcu1@asu.edu
Arizona State University, United States of America

We are proposing to develop a storytelling tool that integrates multimodal mapping for use in language classrooms. Through a Digital Humanities approach on Storytelling Labs, we will be integrating the App StoryMapJS in order to create a storymap of their cultural heritages. This DH tool is very accessible and it will allow the student to engage mapping narrative through images, videos, music, writing and maps; so the heritage learners will interpret space by their personal print, and it will let to other readers from the course or outside the course, to confront other sociopolitical contexts.

The DH Storytelling Heritage Lab will reforge the spatial, and emotional relation from our heritage learners as individuals that can create their own mapping. In a pedagogical perspective the heritage learner will improve their writing, oral, listening and reading skills in Spanish. In a

linguistic research approach we will analyze the outcome of the students, a qualitative discourse analysis.

The workshop will be divided in two sections: the narrative without the DH tool: the student will engaged their narratives through family albums, objects, drawings, and recordings. The second part is to transform the storytelling into a digital narrative with the StoryMapJS. At the end of the Lab the student will have the opportunity to exhibit their narratives maps. The final stories will be compiled in a single web page for their distribution in different areas.

The idea to expand the personal stories and experience in the US of the heritage learners is essential for the course; so the learners engaged Spanish in the sociopolitical context of bilingualism of their own families and community. Their narratives, our narratives, will enrich the course.

Digital Humanities Under Your Fingertips: Tone Perfect as a Pedagogical Tool in Mandarin Chinese Second Language Studies and an Adaptable

Catherine Youngkyung Ryu
ryuc@msu.edu
Michigan State University, United States of America

Learning Chinese, now one of the most widely studied foreign languages in the United States and worldwide, can be challenging, especially for those without any prior exposure to the Chinese tonal system. Mandarin Chinese has four main tones, and one sound carries four different meanings, each tied to a particular tone. For example, "ma" in tone 1 means a "mother"; in tone 2, "hemp"; in tone 3, a "horse"; and in tone 4, a verb to "yell." Chinese as a tonal language thus differs fundamentally from how English speakers often use tone, pitch, and volume to add personal texture to communication. Novice Chinese learners are in great need of sustained and rigorous tonal training with multiple native speakers to develop and sharpen their tonal perception. However, it is usually not feasible to receive such training through in-class or online instruction constrained by time. Digital resources or tools designed for self-guided tone training can help remove such barriers and make tone learning more widely accessible to novice learners in particular.

How does Tone Perfect as a multimodal database render Mandarin Chinese (MC) tone learning accessible?

To create an optimal digital space of learning for each user with different backgrounds, skill sets, and learning styles, Tone Perfect includes by design multiple channels through which the users can synergistically integrate "seeing" and "hearing" into tone learning. Such multiple channels include: (1) a novel color-coded tone visualization (tone 1-yellow; tone 2-green; tone 3-blue; tone 4- red) to enable users to associate the tones with specific colors; (2) a waveform accompanying each sound file to enable the users to see how each of six native speakers produces the same target sound with a particular tone differently, which is also inflected by gender; (3) an additional conventional method of visualizing the tonal information with numbers, so as to aid users with color blindness and to reinforce what the user may have learned through formal instruction; and (4), both simplified and traditional Chinese characters together with a Romanization system (*pinyin*) to enable users to learn tone, sound, and character simultaneously.

How does Tone Perfect maximize its potential as a digital open source?

Tone Perfect is comprised of 9,864 audio assets representing an exhaustive set of monosyllabic sounds in Mandarin Chinese produced by six native speakers (3 female; 3 male). These audio files were produced at MSU to develop a Mandarin tone learning app game, Picky Birds (scheduled to be released in summer 2018). This app game, a digital tool for self-guided tone learning, is an outshoot of a 100% web-based experiment on the efficacy of different methods of visualizing the Mandarin tonal information (i.e., tone-number, tone-pitch contour, tone-color). The app itself was also subsequently utilized as an innovative experiment instrument for another Mandarin tone perception empirical experiment. That is to say, Tone Perfect now serve as an active digital repository that can be accessed by users from various backgrounds for different purposes, for example, as the audio resources for Mandarin Chinese sound tables, computer musical compositions, acoustic analysis, Mandarin linguistics experiments, etc. All audio files can be downloaded directly from the website to enable a wide range of applications of this resource.

Overview

This poster presentation features a multidisciplinary project, Tone Perfect—an interactive audio database—as an example of a multimodal approach to optimizing accessible learning in second language acquisition, specifically for Mandarin tone learning. Tone Perfect also serves as an example of an adaptable multipurpose database that simultaneously functions as an active repository, maximizing the preservation of existing digital materials and amplifying their full potential as digital resources.

Through a hands-on demonstration of how to navigate this database, as well as its metadata structure, this presentation aims to solicit feedback from the audiences

from various backgrounds attending the digital humanities conference. This will enable our team to further enhance the usability of Tone Perfect so as to build an inclusive and accessible space of optimal learning.

References

Godfroid, A., Lin, C., and Ryu, C. (2017). Hearing and seeing tone through color: an efficacy study of web-based, multimodal Chinese tone perception training. *Language Learning*, 76: 819-857.

Grimes, Ryan (2016). With colors and tones, MSU researcher's game gives your brain the tools to learn Mandarin. *The Next Idea*. Michigan Radio. April 14, 2016. http://michiganradio.org/post/colors-and-tones-msu-researcher-s-game-gives-your-brain-tools-learn-mandarin (accessed April 27, 2018).

Ryu, C. and Michigan State University Libraries (2017). Tone Perfect: Multimodal Database for Mandarin Chinese. Michigan State University. East Lansing, Michigan https://tone.lib.msu.edu/ (accessed April 27, 2018).

Codicological Study of pre High Tang Documents from Dunhuang : An Approach using Scientific Analysis Data

Shouji Sakamoto
sakamoto@mac.com
Ryukoku University, Japan / Centre de Recherche sur la Conservation des Collections (CRCC), France

Léon-Bavi Vilmont
leon-bavi.vilmont@mnhn.fr
Centre de Recherche sur la Conservation des Collections (CRCC), France

Yasuhiko Watanabe
watanabe@rins.ryukoku.ac.jp
Ryukoku University, Japan / Centre de Recherche sur la Conservation des Collections (CRCC), France

Dunhuang documents consist of about 40 thousand documents from 5th to 11th century. The documents were discovered in Mogao Cave 17, in Dunhuang, China, by the Daoist monk Wang Yuanlu in 1900. At that time, many foreign explorers visited Central Asia and especially Dunhuang: the Hungarian-British archaeologist Aurel Stein in 1907, followed by the French Sinologist Paul Pelliot in 1908. Both brought back to Europe thousands of documents that they bought from the monk. Although scattered in many countries, the documents are available on the International Dunhuang Project website and the Gallica website of the French National Library; however, except the digital images and bibliographic data, there is no further information on the constituent materials (paper, ink, dyes etc.)

This priceless treasure represents an invaluable resource that led to the creation of a new research field called Dunhuang studies, in order to contribute to a better knowledge of the evolution of paper over 6 centuries. To date the manuscripts, in the 1990s, Prof. Akira Fujieda, a Japanese scholar adopted codicological analysis, focusing both on paper and morphology of the manuscripts, and on the shape of characters written (e.g. Clerical script (□書), Regular script (楷書), etc.). Only preeminent documents were deliberately taken into account by him and thus ignoring many manuscripts. As a result, 5 classes were determined according to the historical periods, that is Northern dynasty (386-581 CE), Sui dynasty (581-618), Early and High Tang dynasty (618-765), 765-786 and Tibetan Empire and Guiyi Circuit (786-1036) (Fujieda, 1999).

In addition to Prof. Fujieda's study, we investigate more details of paper using nondestructive scientific analysis on more than 400 Chinese Dunhuang manuscripts from the Pelliot collection and the Stein collection, and collected various data using a high-resolution digital microscope (Keyence VHX-1000) together with visual checks. Information contained in colophons (date, title of manuscripts) was also collected. A manuscript title is useful for categorization of the manuscripts. Analysis results show differences that can be criteria for differentiate paper. We developed the database (http://www.afc.ryukoku.ac.jp/pelliot/index.html), including scientific analysis data such that microscopic images from the Pelliot collection, as part of new digital archives for old documents.

As we obtained new data by scientific analysis and visual check, we can define new classes, A2, A3, B1 and C1, besides Fujieda's classes, A1, B2 and C2, as follows; A1: Fujieda's class from Northern Wei (北魏 (386-534)) and Western Wei (西魏 (535-556)). Paper is Ma-shi (麻紙) including hemp or ramie with 4~6 laid lines/cm, and with clerical like script of northern dynasty style. On the other hand, A2: paper from southern dynasty, Liang (梁 (502-557)) and Chen (陳 (557-589)), is high quality Ma-shi, and have finer laid lines, 8~9 laid lines/cm, paper width is 49~50 cm and well dyed, and is written sutra with clerical like script of southern dynasty style. Moreover, A3: new class paper from Northern Zhou (北周 (556-581)) is not Ma-shi but Cho-shi (楮紙) including mulberry paper (B. papyrifera, M. alba, etc.), and they have around 6 laid lines/cm, and with clerical like script. B1: new defined class paper from Sui (隋 (581-618)) is Cho-shi with 6~8 laid lines/cm and paper width is narrow, 41~43 cm. Few paper include rice starch. But B2: Fujieda's class from Sui. Paper is Cho-shi with about 6~7 laid lines/cm, paper width is wide, 50~53 cm, and well dyed, and is written sutra with clerical like or regular script of southern dynasty style. C1: paper in this new small class from early and high Tang (初

唐 (618-712), 盛唐 (712-765)) is similar to the ones in B1, that is Cho-shi with 6~8 laid lines/cm and paper width is 37~44 cm. Some paper include rice or millet starch. C2 is also Fujieda's class from early and high Tang. Paper in C2 is Cho-shi and high quality Ma-shi with fine laid line, about 8~10 laid lines/cm, paper width is wider than the ones in C1, 45~51 cm, and well dyed, and is written sutra with regular script.

As mentioned above, scientific analysis data is very useful for Dunhuang studies, for example, the data improved Fujieda's classification. We developed the database, Scientific Analysis of Pelliot Collection, digital archives, including such data

References

Fujieda, A. (1999). *Dunhuang Study and Related Topics*. *Brain Center*, pp.24-56. (in Japanese)

Connecting Gaming Communities and Corporations to their History: The Gen Con Program Database

Matt Shoemaker
mshoemaker@temple.edu
Temple University Libraries, United States of America

2017 saw the 50th anniversary of the Gen Con gaming convention, the oldest and largest continuously running gaming convention in the United States. Started in 1967 as a wargaming convention, Gen Con faced exponential growth following the 1974 creation of Dungeons & Dragons by one of its founders, Gary Gygax. Since then, Gen Con has seen a wealth of change. Evolving from a wargaming convention to a roleplaying game convention, growing to encompass video games and board games and finally reaching its current state of a gaming convention with close ties to popular culture. Aside from the content Gen Con has covered, the convention has also seen fluctuations in the populations that attend the event. All of these factors make Gen Con a prime target for scholarly study in areas of popular culture, games, gender in games studies, and the impact of Dungeons & Dragons. Scholars in media studies, history, material culture and gender studies, to name a few, would all be interested in data related to Gen Con.

Though Gen Con offers a wealth of possibility for scholarship, the information about the convention has largely remained inaccessible to scholars. As a corporate entity, Gen Con LLC, the company that currently runs Gen Con, keeps the majority of their records confidential. One resource that is publically available, however, are the programs from each year of the convention. The quality of the data within the programs varies from year to year, but they generally contain information pertaining to events that were run, who ran them, and descriptions of those events along with other information. Another barrier regarding these programs is that the vast majority of them exist only in physical form, with no digital counterparts. Many of these paper programs are also quite rare, particularly from the conventions that took place in the

1960s and 1970s. An additional resource that is dwindling is those who attended and organized the convention during its early years. Gen Con's most famous founder, Gary Gygax, passed away in 2008. Many of the others involved with the convention from its inception are approaching an advanced age and part of an insular group within gaming culture that few outside of it have approached. These barriers to access have, thus far, limited the scholarship that could be conducted on the Gen Con game convention.

With the above in mind and the 50th anniversary of the convention quickly approaching, we took the opportunity to undertake a project to make resources related to Gen Con more accessible to scholars. The primary work for the first phase of this project took place during 2016 and the first 3 quarters of 2017. We set out to first collect digital and physical copies of all 50 years of Gen Con which we were successful in doing. Second, we converted all event data from these programs into a database of more than 150,000 records which scholars and members of the gaming and Gen Con communities could access online via a Black Light discovery layer. Third, we conducted oral history interviews of several people involved in the history of Gen Con's past and present and transcribed them. Fourth, we conducted some preliminary research using textual analysis and data analysis methods to showcase some of the research that could be conducted using this data and other resources. Finally, we created an Omeka instance and Neatline timeline to both house these resources and make them available for others to use. All of this information can be found at http://best50yearsingaming.com/

We are continuing to conduct research with this dataset and are creating workshops that utilize the dataset in order to educate students in how to use large datasets. We also would like to increase awareness of this open dataset in order to connect more scholars to the resource so they can utilize it in their own research. This project has been able to connect the gaming community, the Gen Con community, and the Gen Con LLC community over a dataset they all have interest in, and we would like to see them connected with more scholars as well. The work we conducted for this project and knowledge of the availability of this dataset is something that attendees of DH2018 would be interested in, particularly those looking for a 20th and 21st century data set suitable for textual and other forms of data analysis, and we hope you will allow us to present it to them.

References

Best 50 Years in Gaming Project Website. http://best-50yearsingaming.com/

Resolving South Asian Orthographic Indeterminacy In Colonial-Era Archives

Amardeep Singh
amsp@lehigh.edu
Lehigh University, United States of America

One of the challenges of doing archival research with respect to colonial-era Indian print archives is orthography. A substantial number of Indian newspapers produced under have now been digitized, and are accessible through services such as Readex's "South Asian Newspapers" archive, the Digital Library of India, the Panjab Digital Library, and others.

Within the English-language archive, the searchability of these archives is limited, in large part due to idiosyncratic choices made by editors and authors in rendering words from South Asian languages in Roman script. Thus, the pioneering feminist doctor whose name is usually rendered as "Rukhmabai" by present-day scholars was quite often represented as "Rukmabai," "Rukmibai" and "Rukhmibai" in English-language newspapers from the British colonial era. The Roman rendering of Bengali-language names such as "Chatterjee" and "Tagore" also have similar indeterminacy (Chatterjee could be rendered in Indian print archives as "Chatterji," "Chaterjee," or "Chattopadhyay"; "Tagore" could be "Thakur").

The orthographic indeterminacies also proliferate beyond how authors' names are rendered; indeed, we see the issue occurring with reference especially to the representation of South Asian vowel forms ("i" vs "ee"; "u" vs. "oo"), aspirated consonants ("d vs" "dh"; "t" vs "th"; "b" vs. "bh"), and labials ("b" vs. "v"). Given that these archives tend to have simple search features that do not feature intelligent spelling correction, searching for topics of historical interest ("sati" or "satee" or "suttee"?) can lead to highly incomplete results.

Finally, orthographic indeterminacy can be an issue within and across South Asian languages themselves. "V" sounds in the Punjabi language, for instance, are frequently pronounced and spelled with "b" or "bh" in Hindi. The "ā" vowel sound common in many north Indian languages is rendered as "p" (that is to say, a soft "o" sound) in Bengali.

A possible solution to the South Asian orthographic indeterminacy problem might be found by appropriating tools developed by digital humanists in Early Modern studies. A team at Newcastle University, led by pioneering DH scholar Hugh Craig, has developed a tool called Corella, which is designed to help resolve orthographic indeterminacies in early modern English corpora (Craig 2010). Here, we propose to use a limited corpus from an existing archive of texts by British authors in India (the Kipling family) as well as a series of Indian authors (the afore-mentioned Rukhmabai as well as several others). We will aim to train Craig's Corella tool to work with Indian languages rather than with early modern orthography. This will allow us to address linguistic indeterminacies in the Roman rendering of Indian languages along the lines of those mentioned above. Can the searchability of these archives be improved via the use of such tools? What are the prospects of training tools such as Corella to work with larger corpora?

References

Hugh Craig, R. Whipp, "Old spellings, new methods: automated procedures for indeterminate linguistic data." *Literary and Linguistic Computing*, Volume 25, Issue 1, 1 April 2010, Pages 37–52

Brâncuși's Metadata: Turning a Graduate Humanities Course Curriculum Digital

Stephen Craig Sturgeon
stephen.sturgeon@bc.edu
Boston College, United States of America

This poster outlines the planning stages for introducing a substantial digital assignment to a paper-based graduate Humanities course and describes techniques for making metadata interesting to graduate students who have never had occasion to give much thought to it. It also details the experience of a librarian co-teaching a graduate seminar, and may provide a basis for reflection on where the particular types of bridges that get built in these activities lead: are they bridges that well-prepared students will take into a competitive job market? Bridges that subject librarians and faculty members will use to traverse a new collaborative environment? Bridges that students will send their scholarly ideas and projects across to a web-based public? Or bridges for university administrators to point to for the comparison of their respective bridges?

A Style Comparative Study of Japanese Pictorial Manuscripts by "Cut, Paste and Share" on IIIF Curation Viewer

Chikahiko Suzuki

ch_suzuki@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan

Akira Takagishi
taka@l.u-tokyo.ac.jp
University of Tokyo, Japan

Asanobu Kitamoto
kitamoto@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems; National Institute of informatics, Japan

Introduction

Today, many institutions provide digital image data for their collections. Easy access to high-quality images not only improves efficiency in art history research but also changes how research is conducted. Our approach to a style-comparative study makes use of this trend with a web-based tool called the "IIIF Curation Viewer," built using IIIF (International Image Interoperability Framework), to change the input and output of research.

We studied pictorial manuscripts called "Emaki," "Eiribon," or "Nara Ehon" (illustrated scrolls and books with calligraphy) from the Edo period in Japan through the IIIF Curation Viewer, then discussed the efficiency and shareability of this approach.

Tools and materials

Composing lists of notable elements from target materials is a fundamental step in style comparison in art history research. The IIIF Curation viewer, developed by the Center for Open Data in the Humanities (CODH), is a useful tool for IIIF-compliant image resources. It has a function called "curation" that creates a list of interesting canvases with metadata. It reduces the effort of using cut and paste for the target material [Figure 1]. The result of cutting and pasting can easily be saved and shared in a JSON format.

Figure 1. Selecting element by mouse drag operation

The "selected thumbnails" function shows a list of 20 curated elements at a time. This function is useful for comparing small details [Figure 2].

Figure 2. Example of the "Selected thumbnail" mode and list of facial expressions

Analysis with the IIIF Curation Viewer

We picked up all facial expressions from four Eiribon and compared lists of the facial expressions using the IIIF Curation Viewer. Comparison suggests that pictures in each Eiribon were painted by different painters, but the same calligrapher wrote the texts. It also suggests that these Eiribon were created by a workgroup of artists.

We further analyzed using the IIIF Curation Viewer by comparing the above-mentioned curation with other Eiribon created by anonymous painters and calligraphers. We found that two anonymous works have the same drawing style as pictures in Asakura's Eiribon [Figure 3].

Figure 3. Comparing facial expressions in Eiribon: Asakura's text (above) and an anonymous work (below)

We found that our approach was useful for both the style comparing process and the sharing process. It is helpful to share pictures as evidence that supports the conclusion of a paper, but many journals did not allow it because of space limitations. Sharing the curation and citing it from the paper can solve this problem. For example, the evidence used in this paper is accessible at the CODH web site, as shown in the reference, so that other researchers can easily verify the results. Curated data can be increasingly promoted, due to its shareablity and reusability, by publishing them in repositories with persistent identifiers.

Conclusion

The IIIF Curation Viewer is important not only for making the entrance process easy through its cut and paste function, but also for making the output process useful through its sharing function. Both insertion and output is useful to art history research, in particular, in Eiribon research. There are many remaining unexamined Eiribon ; each Eiribon has many facial expressions. An easy cut, paste and share tool has been long awaited, and we hope it will enable the creation of a comprehensive facial expression database of Eiribon and Emaki.

We focused on Japanese art in this paper, but we can use this tool for any artwork as long as the images are served in IIIF. For example, we picked up facial expressions from portraits in the Yale Center for British Art and grouped them by century. The increased reusability of research extends possibilities for art historians terms of education and machine learning. Curated data can be reused as training data for machine learning.

Two issues remain for futures study. First, we need to increase IIIF-compliant image services. Especially in Japan, few institutions provide digital images in IIIF. Second, we need an ecosystem for sharing the results of curation, such as correcting metadata, identifier, and a repository for sharing and editing.

References

Center for Open Data in the Humanities (CODH). (2017). IIIF Curation Viewer, http://codh.rois.ac.jp/software/iiif-curation-viewer/ (accessed 20 April 2018a).

Center for Open Data in the Humanities (CODH). (2017). IIIF Global Curation : Facial expression data: British Portraits, http://codh.rois.ac.jp/curation/exhibition/2/index.html.en (accessed 20 April 2018b).

Center for Open Data in the Humanities (CODH). (2017). "Curation" used in this paper (accessed 20 April 2018c). *Daikoku-mai* (Original version provided by National Institute of Japanese Literature, DOI: 10.20730/200006198) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/daikoku.json. *Rashomon* (Original version provided by National Institute of Japanese Literature, DOI: 10.20730/200003096) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/rashou.json. *Tomonaga1/2* (Original version provided by Digital Collection of Keio University Library, ID: 132X@56@2@1) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/tomo.json. *Tomonaga2/2* (Original version provided by Digital Collection of Keio University Library, ID: 132X@56@2@2) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/tomo2.json. *Story of Kumano-Gongen* (Original version provided by Digital Collection of Keio University Library, ID: 11X@31@1) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/kumano.json

Complex Networks of Desire: Fireweed, Fuse, Border/Lines

Felicity Tayler
felicity.tayler@utoronto.ca
University of Toronto, Canada

Tomasz Neugebauer
tomasz.neugebauer@concordia.ca
Concordia University Library

We present ongoing research using data visualization and complex network analysis to historicize the production of three periodicals: *Fireweed, Fuse,* and *Border/Lines.* Computational methods allow for the visualization of metadata describing these magazine issues as a complex network – but what do these visualizations reveal about real social relations involved in the production and circulation of these magazines?

Fireweed, Fuse, and *Border/Lines* emerged between 1976 and 1986 in Toronto, Canada, from a hotbed of lesbian and gay liberation, feminist and cultural race politics, thereby circulating in relation to transnational social, political and cultural movements (Butling and Rudy, Gonosko and Marcellus, Monk, Robertson). Whereas digital art historical scholarship often applies computational methods to the analysis of visual images (Zorich, Manovich), this paper instead applies complex network analysis to bibliographic metadata describing artist-led magazine publishing. We propose that there is a correlation between the magazine as a site of imagined community (as a discursive site where artistic scenes and poetic community are formed) (Allen, 12-17); and the complex networks visualized from metadata describing production teams and content of each printed issue (Knight, Long, Lincoln, Liu).

At this time, we have completed the data gathering stage. Prior to our initiative, *Fireweed* and *Fuse* were not digitized, nor were they comprehensively indexed on digital platforms. A complete data set was created using human cataloguers and a pre-existing metadata schema developed for the e-artexte open repository of publications on contemporary art. *Border/Lines* was previously digitized, and housed in an open journal repository. However, this online collection is not complete, further, it was not possible to extract the metadata from this platform in a consistent format. Contributor names and roles were indexed for each magazine issue (editor, author, translator, etc.). Many of the contributor names and roles

already exist within the e-artexte authority files, and standard indexing protocols were expanded to include roles that are not usually recorded in the metadata (members of editorial committees, designers, typesetters, etc.).

Once indexed in e-artexte, the data became publicly accessible and exportable into various formats, including EPrints XML. A conversion to Graph GML files used Apache Pig Latin scripts (Neugebauer). The resulting Graph GML data was imported into Gephi.

To borrow an expression from Hoyt Long's mapping of literary community, resulting graphs encourage a "sliding back and forth" between the macroscale of the generated graphs and the microscale of the discourse of the artistic and poetic communities represented (316).

A Multi Modal graph (Figure 1) maps relationships between individual magazine issues, contributors (writers, editors, and designers, etc.), artists as subjects of articles, and publishers. Edges were assigned a colour according to magazine title. Node size has been mapped to betweenness centrality, with a filter applied to a range higher than .01.

Lisa Steele and Clive Robertson feature prominently as contributors to *Fuse* magazine, with a high degree of betweenness centrality. This is not a surprise, as both authored multiple articles in the magazine, are founding editors and key figures in the Toronto artistic and activist scenes bridged through the magazine's content (Robertson, Monk). More remarkable is the prominence of Lynne Fernie in the network, best known for later success as the director of documentary films addressing LGBTQ histories. Fernie's high degree of betweenness centrality and position as a connector between the cluster of nodes surrounding *Fuse* magazine and *Fireweed,* provides a bridge between these two magazines as spaces that shared an impulse towards lesbian and feminist liberation. Poet and activist Dionne Brand, who works at the intersection of race and gender, also bridges *Fuse* and *Fireweed*. Cultural policy analyst Jody Berland, and gay activist and environmentalist Alexander Wilson bridge *Fuse* and *Border/Lines*. Feminist cultural historian Rosemary Donegan bridges all three discursive spaces.

A second graph, a Single Mode Contributor Projection will map relationships between individual contributors through their frequency of co-occurence in magazine issues. The graph will be filtered through edge weight, which represents co-occurrence in a minimal number of journal issues. We will colour the graph through community detection on this network of contributor relations using the modularity functionality in Gephi (Blondel et al.).

We anticipate that contributors with a high betweenness centrality will emerge as catalysts for artistic community as it is represented by the discursive spaces of these magazines. Although some of these names may be iconic, "famous" artists and writers, other careers may not have had the same trajectory of visibility. Additional graphs will be generated by publication year to illustrate how the network structure and centrality measures changed over time.

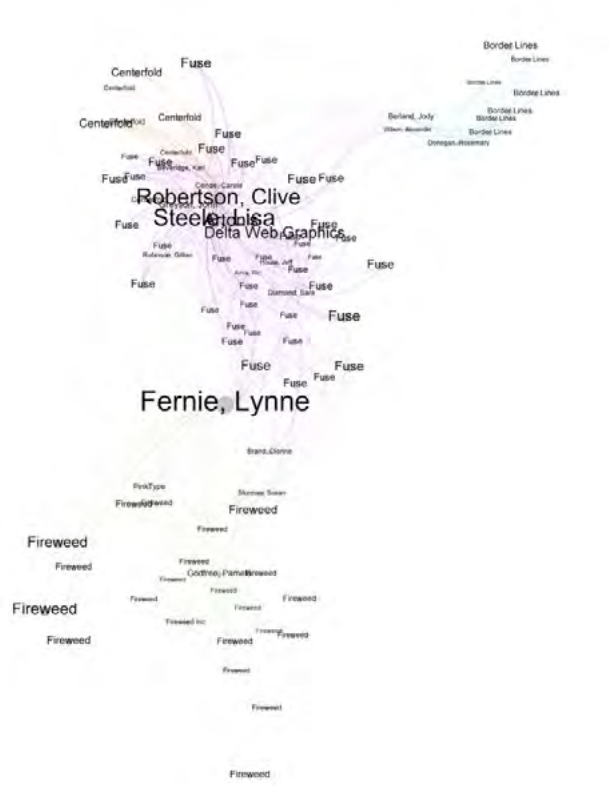


Figure 1. A Multi Modal graph (Figure 1) maps relationships between contributors and issues of magazines with a high degree of betweenness centrality: *Fuse* (purple) previously titled, Centerfold (orange); Fireweed (green), and *Border/Lines* (blue).

References

Allen, G. (2016). *The Magazine.* Cambridge, Mass.: MIT Press.

Butling, P. and Rudy, S. (2005). *Writing in Our Time: Canada's Radical Poetries in English (1957-2003).* Waterloo, Ont.: Wilfred Laurier University Press.

Gonosko, G. and Marcellus, K. (2005). Dead Downtown: Writing the Cultural Obituary of the Alternative Press. *Topia,* 14: 23-35.

Knight, A. R. (2017). Putting them on the map: Mapping the Agents of the Colored Co-operative Publishing Company. https://www.arcgis.com/apps/MapSeries/index.html?appid=665eb933117f4ed-68f0535b4560b5744

Lincoln, M. (2016). "Social Network Centralization Dynamics in Print Production in the Low Countries, 1550-1750" *International Journal of Digital Art History* 2: 134-157.

Long, H. (2015). "Fog and Steel: Mapping Communities of Literary Translation in an Information Age" *The*

Journal of Japanese Studies, 41(2): 281-316. DOI 10.1353/jjs.2015.0062
Liu, A. (2012). "Friending the Humanities Knowledge Base: Exploring Bibliography as Social Network in RoSE." NEH Office of Digital Humanities White Paper. https://rosedocumentation.files.wordpress.com/2012/07/rose-white-paper-as-submitted-to-neh.pdf
Manovich, L. (2015). "Data Science and Digital Art History" *International Journal of Digital Art History* 1:12-34. DOI: 10.11588/dah.2015.1.21631
Neugebauer, T. (2017). "EPrintsData2GML" Eprints Interest Group, 2017 International Conference on Open Repositories. https://github.com/photomedia/EPrintsData2GML
Monk, P. (2016). *Is Toronto Burning*? *Three Years in the Making (and Unmaking) of the Toronto Art Scene.* Toronto: AGYU.
Robertson, C. (2006). *Policy Matters: Administrations of Art and Culture.* Toronto: YYZ Books.
Blondel, V.D. et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10: 10008-10020. DOI 10.1088/1742-5468/2008/10/P10008
Zorich, D. (2012). "Transitioning to a Digital World: Art History, Its Research Centers, and Digital Scholarship," *Kress Foundation.* http://www.kressfoundation.org/news/Article.aspx?id=35338

Locating Place Names at Scale: Using Natural Language Processing to Identify Geographical Information in Text

Lauren Tilton
ltilton@richmond.edu
University of Richmond, United States of America

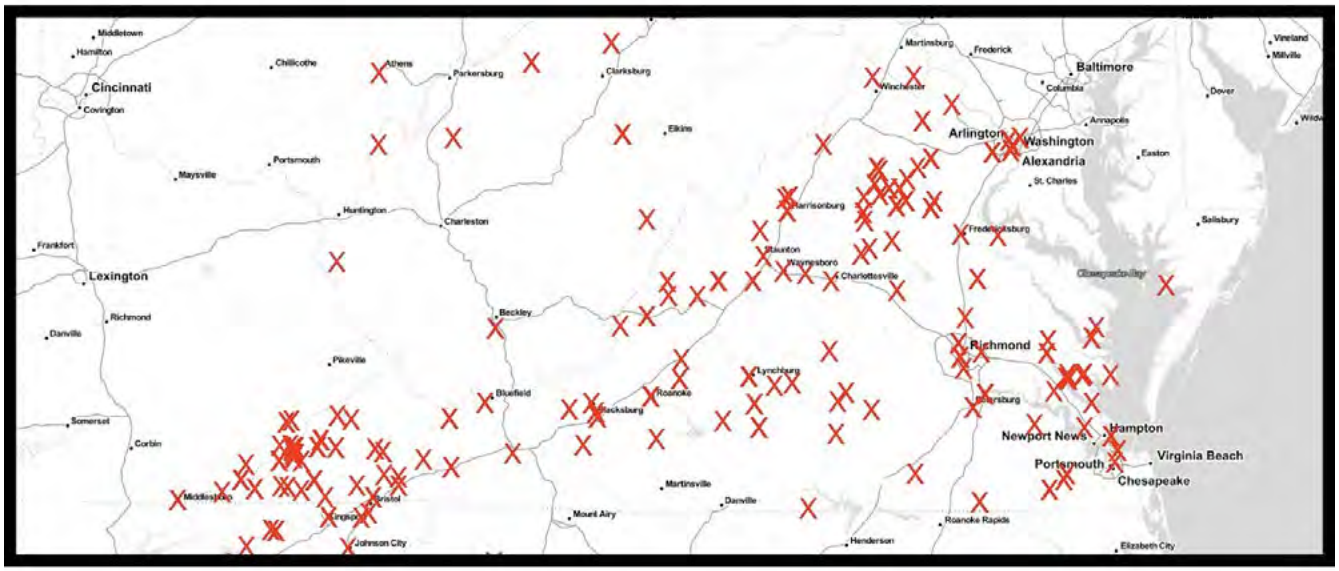
Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Courtney Rivard
crivard@email.unc.edu
University of North Carolina - Chapel Hill, University States of America

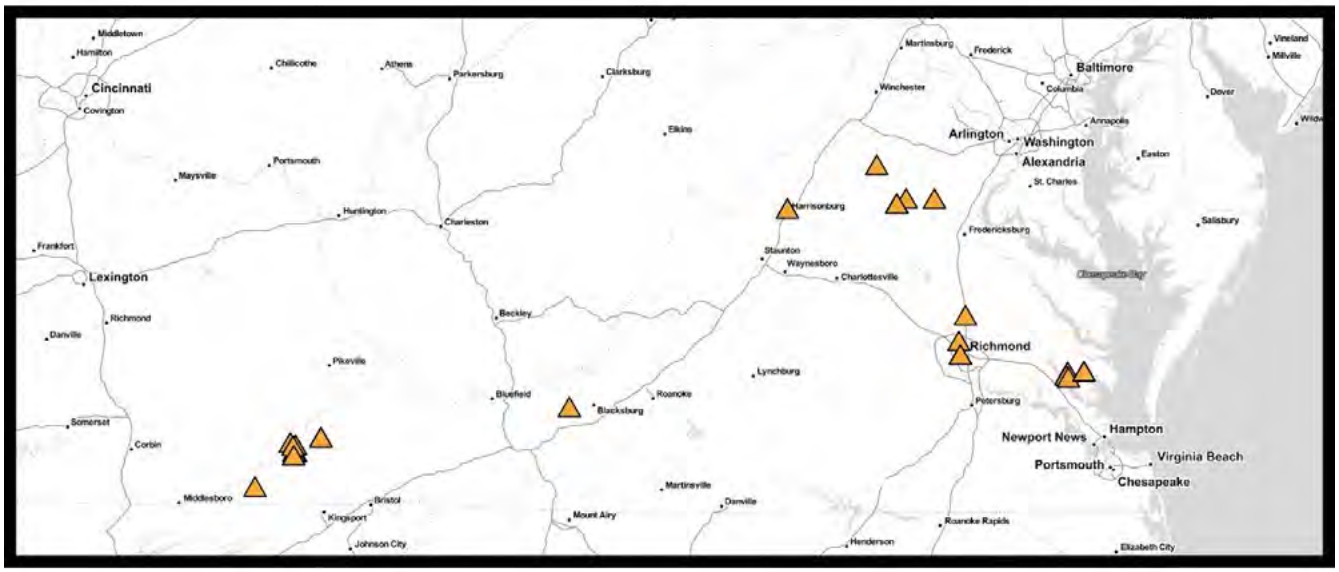
Historical sources are often tagged with metadata about place such as where the object was created, acquired, or stored. Rich latent geographical information is often also mentioned throughout textual documents. A challenge though is how to extract this spatial information at scale. For example, when a text mentions Paris, does the writer mean Paris, Texas, USA or Paris, France? Out of context, most would assume the reference is to more populous capital of France, but it could also be the city in Texas. While close reading would provide an answer, this becomes a challenge when working with hundreds and thousands of documents. How might we be able to more accurately predict the exact location using the broader context?

Our poster „Locating Place Names at Scale: Using Natural Language Processing to Identify Geographical Information in Text" addresses how computational methods can be used to identify and geolocate place-based data. We show how Named Entity Recognition (NER), a natural language processing (NLP) technique, can locate place names using the document's context. We then discuss how to geolocate those places names using a series of computational techniques. Specifically, we start by finding references to specific political divisions (countries, states, and cities), georeferencing them through the Google API. Any political divisions that are uniquely determined become reference points. The reference points are then used to disambiguate terms with multiple results, such as Paris, France and Paris, Texas. Disambiguation is done by appending the political division to the name of the place in order of specificity. If this fails to uniquely determine locations, distances to the closest reference points in the text are used to break ties. This strategy increases proper place name identification and can be applied automatically over a large corpus of digitized texts.

Finally, we turn to an example from our collaborative project on the United States Federal Writers' Project (FWP) entitled Voice of a Nation: Mapping Documentary Expression in New Deal America. During the New Deal, thousands of life histories were written to capture the American experience. While the location of the interviews provides insight into the geographic expanse of the collection (Figure 1), the interviewees consistently spoke about places beyond the location of the physical interview. We apply NER and NLP to identify the place names in the interviews. We are then able to identify and map the many different locations that interviewees mentioned (Figure 2). Across over a thousand interview, what we see is that many of those interviewed spoke of migration - whether their own or their kin - generating a more complex understanding of movement and place during the early 20th century in the United States.



Triangles represent where the metadata identified the interview location in Virginia.



Red "X"s represent locations identified by the use of our algorithm, based on named entity recognition, to the text of the interview referenced in Figure 1

4 Ríos: una construcción transmedia de memoria histórica sobre el conflicto armado en Colombia

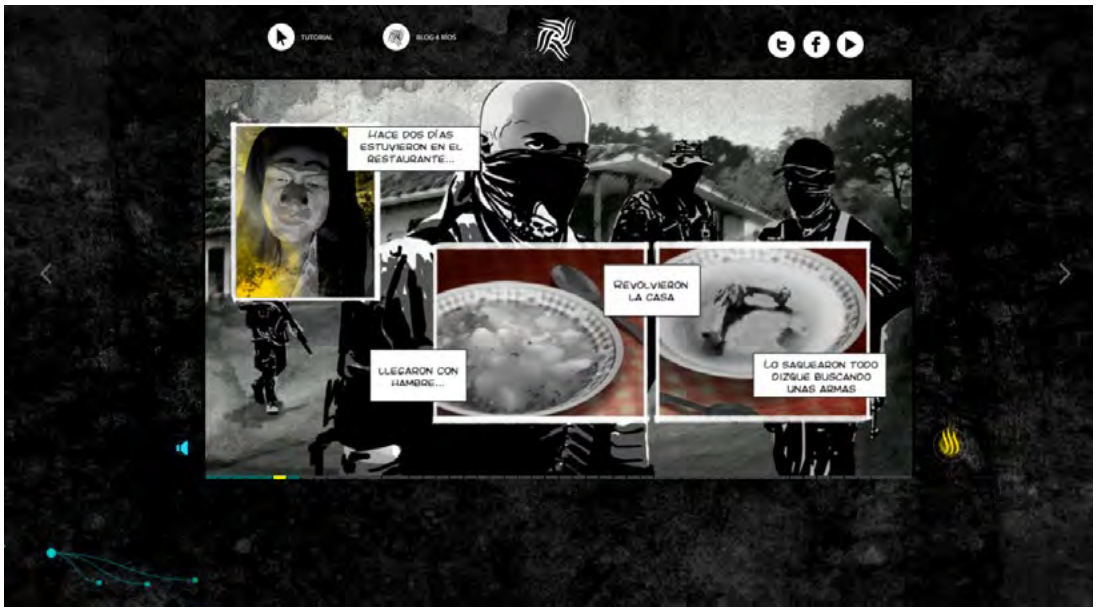
Elder Manuel Tobar Panchoaga
eldertobar@gmail.com
Orgánica Digital; Universidad de los Andes, Colombia

Colombia ha vivido las secuelas del conflicto armado durante más de cinco décadas; el Centro Nacional de Memoria Histórica (CNMH) calcula que existen más de seis millones de desplazados por la violencia, principalmente población civil, campesinos e indígenas.

Desde el inicio del conflicto las víctimas han manifestado de múltiples formas sus sentires, vivencias, recuerdo, testimonios acerca del conflicto y sus consecuencias. Prueba de ello son las más de 177 iniciativas no gubernamentales llevadas a cabo por población víctima de la violencia armada, las cuales giran alrededor de la memoria histórica del conflicto y que fueron registradas en una investigación a fondo realizada por el CNMH (Centro de Memoria Histórica, 2013)


abuelo...
tenemos que irnos
Armas...
...creen que somos Guerrilleros


A
ABUELO...
B1
TENEMOS QUE IRNOS
B2
B3
ARMAS...
(REEN QUE SOMOS GUERRILLEROS


TUTORIAL
BLOG 4 MOS
HACE DOS DÍAS ESTUVIERON EN EL RESTAURANTE...
REVOLVIERON LA CASA
LLEGARON CON HAMBRE...
LO SAQUEARON TODO DIZQUE BUSCANDO UNAS ARMAS

FOGRAFÍA 01

DATOS OFICIALES DE LA MASACRE

Asesinatos y cifras desde la perspectiva de sus protagonistas.

"En los mismos expedientes judiciales se habla de que no hay certeza del número, debido a que varios de los cadáveres fueron arrojados a abismos."[1]

SEGÚN HH[2]

24 o 25
personas asesinadas entre ellas 2 menores de edad y 3 mujeres.

SEGÚN "AUTORIDADES JUDICIALES"[3]

220 Paramilitares

32 Personas asesinadas

10 Personas desaparecidas

SEGÚN JUSTICIA Y PAZ[4]

24 Asesinatos

10 Desapariciones forzadas

67 paramilitares del Bloque Calima fueron condenados a 40 años de cárcel en febrero de 2005. En octubre de 2008, el Consejo de Estado condenó a la Nación por omisión y falla en el servicio, y ordenó indemnizar con seis mil millones de pesos a varias víctimas por el delito de desplazamiento.

500 Hombres del Bloque Calima (Aproximadamente)[5]

PARTE DE VICTORIA DE CARLOS CASTAÑO[6]

"después de combatir durante 72 horas lograron incursionar en el Alto Naya y dar de baja 42 narcoterroristas del ELN y las FARC"

DATOS DE VÍCTIMAS (KITEK KIWE, ONIC, ACNUR, ETC)[7]

400 Paramilitares (Aproximadamente)

más de 40 indígenas, afrodescendientes y campesinos fueron asesinados

60 Personas siguen desaparecidas

más de 1800 Desplazados

COMANDANTES DE LAS AUC

Que participaron en el operativo o la masacre

Elkin Casarrubia, **alias "El Cura"**
Jair Alexander Muñoz Borja, **alias 'Sisas'**
Armando Lugo, **alias 'El Cabezón'**
Luis Fernando Arce Martínez, **alias "Chilapo"**

KITEK KIWE[8]

más de 500 Hombres de las AUC.

3.500 Desplazados o más

Operativo inicia el 6 de abril, el 8 de abril se arman 2 retenes e inicia la masacre con el asesinato de una persona; los asesinatos continúan hasta el 17 de abril.

"... la Fiscalía General de la Nación registró en Abril de 2001 el levantamiento de veintisiete cuerpos en el Alto Naya, y reconoció la existencia de catorce cuerpos más que yacen en fosas comunes en San Antonio, bajo Naya.

...El Cabildo Kitek Kiwe denuncia que en el contexto de la masacre del Naya se han presentado más de cien muertes ocasionadas por los grupos armados ilegales."

Invitación a compartir el contenido para así vivir la experiencia 4 ríos y expanderla a varios públicos. Está en nosotros no olvidar la historia.

COMPARTE ESTE CONTENIDO

1. http://terranova.uniandes.edu.co/noticias/octubre/justicieverdadyreparacionoctubre.pdf
2. http://www.youtube.com/watch?v=qbDy-0BkHDc&list=PL0CB88CC756A4D40A
3. http://www.verdadabierta.com/masacres-seccion/3157-la-fuerza-publica-y-la-masacre-del-naya
4. http://www.verdadabierta.com/masacres-seccion/3187-las-deudas-con-la-comunidad-de-el-naya
5. http://www.wdh.ed.cr/BibliotecaWeb/Varios/Documentos/BD_438003671/Resoluciones-Colombia/Resolucion%20009.htm
6. http://ibs.org.co:81/biblioteca/dwnlds/taq/taq11/taq11-01.pdf (p 33)
7. Rios Aparicio, L., Floresmiro (2001). Caracterización del desplazamiento indígena en el departamento del Cauca. Popayán CNUR, ONIC, RSS. http://servindi.org/pdf/ElCasodelNaya.pdf
8. http://www.humanas.unal.edu.co/colantropos/documentos/Cartilla%20Kitek%20Kiwe%20FINAL%20version%20digital.pdf

El CNMH ha identificado por lo menos tres usos de la memoria dentro de estas acciones comunitarias y sociales, en la primera la memoria es expuesta en búsqueda del esclarecimiento de los hechos sucedidos para exigir justicia por parte del Estado y las instituciones encargadas. En la segunda, la memoria sirve como elemento pedagógico de lo acontecido en búsqueda de la no repetición de estos hechos; mientras que en su tercer uso, la memoria apunta al duelo, a la dimensión reparadora, a proponer 'una oportunidad para restablecer los vínculos sociales y un horizonte para la reconstrucción de lo que se perdió' (Centro de Memoria Histórica, 2013).

El cambio de percepción acerca del conflicto armado a partir de sucesos como la firma del acuerdo de Paz con

la guerrilla de las Farc, la creación del Centro de Memoria Histórica Nacional o la promulgación de la Ley de Víctimas y Restitución de Tierras o Ley 1448 de 2011, ha revitalizado el interés en la construcción y recuperación de las memorias de la violencia, lo que ha repercutido en la producción de múltiples productos artísticos, periodísticos y comunicativos relacionados con este tema.

En ese contexto surge '4 Ríos', un proyecto transmedia que narra historias del conflicto armado en Colombia a través de distintos medios, entre ellos un cómic interactivo, un aplicativo web de memoria además de una exposición interactiva compuesta de maquetas con realidad aumentada.

La primera historia producida está basada en la masacre del Naya, perpetrada durante la Semana Santa del año 2001 en la región del Naya donde alrededor de 300 paramilitares asesinaron a más de 30 personas, lo que provocó el desplazamiento de miles de habitantes de la región.

En el inicio del proceso investigativo, el proyecto se puso en contacto con la población desplazada de la masacre, sin embargo, luego de meses de charlas y reuniones telefónicas, la comunidad manifestó que no estaba interesada en trabajar en nuevos procesos alrededor del tema lo que impidió realizar un trabajo de campo con las víctimas, en cambio las autoridades del cabildo autorizaron el uso de fuentes de archivo donde habían contribuido de forma activa. De esta forma el proceso investigativo se enfocó en la búsqueda, clasificación y curaduría de diversos archivos, investigaciones, tesis, fotografías y noticias de medios públicos.

Una vez establecida la orientación curatorial de la información, se consolidó un equipo de trabajo multidisciplinario integrado por diseñadores gráficos, artistas plásticos, dibujantes de cómic, desarrolladores de software, programadores y animadores que trabajaron en la producción total de todas las plataformas: un cómic interactivo que mezcla una narración ficcional basada en la masacre (disponible en www.4rios.co), acompañado de un aplicativo web que permite a los usuarios dejar mensajes en forma de texto, gráfico o un archivo de audio, además de un corto animado que se complementa con una exposición interactiva compuesta de 3 maquetas que fusionan elementos materiales con animaciones en Realidad Aumentada.

Posterior a su lanzamiento en el año 2016, el proyecto ha interactuado con más de 5.000 usuarios a través de sus distintas plataformas, explorando temas como la narración del conflicto armado a través de Realidad Aumentada. En Internet, el cómic y el Flujo de Memoria han permitido la visualización de historias y documentación además de recibir mensajes, dibujos y audios que reflexionan sobre temas relacionados a las consecuencias del conflicto armado en el territorio nacional.

Así, 4 Ríos busca aportar a la construcción de memoria histórica alrededor del conflicto armado en Colombia, citando a Paloma Aguilar, a través de "una 'memoria prestada' que el sujeto no ha experimentado personalmente, y a la que llega por medio de documentos de diverso tipo" (Aguilar,1996) en donde el trabajo interdisciplinario busca proponer nuevas experiencias que reúnan otras formas de narrar, investigar, crear y construir a través de arte y tecnología.

References

Aguilar Fernández, P. (2008). *Políticas de La Memoria y Memorias de La Política: El Caso Español En Perspectiva Comparada*. Madrid: Alianza Editorial.

Comisión Nacional de Reparación y Reconciliación (Colombia) (ed). (2013). ¡*Basta Ya! Colombia, Memorias de Guerra y Dignidad: Informe General*. Segunda edición corregida. Bogotá: Centro Nacional de Memoria Histórica.

Building a Bridge to Next Generation DH Services in Libraries with a Campus Needs Assessment

Harriett Green
green19@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Eleanor Dickson
dicksone@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Daniel G. Tracy
dtracy@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Sarah Christensen
schrstn@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Melanie Emerson
memerson@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

JoAnn Jacoby
jjacoby@ColoradoCollege.edu
Colorado College, United States of America

This poster reports on a needs assessment for digital humanities library services undertaken at large research university in order to provide a basis for transition to a

next phase of Digital Humanities (DH) support at a library supporting a growing amount of DH work on campus. It reports key findings and how the library services will evolve to meet needs identified on campus.

A recent survey tallied over ninety research centers and initiatives around the world that support DH research, and the majority are associated with university campuses. The recent ARL report *Supporting Digital Scholarship* (Mulligan 2016) observed the trend for digital scholarship support to be centered in a single department, sometimes in a dedicated digital scholarship center, but with support for digital scholarship extending throughout the library. Despite the growing number of DH initiatives and support models for digital scholarship at institutions of higher education around the U.S. and world, few have conducted formal needs assessments on their campuses to ascertain the needs of researchers and other stakeholders. The professional literature that provides a strong guiding framework for this study includes the report on the University of Colorado's recent digital humanities needs assessment (Lindquist et al., 2013) and the Ithaka S+R Sustaining Digital Humanities study and Implementation Toolkit (Maron and Pickle, 2014).

The members of the working group conducting this needs assessment sought to use the study to provide a bridge to the next generation of DH services in the library. The timing was opportune, as there were several features of the library and campus environment at that moment made this a good time to asses DH needs. First, the library's digital scholarship public service space had entered its fifth year and had begun planning to move to a new, larger, and more visible space. Researchers' and instructors' interest in DH collaborations with the library had steadily grown since the foundation of DH services in 2010. The library had grown support for digital scholarship and communication in recent years and, like many peer institutions, sought to increase capacity by involving more librarians in DH services. All of this planning required updated knowledge of campus DH activity in order to evolve services appropriately.

For the first phase of the study, two members of the team conducted a total of 15 interviews with faculty, administrators, academic professionals, and graduate students from multiple colleges and campus units with interest or active involvement in digital humanities research and teaching. The group also reviewed recent dissertations across a range of arts, humanities, and humanistic social science fields to identify recent DH related work and the advisors for those projects.

From the interview responses, the working group developed a survey protocol for the second stage of the study. The group administered a survey that was sent to a random sample of 5% of faculty and graduate students from the colleges and units of Liberal Arts and Sciences, Fine and Applied Arts, College of Media, and School of Information Sciences; as well as targeted sampling of known practitioners of digital scholarship on campus. The survey was open for two months from November 2016 through early January 2017, and gathered 55 responses.

The group identified several areas of need expressed by researchers. These included access to collections and data, funding, networks of research and community, education, and infrastructure and research support. The study showed some differences between needs of graduate students and researchers. For example, graduate students saw a greater urgency around library support for tools and software. Faculty and staff saw greater urgency across all other areas including access to library expertise, assistance with access to digital content, and data storage. Access to digital collections as data appeared as a key barrier to researchers pursuing projects.

Based on these needs, the group developed six broad recommendations for library services: (1) provide opportunities for in-depth training; (2) connect the library's role in research data curation to digital scholarship creation; (3) expand the library's strengths in discovery and access to digital collections; (4) build space and opportunities for people to form communities of practice, (5) act as a key node in the network of digital scholarship research initiatives, and (6) build library personnel capacity for digital scholarship services. Each of these recommendations had specific associated action items.

This poster will provide an opportunity to discuss these findings, the steps being taken by the library to accomplish the goals identified, and the general landscape of next generation DH services in libraries.

References

Lindquist, T., et al. (2013). *dh+CU: Future Directions for Digital Humanities at CU Boulder*. Boulder, CO: University of Colorado, University Libraries Digital Humanities Task Force. http://scholar.colorado.edu/libr_facpapers/32/

Maron, N., and Pickle, S. (2014). *Sustaining the Digital Humanities: Host Institution Support beyond the Start-Up Phase*. Ithaka S+R. https://doi.org/10.18665/sr.22548

Mulligan, R. (2016). *Supporting Digital Scholarship*. SPEC Kit 350. Washington, DC: Association of Research Libraries, May 2016. https://doi.org/10.29242/spec.350

Chromatic Structure and Family Resemblance in Large Art Collections — Exemplary Quantification and Visualizations

Loan T Tran
lxt110930@utdallas.edu
The University of Texas at Dallas, Richardson, TX, United States of America

Kelly Park
kelly.park@utdallas.edu
The University of Texas at Dallas, Richardson, TX, United States of America

Poshen Lee
sephonlee@gmail.com
The University of Washington, Seattle, WA, United States of America

Jevin West
jevinw@uw.edu
The University of Washington, Seattle, WA, United States of America

Maximilian Schich
maximilian.schich@utdallas.edu
The University of Texas at Dallas, Richardson, TX, United States of America

Computational pattern recognition has made ground-breaking progress in recent years by combining advanced methods of machine learning with ever increasing amounts of visual data. Algorithms that learn to learn, combined with massive parallel computation in so-called GPU clusters, and billions of images a day acquired via sensors, or uploaded by Web users, have led to a situation where computers are able to recognize faces, spot cats in any body-configuration, and even drive cars without human interaction. In Art History such advanced methods of pattern recognition increasingly aim to compete with human connoisseurship. Relevant studies, for example, successfully identify duplicate photos in image archives (Resig, 2013), quickly find artworks given a certain object (Crowley and Zisserman, 2014), quantify the innovativeness of paintings (Elgammal and Saleh, 2015), convincingly discern and date architectural styles at a mega-city scale (Lee et al., 2015), and track the evolution of color contrast in Western Art from *chiaroscuro* to landscape painting (Kim et al., 2014 and Lee et al., 2017). What is missing is a rigorous reconciliation between state-of-the-art computer science techniques and established art historical standards based on trained observation and hermeneutic interpretation. Such a reconciliation is hard due to both the so-called "curse of dimensionality" in machine learning, and the cognitive limit of individual researchers confronted with potentially millions of images.

Our project aims to work towards a reconciliation of the computational and hermeneutic perspectives via two pathways. First, through visualizing the chromatic structure of paintings up to entire collections by consistently sorting color pixels, we uncover hidden color patterns of individual paintings, artist oeuvres, periods, and museum collections. Here, we also deal with a well-known multidimensional phenomenon, i.e. color, which could be a starting point to deal with hidden dimensions in machine learning using a traditional hermeneutic approach. Second, using cutting-edge deep learning algorithms and dimension reduction techniques that reduce the high dimensions of the machine learning results to a human-digestible level, we calculate visual family resemblance, generate a variety of clustering possibilities, and produce different visualizations. Combining both pathways, while performing these analyses on three different art collections, we will be able to evaluate the machine learning results, from both an art historian's and a computer scientist's perspective, in a manner that is understandable by a broad audience.

We work with three datasets: the Dallas Museum of Art, a "universal" art collection, circa 18,000 artworks; the Barrett collection, a comprehensive private collection of Swiss art, circa 400 paintings; and WikiArt, an encyclopedic online collection of circa 75,000 paintings. The DMA data is particularly strong in its six-thousand-year coverage, well in line with the exponential growth of world population and cultural production. The Swiss art collection, including high resolution images taken under controlled lighting conditions, is strong in its topical coherence. The WikiArt dataset, though subjects to shortcomings in lighting conditions and temporal coverage, is widely used as a de facto benchmark among machine learning community, and is therefore used for comparative analysis with the other collections.

(**A**) Color palette of Claude Monet at DMA

(**B**) Color palette of Piet Mondrian at DMA

(**C**) Color palette of Georgia O'Keefe at DMA

(**D**) Color palette of Mark Rothko at DMA

(**E**) Color palette of Gerhard Richter at DMA

Fig. 1. Colors in the oeuvre of individual artists. Colors are consistently sorted by luminosity, indicating color frequency (number of pixels), equivalent to area coverage. The strips for (**A**) Monet, (**B**) Mondrian, (**C**) O'Keefe, (**D**) Rothko, and (**E**) Richter reveal striking individual differences between artists. Similar sense-making comparison can be used to differentiate collection coverage as well as canonicity of artists, departments, and other sub-selections of works across museums.

References

Crowley, Elliot J., and Andrew Zisserman. (2014). In Search of Art. *Workshop at the European Conference on Computer Vision*. Springer International Publishing, pp. 54-70. https://www.robots.ox.ac.uk/~vgg/publications/2014/Crowley14a/crowley14a.pdf

Elgammal, Ahmed, and Babak Saleh. (2015). Quantifying Creativity in Art Networks. *arXiv preprint arXiv:1506.00711*. http://arxiv.org/abs/1506.00711

Kim, Daniel, Seung-Woo Son, and Hawoong Jeong. (2014) Large Scale Quantitative Analysis of Painting Arts. *Scientific Reports* 4: 7370. https://www.nature.com/articles/srep07370

Lee, Byunghwee, Daniel Kim, Hawoong Jeong, Seunghye Sun, and Juyong Park. (2017). Understanding the Historic Emergence of Diversity in Painting via Color Contrast. *arXiv preprint arXiv:1701.07164*. https://arxiv.org/pdf/1701.07164.pdf

Lee, Stefan, Nicolas Maisonneuve, David Crandall, Alexei A. Efros, and Josef Sivic. (2015). Linking Past to Present: Discovering Style in Two Centuries of Architecture. *IEEE International Conference on Computational Photography*. http://dx.doi.org/10.1109/ICCPHOT.2015.7168368

Resig, John. (2013). Using Computer Vision to Increase the Research Potential of Photo Archives. *Journal of Digital Humanities* 3: 3-2. http://journalofdigitalhumanities.org/wp-content/uploads/2014/07/Using-Computer-Vision-to-Increase-the-Rese-John-Resig.pdf

Ethical Constraints in Digital Humanities and Computational Social Science

Anagha Uppal
auppal@vols.utk.edu
University of Tennessee, United States of America

As it developed, the field of Digital Humanities has had a particular set of advantages in making advancements and gaining approval among the scientific community, allowing it to serve as a "means to revitalize the humanities" in the face of decreased funding and appreciation for its contributions (Reid 2011, pp. 352-353). Both for Digital Humanities and Computational Social Science, principal among these advantages are:

- Easy and fast access, via the Internet, to data resources and databases.
- Inexpensive computational power, including large amounts of inexpensive memory and physical storage.
- New forms of data (especially text) that can be easily obtained from many sources, particularly social media and blogs.

- Open-source software and a culture of code-sharing
- Modern advocacy and acceptance of interdisciplinary and multidisciplinary research (Alvarez 2016, pp. 3-4)

Watts (2013, p. 7) adds to this list a shorter timescale and lower cost for experiments in theory.

But alongside these advantages come challenges in the use of such data and methods that, if ignored, have the capacity to harm the public and the advancement of knowledge. From the perspective of the researcher, the necessary combination of tools and applications required, often from "multiple research traditions," are not all familiar to any individual researcher (Watts, 2013, pp. 5-6). Data acquisition is becoming more and more difficult, with much proprietary big data (such as the Social Security Administration database or IRS database that would be useful for the study of job networks and the economy) locked away and expensive. Data, once made available, is also messy, unreliable and easily falsified. In order to be usable, it must be grounded with offline findings or other web data. When decentralized online data is found to be false, there is no system of institutional accountability, further increasing uncertainty and eroding trust in the use of the web to crowdsource the production of data and knowledge (Conte et. al, 2012, p. 336). Additionally, now that the use of social network sites is becoming more common, users become more adept at toggling privacy controls and choosing which content to share publicly and which to keep hidden, and the availability of social media data decreases (Giglietto & Rossi, 2012, p. 25).

For study participants, the concerns of weight particularly relate to data acquisition, and its privacy and confidentiality, security and reliability. As social media data is extensively used in DH studies, we demarcate the line at which it is appropriate to use such information without users' consent by confronting extant questions of public/private arenas of publishing and accountholder motivation. Although it is important to retain the approval of users and collect private data ethically, failure to do so has its most damaging consequences when those who have access once it is collected are able to identify users and withdraw participants' privacy, and therefore, we discuss individual-level data and ways to retain people's confidentiality.

We also review ways of benefiting from data that comes from online sources, despite its inherent exclusion of those of low income and low socioeconomic status throughout much of the world, including the U.S. Also excluded are independent researchers, students and those associated with small organizations – especially interdisciplinarians – conducting this work often requires special supercomputers, and many humanities researchers do not have access to such resources or the skillset to use them. A number of papers have been written about data use ethics in other fields of research. This paper attempts to review and combine these needs for the specific purposes of Digital Humanities and Computational Social Science. Through an extended literature review, it collects ethical questions surrounding data use, and applies them to two infamous case studies: that of AOL's release of search data in 2006 and of Facebook's emotional contagion study published in 2014.

It is feasible to imagine that computational advantages, and the promise of DH and CSS, lead to a world of the analysis of not only text, but also sound, images and video, of richly-visualized data so that a maximum number of people can overcome confirmation bias and understand complex research results and contribute, and large-scale undertaking of crowd-sourced data and sophisticated citizen science is commonplace enough to allow us to solve high-impact questions. As we move towards such a world, a periodic reconsideration of ethics is judicious; it remains ever a timely topic with violations resulting in vast scandals and increasing public distrust (most recently the bout of data breaches, such as Uber's - Shaban, 2017).

References

Alvarez, R. M. (2016b). Introduction. In R. M. Alvarez (Ed.), *Computational Social Science: Discovery and Prediction* (pp. 1-24): Cambridge University Press.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Helbing, D. (2012). Manifesto of computational social science. *European Physical Journal-Special Topics, 214*(1), 325-346. doi:10.1140/epjst/e2012-01697-8

Giglietto, F., & Rossi, L. (2012). Ethics and Interdisciplinarity in Computational Social Science. *Methodological Innovations Online, 7*(1), 25-36. doi:10.4256/mio.2012.003

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (1 ed., Vol. 1, pp. 460-475): University of Minnesota Press.

Reid, A. (2011). Graduate Education and the Ethics of the Digital Humanities. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (1 ed., Vol. 1, pp. 350-367): University of Minnesota Press.

Shaban, H. (2017). *Uber is sued over massive data breach after paying hackers to keep quiet*. *The Washington Post*. Retrieved 28 November 2017, from https://www.washingtonpost.com/news/the-switch/wp/2017/11/24/uber-is-sued-over-massive-data-breach-after-paying-hackers-to-keep-quiet/

Watts, D. J. (2013). Computational Social Science: Exciting Progress and Future Directions. *The Bridge: Linking Engineering and Society, 43*(4), 5-10.

Bridging the Gap: Digital Humanities and the Arabic-Islamic Corpus

Dafne Erica van Kuppevelt
d.vankuppevelt@esciencecenter.nl
Netherlands eScience Center, The Netherlands

E.G. Patrick Bos
p.bos@esciencecenter.nl
Netherlands eScience Center, The Netherlands

A. Melle Lyklema
a.m.lyklema@uu.nl
Utrecht University, The Netherlands

Umar Ryad
amr.ryad@kuleuven.be
University of Leuven, Belgium

Christian R. Lange
C.R.Lange@uu.nl
Utrecht University, The Netherlands

Janneke van der Zwaan
j.vanderzwaan@esciencecenter.nl
Netherlands eScience Center, The Netherlands

Introduction

Despite some pioneering efforts in recent times, the *longue durée* analysis of conceptual history in the Islamic world remains largely unexplored. Researchers of Islamic intellectual history still tend to study a certain canon of texts, made available by previous Western researchers of the Islamic world largely based on considerations of the relevance of these texts for Western theories, concepts and ideas. Indigenous conceptual developments and innovations are therefore insufficiently understood, particularly as concerns the transition from premodern to modern thought in Islam.

What, then, are the silenced continuities, transformations and major fault lines in Arabic-Islamic discourses? The Islamic tradition offers a vast textual corpus for exploring this question from a *longue durée* perspective, but its very breadth poses substantial problems for the individual scholar seeking to survey the literature by traditional methods. In the last decade, vast collections of digitized classical Arabic texts have become available online (Muhanna 2016, pp. 11-64). This marks the "beginning of what could become a methodological revolution in the fields of Arabic and Islamic Studies", as noted by Peralta and Verkinderen in the very first edited volume on Digital Humanities and the Arabic-Islamic corpus (Muhanna 2016, pp. 199).

This paper presents ongoing research to use state-of-the art Digital Humanities approaches and technologies to make pioneering forays into the vast corpus of digitized Arabic. This is done along the lines of three case studies, each of which examines a separate genre of Arabic and Islamic literary history.

Case studies

(1) Islamic law: This case study analyzes the corpus of digitally available (Sunni) legal works (furu' al-fiqh) from premodern to modern times (ca. 150 digitized works with ca. 75 million words, extracted from the OpenITI corpus[1] to investigate *longue durée* shifts in concepts and idioms employed in Muslim juridical discourse. The scholarly questions pursued relate to the history of the senses and of sense perception in the Islamic world, and of the human body more broadly speaking. Digital humanities methods applied to this corpus will include topic modelling (around the five senses) and computer-supported statistical analysis in historical perspective, that is, by comparing legal teachings throughout the fourteen centuries of Islamic law.

(2) Modern Islamic proselytizing literature: This case study analyses a largely neglected corpus of Arabic texts written between the 19th and 21st centuries (approx. 500 titles) on Islamic missionary activities (da'wa). The focus of the analysis will be to identify continuities and changes regarding the key concept of da'wa and the discursive idioms used to express them, and identify, graph and visualize the transnational networks involved with the discourses on da'wa.

(3) Arabic poetry: This case study will investigate the digital corpus of Arabic poetry (estimated 2,5 billion words, extracted from the OpenITI corpus). Poetry is an especially apt corpus to study the history of the senses and of sense perception in the Islamic world. What senses were favored by Arabic poets over the course of centuries? What kind of semantic fields are constructed in Arabic poetry around, for example, the sense of vision, and how does this contrast with, for example, legal constructions of vision?

Method

Most of the research projects in Digital Humanities have focused on Western Europe and the Americas, leaving a gap between state-of-the-art Digital Humanities tools and the Arabic text corpus. Many current initiatives in Arabic Digital Humanities seek to teach programming languages to humanities scholars. We pursue a different strategy to move Arabic Digital Humanities forward, by developing a freely accessible, user friendly interface to Digital Humanities technology, based on existing software.

The development of the technology is at an early stage, and we aim to present a first version of an Arabic-specific Digital Humanities toolkit at the conference.

1 Romanov, M, OpenITI. http://alraqmiyyat.github.io/OpenITI/

The toolkit integrates existing tools for stemming and morphological analysis in Arabic, as such as the Khoja stemmer (Khoja, Garside and Knowles, 2001), Tashaphyne stemmer[2] and the AlKhalil morphological analyzer (Boudchiche *et al.*, 2017). We will use the SAFAR software (Jaafar and Bouzoubaa, 2015) to compare these libraries and integrate the most relevant tools in a pipeline for humanities research. The resulting tagged datasets will be made available in an existing search engine, such as BlackLab[3]. All software developed for this paper is published open source[4].

We will present the development of the Arabic-specific Digital Humanities toolkit, including challenges that emerge from developing text mining tools specific for Arabic, with proposed solutions. It will also present early findings from the three case studies.

References

Boudchiche, M., Mazroui, A., Ould Abdallahi Ould Bebah, M., Lakhouaja, A. and Boudlal, A. (2017) 'AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer', *Journal of King Saud University - Computer and Information Sciences*. Elsevier, 29(2), pp. 141–146. doi: 10.1016/J.JKSUCI.2016.05.002.

Jaafar, Y. and Bouzoubaa, K. (2015) 'Arabic Natural Language Processing from Software Engineering to Complex Pipeline', in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. IEEE, pp. 29–36. doi: 10.1109/ACLing.2015.11.

Khoja, S., Garside, R. and Knowles, G. (2001) 'An Arabic tagset for the morphosyntactic tagging of Arabic', in *Corpus Linguistics*. Lancaster University. Available at: http://eprints.lancs.ac.uk/11985/ (Accessed: 24 April 2018).

Muhanna, E. (ed.) (2016) *The Digital Humanities and Islamic & Middle East Studies*. Berlin, Boston: De Gruyter. doi: 10.1515/9783110376517.

Off-line sStrategies for On-line Publications: Preparing the Shelley-Godwin Archive for Off-line Use

Raffaele Viglianti
rviglian@umd.edu
Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

Digital scholarly editions and archives are typically published on the web, which makes it possible to create interactive and reading experiences with the potential of reaching worldwide audiences. When text is encoded with care, for example by adopting the Text Encoding Initiative standard, it becomes possible for the same encoded content to be delivered in other formats and media, such as e-book and PDF for print. Web-based interactive digital editions, however, are the most efficient in utilizing the interactive and interconnected features of the web for presenting both text and the editorial scholarship that produced it. Ongoing scholarship around minimal computing and minimal editions has pointed out some important, yet addressable, flaws of many TEI digital editions. Bloatedness of infrastructure, for example, particularly when paired with rapid technical obsolescence and changes in funding, can hamper long-term preservation efforts; weighty resources may not be easily accessible from slower connections; and online-only access to a digital edition can be an obstacle to the world-wide access potential highlighted earlier.

The Shelley-Godwin Archive (S-GA) has taken steps to reduce its infrastructure footprint by generating a static site: in its production form, with the exception of its search index, S-GA is a collection of TEI, HTML, CSS, and JavaScript that can be hosted on any server without needing to set-up any server-side component (see Fig. 1).

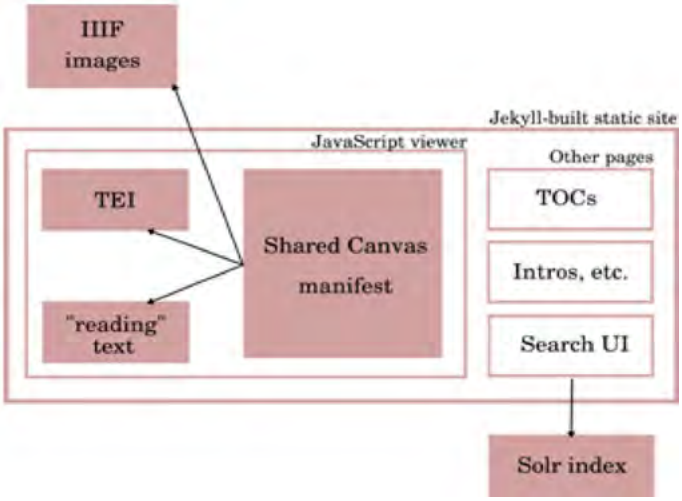


Fig. 1 The architecture of the S-GA website, built using Jekyll and static assets. Images are served primarily by the Oxford Bodleian Library using the International Image Interoperability Framework API.

This approach also makes it possible to bundle resources together for off-line use. This poster will show three potential approaches to creating off-line resources for an on-line publication: a one-document HTML bundle, a compressed archive of resources and an Electron desktop application. Unlike a PDF or e-book version, these downloadable resources will preserve the functionality of S-GA's website (with the exception, for now, of full text search), thus making the archive more usable in a poten-

2 T. Zerrouki, Tashaphyne, Arabic light stemmer, https://pypi.python.org/pypi/Tashaphyne/0.2
3 http://inl.github.io/BlackLab/
4 https://github.com/arabic-digital-humanities

tially greater number of cases, including increasing access for users with slow or no internet connections.

Academy of Finland Research Programme "Digital Humanities" (DIGIHUM)

Risto Pekka Vilkko
risto.vilkko@aka.fi
Academy of Finland, Finland

Digital Humanities (DIGIHUM) is a four-year research programme funded by the Academy of Finland. Its aim is to address novel methods and techniques in which digital technology and state-of-the-art computational science are used for collecting, managing and analysing data in humanities and social sciences research as well as for modelling humanities and social science phenomena.

Finland has a strong tradition in digital humanities. By bringing together the existing best knowledge and skills in digital humanities, Finland aims to put itself in a strong position to become a world leader in this rapidly evolving field. The programme is grounded in the needs of basic research, but technological advances in this area also have great potential for practical applications that warrant research.

The development of research in this area is based on broad collaboration involving not only researchers in the field but also technology experts, representatives of memory organisations (libraries, archives) and database administrators and developers. One aspect of the programme is to examine digitalisation as a cultural and social phenomenon.

The programme has three thematic areas:

1. Research into digital interaction and digital services
2. Employing open, multiform and/or real-time data in research
3. Data-based analysis and modelling of humanities and social sciences phenomena.

The programme produces new and more comprehensive knowledge and understanding about the themes under investigation. It fosters dialogue and exchange between a wide variety of scientific fields and disciplines, for example, by integrating methodologies and networking at national and international level. The programme encourages interdisciplinary or multidisciplinary projects that combine two or more fields of scientific research employing different methodologies and approaches. The aim is to promote:

1. collaboration among producers, processors and users of humanities and social sciences data
2. the development of research methods
3. ethical examination of the research field
4. the usability and awareness of datasets.

The poster will interactively introduce the programme's themes and objectives as well as the six research consortia that form the core of the programme:

- *Profiling Premodern Authors* (Prof. Marjo Kaartinen et al., University of Turku)
- *Interfacing Structured and Unstructured Data in Sociolinguistic Research on Language Change* (Prof. Terttu Nevalainen et al., University of Helsinki)
- *Citizen Mindscapes – Detecting Social, Emotional and National Dynamics in Social Media* (Prof. Jussi Pakkasvirta et al., University of Helsinki)
- *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (Prof. Hannu Salmi et al., University of Turku)
- *Digital Face* (Prof. Janne Seppänen et al., University of Tampere)
- *Digital Language Typology: Mining from the Surface to the Core* (Prof. Martti Vainio et al., University of Helsinki).

The poster will also introduce the four additional projects related to the Trans-Atlantic (T-AP) Platform *Digging into Data Challenge*:

- Digging into Manuscript Data (Prof. Eero Hyvönen, University of Helsinki)
- Analyzing Child Language Experiences Around the World (Prof. Okko Räsänen, Aalto University)
- Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914 (Prof. Hannu Salmi, University of Turku)
- Digging into High Frequency Data: Present and Future Risks and Opportunities (Prof. Peter Sarlin, Hanken School of Economics).

The Academy of Finland is a government agency within the administrative branch of the Finnish Ministry of Education, Science and Culture. Its mission is to fund high-quality scientific research, provide expertise in science and science policy, and strengthen the position of science and research.

Modeling the Genealogy of Imagetexts: Studying Images and Texts in Conjunction using Computational Methods

Melvin Wevers
melvin.wevers@huygens.knaw.nl
DH Group, KNAW Humanities Cluster, The Netherlands

Thomas Smits
t.smits@let.ru.nl
Institute for Historical, Literary and Cultural Studies, Radboud University, The Netherlands

Leonardo Impett
leonardo.impett@epfl.ch
Digital Humanities Institute, EPFL, Switzerland

In his influential article "There are no Visual Media", theorist of visual culture W.J.T Mitchell argues that "all media are mixed media" (Mitchell, 2005). In earlier work, Mitchell already noted that composite works—media formats that consist of both image and text—cannot be adequately studied by comparing the meaning of these two forms of representation separately (Mitchell, 1994, p. 89). The subject matter of these "imagetexts", is, rather, the "whole ensemble of relations between media" (Mitchell, 1994, p. 89). In other words, the meaning of one of the components of an imagetext, be it either the image or the text, can only be understood in relation to the other. This paper combines methods from text mining, computer vision, and information theory to increase our understanding of this relationship throughout several historical datasets.

Several scholars have observed that Digital Humanities research mainly focuses on (large-scale) textual analysis (Champion, 2017; Meeks, 2013). Erik Champion, for instance, notes that the influential definition of Digital Humanities by the University of Oxford is entirely "text based and desk based" (Champion, 2017, p. 25). While he rightly claims that research in the Digital Humanities is centered on text, in recent years an increasing number of researchers have started studying visual material, in which has been called "visual big data" (Ordelman et al., 2014; Smith, 2013). Scholars increasingly rely on computational methods to analyze these large digitized visual datasets in innovative ways. Important examples are the work of Seguin (Seguin et al., 2017) on visual pattern discovery in large databases of paintings, Impett and Moretti's (Impett and Moretti, 2017) large-scale analysis of body postures in Aby Warburg's Atlas Mnemosyne, and Wevers' (Wevers and Lonij, 2017) and Smits's (Smits, 2017; Smits and Faber, 2018) analysis of visual trends in advertisements and images in newspapers. These projects were all presented at DH2017, some during the well-attended pre-conference workshop of the Special Interest Group AudioVisual Material in Digital Humanities (AVinDH).

The recent upsurge of large-scale analysis of visual material shifts the focus in Digital Humanities research away from texts. However, this has also led researchers to approach text and images as disjointed entities. Computational techniques can analyze similarity and change in both textual and visual discourse. Our project applies techniques from both textual and visual computational analysis to a dataset of advertisements for cars extracted from the widely-read Dutch newspaper *De Volkskrant* between 1945 and 1995, which we extracted from the large collection of digitized newspapers maintained by the National Library of the Netherlands. By juxtaposing change points in text and visual material, we show that the meaning of imagetexts can be studied by looking at the relation between the two forms of representation. Put differently, how does change and continuity in the visual correspond to changes in the textual and vice versa?

Using Kleinberg's burst algorithm, we detected bursty words in the textual content of advertisements (Kleinberg, 2002). These bursts indicate possible change points in advertising discourse that call for closer examination of the advertisements and can be cross-examined with possible changes in the visual content. Also, topic modeling (LDA) was used to detect clusters of advertisements based on textual context. These clusters were compared to cluster based on visual aspects.

Trends, similarities, and points of inflection in the image sets will be traced using a subspace learned by training a Generative Adversarial Network (GAN; see Goodfellow et al. 2016), which has been shown to generate semantically-meaningful vector subspaces. GANs work best with regular sets of images - our visual analysis process is thus twofold. First, we use a pretrained Mobilenet CNN (Howard et al. 2017) to detect objects (cars, trucks, people, etc), and then train individual GANs to explore the visual-semantic space of each object through time.

Whereas a traditional CNN can only encode from image to vector, a GAN can also decode from any vector to generate artificial images; trends or clusters hypothesized in a vectorial subspace can therefore be subjected to a 'close reading' of the corresponding artificial images. This generative hermeneutic avoids the 'black box' nature of traditional neural network image analysis.

The ability to detect how changes and continuity between text and images correlate increases our understanding of the function of imagetexts in modern culture. It also helps us understand whether the relationship between the two forms of representation became more entangled over time, or whether this entanglement is specific to particular products or specific periods.

References

Champion, E.M. (2017), "Digital humanities is text heavy, visualization light, and simulation poor", *Digital Scholarship in the Humanities*, Vol. 32 No. 1 sup, pp. i25–i32.

Howard, A.., et al. (2017) "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861.

Impett, L. and Moretti, F. (2017), "Totentanz", *New Left Review*, No. 107, pp. 68–97.

Kleinberg, J. (2002), "Bursty and Hierarchical Structure in Streams", P*roceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Edmonton, Canada (2002), pp. 91–101.

Meeks, E. (2013), "Is Digital Humanities Too Text-Heavy?", *Digital Humanities Specialist*.
Mitchell, W., 1994. Picture Theory. Essays on Verbal and Visual Representation, University of Chicago Press, Chicago.
Mitchell, W.J.T. (2005), "There Are No Visual Media", *Journal of Visual Culture*, Vol. 4 No. 2, pp.257–266.
Ordelman, R., Kleppe, M., Kemman, M. and De Jong, F. (2014), "Sound and (moving images) in focus – How to integrate audiovisual material in Digital Humanities research", ADHO 2014.
Seguin, B., di Leonardo, I. and Kaplan, F. (2017), "Tracking Transmission of Details in Paintings", ADHO 2017.
Smith, J.R. (2013), "Riding the Multimedia Big Data Wave", *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, USA, pp. 1–2.
Smits, T. (2017), "Illustrations to Photographs: using computer vision to analyse news pictures in Dutch newspapers, 1860-1940", ADHO 2017.
Smits, T., Faber, W.J. (2018), "*CHRONIC* (Classified Historical Newspaper Images)", *KB Lab*, 21 March, http://lab.kb.nl/dataset/chronic-classified-historical-newspaper-images.
Wevers, M. and Lonij, J. (2017), "Siamese", *KB Lab*, 15 October, http://lab.kb.nl/tool/siamese.

History for Everyone/Historia para todos: Ancient History Encyclopedia

James Blake Wiener
james.wiener@ancient.eu
Ancient History Encyclopedia (AHE), United Kingdom

Gimena del Rio Riande
gdelrio@conicet.gov.ar
IIBICRIT, CONICET, Argentina

The most important publication to emerge from 18th-century Europe was arguably the *Encyclopedia* by Denis Diderot (1713-1784 CE). The *Encyclopedia* sought to "change the general way of thinking," challenging all forms of bigotry, repression, fanaticism, and misinformation (Fowler, 2011). Through his commission of articles on a variety of topics, Diderot endeavored to summarize and disseminate the world's information in order to help human society progress to new heights of accomplishment while also mitigating the sufferings of civilization. Helped by the mathematician Jean Le Rond d'Alembert (1717-1783) as well as Montesquieu (1689-1755 CE) and Voltaire (1694-1778), Diderot's *Encyclopedia* was very much a collaborative project, which reflected the "party of humanity" in a new age of international and informational exchange (Micale and Dietle, 2000).

Ancient History Encyclopedia (http://www.ancient.eu) was founded in 2009, in the spirit of the Enlightenment, with the mission to improve education through the creation of the most complete, freely accessible, and reliable online, historical resource in the world. As scholars and supporters of the digital humanities, the contributors at Ancient History Encyclopedia felt a responsibility to construct a site in which users not only found what they were looking for, but also one which stressed the importance of global cultural heritage and world history. Our knowledge and interpretation of history shapes how we define ourselves as nations and as cultures, and it influences how we see other cultures as well. Whether through its interactive map of the ancient world, online videos, or its carefully curated articles and definitions, Ancient History Encyclopedia digitally imparts knowledge in new and creative ways.

Before Ancient History Encyclopedia's inception, most of what was available online pertaining to ancient history was scattered across various websites, illegible due to poor presentations, targeted exclusively at academic audiences and hidden behind paywalls, or tainted by a distinct nationalistic agendas. While Wikipedia undeniably advanced and pushed the aims of the Open Access movement, it sometimes remains riddled with inaccuracies and occasional bias. Omnipresent too is the lack of available content in major world languages like Spanish, Russian, Mandarin Chinese, Arabic, Portuguese, and Hindi. Other sites, like La guía de historia (https://www.laguia2000.com/), do not afford proper attributions to sources and lack curated multimedia libraries of pictures, videos, and other interactive learning tools. Over the last two decades, open access publishing has become increasingly widespread with the help of the Internet. The Open Access movement helps researchers, students, and educators access the latest research and data without restrictions. It is a movement defined by high standards, the exchange of information, the development and synchronization of models, and the promotion of innovation in technology and research methodology.

Through a shared commitment to Open Access Education, Ancient History Encyclopedia and its partners create interactive tools that facilitate historical and media literacy, build models of data exchange, and serve a broader community rather than solely those in academia. In this sense, Ancient History Encyclopedia is acting in unison with the principles of Open Access, Open Education, and Open Research. These are positive developments, but it is not nearly enough: historians, researchers, publishers, museums, and other institutional bodies must move beyond the paradigm of simply making it free or available only in English. Ancient History Encyclopedia is an international project with contributors from Germany, the United States, Hungary, India, Argentina, the United Kingdom, and Australia. Through Ancient History Encyclopedia's collaborations with other digital humanities pro-

jects and organizations, including the Pelagios Commons, Europeana's Eagle Project on ancient Roman epigraphy, Humanidades Digitales del Centro Argentino de Información Científica y Tecnológica (HD CAICYT-CONICET), and Laboratorio de Innovacíón en Humanidades Digitales at Madrid's Universidad Nacional de Educación a Distancia (LINHD-UNED), Ancient History Encyclopedia has aided in making important academic research and datasets available and digestible to Anglophone audiences.

In this poster, Ancient History Encyclopedia and Humanidades Digitales CAICYT-CONICET (http://www.caicyt-conicet.gov.ar/micrositios/hd/) review Ancient History Encyclopedia's encyclopedic model and successes, while also sharing plans for future projects that will include the translation and publication texts at CAICYT and the joint use of map data from Pelagios Commons (http://commons.pelagios.org/

References

Fowler, J. (2011). *New essays on Diderot*. Cambridge: Cambridge University Press.

Micale, M. S., and Dietle, R. L. (2000). E*nlightenment, Passion, Modernity. Historical Essays in European Thought and Culture*. Redwood City: Stanford University Press.

Princeton Prosody Archive: Rebuilding the Collection and User Interface

Meredith Martin
mm4@princeton.edu
Princeton Prosody Archive, Princeton University Center for Digital Humanities, United States of America

Meagan Wilson
mrwilson@princeton.edu
Princeton Prosody Archive, Princeton University Center for Digital Humanities, United States of America

Mary Naydan
mnaydan@princeton.edu
Princeton Prosody Archive, Princeton University Center for Digital Humanities, United States of America

The PPA collects and displays historical documents prior to 1923, bringing to light little-known texts about the study of language, the study of poetry, and where and how these intersect and diverge. By gathering these documents into one place, the PPA tracks the development of English poetry as a subject of study and shows how this development bridges a variety of discourses, most prominently the rise of linguistic nationalism and linguistic imperialism, but also the advent of stadial history and historiography, the rise of phonetic science and the beginnings of historical linguistics, and a variety of related pedagogical movements that evolve from rhetoric through to elocution and the study of "speech." The PPA is the only large-scale corpus focused specifically on the study of poetry in the English language. Materials in the archive include grammar handbooks, poetic treatises, versification manuals, elocution guides, histories of literature, editorial introductions, phonetic tracts, and journal articles pertaining to the measure and pronunciation of poetry. By viewing prosody broadly and collecting these materials into one archive, scholars can finally see how the histories of English poetics and linguistics are intertwined, and how the story of English poetic development, alongside the development of historical linguistics, increasingly borrowed, co-opted, imitated, erased, or "civilized" poetic forms from other languages.

Critical attention to these poetic histories and debates are the foundation of Historical Poetics. In addition to scholars of Historical Poetics, the PPA's audience is teachers of poetry, scholars of poetry, linguists, practicing poets, historians of language, historians of pedagogy, scholars of sound studies, scholars of rhetoric, and lexicographers—all of whom can use the PPA to discover the emergence of a disciplinary term, trace its evolution, or determine its ties to national or political debates. Finally, computer scientists and digital humanists are eager to run textual analytic algorithms on a curated data set that might reveal previously unknown or unexpected results such as the most frequently reprinted poetic example or the most frequently repeated (perhaps without attribution) definition of a particular term.

"Rebuilding the Collection and User Interface," the PPA's poster and interactive demonstration for DH2018, showcases the immense data-refinement and metadata-cleaning performed by the PPA since its DH2014 poster session. After launching our new website in May 2018, we are well-positioned to discuss the strengths and struggles of curating and designing an interactive website that relies on HathiTrust Digital Library content. In this way, the PPA sees itself as a project similar to Early American Cookbooks, recently published as a HathiTrust case study in *Code4Lib*. "Legacy MARC data for early books held in special collections presents particular challenges," Gioia Stevens writes; "Cleaning and standardizing this legacy data is an essential step in analyzing special collections metadata as a dataset rather than as individual records" (Stevens, 2017). This has proven especially germane to the PPA. From 2015 to 2017, the PPA refined its core collection by eliminating 3,729 duplicate works through a complex and painstaking metadata cleaning process. These duplications were the result of our initial file transfer from HathiTrust and the replicas were skewing users' search results. The PPA offers a case study in the challenges posed by working with unstandardized metadata. In addition to addressing the benefits and drawbacks of our collaboration with HathiTrust, our poster session aims to highlight how our new interface

guides users toward the database's implicit and explicit arguments, highlights unusual content, and provides pathways for discovery.

References

Stevens, Gioia. (2017). "New Metadata Recipes for Old Cookbooks: Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal." *Code4Lib* **37**, http://journal.code4lib.org/articles/12548 (accessed 1 May 2018).

ELEXIS: Yet Another Research Infrastructure. Or Why We Need An Special Infrastructure for E-Lexicography In The Digital Humanities

Tanja Wissik
tanja.wissik@oeaw.ac.at
Austrian Academy of Sciences, Austria

Ksenia Zaytseva
ksenia.zaytseva@oeaw.ac.at
Austrian Academy of Sciences, Austria

Thierry Declerck
declerck@dfki.de
Austrian Academy of Sciences, Austria

In this presentation, we will discuss the recently started European project ELEXIS – European Lexicographic Infrastructure and its potential in the context of digital humanities.

The use of the computer in modern lexicography is intertwined with the history of the digital humanities (c.f. Schreibmann et al. 2004) and the lexical data have grown to be is indispensable in more and more DH projects, especially with the rise of the Semantic Web and Linked Open Data (c.f. Oldman et al. 2016).

However, current lexicographic resources, both modern and historical, have different levels of structuring and are not equally suitable for the application in other fields, such as Natural Language Processing, and thus not directly usable in DH projects for Semantic Web applications and methods.

Therefore, ELEXIS will develop strategies, tools and standards for extracting, structuring and linking lexicographic resources to unlock their full potential for Linked Open Data and the Semantic Web, as well as in the context of digital humanities.

The ELEXIS project is carried out by a consortium of partners from various fields (e.g. lexicography, computational linguistics, natural language processing, digital humanities, and artificial intelligence). The consortium consists of the following scientific institutions, language institutes, standardisation bodies, and publishing houses: "Jožef Stefan" Institute (Slovenia), Lexical Computing CZ s.r.o. (Czech Republic), Instituut voor de Nederlandse Taal (Netherlands), La Sapienza University of Rome (Italy), National University of Ireland, Galway (Ireland), Austrian Academy of Sciences (Austria), Belgrade Center for Digital Humanities (Serbia), Hungarian Academy of Sciences, Research Institute for Linguistics (Hungary), Institute for Bulgarian Language »Prof Lyubomir Andreychin« (Bulgaria), Universidade Nova de Lisboa (Portugal), K Dictionaries (Israel), Istituto di Linguistica Computazionale "A. Zampolli" (Italy), The Society for Danish Language and Literature (Denmark), University of Copenhagen, Centre for Language Technology (Denmark), Trier University, Center for Digital Humanities (Germany), Institute of the Estonian Language (Estonia), Real Academia Española (Spain).

The ELEXIS project aims to integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure which will enable efficient access to high quality lexical data in the digital age, and bridge the gap between more advanced and lesser-resourced scholarly communities working on lexicographic resources.

ELEXIS intends to take an innovative approach of production and development of lexico-semantic resources by creating intelligent applications for crucial tasks such as linking lexical resources, word sense disambiguation and cross-lingual mapping on the basis of applied methods and techniques in the fields of NLP and Artificial Intelligence fields.

The ELEXIS infrastructure will help researchers create, access, share, link, analyse, and interpret heterogeneous lexicographic data across national borders, paving the way for ambitious, trans-national, data-driven advancements in the field, while significantly reducing the duplication of efforts across disciplinary boundaries. In order to ensure the sustainability of the technical infrastructure after the end of the project, the created infrastructure will be integrated into the already existing infrastructures CLARIN and DARIAH, since most of the partners are members of CLARIN and DARIAH national consortia.

Besides the technical infrastructure, ELEXIS will establish a network for knowledge exchange and will develop and implement free online training courses for lexicography. Furthermore, ELEXIS will give researchers and research teams trans-national access to research facilities and lexicographical resources which are not fully accessible online or where professional on the spot expertise is needed in order to ensure and optimise mutual knowledge exchange. The trans-national access will have impact especially for under-resourced languages and will

all in all strengthen the infrastructure and collaborative network provided by ELEXIS.

Even though the infrastructure is at the moment planned as a European infrastructure, there are thoughts to expand it beyond Europe in order to cater for the needs of DH researchers around the globe.

References

Schreibman, S., Siemens, R. and Unsworth, J. (eds.) (2004). *A Companion to Digital Humanities. Oxford: Blackwell*. http://www.digitalhumanities.org/companion/

Oldman, D., Doerr, M. and Gradmann, S. (2016). Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge. In Schreibman S. et al. (eds.) (2016). *A New Companion to Digital Humanities*, 2nd Edition. Oxford: Wiley-Blackwell.

"Moon:" A Spatial Analysis of the Gumar Corpus of Gulf Arabic Internet Fiction

David Joseph Wrisley
djw12@nyu.edu
New York University Abu Dhabi, United Arab Emirates

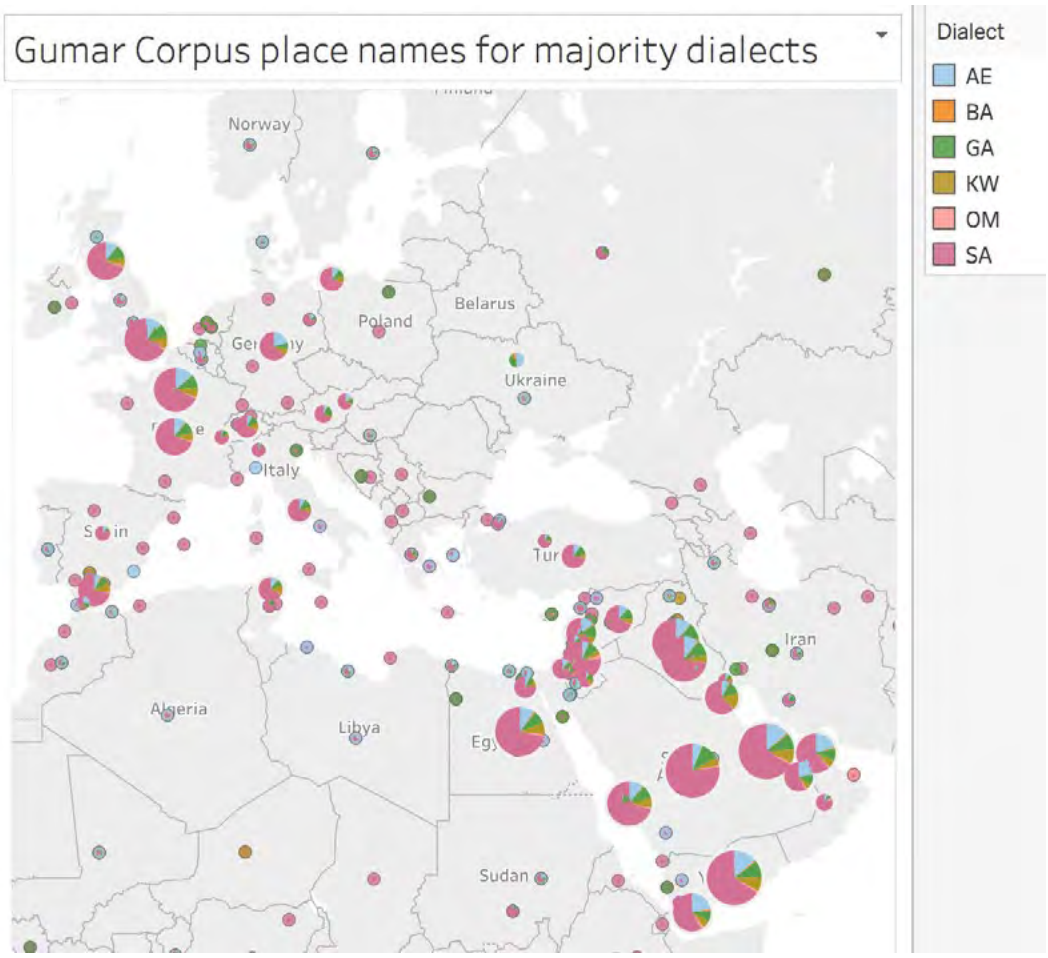
Hind Saddiki
hind.saddiki@nyu.edu
Mohammadia School of Engineering,
Mohammed V University in Rabat, Morocco; Computational Approaches to Modeling Language Lab,
New York University Abu Dhabi, United Arab Emirates

The Gumar Corpus (https://camel.abudhabi.nyu.edu/gumar/) consists of 110 million words from 1,200+ Internet forum novels written in a conversational style about romantic topics. Whereas the corpus was originally harvested and annotated for use within the context of dialectal Arabic (DA) natural language processing, the material is also of cultural and sociological significance concerning popular culture in the Gulf Arab region. The corpus' name comes from the Gulf Arabic word for moon {gumɛr}, a popular Arabic term of endearment. Whereas the genre is all but unknown outside of the Arab world, the Arabic blogosphere and social media are full of discussions about these "net novels," the authors of which are purportedly young women. In addition to Modern Standard Arabic (MSA), we have five dialect varieties mapping to roughly 12 national sub-varieties of dialectal Arabic--usually only one tag is assigned to each internet novel. Our poster is a very first attempt to tap into the cultural richness of the corpus using methods adapted to the Arabic language, in particular from the angle of spatial analysis of corpora.

The internet novels sometimes identify their country of origin in a short prologue, but there are additional clues as to their provenance including the fact that they are all written in DA, which is not necessarily the native dialect of the author. Much progress has been made in information extraction and NLP for Arabic in the last decade, but in dialectal forms much work remains to be done to catch up to Western languages. Even though we would not expect significant variance in toponyms in DA, initial attempts at extracting place names directly from the Arabic corpus posed a methodological challenge, particularly for disambiguation. Practical workarounds, often translingual and through English, are sometimes adopted in such cases with Arabic, as in the case of BetaCode that uses English script to deal with the vocalization, or partial vocalization of texts (Romanov, 2015).

With the goal of extracting place name entities from Gumar, our pilot study carried out morphological analysis and disambiguation on the texts using MADAMIRA (Pasha et al., 2014), a tool that currently functions for both for MSA and Egyptian Arabic. The configuration for Egyptian has been shown to outperform the MSA setting when compared to a manually annotated sample of 4K words from the Gumar corpus (Khalifa et al., 2015). Since the MADAMIRA morphological annotation provides both the lemma of a word and the English translation of the lemma, we build an English approximation of the novels and run them through Stanford Named Entity Recognizer to detect locations (Finkel et al., 2005).

Using Stanford NER, 19000+ occurrences of some 400+ distinct locations were extracted from the aggregate of the novels. Having English versions of the place names made the geocoding a relatively straightforward process. Geovisualization shows that the highest frequency of place names are found in the Arabian Gulf, Iraq, *bilad as-Sham* and Al-Andalus (southern Spain), as well as in England, France and Germany. Given that about sixty percent of the novels are identified as the dialect of Saudi Arabia, the high frequency of mentions of the Kingdom seems predictable. On the other hand, the places are not specific locales, as in the case of the city-level geographies of Palestine and Iraq. Other more detailed analysis about such specificity of place needs to be carried out through subsequent close reading of the corpus.



While the corpus is a rare opportunity both to work with contemporary popular culture Arabic in the textual digital humanities and to experiment with named entity recognition methods for non-Western contexts, caution must be exercised in our interpretations since the methods which work well for western languages are much more tentative in the (regional) Arabic case. For example, some cross checking was done against the Arabic texts in the corpus and revealed errors where DA colloquialisms {kif} ("what?") and {bliz} ("please"), generated some high frequency false locations "Kiev" and "Belize." As our research evolves, we intend to benchmark other Arabic-only tools for entity recognition to test their stability and performance on the set of materials in question (Gahbiche-Braham et al. 2013; Shaalan 2014). Time permitting, we would like to begin to do some correlations between topic and geography, what has been recently labelled a "geospatial semantics" (Gavin/Gidal, 2017) but for the transnational, multiregional context of Arabic. Our hope is to use the Gumar corpus to take on more in-depth analysis of a Gulf Arabic geopoetics of romance.

References

Finkel, J. R., Grenager, T. and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), pp. 363-370.

Gahbiche-Braham, S., Bonneau-Maynard, H., and Yvon, F. (2013). Traitement automatique des entités nommées en arabe: détection et traduction. *Traitement Automatique des Langues*, 54(2): 101-32.

Gavin, M., Gidal, E. (2017). Scotland's Poetics of Space: An Experiment in Geospatial Semantics, *Cultural Analytics*. http://culturalanalytics.org/2017/11/scotlands-poetics-of-space-an-experiment-in-geospatial-semantics/ (accessed 30 April 2018).

Khalifa, S. et al. (2016). A Large Scale Corpus of Gulf Arabic, In *Proceedings of the Language Resources and Evaluation Conference* (LREC), Portorož, Slovenia, pp. 4282-89.

Pasha, A. et al. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference* (LREC), Reykjavík, Iceland, pp. 1094-1101.

Romanov, M. (2015). BetaCode for Arabic, *Al-Raqmiyyat*, https://maximromanov.github.io/2015/02-07.html (accessed 30 April 2018).

Shaalan, K. (2014) A Survey of Arabic Named Entity Recognition and Classification, *Computational Linguistics*. 40(2): 469-510.

A New Methodology for Error Detection and Data Completion in a Large Historical Catalogue Based on an Event Ontology and Network Analysis

Gila Prebor
gila.prebor@biu.ac.il
Bar Ilan University, Israel

Maayan Zhitomirsky-Geffet
maayan.geffet@gmail.com
Bar Ilan University, Israel

Olha Buchel
obuchel@gmail.com
Faculty of Information and Media Studies, University of Western Ontario, Canada

Dan Bouhnik
dan.bouhnik@gmail.com
Bar Ilan University, Israel; Jerusalem College of Technology, Israel

Introduction

The catalogue of Historical Hebrew Manuscripts, curated by the National Library of Israel, represents the largest collection in the world of over 130,000 Hebrew manuscripts that survived through the last millennium and are currently spread off in a variety of institutions all over the globe. The catalogue was created by many different classifiers during the long period of some 70 years. As a result, many of the fields are inconsistent and unorganized (Zhitomirsky-Geffet and Prebor, 2016). Moreover, a deeper examination of the data reveals missing and incorrect information (e.g. manuscripts with unknown date and place of writing). This missing and incorrect information poses a great pitfall for researchers who need reliable data to base their research on (Hric et al., 2016).

In this paper we present a novel approach for completion and correction of historical data from a large manuscript catalogue based on an event-based ontology and network communities' analysis. To resolve data inconsistencies in the catalogue, in the previous study we proposed an event-based ontology model (Zhitomirsky-Geffet and Prebor, 2016). The ontology model is shown in Figure 1.

Figure 1: Ontology model of the manuscript data.

Approach

- The proposed methodology comprises the following stages:
- Extraction of ontological entities from the catalogue data and ontology construction;
- Building networks of ontological entities based on direct and indirect ontological relationships between these entities, e.g. a network of censors who participated in the common Manuscript Censorship Events, or a bipartite network of manuscripts and people related to them through some events;
- Automatic community identification in the constructed networks (Blondel et al., 2008);
- Outlier detection among the related events in the network or in the closest community, i.e. if the manuscript creation event's date is later than its censorship event's date;
- Semi-automatic inference of missing data based on the ontological relationships and communities in the network, e.g. inferring a censor/author's missing time and place of living from the corresponding data of his peers in the community.

Results

Here we present preliminary results of the proposed approach applied on the case of Censorship Events of Hebrew manuscripts in medieval Italy. In the context of the Counter-Reformation, during the 16th-18th centuries, the Catholic Church closely supervised written and printed literature. The Church appointed censors (most of them apostates and experts in the Hebrew language) to censor and approve the Hebrew books.

The diagram in Figure 2 emphasizes the most influential censors and demonstrates the strengths of collaborations.

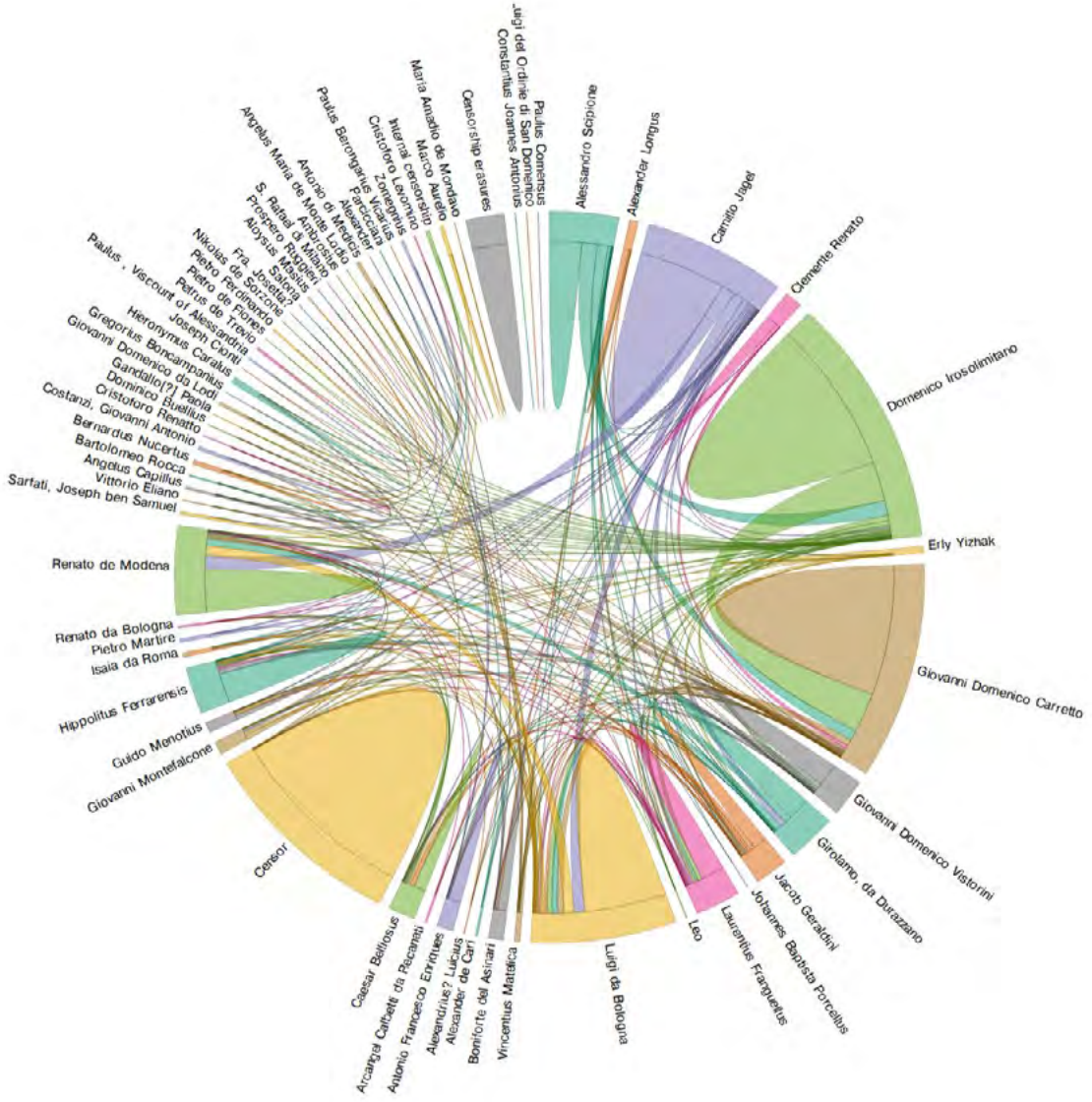


Figure 2: A chord diagram of censors related by common manuscripts.

The total number of relationships in the network depicted in Figure 3 is 2,037, and the number of nodes (representing Italian censors) is 62. For 37% of the censors we can observe the discrepancies between explicitly mentioned in the catalogue and automatically inferred periods of activity, the inference was based on the dates of individual censorship events in the ontology. In 5% of the cases the dates of their activity in the catalogue were incorrect (e.g. in cases where the activity period of 50 and more years). In addition, given Censorship Events' related entities, such as, a censor name and date and manuscripts censored by him and their related dates and locations we could infer its location and as a result, we obtained places of censorship for 53.9% out of all the Censorship Events in the corpus. Eight of the inferred places did not appear in the original list of 12 places that have been recorded in the catalogue.

Grey links in Figure 3 show that the most influential censors were censoring manuscripts one after another. Such collaborations can probably be regarded as waves of censorship. Two timelines under the map display which locations were inferred and which ones were specified explicitly in the catalogue. Comparing these two timelines allows researchers to identify mistakes about time and gives suggestions for unknown locations.

To conclude, our preliminary findings show that ontology-based network analysis and detection of communities provide an effective tool for correction and completion of missing or incorrect historical data.

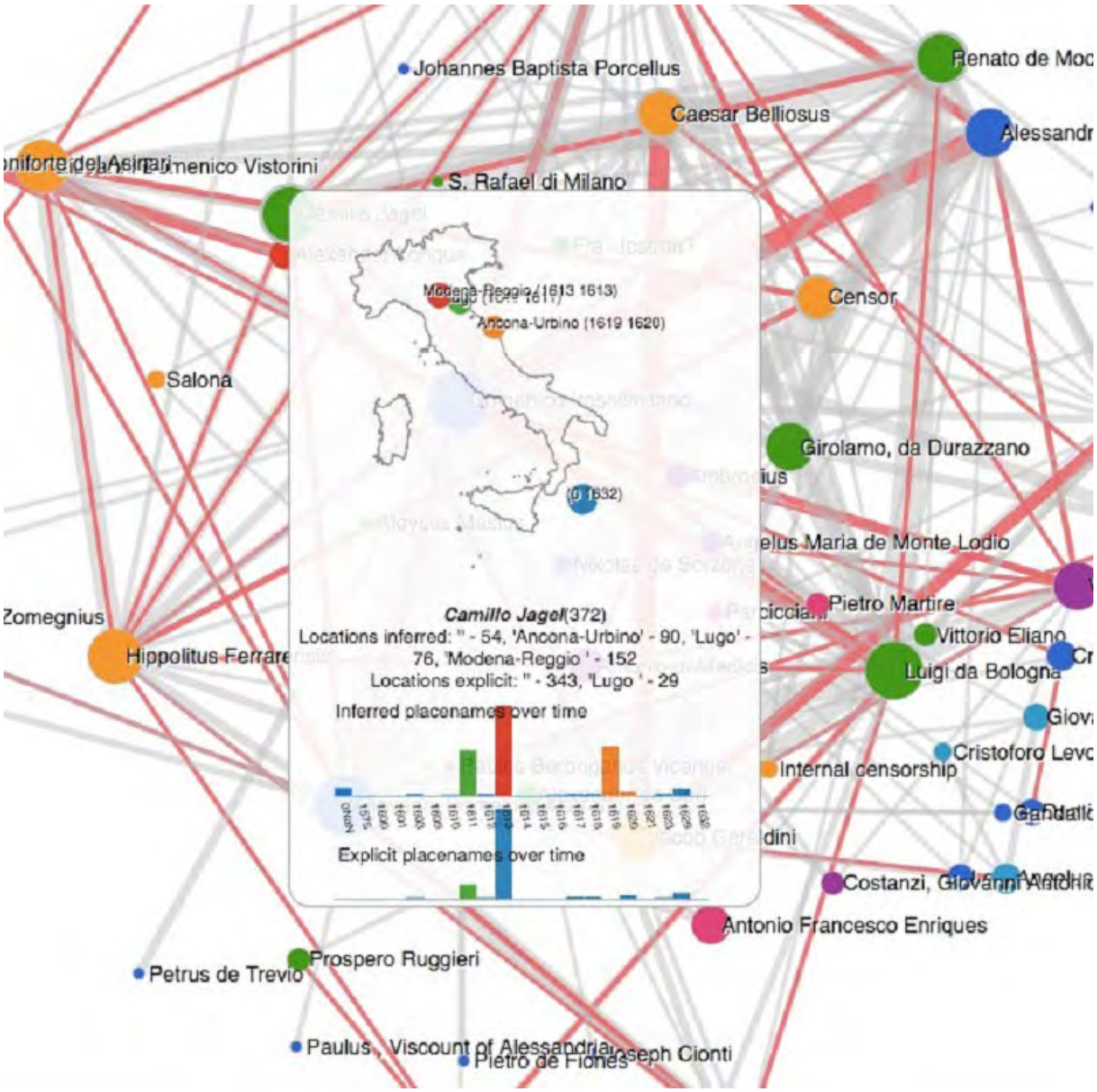


Figure 3: A net*work of censors in Italy with two relationship types – censors who worked on the same manuscript at the same time and censors who worked on the same manuscripts at different time periods (represented by red and grey links, correspondingly). Line thickness represents the number of joint manuscripts for a pair of censors. The censors were divided into seven communities by the Louvain algorithm (Blondel et al., 2008) (represented by different colours of nodes). Clicking on nodes shows time maps of censors.*

References

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, & E. Lefebvre. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, Oct. 2008.

Hric, D., Peixoto, T. P., & Fortunato, S. (2016). Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, *6*(3), 031038.

Zhitomirsky-Geffet M. & Prebor G. (2016). Towards an ontopedia for historical Hebrew manuscripts. *Frontiers in Digital Humanities, section of Digital Paleography and Book History*, *3*, 3. http://dx.doi.org/10.3389/fdigh.2016.00003.

Preconference Workshops

Jumpstarting Digital Humanities Projects

Amanda French
amandafrench@gwu.edu
George Washington University, United States of America

Anne Chao
annechao@rice.edu
Rice University, United States of America

Marco Robinson
mtrobinson@pvamu.edu
Prairie View A&M University, United States of America

Brian Riedel
riedelbs@rice.edu
Rice University, United States of America

Brief Summary

"Jumpstarting Digital Humanities Projects" is a half-day pre-conference workshop on various aspects of beginning a digital humanities project: scoping and planning a sizable project; determining when to use institutional infrastructure and when to go beyond the institution; winning cooperation from institutional authorities and collaborators; collecting and digitizing materials; and designing for iterative development and efficient feedback loops. Our sessions will focus on the common type of digital humanities project that consists of a assembling a database of source material and generating interactive interpretations such as maps and visualizations from that database. Five scholars from different disciplines and institutions, each a participant in the Mellon-funded Resilient Networks for Inclusive Digital Humanities initiative, will give short tutorials, and workshop attendees will spend an hour on exercises in which they can begin planning a digital humanities project with help from the instructors.

Description of Content

"Jumpstarting Digital Humanities Projects" is a half-day pre-conference workshop on various aspects of beginning a digital humanities project: scoping and planning a sizable project; determining when to use institutional infrastructure and when to go beyond the institution; winning cooperation from institutional authorities and collaborators; collecting and digitizing materials; hiring students and technologists; and designing for iterative development and efficient feedback loops. Our sessions will focus on the common type of digital humanities project that consists of a assembling a database of source material and generating interactive interpretations such as maps and visualizations from that database. Five scholars from different disciplines and institutions, each a participant in the Mellon-funded Resilient Networks for Inclusive Digital Humanities initiative, will give presentations apiece of 30-45 minutes, and workshop attendees will spend an hour on exercises in which they themselves can begin planning their own digital humanities project with individualized help from the instructors. We will end the day with a brief group discussion on how humanities scholars at institutions without digital humanities centers can best form networks and advocate for infrastructure at their own institutions to support digital scholarship.

Scoping and Planning

Workshop leaders will discuss the collaborative and creative processes by which they determine what is achievable in a given project, and how they found the most optimal paths towards achieving their goals. These presentations will not be didactic but exploratory, the "leaders" having at this stage, on average, only begun to execute their workflows. This will provide an ideal space for attendees at various stages in their projects to feel invited to ask questions and contribute to strategies for determining what can be achieved within the specific constraints of budget, time, skills, and archival resources.

Institutional and Extra-Institutional Infrastructure

One of the major decisions projects have to make in their beginning stages is where to host content. Digital humanities projects of the type we are discussing in this workshop require a website, yet many if not most institutions do not provide server space for humanities scholars. Increasingly, libraries will host and manage digital humanities projects, but not all libraries provide this service, and those that have provided it in the past often find that as software and systems age, the cost in labor of maintaining digital humanities projects is a disincentive to provide such services for future projects. Commercial hosts such as GoDaddy and HostGator are one option, and an increasingly well-known option is Reclaim Hosting, founded by instructional technologists by and for educators, but many humanities faculty members are either not aware of these options or do not know how to choose between them. Workshop leaders will discuss their own choices and the relative advantages and disadvantages of each, balancing speed, efficiency, cost, support, sustainability, and longevity.

Feedback loops & iterative design

Collaborative humanities projects depend on the gathering of diverse skills in the pursuit of complex goals. While it is difficult in institutional settings to achieve appropriate parity, this sort of cross-department and cross-strata project work can form alternative modes of collecti-

ve intellectual labor that takes seriously the input of all stakeholders. The appropriate site for this integration of viewpoints in the context of project work is what we call "design." By negotiating over what a thing does and how, a team comes to understand better what it is they are doing in the first place. A project often looks different at the end than it did in the earliest planning stages, and this aspect of the discussion will invite participants to think more creatively about the possibilities of interdisciplinary and inter-departmental collaboration.

Achieving and Maintaining Buy-in

The differences in institutional situations between the different groups represented by collaborating members in an interdisciplinary project necessarily create communicative friction and potential divergences in goals and perceptions. While this on some level represents differences in commitments, the perceived shared goal of any project is what brings collaborators to the table in the first place, and a flexible orientated-ness is what maintains buy-in. Workshop leaders will lead open-ended discussions about experiences in this process.

Collecting and Digitizing Materials

Many digital projects in the humanities begin with non-digital materials, such as the images and documents in the county archives of Waller County, Texas. Projects that include oral histories such as the Houston Asian American Archive now usually capture recordings in born-digital formats, but comprehensive archives of this nature may also need to convert analog audio and video materials from earlier eras. Libraries and archives have a great deal of knowledge about digitization and metadata standards and conversion and migration technologies that can be of use to humanities scholars, so partnering with library and archives professionals early on can be of great benefit. Workshop leaders in this section will discuss their practices with digitizing and collecting materials, especially in partnership with librarians.

Description of Audience

Humanities scholars in the early planning stages of large projects that require a broad array of technical and scholarly competencies. While Digital Humanities is of course a conference for advanced practitioners, we hope in this session both to entice "analog" humanities scholars to commingle with more experienced digital humanities scholars and to encourage experienced digital humanities scholars to think about how best to foster the spread of their methods.

Technical Requirements

This workshop requires a digital projector with audio capabilities, preferably one that can be used with instructor laptops: it requires no special software or hardware. We will expect attendees to bring laptops, and we hope that the workshop room will have sufficient power outlets for attendees.

Length, Format, and Budget

"Jumpstarting Digital Humanities Projects" will be a one-day workshop on the following schedule:

9am-12:30pm: Presentations of 20 to 30 minutes by course instructors
12:30pm-1:30pm: Lunch
1:30pm-3:30pm: Guided exercises in digital humanities project planning
3:30pm-4:15pm: Reflections on the day and discussion of institutional support needs for digital humanities projects

The Resilient Networks for Inclusive Digital Humanities project can fund the registration and travel of instructors. We would prefer a cost of no more than $25 USD for participants, especially since this workshop is meant to appeal chiefly to relative beginners in digital humanities.

Workshop Leaders

Anne Chao
Title: Manager, Houston Asian American Archive
Email: annechao@rice.edu
Phone: 713-202-5599
Address: 3970 Inverness Dr., Houston, TX 77019

Anne Chao is manager of the Houston Asian American Archive at Rice University. She oversees Rice student interns to conduct interviews with Asian Americans in Houston and the greater metropolitan area. Since 2010, HAAA has accumulated over 160 oral history interviews spanning diverse ethnicities from East, to Southeast, and South Asian-Americans. The collection of primary source materials details the contribution of Asian Americans in the building of greater Houston since the Jim Crow era, and provides new insight into the history of the region. Working with the archivist at the Fondren Library, HAAA uses the Omeka platform and includes GIS mapping to plot the life trajectories of the interviewees. The interviews are fully transcribed and time-stamped, synchronized, indexed with key words through the use of the Oral History Metadata Synchronizer (OHMS).

Amanda French
Title: Director, Resilient Networks for Inclusive Digital Humanities

Email: amandafrench@gwu.edu
Phone: 720-530-7515
Address: GWU Libraries, 2130 H Street NW, Washington, DC 20052

Amanda French's particular expertise consists of making humanities content (both cultural content and scholarly interpretation of that content) openly available online, as well as introducing scholars to the various methods of and issues with making humanities content openly available online. She held the CLIR Postdoctoral Research Fellowship at NCSU Libraries from 2004-2006. From 2010-2014, she was first Coordinator and later Principal Investigator for the Mellon-funded initiative THATCamp (The Humanities and Technology Camp), an international unconference that has seen more than 300 events to date attended by more than 7000 people. She often speaks and sometimes writes about open access, the scholarly publication landscape, Omeka, Scalar, Hypothes.is, THATCamp, the Digital Public Library of America, Wikipedia, grant-writing, and alternative careers for humanities PhDs. Her most recent digital research project is a catalog with accompanying exhibits of the personal library of the American poet Edna St. Vincent Millay, available at http://steepletoplibrary.org.

Brian Riedel
Title: Professor in the Practice of Humanities; Associate Director, Center for the Study of Women, Gender, and Sexuality – Rice University
Email: riedelbs@rice.edu
Phone: 713-348-2162
Address: CSWGS, MS-38 | 6100 Main St | Houston, TX | 77005-1892

Brian Riedel received his Ph.D. in Anthropology from Rice University. His research and teaching focus on engaged research and lesbian, gay, bisexual, transgender, and queer social movements, particularly in Greece and the United States. Two of his current projects use GIS to examine the historical connections of place and sexuality. One project examines the histories of the Montrose neighborhood of Houston, Texas, and the uses to which they are put. A core component of that project is a GIS visualization of Houston's LGBT-centered businesses from 1945 to 2015. The other project, conducted in collaboration with the African American Library at the Gregory School (part of Houston Public Library) and Rice Century Scholar Cameron Wallace, documents Houston's formal red-light district known as the "reservation," which operated from 1908 to 1917. Although freed slaves had settled on that land since Emancipation, the city claimed the area held "only a few Negro huts." The project uses GIS and StoryMaps to meld primary resources like census, city directory, and tax record data.

Marco Robinson
Title: Assistant Professor of History, Prairie View A & M University, Prairie View, Texas
Email: mtrobinson@pvamu.edu
Phone: 936-261-3219
Address: Division of Social Work, Behavioral, and Political Sciences, Prairie View A&M University, P.O. Box 519; MS 2203, Prairie View, TX 77446-2203

Marco Robinson is an Assistant Professor of History at Prairie View A & M University, Prairie View, Texas. Marco's research is centered around capturing the social, political, economic, and cultural histories of communities in the American South through collecting, preserving, and analyzing archival and oral history data. As it relates to digital humanities, Dr. Robinson uses this data to tell digital stories, for mapping using GIS and the digitization of historical artifacts. His most recent publication and project are "Telling the Stories of Forgotten Communities: Oral History, Public Memory, and Black Communities in the American South" (Collections: A Journal for Museum and Archives Professionals, Volume 13, Number 2, (Spring 2017): 171- 184.) and Using Interactive Maps and Apps to Preserve Local History: Digitizing the Black Experience in Waller County, Texas.

New Scholars Seminar

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca
University of Alberta, Canada

Rachel Hendery
r.hendery@westernsydney.edu.au
Western Sydney University, Australia

Juan Steyn
juan.steyn@nwu.ac.za
South African Centre for Digital Language Resources, South Africa

Elise Bohan
elise.bohan@gmail.com
Edith Cowan University, Australia

The New Scholars Symposium has been running since DH2015. It brings together graduate students and recent graduates in a one day "unconference" where they can develop their own research agenda and prepare for the conference. The NSS also includes an opportunity to meet with digital humanities leaders and a mentoring opportunity for the new scholars.

In the last three years centerNet has supported the NSS along with CHCI. The CHCI funding has come to an end, which is why we are applying as a workshop. The Kule Institute for Advanced Study at the University of Al-

berta (CHCI member) has and will provided support for organizing this seminar on behalf of centerNet. Rachel Hendery (Associate Professor of Digital Humanities, Western Sydney University) and Geoffrey Rockwell (Director, Kule Institute for Advanced Study, University of Alberta, Canada) have acted as conveners of the New Scholars Seminar. We propose to build on our experience with this format but add new workshop leaders including Juan Steyne from North-West University, South Africa. CenterNet will also be more directly involved with running of the symposium this year through the assistance of the CenterNet secretary, Elise Bohan.

Target Audience

For the purposes of the Seminar a "new scholar" is defined as someone who is either a graduate student or someone who has received their PhD within the last 5 years (or longer if a case is made for career interruption). Postdoctoral fellows and people in alternative academic positions are welcome to apply.

Participation is by reviewed application and participation is limited to a maximum of 20 people. Typically we support 10 from outside the target continent and 10 from inside, many of whom are students at the hosting university.

Deadline and application process

Applications have usually been due in April. We intend this to be the case this year too, if we the workshop is accepted in time for this to be feasible. Otherwise we will select the earliest feasible deadline. Applications include i) a Statement of Research that outlines their research interests in digital humanities; ii) a letter of support from a centerNet centre/institute director if applicable; and iii) a short two-page CV. Applications are sent to the Kule Institute for Advanced Study <kias@ualberta.ca> at the University of Alberta, a centerNet member. The applications will be reviewed by the following committee:

Geoffrey Rockwell (University of Alberta, Canada)
Rachel Hendery (Western Sydney University, Australia)
Juan Steyn (Northwest University, South Africa)
Elise Bohan (Macquarie University, Australia)
Adam Dombovari (University of Alberta, Canada)

Brief Outline: Intended length and format

The programme for the seminar is developed by the participants once accepted and coordinated by the workshop leaders. The idea is to empower new scholars to develop their own research directions and collaborations. This has previously been very successful, developing a program with a diversity of themes that could not have been anticipated by the workshop leaders. There are therefore typically two phases:

Before the Seminar there is an online gathering component using the University of Alberta eClass (Moodle) platform. Participants share their Statements and discuss what they are interested in discussing together. Clusters of research interests emerge which form the intellectual backbone of the Seminar. We encourage leadership to emerge from within the group so that the actual structure of the on-site days will be primarily organized by the participants.

The on-site portion of the Seminar then takes place in the days before the DH conference. Ideally we would have a day and a half for this, but it could be reduced to one day if necessary. The program that we find works includes three components:

Short presentations by participants of their research and interests followed by a social event the evening before the unconference. This helps break the ice and introduce everyone.

The unconference where we spend an initial hour identifying the key issues/sessions that participants want to organize followed by breakout sessions. The sessions are participant-designed and facilitated. When we reconvene, reporters from the sessions report back to the whole group. This can be structured to fit the time available by increasing or decreasing the number of sessions and running more or fewer of them parallel to each other.

Topics for these small sessions on the unconference day in previous years have included:

- DH pedagogy
- Amplifying diverse voices in DH
- Working with archival materials
- Working with databases
- Quantitative vs qualitative data
- Artificial Intelligence
- Crowdsourcing
- Web scraping
- Creating Twitterbots

One of the sessions from 2016 produced a Manifesto on Student-Driven Research that has since been further developed by the participants and submitted to the *Debates in the Digital Humanities* new series on 'Institutions, Infrastructures at the Interstices'.

c. Mentoring during the DH conference around careers or opportunities in the digital humanities. This last year (2017) we organized mentoring with senior scholars in the field of digital humanities. Before the Seminar participants identified the sort of mentoring they would like and we (Rockwell and Hendery) then contacted people we knew would be at the DH conference and asked them if they were willing to meet for coffee or lunch with a new scholar. The participant and mentor then arranged to meet at their convenience. This was a new feature of the NSS this last year and those that took advantage of it reported that they appreciated the

opportunity. In previous years it has taken the form of e.g. a panel discussion about careers with senior DHers, or small group discussion time with such people. This year we propose to connect the New Scholars with leaders from the centerNet community, both through a networking event sponsored by centerNet, and through one-to-one mentoring opportunities.

Budget

The NSS has secured support from centerNet and SADiLaR. CenterNet will provide catering for breaks and lunch. CenterNet will also provide support for the mentoring component and invite the NSS participants to a networking event with centerNet leaders. SADiLaR has provided assistance with organisation via Juan Steyn and will further provide full support for one participant from South Africa.

In previous years thanks to CHCI funding we have been able to offer participants a significant funding package to assist them to attend. Many of our students and ECRs have said in evaluations they would not have been able to get to the ADHO conferences without this. As we are unable to offer that this year, we would like to find other ways to lessen the burden on these participants. We do not charge any registration fee for this workshop. We also hope that the conference organizers and/or ADHO might provide discounts on registration for our New Scholars. We will also work with participants from outside North America to find travel and conference support for them from other sources where possible.

Special requirements for technical support

We would need space for parallel break-out sessions – usually a total of three spaces is sufficient. A single room can work if it is large enough that small groups can sit in separate corners and hold discussions without disturbing each other too much. Apart from this we only need a projector and a whiteboard.

Getting to Grips with Semantic and Geo-annotation using Recogito 2

Leif Isaksen
l.isaksen@exeter.ac.uk
University of Exeter, United Kingdom

Gimena del Río Riande
gdelrio.riande@gmail.com
CONICET, Argentina

Romina De León
rdeleon@conicet.gov.ar
CONICET, Argentina

Nidia Hernández
nidiahernandez@conicet.gov.ar
CONICET, Argentina

This workshop introduces *Recogito 2*, a tool developed by Pelagios Commons that enables annotation of geographic place references in text, images and data through a user-friendly online platform. Perhaps the most notable feature of Recogito 2 is the ability to produce semantic data without the need to work with formal languages directly, while at the same time allowing the user to export the annotations produced as valid RDF, XML and GeoJSON formats.

The availability of born digital data as well as digitised collections, is changing the way we study and understand the humanities. This amount of information has even greater potential for research when semantic links can be established, and relationships between entities highlighted. The work of Pelagios Commons has shown that connecting historical data according to their common reference to places (expressed via URIs stored in gazetteers) is a particularly powerful approach: information about material culture, archaeological excavations, ancient texts and related scholarship can be connected and cross referenced through the geodata.

Producing semantic annotations usually requires a certain amount of knowledge of digital technologies such as RDF, ontologies and/or text encoding. These techniques can sometimes act as a barrier for users that are not already familiar with Semantic Web theory. The Recogito annotation tool aims to facilitate the creation and publication of Linked Open Data by dramatically reducing some commonly encountered obstacles. First developed in 2014, the community-oriented philosophy behind Pelagios Commons has made users an active agent in shaping its functionality and interface. A dedicated forum on the Pelagios Commons website gathers feedback and suggestions. Recogito code is Open Access and available through GitHub where discussions of Recogito's more technical aspects are held. After a year of intensive redevelopment from the ground up, Recogito 2 was launched in December 2016 and now has almost 1,500 registered users. Introductory documentation is available in English, Spanish, German and Italian with the interface itself being translated into multiple languages in February 2018.

Recogito now supports both additional image standards (such as IIIF) and text standards (TEI export). This allows researchers to use the annotation tool as either a starting or intermediate point for their workflow in the production of semantic annotations that can be then built upon with other technologies. While the initial release already enabled collaboration among users, Recogito 2 features a more refined series of options to manage degrees of collaboration, from private annotations that can only be accessed by their creator, to collaborative and public ones that anyone can see and download. These options offer the opportunity to collaborate, but leaves users free to choose the degree of openness that best suits their materials at different stages of research.

Originally conceived for data related to the ancient world, Recogito 2 has become a valuable tool for annotating many other kinds of historical and modern sources, especially (but not confined to) those containing geographical information. Recogito 2 facilitates the annotation of any named entity. Where applicable, they can be resolved against a number of aligned digital gazetteers, including the ancient world (Pleiades) and modern (Geonames). Although the annotation of geographical information is its most principal focus, Recogito 2 also allows "people" and "event" references to be annotated (currently without semantic resolution), and the opportunity to add tags and comments to disambiguate and refine later searches. Two different colour-coding options makes it easy to identify the different kind of annotations (places, people or events) or different status of the geographic annotations.

This workshop walks participants through all stages of using Recogito 2 to annotate different types of source documents: from uploading a file to the online platform, through annotation, to the download of the annotations in the available data formats. More specifically, the workshop will show practical examples of:

Annotation of sources in text format

Attendees will learn how to benefit from Recogito's automatic recognition of named entities, and how to refine it manually. They will create annotations ex novo, and check or modify those identified by Recogito. The geo-annotations produced on the text can then be plotted on a digital map, through a user-friendly visualisation mode. The relevance of each place is displayed on the map proportionally to the number of annotations that the place has received. Places are linked, via a pop-up window, to all their annotations in the same document, and users are able to browse each annotation in a short, essential context, or to see them in the full text.

Annotation of images and tables

After beginning with text files, attendees will work on the semantic annotations of images. Maps are especially well suited to geo-annotation but Recogito 2 can also be used for the annotation of other types of image, such as photographs or even textual sources in the form of digitised manuscripts. Users will upload images to the Recogito platform and be able to select, transcribe, annotate and, georesolve toponyms within the image. Workshop attendees will also see how Recogito can import and annotate or align tabular (CSV) data such as that derived from spreadsheets, databases or gazetteers.

Exporting data from Recogito

Finally, participants will learn how to export data from Recogito in a variety of formats suitable for visualizing and analysis in other tools, such as spreadsheets, databases or GIS.

To maximise the benefit of the workshop, participants are invited to bring their own data and documents to annotate. Recogito currently has greatest support for ancient and modern sources (including most languages). Materials from other periods can also be annotated but the level and quality of georesolution may vary. The workshop will provide sample texts, imagery and data for attendees without their own datasets. The workshop will show examples of annotations of different kind of sources, and discuss their specific challenges. Throughout the workshop there will be opportunities for participants to discuss how Recogito 2 might be used to support their own research.

Visualising and contextualizing geographical information within documents can be an important step in reaching a deeper understanding of their content, potentially highlighting phenomena that would have been otherwise difficult to identify. It is also an effective tool for engaging students when encountering historical texts and collections. The design of Recogito 2 is intended to make the production of semantic annotations easy and intuitive, opening the door of the Linked Open Data ecosystem to a wide range of users, including without prior experience of semantic technologies.

Semi-automated Alignment of Text Versions with iteal

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

David Joseph Wrisley
djw12@nyu.edu
New York University Abu Dhabi, United Arab Emirates

Overview

Our half-day tutorial proposed for DH2018 concerns the semi-automatic alignment of different witnesses in complex textual traditions, with demonstration of specific use cases, a discussion of the relevance of the implemented system to particular textual problems relevant to the participants as well as a hands on discovery of the system. Alignment is a relatively simple task for modern languages with orthographic stability and relatively similar texts, but when there is a degree of instability of textual transmission as in oral literatures, popular music or poetry, or other complex texts with partial repetition the task becomes more difficult. Whereas methods of hand aligning and visualizing texts exists in TEI, we focus on the possibility of computational alignment for the purpose of exploratory textual visualization. Scholars who are

interested in visualizing scaled forms of reading will be interested in this tutorial.

Our visual analytics environment *iteal* supports the computational alignment of textual similarities and is not English-specific. It was originally implemented using orally inflected medieval French poetic texts (with test cases of the fabliaux and epic) and so is known to work on texts in Latin alphabets with inconsistent orthography.

This half-day tutorial aims at introducing *iteal* to the DH community for which the questions of multi-text problems, spelling variance and debates about distant forms of reading are currently quite salient. Many language processing and visualization tools do not work well with languages beyond English. Our environment is known to work with languages beyond English will be of interest those interested in expanding innovative techniques in the textual humanities across the North/South divide. Participants of the tutorial will be led in a step-by-step, hands-on approach through the full cycle of an *iteal*-based text alignment workflow, and they will finally have the opportunity of testing the tool with their own data. Although proven to be effectively useful for text variants of medieval poetry, we will not focus only on this type of text as *iteal* can be used to determine alignments among texts of a different kind in any language and in multiple genres. Currently, *iteal* works with plain text in utf8.

iteal consists of two major modules:

First, it automatically determines line-to-line alignments pairwise between all given text editions based on user-configurable parameters including:

- **Edit distance:** Variant spellings are taken into account by this function. We define two words as spelling variants if they have the same first letter, and if the string similarity of the remaining substrings is higher than a user-configurable threshold.
- **Coverage:** In order to ensure that a specific proportion of words of both lines are aligned, the user can configure a minimum coverage value of the line.
- **N-grams:** The user can configure the minimum required n-gram size n that is the largest number of subsequent word matches of both lines.
- **Broken n-grams:** Quite often, the only difference between two lines is a single word in the middle of a line that is either inserted, synonymous, or a transposed stopword. Large n-grams, from this perspective are not achieved. Thus, we allow the user for considering broken n-grams, which is the total number of word matches among both lines.

Second, for the purpose of analyzing the determined alignment we provide interactive visualizations for different text hierarchy levels (examples for all three views can be found in Figures 1, 2 and 3, and a teaser outlining a brief workflow with *iteal* can be found at https://vimeo.com/230829975):

- **Distant Reading:** In order to get a rough overview of alignment patterns throughout the observed text versions, we draw a miniature representation for each version in the form of a vertical bar reflecting its number of verse lines in contrast to the other shown versions. For us, this is the most distant form of reading, where the text itself is not visualized, but rather abstract depictions of textual similarity point to patterns worth discovering.
- **Meso Reading:** Since multiple texts are displayed in synoptic views, the visualization is able to convey more complex patterns of textual relationship. We call this a meso reading that might be said to connect multiple close readings all the while transmitting information that lies beyond the scope of a close reading. Here, we use the intuitivity of stream graphs to connect aligned verse lines among different versions. For a more detailed inspection of an individual alignment, clicking on a stream opens a popup window for line-level close reading.
- **Close Reading:** Next to plain text, the close reading view provides word level alignments for the corresponding verse lines in the form of two Variant Graph visualizations. Within the close reading view, individual alignments can be confirmed with user input, so that it gets persistently stored in the backend.

Target audience: Anyone studying variance in the textual digital humanities and its visualization would be interested in our tutorial. It will be offered in English, but can accommodate data in a variety of languages. Potential participants in the tutorial are encouraged to be in touch with the presenters in advance of DH2018 to provide some sample data that can used to provide a mashup. Required for this step is a version of at least two documents sharing some text in common, of at least 20 lines.

Tutorial Schedule

Part I (1 hour + break time)

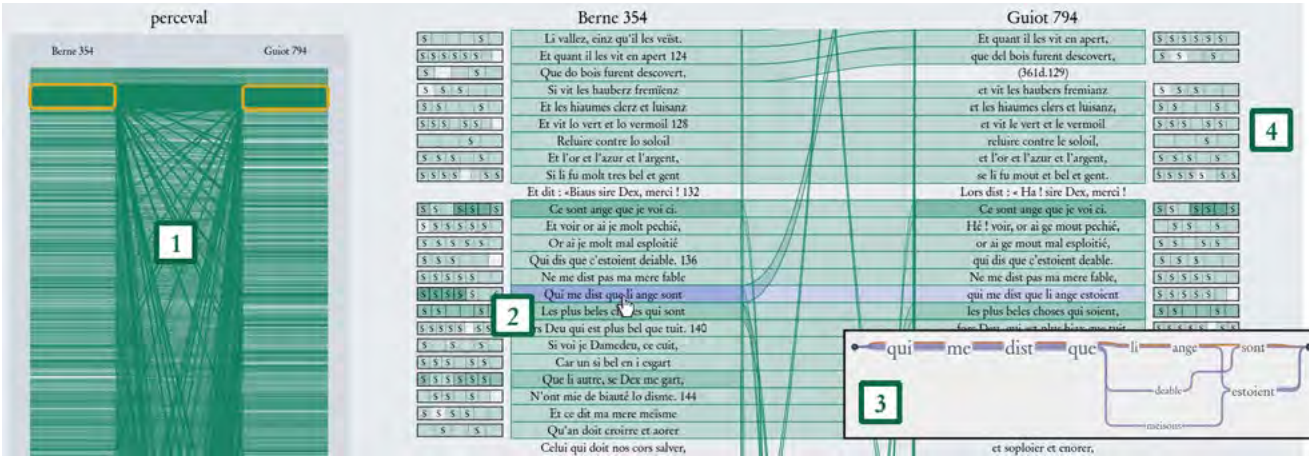
– *iteal* introduction: purpose, functionality, configuration, visualization (Stefan Jänicke)
– Medieval French poetry as an *iteal* use case (David J. Wrisley)
– Further use cases, future work, questions (Stefan Jänicke & David J. Wrisley)

Break

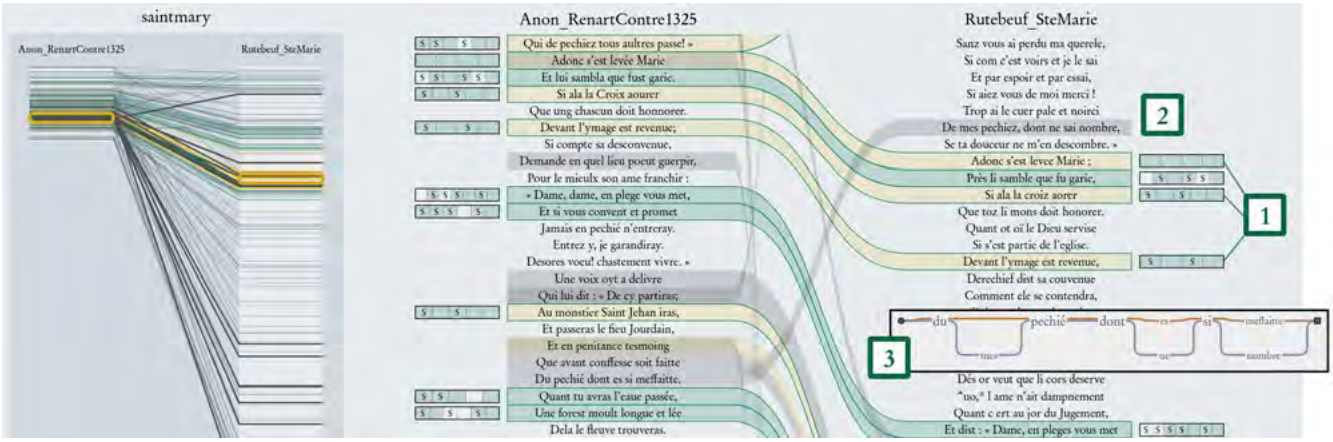
Part II (2 hours - break time)

– Step-by-step hands-on session with texts brought by tutorial participants
– wrap up, feedback and steps forward

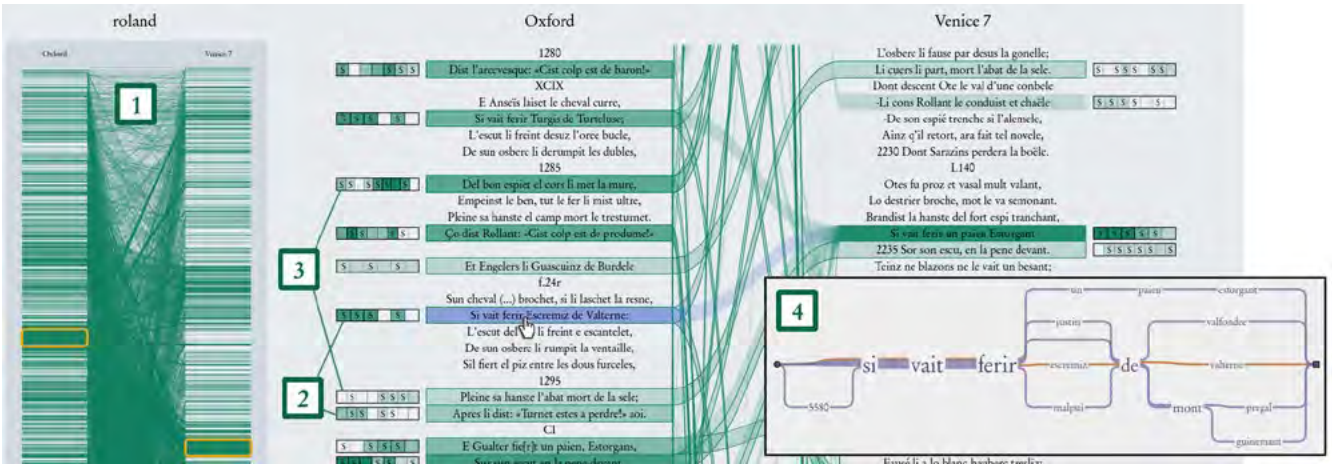
Sample images from iteal



Aligning two editions of Perceval with iteal



Aligning two editions of La vie de saint Marie l'Egyptienne with iteal



Aligning two editions of the Chanson de Roland with iteal

Biographies of the Tutorial Organizers

Stefan Jänicke (stjaenicke@informatik.uni-leipzig.de): Dr. Stefan Jänicke is a post-doctoral researcher at the Image and Signal Processing Group at Leipzig University, Germany, where he leads a text visualization group focusing on applications in the digital humanities. Over the last years, he has gained experience in developing information visualization and visual analytics techniques within a number of digital humanities projects. His PhD thesis investigates the utility of visualization techniques to support the comparative analysis of digital humanities data, and his current research relates to information visualization with a focus on applications for text- and geovisualization in digital humanities. *Homepage:* http://stjaenicke.vizcovery.de

David Joseph Wrisley (djw12@nyu.edu): Dr. David Joseph Wrisley is Associate Professor of Digital Humanities at New York University Abu Dhabi. His research interests include the creation of open, inclusive corpora in medieval studies, corpus-based geovisualization as well as visual exploration of variance in poetic traditions. Furthermore, he is interested in the challenges in humanities data stemming from both multilingual environments and social data creation. *Homepage:* http://djwrisley.com

References

S. Jänicke, A. Geßner, G. Franzini, M. Terras, S. Mahony and G. Scheuermann (2015). TRAViz: A Visualization for Variant Graphs. In: *Digital Scholarship in the Humanities* 30, suppl 1, pp i83–i99.

S. Jänicke, G. Franzini, M. F. Cheema and G. Scheuermann (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In: Eurographics Conference on *Visualization (EuroVis) - STARs*. The Eurographics Association.

S. Jänicke and D. J. Wrisley (2017). Visualizing Mouvance: Towards a Visual Analysis of Variant Medieval Text Traditions. In: *Digital Scholarship in the Humanities* 32, suppl 2, pp ii106–ii123.

S. Jänicke, A. Geßner, M. Büchler and G. Scheuermann (2014). Visualizations for Text Re-use. In: Proceedings of the *5th International Conference on Information Visualization Theory and Applications (VISIGRAPP 2014)*, pp 59–70.

S. Jänicke and D. J. Wrisley (2017). Interactive Visual Alignment of Medieval Text Versions. In: *IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2017*.

S. Jänicke, A. Geßner, M. Büchler and G. Scheuermann (2014). 5 Design Rules for Visualizing Text Variant Graphs. In: *Conference Abstracts of the Digital Humanities 2014*.

S. Jänicke and D. J. Wrisley (2016). Visualizing Mouvance: Towards an Alignment of Medieval Vernacular Text Traditions. In: *Conference Abstracts of the Digital Humanities 2016*.

Innovations in Digital Humanities Pedagogy: Local, National, and International Training

Diane Katherine Jakacki
diane.jakacki@bucknell.edu
Bucknell University, United States of America

Raymond George Siemens
siemens@uvic.ca
University of Victoria, Canada

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Angelica Huizar
ahuizar@odu.edu
Old Dominion University, United States of America

Esteban Romero-Frías
erf@ugr.es
University of Granada, Spain

Brian Croxall
brian.croxall@byu.edu
Brigham Young University, United States of America

Tanja Wissik
tanja.wissik@oeaw.ac.at
Austrian Academy of Sciences, Austria

Walter Scholger
walter.scholger@uni-graz-at
University of Graz, Austria

Erik Simpson
simpsone@grinnell.edu
Grinnell College, United States of America

Elisabeth Burr
elisabeth.burr@uni-leipzig.de
Universität Leipzig, Germany

Context: as the digital humanities take firm root in the humanities curriculum, institutions around the world are now committing significant resources toward developing DH and integrating it in standalone courses, graduate degrees and undergraduate majors and minors within and across departments. With this commitment comes the realization that such formal implementation of DH and its siblings (e.g. digital social sciences, digital media, etc.) at a degree-granting level requires articulation of core requirements and competencies, identification and hiring of faculty who are capable of teaching DH in a variety of learning environments (coding, systems, application of methods), evaluating a broad spectrum of student work,

and beyond. It also changes the foundational principles of the work of those in our network, as training increasingly involves learning how to teach competencies at the same time as we ourselves develop and maintain them in light of fast-paced advances.

2018 Focus, and Call for Proposals: at the 2017 mini-conference, attendees reached consensus about forming an ADHO Special Interest Group (SIG) dedicated to DH Pedagogy in all its forms. In support of this, for our 2018 mini-conference and meeting, we continue in inviting proposals for lightning talks on all topics relating to digital pedagogy and training -- and especially this year for those that will lead us to substantial discussion about how a SIG could support instructors, students, practitioners, and administrators. Mini-conference talks will take place in the morning, and the afternoon member meeting will be dedicated to work on a collaborative draft of the SIG proposal. In particular, we welcome proposals with a focus on:

- Ways in which individual universities, colleges, and other educational institutions are extending DH in the classroom.
- Implementing DH pedagogical frameworks locally and working across institutions and training institutes to develop and collaborate on materials that can inform ways in which DH offerings and programs are formalized.
- Assessment techniques in DH curriculum. What types of assessment should occur in digital humanities courses? And, significantly, how might these assessment practices challenge existing university or community-based outcomes? We particularly desire talks that include involvement of students who have been assessed.
- DH training in an international context-how do we articulate/coordinate/collaborate across international boundaries? What can we learn from our differences?
- Developing a multilingual lexicon for teaching DH.
- Discussion of pedagogical materials, pre-circulated for critique and consideration. We are particularly interested in the submission of specific syllabi, tutorials, exercises, learning outcomes, assessment and rubrics that attendees might complete during the workgroup portion of the mini-conference.
- Any topics that might further inform our discussion about DH training.

Machine Reading Part II: Advanced Topics in Word Vectors

Eun Seo Jo
eunseo@stanford.edu
Stanford University, History Department, United States of America

Javier de la Rosa Pérez
versae@stanford.edu
Stanford University, Center for Interdisciplinary Digital Research, United States of America

Scott Bailey
scottbailey@stanford.edu
Stanford University, Center for Interdisciplinary Digital Research, United States of America

Fernando Sancho
fsancho@us.es
Dept. of Computational Sciences and Artificial Intelligence, University of Seville, Spain

Description

This half day workshop is an introduction to word vectors and text vectorization broadly. We will focus on building intuition of how word vectors work, incorporating visualization methods, using pre-trained vectors, and exploring applications of word embeddings. We will teach you both the high-level concepts and the practical usages of these widely used analytical tools for text analysis in digital humanities (DH). It is a hands-on workshop with practical activities for the participants starting with a review of word vectors by way of visualization, an overview of downloadable word vectors, and examining the potential pitfalls of using word vectors in humanistic analysis and the methods for mitigating these issues. Given the general applicability of machine learning models in real life, addressing issues concerning biased models, datasets, and algorithms, is of vital importance for correct interpretation of their applications.

We will provide a Python Jupyter Notebook and an accompanying text corpus that we will work through as a group. By the end of the workshop, the participants will have working knowledge of how and where to download or train word embeddings and the caveats of using them.

Relevance to the DH Community

Since the apparition of analytical approaches to distant reading and macro-analysis, popularized by Moretti and Jockers, and the possibility of access to huge amounts of textual data and long-term studies such as Culturomics, new tools were needed to tackle the increasing complexity of large corpora. Borrowing from advances in machine learning and computational linguistics, digital humanists have experimented with various methods of text quantification for interpreting macro contours of culture and language. In particular, word vectors have gained recognition for their versatility in DH studies. Scholars have used word vectors in a variety of tasks such as measuring similarity in word meaning (Caliskan et al., 2017), authorship attribution (Kocher et al., 2017), or dialogism in novels (Muzny et al., 2017).

This workshop is both a theoretical and practical introduction to humanist applications of these methods.

Those interested in large scale text-analysis of any corpora will learn the basics of transforming textual data into numerical form.

Instructors

Eun Seo Jo researches the language of American foreign relations in historical contexts and applications of NLP and ML in history. She is a PhD candidate in history at Stanford University where she is also a member of the Literary Lab and a Digital Humanities Fellow. She has presented at various DH conferences and is a DH methodology consultant at Stanford.

Scott Bailey is a Research Developer in the Center for Interdisciplinary Digital Research in the Stanford University Libraries. He collaborates and consults on research projects across the humanities and social sciences, and teaches workshops on tools and methods in digital scholarship, such as natural language processing. His research ranges from vulnerability in the context of theological anthropology to computational approaches to systematic and historical theological works, such as Karl Barth's *Church Dogmatics*.

Javier de la Rosa is a Research Engineer at the Center for Interdisciplinary Digital Research, a unit at the Stanford University Libraries focused on digital scholarship. He is an active member of the DH scholarly community at Stanford and regularly participates in conferences, professional organizations, and teaches workshops and tutorials to faculty and graduate students. He holds a Post-doctorate research fellowship and a PhD in Hispanic Studies at Western University, Ontario, where he also served as Tech Lead for the CulturePlex Lab. He completed both his MSc. in Artificial Intelligence and BSc. in Computer Engineering at University of Seville, Spain. His work and interests span from cultural network analysis and computer vision, to text mining and authorship attribution in the Spanish Golden Age of literature.

Fernando Sancho is an Associate Professor at the Dept. of Computational Sciences and Artificial Intelligence at the University of Seville, and holds a PhD by the same university. He has worked in topics ranging from complex systems, and data analysis to cultural objects studies. He has regularly collaborated with the CulturePlex Lab at the University of Western Ontario, and the Complex Systems Modeling Group at University of Central Ecuador.

Target Audience and Prereqs

Post-docs, faculty, and advanced graduate students with Python prerequisites. Although the main concepts will be overviewed, knowledge of basic word embeddings and word2vec specifically would be desirable. In order to participate fully in all activities, participants must have working knowledge of basic programming concepts, the Python language, data structures, and the Numpy library.

- Technical Support: Microphones and Projector
- Proposed Length: Half-day (4 hours; 4 sessions)
- Medium: Notebook (Jupyter)
- Libraries: Numpy, Pandas, Textacy, SpaCy, Gensim, scikit-learn, matplotlib

Workshop Outline

The workshop is split into four 50 min sessions with 10 minutes breaks in-between. We teach several methods in each unit with increasing difficulty. The schedule is broken down below:

Understanding Word Vectors with Visualization

This unit will give a brief introduction of word vectors and word embeddings. Concepts needed to understand the internal mechanics of how they work will also be explained, with the help of plots and visualizations that are commonly used when working with them.

- 0:00 - 0:20 From word counts to ML-derived Word Vectors (SVD, PMI, etc.)
- 0:20 - 0:35 Clustering, Vector Math, Vector Space Theory (Euclidean Distance, etc.)
- 0:35 - 0:50 [Activity 1] Visualizations (Clustering, PCA, t-SNE) [We provide vectors]

Word Vectors via Word2Vec

This unit will focus on Word2Vec as an example of neural net-based approaches of vector encodings, starting with a conceptual overview of the algorithm itself and end with an activity to train participants' own vectors.

- 0:00 - 0:15 Conceptual explanation of Word2Vec
- 0:15 - 0:30 Word2Vec Visualization and Vectorial Features and Math
- 0:30 - 0:50 [Activity 2] Word2Vec Construction [using Gensim] and Visualization(from part 1) [We provide corpus]
-

Extended Vector Algorithms and Pre-trained Models

This unit will explore the various flavors of word embeddings specifically tailored to sentences, word meaning, paragraph, or entire documents. We will give an overview of pre-trained embeddings including where they can be found and how to use them.

- 0:00 - 0:20 Overview of other 2Vecs & other vector engineering: Paragraph2Vec, Sense2Vec, Doc2Vec, etc.
- 0:20 - 0:35 Pre-trained word embeddings (where to find them, which are good, configurations, trained corpus, etc.)
- 0:35 - 0:50 [Activity 3] Choose, download, and use a pre-trained model

Role of Bias in Word Embeddings

In this unit, we will explore an application and caveat of using word embeddings -- cultural bias. Presenting methods and results from recent articles, we will show how word embeddings can carry historical bias of the corpora trained on and lead an activity that shows these human-biases on vectors and how they can be mitigated.

- 0:00 - 0:10 Algorithmic bias vs human bias
- 0:10 - 0:40 [Activity 4] Identifying bias in corpora (occupations, gender, ...) [GloVe] (Caliskan et al., 2017)
- 0:40 - 0:50 Towards unbiased embeddings; Examine "debiased" embeddings
- 0:50 - 0:60 Conclusion remarks and debate

References

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. https://doi.org/10.1126/science.aal4230

Kocher, M., Savoy, J., 2017. Distributed language representation for authorship attribution. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx046

Nanni, F., Dietz, L., Ponzetto, S.P., 2017. Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx062

Interactions: Platforms for Working with Linked Data

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Kim Martin
kimberleymartin@gmail.com
University of Guelph, Canada

Following on from a successful LOD workshop in Montreal that saw 30 plus people come together and discuss the potential for linked data in the humanities, we propose a workshop that focuses more specifically on interacting with Linked Data. There are many different platforms for working with linked data — for visualizing, creating, reconciling, cleaning, and analyzing it. Some of these tools have been developed from within the Digital Humanities community, and others have been developed beyond it but adapted to our purposes. We hope to create the opportunity for fruitful exchange by providing time for hands-on demonstration and discussion.

All participants will have the opportunity to submit the following in advance of the workshop:

1. Answering an online form that indicates the type of LOD tools or platforms, or features within these, that they wish to see discussed at the workshop.
2. Details on where their tool (if they have one) fits in, and a description of their work with LOD.
3. A description (1 page max) of their LOD platform with features that they wish to showcase during the workshop.

Outline/Schedule (Based on 6 hrs – Full Day)

Introductions	20 mins
Featured Tool Demos x4	30 mins
Coffee break + discussion	15 mins
Lightning Tool Demos x12	60 mins
Poster session and lunch	100 mins
(Participants wanting to do an afternoon workshop could shift to that at this point)	
Discussion of challenges, desiderata, next steps etc.	60 mins
Coffee break	15 mins
Time for breakout discussions (re: possible collaborations etc.)	60 mins

Workshop leaders:

Susan Brown (sbrown@uoguelph.ca) is a Canada Research Chair in Collaborative Digital Scholarship and Professor of English at the University of Guelph, and Visiting Professor at the University of Alberta. She researches Victorian literature, women's writing, and digital humanities. All of these interests inform *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*, an ongoing experiment in digital literary history published by Cambridge UP since 2006 that she co-directs. She directs the Canadian Writing Research Collaboratory, an online repository and research environment for literary studies in and about. Her current research touches on a range of topics in the digital humanities including interface design and usability, visualization and data mining, semantic technologies, and humanist-centered tool development. She is increasingly engaged with inquiry into how linked open data can serve humanities research. She also works on the impact of new technologies in the literature of the Victorian period. Brown is President of the Canadian Society for Digital Humanities/Société canadienne des humanités numériques.

Kim Martin is the Ridley Post-doc in Digital Humanities and the Associate Director of THINC Lab at the University of Guelph. Her PhD thesis focussed on the serendipitous experiences of historians during their research

process. She is currently working on linked data projects for Canadian Writing Research Collaboratory (CWRC) and on the Mellon-funded Records of Early English Drama - London (REED-London) project. Kim's interests in serendipity and linked data tie neatly together with her work on HuViz — the humanities visualizer, a tool developed by the CWRC team for interacting with linked open datasets.

Target Audience/Expected number of participants: 40
Special requirements for technical support:
Projector and screen. Ideally also boards of some kind (black, white, paper) on which to write, though we can bring something if need be. Decent internet access.
We will need one room that is big enough for 40 people.
Proposed budget:
The organizers will provide funds for lunch. Estimated cost: $300-400.

Call for Participation: We will put out a call for participation based upon this proposal within two weeks of confirmation that the workshop will run.

Deadline for submissions: We will ask for submissions by April 30th. We will make the papers available through a Dynamic Table of Contexts edition that expands the one from last year to other prospective participants and mount the ranking poll by May 5th and run the poll until May 31st. Applicants will be informed if they will be presenting and/or participating in the workshop by June 1st.

Program Committee:

Susan Brown, Professor of English
University of Guelph

Sharon Farnel, Metadata Coordinator
University of Alberta Libraries

Lisa Goddard, Academic Systems Librarian
University of Victoria

Karl Grossner, Geographer, DH Researcher
University of Pittsburgh

Abigel Lemak, PhD Student, English
University of Guelph

Kim Martin, Postdoc in DH
University of Guelph

Deb Stacey, Professor, Computer Science
University of Guelph

Building International Bridges Through Digital Scholarship: The Trans-Atlantic Platform Digging Into Data Challenge Experience

Elizabeth Tran
etran@neh.gov
National Endowment for the Humanities, United States of America

Crystal Sissons
crystal.sissons@sshrc-crsh.gc.ca
Social Sciences and Humanities Research Council, Canada

Nicolas Parker
nicolas.parker@sshrc-crsh.gc.ca
Social Sciences and Humanities Research Council, Canada

Mika Oehling
mika.oehling@sshrc-crsh.gc.ca
Social Sciences and Humanities Research Council, Canada

This workshop will focus on how international partnership can benefit large-scale research projects in digital scholarship. During the workshop, participants will learn about the Digging into Data Challenge 4, an initiative of the Trans-Atlantic Platform (T-AP) for Social Sciences and Humanities, a network of public funders representing countries in Europe, North America, and South America. The Digging into Data Challenge invited international teams to undertake multidisciplinary projects that use techniques of large-scale data analysis and demonstrate how these can lead to new insights. The Digging into Data Challenge has had four rounds of funding, and offers an valuable opportunity to (1) see how the international dimension benefits the scholarship; (2) understand the challenges of working internationally on big data projects addressing questions in the humanities and social sciences; (3) understand how international funding initiatives might enable research in ways that domestic funding cannot.

This workshop is targeted at (1) individuals who are interested in "scaling up" their research efforts to include an international dimension and (2) funders who are interested in launching or joining international funding opportunities. The workshop will touch on various themes that impact digital researchers and international collaboration, including:

- legal considerations,
- the intellectual challenges for large scale research,
- big data skills,
- funding policies and processes, and
- the challenges to international research collaboration for researchers from both small and large countries.

The workshop is scheduled as a full-day event so as to allow ample time for conversation and networking.

In order to better incorporate and interests of workshop participants and foster dialog and discussion, participants may provide a brief one-page synopsis outlining their interest in international collaboration and what they hope to gain from the workshop. The synopsis should be sent to: odh@neh.gov.

Herramientas para los usuarios: colecciones y anotaciones digitales

Amelia Sanz
amsanz@filol.ucm.es
Complutense University, Spain

Alckmar Dos Santos
alckmar@gmail.com
Federal University of Santa Catarina, Brazil

Ana Fernández-Pampillón
apampi@ucm.es
Complutense University, Spain

Oscar García-Rama
ogarcia@supportfactory.net
Support Factory, Spain

Joaquin Gayoso
jgayoso@ucm.es
Complutense University, Spain

María Goicoechea
mgoico@filol.ucm.es
Complutense University, Spain

Dolores Romero
dromero@filol.ucm.es
Complutense University, Spain

José Luis Sierra
jlsierra@ucm.es
Complutense University, Spain

Desde el año 2010, el Grupo de investigación LEETHI (Literaturas Europeas del tExto al Hipermedia) desde la Facultad de Filología y el grupo ILSA (Implementation of Language-Driven Software and Applications) desde la Facultad de Informática de la Universidad Complutense de Madrid trabajan juntos afin de dar respuesta concreta a necesidades de la docencia y la investigación en Humanidades en la UCM. Han diseñado y probado sistemas y herramientas que permiten tanto la construcción de repertorios digitales para colecciones propias como el comentario detallado y a la edición creativa; han desarrollado repertorios como las bibliotecas Mnemosine con el grupo LOEP o Ciberia y herramientas de anotación como @Note que mereció uno de los 12 premios otorgados por Google en 2010 dentro de su *Google's Digital Humanities Award Program*. LEETHI e ILSA han colaborado en proyectos de carácter europeo como la COST Action INTEREDITION o el grupo de trabajo en DARIAH, Women Writers in History; han participado con "short papers" en las DH Conferences en Hambourg, Lausanne y Montréal.

Ambos grupos colaboran desde el año 2006 con el grupo NUPILL (Nucleo de Pesquisas con Informatica, Literatura i Lingüística) de la Universidad Federal de Santa Catarina (Brasil). El grupo tiene como vocación la exploración de las posibilidades que las tecnologías ofrecían al desarrollo de la investigación sobrela lectura y la escritura literaria en el medio electrónico. Así ha desarrollado la Biblioteca de Literaturas de Lingua Portuguesa, revistas como Texto Digital, ediciones como la de Machado de Asís. Estos grupos han organizado seminarios conjuntos en Florianópolis y en Madrid, y han participado en congresos en Brasil, Francia y España.

Support Factory es una empresa dedicada a la investigación y creación de software seguro y estable para entornos educativos a partir de un acompañamiento del usuario durante todo el proceso de configuración, prueba y utilización. Colabora con el Grupo LEETHI en el diseño de software para la edición electrónica amigable desde 2017.

Todos comparten una búsqueda de:

- estrategias de desarrollo de utilidades, sistemas y competencias de abajo arriba: desde las necesidades de los usuarios expertos y para los usuarios en la docencia y la investigación en Humanidades;
- cooperación horizontal que permita mutualizar y mancomunar tecnologías digitales en español y en portugués;
- desarrollos que generen masa crítica científica, pero también innovación social y beneficio comerical;
- respuestas a necesidades de investigación y de enseñanza localizadas: una solución tecnológica global no funciona necesariamente en todos los espacios, cuando sus modelos no corresponden con prácticas culturales y soberanía epistemológica de los lugares.

Los grupos están en condiciones de ser actantes en el campo de las Humanidades Digitales globales y de visibilizar sus prácticas, con el fin de contrastarlas, ponerlas al servicio de la comunidad investigadora que habla español y portugués y pluralizar así sus funcionalidades.

Propopen presentar las siguientes herramientas:

- CLAVY (http://clavy.fdi.ucm.es), desarrollada por ILSA, permite la administración de colecciones digitales heterogéneas, para:
  - agregar colecciones externas gracias a una potente arquitectura de importación;
  - integrar y unificar estas colecciones a través de una representación explícita de sus estructuras

mediante ontologías reconfigurables en un entorno amigable para expertos en Humanidades;
- transformar la arquitectura del plug-in a un nivel más profundo con la intervención de programadores;
- exportar las colecciones a plataformas externas gracias a una arquitectura de exportación para ello.

La plataforma también permite gestionar un sistema extensible para alimentar y refinar las colecciones, de forma que la arquitectura para integrar los plug-ins específicos de edición permita adaptar el flujo de autor a las necesidades específicas de cada campo. Así resulta posible integrar de manera sencilla, por ejemplo, la geolocalización de los objetos o la anotación de recursos.

La plataforma ha sido ya utilizada en Mnemosine y Ciberia. Está especialmente destinada a investigadores-profesores que quieren diseñar su propio repertorio y adaptarlo a sus propias necesidades según avanza su construcción.

- *@Note* (http://anote.fdi.ucm.es), desarrollada por ILSA/LEETHI, permite la anotación de textos digitalizados. Es el resultado del *Google's Digital Humanities Award Program* 2010. La herramienta permite a los profesores-investigadores crear actividades de anotación de textos en modo imagen apoyándose en ontologías con las que clasificar las anotaciones de forma colaborativa, tanto para definir las ontologías, como para realizar las anotaciones. La herramienta permite incluir notas multimedia e hilos de discusión asociados a cada anotación, así como la utilización de esas anotaciones para escribir ensayos, comentarios o realizar análisis. Está especialmente diseñada para el análisis de textos en la enseñanza universitaria y secundaria, al alcance de cualquier profesor con sus estudiantes.
- DLNOTES2 (http://www.dlnotese2.ufsc.br), creada por NUPILL, permite hacer anotaciones libres y semánticas en obras digitalizadas y en modo texto. Se proporcione a los alumnos la posibilidad de 1) hacer comentarios en cualquier extracto de un texto; 2) gestionar las anotaciones para análisis de la lectura (por parte del profesor o incluso del estudiante) o para elaboración de revisiones de la obra leída. En cuanto a las anotaciones semánticas („semántica" en el sentido que le atribuyen las Ciencias de la Computación), el propósito es asociar conceptos de la teoría literaria a cualquier secuencia de caracteres de la obra, lo que ayuda a los estudiantes a comprender esos conceptos y a desarrollar lecturas en perspectivas diferentes (y complementarias) de la lectura tradicional.
- AOIDOS(http://aoidos.ufsc.br), desarrollada por NUPILL, realiza escansiones automáticas en corpora textuales poéticos, señalando todos los fenómenos fonéticos necesarios a la realización rítmica de los versos, y posibilitando, además, que los lectores tengan acceso a todos los datos cuantificados y organizados de varias maneras. Su funcionamiento en portugués está ya demostrado y se mostrará una versión en español para este taller.
- CONTENT-AWAY: es una herramienta que permite a alumnos y profesores crear sus libros a medida, tanto físicos como digitales, mejorar el flujo de información entre profesores y alumnos, facilitar el acceso a materiales de estudio en differentes formatos y soportes. Partiremos del desarrollo de Enclave realizado para la Real Academia Española y mostraremos a los participantes como adaptar y comenzar a utilizar la herramienta para sus propias necesidades docentes.

El taller persigue dos objetivos:

- los usuarios podrán calibrar la posibilidad de elaborar su propio repertorio digital con sus propias categorías, su propio sistema de anotación de textos;
- intercambiaremos modelos y propuestas con otros investigadores y expertos de forma que sea posible compartir y enriquecer sistemas y herramientas.

Descripción de la audiencia:
- Por un lado, profesores-investigadores sobre literaturas en soporte electrónico que quieran conocer las funcionalidades de estas herramientas y adaptarlas a su propio uso; por otro, programadores que quieran conocer las claves de las herramientas para adaptarlas a su entorno.
- Podemos asumir hasta 60 participantes solo para presentaciones.

Necesidades para la celebración:
- Se requerirá una sala con ordenadores o, en su defecto, se requerirá a los participantes que aporten sus propios equipos. En cualquier caso, será necesaria una buena conexión a Internet.
- Todas las sesiones se desarrollarán en portugués y en español, según el equipo que presente cada herramienta.

References

Mittmana, A.; Samanta R. M. ; dos Santos A. L. Análise comparativa entre as escansões manual e automática dos versos de Gregório de Matos A comparative analysis between automatic and manual scansions of Gregório de Matos' verses. Texto Digital, Florianópolis, v. 1, n. 1, p. 157-179, jan./jun. 2017.

Gayoso-Cabada, J., Rodríguez-Cerezo D., Sierra, J.L. Browsing Digital Collections with Reconfigurable Faceted Thesauri. En J. Gołuchowski, M. Pańkowska, H. Linger, C. Barry, M. Lang & C. Schneider (Eds.), *Information Systems Development: Complexity in Information Systems Development (ISD2016 Selected and Extended Papers)*. Lecture Notes in Information Systems and Organisation Vol. 22, pp. 69-86.

Gayoso-Cabada, J., Rodríguez-Cerezo D., Sierra, J.L. Multilevel Browsing of Folksonomy-Based Digital

Collections. In Wojciech Cellary, Mohamed F. Mokbel, Jianmin Wang, Hua Wang, Rui Zhou, Yanchun Zhang (eds): *Web Information Systems Engineering - WISE 2016 - 17th International Conference, Shanghai, China, November 8-10, 2016, Proceedings, Part II. Lecture Notes in Computer Science 10042,* pg. 43-51. 2016.

Where is the Open in DH?

Wouter Schallier
wouter.schallier@un.org
UN/ECLAC, Chile

Gimena del Rio Riande
gdelrio@conicet.gov.ar
CONICET, Argentina

April M. Hathcock
april.hathcock@nyu.edu
New York University, United States of America

Daniel O'Donnell
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

When it comes to promoting the importance of open scholarship, Latin America and the Caribbean stand out in a sense that the concept of "openness" is generally accepted all over the region. Several countries, such as Peru, Argentina, Brazil and Mexico, have shown real advances in terms of national laws that seek to make knowledge produced with public funds a common good, managed by the academic community. We can also highlight regional projects such as SciELO and redalyc.org that have played a unique role to make the production published in Ibero American and Latin American journals available free of charge. Open access is now established in Latin America and the Caribbean as the most extended communication model in the academic community, giving visibility and value to scientific production at a regional and global level.

Nevertheless, the question remains to what extent this wide acceptance of openness has influenced the work of digital humanists in Latin America and the Caribbean and beyond. Much of the most well-known digital humanities (DH) work in the world tends to focus on projects coming out of North America and Western Europe. And despite efforts by groups such as Global Outlook::Digital Humanities (GO::DH) and the Alliance of Digital Humanities Organisations more broadly, DH still remains a very English language centric interdiscipline (Fiormonte y del Rio Riande, 2017).

What would it take to bring DH into a more global openness, not only in terms of access but also in terms of methods, best practices and opportunities for collaboration? And what could this openness look like set against the backdrop of the long-standing and highly developed open access movement in Latin America and the Caribbean?

The workshop will analyse these challenges, as well as highlight initiatives and explore options to advance open in DH in Latin America and the Caribbean. It will begin by examining the aforementioned national laws and specific cases that illustrate the progress and challenges of open access as a movement in Latin America and the Caribbean, as well as in the global context and present a practical approach to deal with the „different open accesses in the world" (Curry, 2017; Babini, 2013). The workshop will then shift to focus on the ways these various infrastructures for open can be deployed to build a more globally open DH.

Furthermore, the workshop will highlight particular existing DH projects that have begun building openness, in access, methods, and collaboration. Instructors and facilitators will help attendees to explore examples from the Global North and South, such as the LEARN project (http://www.learn-rdm.eu/), CLACSO's activities (https://clacso.org.ar/), Red Argentina de Educación Abierta (AREA. http://a-rea.org/), Cientópolis (https://www.cientopolis.org/), Acta Académica (https://www.aacademica.org/), Humanities Commons (https://hcommons.org) , OpenCon (http://www.opencon2017.org/), FORCE11 (https://www.force11.org/), DARIAH (https://www.dariah.eu/), among others, to begin building a set of good practices, including examples of institutional policies and practical recommendations from Europe and Latin America and the Caribbean devoted specifically to DH projects. We will give examples of Open projects in DH, in this set of good practices, institutional policies and practical recommendations that will address project work, digital objects, Open Access publishing and research collaboration.

Finally, the workshop will place DH output modes, from collaborative web projects to traditional publications to research data, in the context of the larger open access movement, which is changing the face of academic research and society in a very profound way. This vision of open access is creating a global environment where researchers, innovators, and citizens can publish, find, use and reuse each other's data, tools, publications and other outputs for research, innovation and educational purposes.

In addition to people interested specifically in the case of Latin America and the Caribbean, this course will be of comparative interest to people working in other regions in both the Global South and the Global North. We will encourage participants to engage reflectively with the material, bringing the own experiences to bear.

References

Arévalo, A. J. (2016). Análisis de los estudios sobre las ventajas del acceso abierto y la ventaja de cita. *Blog de la biblioteca de Traducción y Documentación de la Universidad de Salamanca.* https://universoabierto.org/2016/05/29/analisis-de-los-estudios-sobre-las-ventajas-del-acceso-abierto-y-la-venta-

ja-de-cita/ (last visit: 27 April 2018)
Alperin, J.P. (2015). The Public Impact of Latin American's Approach to Open Access. https://stacks.stanford.edu/file/druid:jr256tk1194/AlperinDissertationFinalPublicImpact-augmented.pdf (last visit: 27 April 2018)
Babini, D. (2013). Open access initiatives in the Global South affirm the lasting value of a shared scholarly communications system. *London School of Economics and Political Science Impact Blog* http://blogs.lse.ac.uk/impactofsocialsciences/2013/10/23/global-south-open-access-initiatives/ (last visit: 27 April 2018)
Curry, S. (2017). Why I don't share Elsevier's vision of the transition to open access http://occamstypewriter.org/scurry/2017/10/03/why-i-dont-share-elseviers-vision-of-the-transition-to-open-access/ (last visit: 27 April 2018)
Fernándes, P. and Vos, R. A. (2017). Open Science, Open Data, Open Source. https://pfern.github.io/OSOD-OS/gitbook/ (last visit: 27 April 2018)
Fiormonte, D. and del Rio Riande, G. (2017). Por unas Humanidades Digitales Globales. https://infolet.it/2017/10/09/humanidades-digitales-globales/ (last visit: 27 April 2018)
Packer, A. L. et al. (2018). Los criterios de Indexación de SciELO se alinean con la comunicación en la ciencia abierta. *SciELO en Perspectiva.* http://blog.scielo.org/es/2018/01/10/los-criterios-de-indexacion-de-scielo-se-alinean-con-la-comunicacion-en-la-ciencia-abierta/#.WocMbujwaM9 (last visit: 27 April 2018)
Suárez, A. V. and McGlynn, T. (2017). The fallacy of Open-Access Publication. *The Chronicle of Higher Education.* https://www.chronicle.com/article/the-fallacy-of-open-access/241786

Indexing Multilingual Content with the Oral History Metadata Synchronizer (OHMS)

Teague Schneiter
tschneiter@oscars.org
Academy of Motion Picture Arts & Sciences, United States of America

Brendan Coates
bcoates@oscars.org
Academy of Motion Picture Arts & Sciences, United States of America

Brief Description:

Are you in need of a way to provide access to oral histories not recorded in English? Do you have dreams of creating multilingual metadata for interviews recorded in English to non-English speaking communities/users? In late 2016 The University of Kentucky Nunn Center updated the OHMS application and viewer to have multilingual functionalities, creating the capability to synchronize both a transcript/translation, as well as to create a bilingual index, making all of these searchable and synchronized to the corresponding moment in the audio or video. In this mini-workshop OHMS power users Teague Schneiter and Brendan Coates will demonstrate the multilingual functionalities of OHMS. Through demonstration of a bilingual use case, instructors will walk attendees through each step of the indexing process to prepare a sample Spanish-English index. Instructors will also guide attendees to develop workflows to support multilingual indexing.

Workshop outline:

1. OHMS intro
    a. History
    b. Development
    c. Multilingual Functionality

OHMS basics - in conjunction with a worksheet sent prior to conference setting up an account
    a. Linking a video
    b. adding thesaurus/ ontology/ data dictionary
    c. End user functionality, switching languages, etc.

Basic indexing
    a. Instructors lead group in indexing a video
    b. in small groups or pairs, index a video

Multilingual indexing
    a. Instructors lead group in creating a multilingual index of a video

*Note - participants do not need to be bilingual, videos can be indexed as Index1 - Index2 instead of English - Non-English

Instructors

Teague Schneiter is an Audiovisual Archivist, Project Manager and Strategist who is currently working for the Academy of Motion Picture Arts and Sciences as Senior Manager (and founder) of the Oral History Projects department, instituting digital content initiatives around filmmaker oral histories. Her professional experience spans moving image preservation and access infrastructure, with the bulk of her experience in human rights and cultural heritage content. Her work at WITNESS solidified her interest in web knowledge management projects and in people organizing, and concretized a firm belief that communication technologies in the digital age should facilitate openness, innovation, participation among individuals and communities, and should further social change. In the past she has worked as a long-term consultant for indigenous media organization IsumaTV, focused mostly on outreach, strategic planning, knowledge-sharing and social media.

Brendan Coates is an Audiovisual Archivist and Preservationist, currently working at the Academy of Motion Pic-

ture Arts and Sciences, where he oversees the ingest, description, preservation, and dissemination of the Academy's Oral History holdings. Prior to this, he ran the UCSB Library's audiovisual digitization and preservation program, including its Cylinder Audio Archive, and its participation in the Library of Congress National Jukebox project and the Discography of American Historical Recordings (DAHR). His research interests are grounded in workflow automation and quality control, ensuring that video is digitized to appropriate standards and is playable and accessible long into the future.

Target Audience:

Our target audience is anybody working with video assets who would like to make subject/ language/ community specific, time-based metadata to describe them. We're anticipating about 20 people.

Technological support:

The workshop will require a computer with projection, WiFi and participants will need their own workstations or to bring their own laptops.

Sig Endorsed

Distant Viewing with Deep Learning: An Introduction to Analyzing Large Corpora of Images

Taylor Baillie Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Lauren Craig Tilton
ltilton@richmond.edu
University of Richmond, United States of America

Short Description

This tutorial provides a hands-on introduction to the use of deep learning techniques in the study of large image corpora. The TensorFlow and Keras libraries within the Python programming language are used to facilitate this analysis. No prior programming experience is required.

Image analysis tasks covered in the tutorial include object detection, facial recognition, image similarity, and image clustering. We will make three open-access image corpora (historic photographs, still frames from moving images, and scanned works of art) available in order to test these methods. Alternatively, participants may bring and use an image dataset of interest to them. At the conclusion of the tutorial, participants will have created an interactive website running locally on their machines. This website will provide tools for analyzing their selected dataset. Additional instructions for making the website publicly available will be provided.

Audience and Number of Participants

This tutorial is aimed at scholars who work with visual materials who want to integrate DH methods into their analysis of image corpora. Our tutorial is based off of lectures notes used in a non-major, undergraduate-level course at the University of Richmond. It is accessible to participants with little to no programming background. However, as the tutorial will focus on the methods behind image processing rather than low-level coding, it will also be interesting and useful for experienced programmers new to image processing.

Following the large number of participants at the AVinDH SIG sponsored Workshop in Montreal for DH20167 and our popular tutorial at DH2016 in Krakow, we expect the workshop participation to be equally popular with somewhere between 15 and 25 participants.

Presenter Information

Taylor Arnold is Assistant Professor of Statistics at the University of Richmond. A recipient of grants from the NEH and ACLS, Arnold's research focuses on computational statistics, text analysis, image processing, and applications within the humanities. His first book *Humanities Data in R* (Springer, 2015) explores four core analytical areas applicable to data analysis in the humanities: networks, text, geospatial data, and images. His second book, the forthcoming *A Computational Approach to Statistical Learning* (CRC Press 2018), explores connections between modern machine learning techniques with theories of statistical estimation. Numerous journal articles extrapolate on these ideas in the context of particular applications. Arnold has also released several open-source libraries in R, Python, Javascript and C. Visiting appointments have included Invited Professor at Université Paris Diderot and Senior Scientist at AT&T Labs.

Lauren Tilton is Assistant Professor of Digital Humanities in the Department of Rhetoric and Communications at the University of Richmond and a member of Richmond's Digital Scholarship Lab. Her current book project focuses on participatory media in the 1960s and 1970s. She is the Co-PI of the project *Participatory Media*, which interactively engages with and presents participatory community media from the 1960s and 1970s. She is also a director of *Photogrammar*, a web-based platform for organizing, searching and visualizing the 170,000 photographs from 1935 to 1945 created by the United States Farm Security Administration and Office of War Information (FSA-OWI). She is the co-author of *Humanities Data in R* (Springer, 2015). She is co-chair of the American Studies Association's Digital Humanities Caucus.

Detailed Outline

In this three hour tutorial we plan to spend the first 15 minutes getting all participants set up with the software and datasets required for the tutorial. The tutorial participants will we able to work on any reasonably recent version of Windows, macOS, or Linux. All of the software is free and open source. The remained of the workshop will consist of two 75-minutes sessions with a 15 minute break between them.

Each of the two 75-minute sessions will consist of working collectively through "labs" formatted as IPython notebooks. Participants will have the option of using one of three pre-compiled datasets during the workshop depending on their interests:

- historic photographs
- still frames from moving images
- scanned works of art

Alternatively, tutorial participants may alternatively work with their own collection of images.

The first session will focus on describing the potential difficulties of working with image data and explaining how deep learning can be used to address several of the-

se challenges. Working at a conceptual level we will work through the following tasks:

- how to structure a large collection of images as files on a computer
- how to load images into Python as multidimensional arrays
- the concepts behind applying neural networks to image data
- code for projecting images into the penultimate layer of the YOLOv4 neural network
- methods for visualizing the output projects from the neural networks

The second session will focus on how the features detect in the first session can be used to annotate higher level features and measure the similarity between images. Specifically:

- the application of image projections to image similarity metrics
- the application of image projections to object detection
- the application of image projections to face detection

In the final 30 minutes, we will discuss how these techniques ultimately can be used to address humanities questions. This will culminate in running Python code that will output the constructed annotations as an interactive website running locally on each user's computer. This will open up further possibilities for extending the methods of the tutorial without the need for an extensive programming background.

References

Arnold, T. and Tilton, C. (2015). *Humanities Data in R*. New York, NY: Springer.

Arnold, T., Kane, M., and Lewis, B. (2017). *A Computational Approach to Statistical Learning*. New York, NY: CRC Press.

The re-creation of Harry Potter: Tracing style and content across novels, movie scripts and fanfiction

Marco Büchler
mbuechler@etrap.eu
University of Göttingen, Germany

Greta Franzini
gfranzini@etrap.eu
University of Göttingen, Germany

Mike Kestemont
mike.kestemont@uantwerpen.be
University of Antwerp, Belgium

Enrique Manjavacas
enrique.manjavacas@uantwerpen.be
University of Antwerp, Belgium

The tutors

This one-day tutorial will be given by Marco Büchler, Greta Franzini, Mike Kestemont and Enrique Manjavacas.

Endorsement: This workshop is formally endorsed by the Special Interest Group on *Digital Literary Stylistics* (SIG-DLS).

Mike Kestemont (mike.kestemont@uantwerpen.be) is assistant research professor in the department of Literature at the University of Antwerp. He specializes in computational text analysis for the Digital Humanities, in particular stylometry and machine learning, topics on which he has given dozens of hands-on courses. Whereas his work has a strong focus on historical literature, his present research projects cover a wide range of topics in literary history, including classical, medieval, early modern and modernist texts. Mike currently takes a strong interest in representation learning via neural networks.

Marco Büchler (mbuechler@etrap.eu) is a computer scientist and leader of the *Electronic Text Reuse Acquisition Project* (eTRAP) research group at the University of Göttingen. Marco's research interests concern the processing of natural languages with a specialization in the detection of historical text reuse. Furthermore, he is interested in the mining process and the systematization of changes of text reuse. He has worked in this field for over eight years. Together with his eTRAP team, in the past three years he has organized ten text reuse tutorials.

Greta Franzini (gfranzini@etrap.eu) is a Classicist and member of the *Electronic Text Reuse Acquisition Project* (eTRAP) research group at the University of Göttingen. Greta's research interests concern the production of digital editions of texts as well as the combination of quantitative and qualitative methods to advance computational analyses and linguistic resources for Classical literature. Together with her team, Greta has already given eight text reuse tutorials.

Enrique Manjavacas (enrique.manjavacas@uantwerpen.be) is a PhD student at the University of Antwerp. He is associated with the Antwerp Centre for Digital Humanities and Literary Criticism. His current research focuses on sequential methods based on recurrent neural networks to develop semantically-infused models for Stylometry and text reuse detection. He is also interested in Natural Language Generation and has been involved in various projects around the concept of Synthetic Literature.

Description

Computer-assisted text analysis is a core research area in the Digital Humanities. It embraces a wide variety of applications (stylometry, text reuse detection, topic modelling, etc.) and can assist researchers in complex tasks, particularly when it comes to processing large amounts of text. This tutorial brings together two popular and complementary text analysis tasks, stylometry (the quantitative study of writing style) and text reuse detection. While stylometry typically focuses on stylistic similarities between texts (i.e. *how* texts are written), text reuse studies are geared towards the reuse of elements across works (i.e. *what* texts are written about). As such, both methodologies tie into the theoretical notion of *intertextuality* (Orr 2003), albeit in complementary ways.

Creativity and individuality are important phenomena at stake in both fields: are writers at liberty to escape their own 'stylome' - or unique stylistic fingerprint - and to which extent can they free themselves from the many predecessors to which they are intertextually indebted? (Harold Bloom (1973) famously spoke of the 'Anxiety of Influence' in this respect) This leads to interesting theoretical tensions: if authors are stylistically close to one another, does that imply that we can also expect a more elevated level of text reuse between them (and vice versa)? Or can authors frequently reuse textual elements while developing an independent stylistic profile? To which extent is it theoretically possible to oppose style and content?

In this workshop we offer a hands-on introduction to these topics using the case study of Rowling's Harry Potter novels. The vast body of academic scholarship of these writings attests to the relevance of this series, including the highly mediatized stylometric study by Patrick Juola (2013) unmasking Rowling as "Robert Galbraith", the pseudonym under which she temporarily managed to escape her own fame. Intertextuality is also a major concern of Rowling scholarship and scholars as Karin Westman (2007) have meticulously analyzed Rowling's nuanced indebtedness to British authors such as Jane Austen. Rowling herself has invited much intertextual offspring by now too, not in the least in the form of so-called fanfiction (Milli & Bamman 2016), the global phenomenon where (typically non-professional) writers read, reinterpret and expand literary universes (*fandoms*) originally created by acclaimed authors in their own writings (or *fanfics*).

The workshop's tutorial will focus on offering scholars the practical tools and skills to begin to tackle such complex issues. For text reuse detection, participants will learn how to operate TRACER, a language-independent suite of state-of-the-art Natural Language Processing (NLP) algorithms aimed at discovering text reuse in both historical and modern texts, helping users to identify different types of text reuse ranging from verbatim quotations to paraphrase. For the stylometric analyses and visualizations, participants will mainly use custom scripts that exploit the numerous possibilities of the popular Python library *scikit-learn* for Machine Learning. Stylometry with R (Eder et al. 2016), a software package for text analysis in *R*, is another tool that will be used in the introductory sessions.

Data

Participants will practise with data provided by the organizers to better familiarize themselves with the software. The texts under analysis will be the seven English language Harry Potter novels by J. K. Rowling (the so-called core canon of the fandom), a large corpus selection of Harry Potter fanfiction (harvested from *Archive of Our Own*) as well as the Harry Potter movie subtitles.

Objectives

The first objective of the tutorial is to introduce participants to two popular applications of text analysis that tie in closely with intertextuality studies, providing them with an understanding of some of the challenges, methods and strategies proper to this area of research. To this end, we use the illustrative Rowling case study to identify which proportion of the original novels and how much of their style the movies and fanfiction both retain. Additionally, the tutorial seeks to equip participants with the necessary knowledge to independently use the demonstrated software at home (and on their own corpora). Finally, it introduces visualization techniques to display results in an intuitive fashion, provoking new hermeneutic questions.

References

Bloom, H. (1973). *The Anxiety of Influence: A Theory of Poetry*. Oxford, New York: Oxford University Press.

Eder, M., Rybicki, J., Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8: 107–121.

Juola, P. (2013). Rowling and "Galbraith": an authorial analysis. *Language Log*. http://languagelog.ldc.upenn.edu/nll/?p=5315 (accessed 2 May 2018).

Milli, S., Bamman, D. (2016). Beyond Canonical Texts: A Computational Analysis of Fanfiction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2048–2053. https://doi.org/10.18653/v1/D16-1218.

Orr, M. (2003). *Intertextuality: Debates and Contexts*. Polity.

Westman, K.E. (2007). Perspective, Memory, and Moral Authority: The Legacy of Jane Austen in J. K. Rowling's Harry Potter. *Children's Literature*, 35: 145–165. https://doi.org/10.1353/chl.2007.0021.

Archiving Small Twitter Datasets for Text Analysis: A Workshop for Beginners

Ernesto Priego
efpriego@gmail.com
City, University of London, United Kingdom

Abstract

In this workshop for non - coders, participants will be guided through two tasks: the first task will guide participants in creating an application to tap into Twitter's API, in our case to get Twitter data. The second task will guide participants in the use of a Google spreadsheet to capture streaming (live) data from Twitter in order to archive it, download it and perform text analysis, data visualization and other studies. This workshop will include a brief introduction contextualizing social media data collection good practices including user data privacy issues.

Rationale

Twitter data can be very valuable for researchers of perhaps all disciplines, not just DH. Given the difficulties to properly collect and analyse Twitter data as viewable from most Twitter Web and mobile clients (as most people use Twitter) and the very limited short - span of search results, there is the danger of losing huge amounts of valuable historical material.

Tweets are like butterflies – one can only really look at them for long if one pins them down out of their natural environment. The reason why we have access to Twitter in any form is because of Twitter's API, which stands for Application Programming Interface. Free access to historic Twitter search results is limited to the last 7 days. This is due to several reasons, including the incredible amount of data that is requested from Twitter's API, and – this is an educated guess – not disconnected from the fact that Twitter's business model relies on its data being a commodity that can be resold for research. Twitter's data is stored and managed by Twitter's enterprise API platform.

For the researcher interested in researching Twitter data, this means that harvesting needs to be do ne not only through automated means but in real time. It also puts scholars without the required coding and data mining skills at a disadvantage. As a researcher, this basically means that there is no way to do proper research of Twitter data without understanding how it works at API level, and this means understanding the limitations and possibilities this imposes on researchers.

What's a n individual researcher without access to pay corporate access to do? The whole butterfly colony cannot be captured with the nets most of us have available. At small scale, however, and collecting in a timely fashion, it is still possible to capture interesting and more - or - less complete specimens using fairly simply, non - coding required methods. (The Library of Congress h s now 12 years' worth of text - only Tweets. However, as before, the Library of Congress Twitter collection will remain embargoed and there was no projected timetable for providing public access as of 26 December 2017).

Most researchers out there are likely not to benefit from access to huge Twitter data dumps. For researchers without much resources that are trying to do the talk whilst doing the walk, and conduct research *on* Twitter and *about* Twitter, this workshop and tutorial will guide participants into creating a Twitter application in order to tap into the Twitter API, followed

by the setting up of a Twitter Google Archiving Spreadsheet. Once a trial archive or dataset has been collected, we will attempt text analysis and basic visualisations using Excel and Voyant Tools. This workshop will include a brief introduction contextualizing social media data collection good practices including user data privacy and research ethics issues.

References

Priego, E. 2018. #rfringe17: Top 230 Terms in Tweetage. https://epriego.blog/2017/08/05/rfringe17-top 230-terms-in-tweetage/Accessed 30 January 2018]

Priego, E., 2016. Bar Chart: Number of #DH2016 Tweets in Archive per Conference Day (Sunday 10 to Friday 15 July 2016 GMT). Available from: https://figshare.com/articles/Bar_Chart_Number_of_DH2016_Tweets_in_Archive_per_Conf erence_Day_Sunday_10_to_Friday_15_July_2016_GMT_/3490001/1 [Accessed 31 Jan 2018].

Priego, E. 2016. "Stronger In": Looking Into a Sample Archive of 1,005 StrongerIn Tweets. https://epriego.blog/2016/06/21/stronger-in-looking-into-a-sample-archive-of-1005- strongerin-tweets/ [Accessed 30 January 2018]

Priego, E. and Zarate, C., 2014. #MLA14 Twitter Archive, 9 - 12 January 2014. Available from: https://figshare.com/aticles_MLA14_Twitter_Archive_9_12_January_2014/924801/1 [Accessed 31 Jan 2018].

Bridging Justice Based Practices for Archives + Critical DH

T-Kay Sangwand
sangwand@gmail.com
UCLA, United States of America

Caitlin Christian-Lamb
caitlin.christianlamb@gmail.com
University of Maryland, United States of America

Purdom Lindblad
purdom@umd.edu
University of Maryland, United States of America

As scholars and practitioners in digital humanities, we create, analyze, trouble, and reference "the archive," though are often signaling vastly different (mis)understandings of archives, archivists, and archival practices. While both archivists and digital humanists engage critical questions around shared areas of practice (i.e. access, labor, privacy) these conversations often occur in parallel spheres with little recognition of the intellectual contributions in the distinct yet intersecting fields of archives and DH. This workshop aims to bridge the discourse occurring in critical archival studies and critical digital humanities by engaging participants in articulating justice based practices related to appraisal, access, description, pedagogy, privacy, provenance, and system design, as well as collectively contribute these suggested practices to expand existing resources on critical archives and DH (Caswell et al., 2017). At their best, archives and digital humanities center voices that have been obscured through negligence or violently silenced from mainstream narratives. In the face of increased criminalization of and violence towards people of color, immigrants, journalists, mounting militarization, consolidation of media outlets, the political, social, and material impacts of climate change, global capitalism, and white supremacy, we feel a renewed sense of urgency to surface, highlight, and empower narratives from marginalized groups as a tool for social justice and envision new critical archives and digital humanities realities while not recreating oppressive and exploitative power dynamics in the process. This workshop is inspired by Rasheedah Phillips call to articulate "oral futures" and "speaking into existence of what you want to happen" (Phillips, 2017) as well as Michelle Caswell's classroom exercise to "collectively strategize concrete steps to dismantle white supremacy" (Caswell, 2017). The workshop will address the following questions: What are the archival processes of appraisal, accession, description, and access that shape the materials that we can use/collect/analyze as digital scholars and practitioners? How do archivists exercise agency at these various points in an archives' life cycle? What power do researchers/users exercise in their use and (re)presentation of archives? How are communities represented in archives impacted by the use of their archives? What are our collective and individual responsibilities to issues of privacy, description, and access to the materials we collect, analyze, and publish? How can we interrogate archival and scholarly "best practices" and work towards ethical and just practices? How can investigating these overlaps better identify points of collaboration and promote better understandings of cultural heritage across a range of roles, disciplines, and publics?

References

Caswell, M. (2017). Teaching to Dismantle White Supremacy in Archives. *Library Quarterly: Information, Community, Policy*. 87 (3): 222-235. https://doi.org/10.1086/692299.

Caswell, M. et al. (2017). Critical Archival Studies: An Introduction. *Journal of Critical Library and Information Studies Special Issue: Critical Archival Studies*. 1 (2). https://doi.org/10.24242/jclis.v1i2.50.

Phillips, R. (2017). Time, Memory, and Justice in Marginalized Communities. Instagram post. April 23. https://www.instagram.com/p/BTODUEmBZpK/?taken-by=communityfutureslab

Academic Reviewers

Aalberg Trond
Abdul-Rahman Alfie
Adams Robyn Jade
Akbulut Muge
Akça Sümeyye
Albritton Benjamin Long
Alexander Marc
Allés Torrent Susanna
Alpert-Abrams Hannah
Alvarado Rafael
Alzetta Chiara
Anderson Deborah
Anderson Wendy
Anderson Clifford Blake
Andreev Vadim Sergeevich
Andrews Tara Lee
Antonijevic Smiljana
Appleford Simon James
Applegate Matt
Arbuckle Alyssa Emily
Armaselu Florentina
Arneil Stewart
Arora Shaifali
Arriaga Eduard
Arthur Paul
Auddy Purbasha
B Ferronato Priscilla
Babeu Alison L.
Bailey Christopher Scott
Baillot Anne
Baker James William
Bamman David
Bandmann Megyesi Beata
Bangert Daniel Fritz
Barbaresi Adrien
Bardiot Clarisse
Barnett Tully
Barth Florian
Barthonnat Céline
Batjargal Biligsaikhan
Bauer Jean Ann
Baumann Ryan Frederick
Beals M. H.
Beaudouin Valérie
Beelen Kaspar
Bégnis Hélène
Beierle Christoph
Bellandi Andrea
Bellia Angela
Bender Michael
Benedict Nora Christine
Bénel Aurélien
Berens Kathi Inman
Berra Aurélien
Beshero-Bondar Elisa
Bessette Lee
Bhattacharyya Sayan
Bizzoni Yuri
Blümm Mirjam
Bon Bruno
Bonds Elizabeth Leigh
Boot Peter
Borbinha José
Bordalejo Barbara
Borgna Alice
Bornet Philippe
Borovsky Zoe
Bosse Arno
Bouchard Matthew
Bourget Nicolle
Bourgne Gauvain
Boyd Jason Alexander
Boyles Christina
Bozia Eleni
Brando Carmen
Braunstein Laura
Brown Susan
Brumfield Ben
Brumfield Sara
Brussa virginia
Büchler Marco
Burghardt Manuel
Burr Elisabeth
Burrows Toby Nicolas
Cafiero Florian Raphaël
Calvo Tello José
Câmara Alexandra Gago
Campagnolo Alberto
Camps Jean-Baptiste
Cao Ling
Cardillo Elena
Carlton Patricia Lynn
Casarosa Vittore
Casenave Joana
Casties Robert
Caton Paul
Cavanaugh Erica Fallon
Cayless Hugh
Chammas Michel
Charles Valentine
Chartrand Louis
Château-Dutier Emmanuel
Chavez Villa Micaela
Chawla Swati
Cheesman Tom
Chen Shih-Pei
Chen Kuang-hua
Chen Jing
Chiaravalloti Maria Teresa
Childress Dawn
Chuang Tyng-Ruey
Chue Hong Neil

Ciotti Fabio
Ciula Arianna
Clavert Frédéric
Clement Tanya
Clivaz Claire
Cochrane Euan
Cohen Hart
Colavizza Giovanni
Conway Paul
Cooney Charles M.
Cordell Ryan
Cotarelo-Esteban Lucia
Courboulay Vincent
Cowan William
Cowan T.L.
Craig Hugh
Crawford Cole Daniel
Crompton Constance
Croxall Brian
Cummings James
Curado Malta Mariana
Dabbs Thomas Winn
Dadvar Maral
Daengeli Peter
Dahlstrom Mats
Dallachy Fraser James
Dalmau Michelle
Damerow Julia Luise
Davis Rebecca Frost
De la Cruz Fernandez Paula
De la Rosa Pérez Javier
De Roure David
De- Matteis Lorena Marta Amalia
Declerck Thierry
Degaetano-Ortlieb Stefania
Del Grosso Angelo Mario
Del Rio Riande Gimena
Delve Janet
Derven Caleb
Devaney Johanna
Di Bacco Giuliano
Di Cresce Rachel
Di Donato Francesca
Di Ludovico Alessandro
Dilley Paul
Dogruoz Seza
Dombrowski Quinn
Dorn Amelie
Duckett Victoria
Dunst Alexander
Dussault Jessica Valerie
Eccles Kathryn
Eckart Thomas
Eckert Kai
Eder Maciej
Edmond Jennifer C
Ehrmann Maud
Eichmann-Kalwara Nickoal
Eide Øyvind
Elli Tommaso
Endres Bill
Engel Maureen
Escandell-Montiel Daniel
Escobar Varela Miguel
Esteva Maria
Estill Laura
Falk Michael Gregory
Faull Katherine Mary
Fendt Kurt E
Fenlon Katrina Simone
Fernandez Riva Gustavo
Ferschke Oliver
Fields Paul J.
Finn Edward
Fischer Franz
Flanders Julia
Fokkens Antske
Forest Dominic
Forlini Stefania
Fornes Alicia
France Fenella Grace
Franzini Greta
Fredner Erik Christopher
French Amanda
Friedland Nancy E.
Froehlich Heather
Frontini Francesca
Gagarina Dinara
Gairola Rahul Krishna
Galina Russell Isabel
Galleron Ioana
Gallet-Blanchard Liliane
Gao Jin
Garcia-Fernandez Anne
Garfinkel Susan
Garnett Vicky
Gartner Georg
Gautier Laurent
Giannella Julia
Giannetti Francesca
Gil Alexander
Giovannetti Emiliano
Girard Paul
Giroux Amy Larner
Gius Evelyn
Gladstone Clovis
Glass Erin Rose
Gniady Tassie
Goddard Lisa
Gold Matthew K.
Gordea Sergiu
Gordon Tamar

Goto Makoto
Goudarouli Eirini
Grandjean Martin
Grant Katrina Caroline
Griffin Howard Kevin
Griggs Hannah C.
Grincheva Natalia
Grüntgens Max
Guido Daniele
Guiliano Jennifer Elizabeth
Gutiérrez De la Torre Silvia Eunice
Guzman Carina Emilia
Hackney S. E.
Hammond Adam
Han Myung-Ja K.
Heiden Serge
Hendery Rachel Marion
Hennicke Steffen
Henny-Krahmer Ulrike Edith Gerda
Henrich Andreas
Henry Geneva
Heppler Jason A.
Herrmann J. Berenike
Heuser Ryan James
Heuvel Charles van den
Heyer Gerhard
Hicks Benjamin Wesley
Hiebert Matthew
Higgins Devin
Hinrichs Uta
Hladík Radim
Ho Hou Ieong
Hodel Tobias
Hodošček Bor
Hoekstra Rik
Hoenen Armin
Holmes Martin
Homburg Timo
Hoover David L.
Horstmann Jan
Houston Natalie M
Hsiang Jieh
Hswe Patricia
Huculak John Matthew
Huijnen Pim
Huitric Solenn
Hulden Vilja
Hunter Jane
Hunter John
Hurtado Tarazona Alejandra
Hyman Christy
Idmhand Fatiha
Impett Leonardo Laurence
Isaksen Leif
Jacobs Hannah L.
Jakacki Diane Katherine
Jamison Anne
Janco Andrew
Jannidis Fotis
Jensen Thessa
Jett Jacob
Johnson Ian R.
Jones Michael Alastair
Jones Catherine Emma
Jones Madison Percy
Jordanous Anna Katerina
Juola Patrick
Kampkaspar Dario
Kane Julie
Karadkar Unmil
Kaufman Micki
Kawase Akihiro
Kelleher Margaret
Kemman Max
Kenderdine Sarah
Kermes Hannah
Kerr Sara Jane
Kessler Carsten
Khosmood Foaad
Kijas Anna Ewelina
Kim Minhyoung
Kim Evgeny Gamletovitch
King Lindsay
Kitamoto Asanobu
Kizhner Inna
Klein Lauren F.
Kleppe Martijn
Klinger Roman
Koho Mikko Kristian
Koolen Marijn
Körner Fabian
Koumpis Adamantios
Kretzschmar William
Kröger Bärbel
Kumar Ritesh
Kumari Ashanka
Kurlinkus Will
Lach Pamella R
Lahti Leo
Lana Maurizio
Lang Anouk
Lang Matthias
Laubrock Jochen
Lavagnino John
Lavrentiev Alexei
Leavy Susan
Leblay Christophe
Leem Deborah
Lester Connie Lee
Letricot Rosemonde
Levallois Clement
Licastro Amanda Marie

Lincoln Matthew
Lindblad Purdom
Lindquist Thea
Litta Eleonora
Liu Chao-Lin
Liu Jyi-Shane
Lopes Patricia
Lorang Elizabeth M
Losh Elizabeth
Madron Justin
Maeda Akira
Mäkelä Eetu
Makinen Martti
Malm Mats
Malta Joana
Manzanera Silva Norma Aida
Mapes Kristen
Marchetti Andrea
Martin Kim
Martinez-Canton Clara
Martins Bruno Emanuel
Maryl Maciej
Mas Joan
Mathiak Brigitte
Mattock Lindsay Kistler
Mauro Aaron Mathew
McDonald Robert
McGarry Shane Adam
McGrath Jim
Mehler Alexander
Melton Sarah
Mendoza Juan José
Meneses Luis
Menini Stefano
Menon Nirmala
Merritt Don
Meyer Eric T.
Meza Aurelio
Michlowitz Robert
Miller Ben
Milligan Ian
Mimno David
Miyagawa So
Monteiro Vieira Jose Miguel
Morán Ariel
Morgan Paige Courtney
Moritz Maria
Morlock Emmanuelle
Moro Jeffrey Tyler
Motilla José Antonio
Mpouli Suzanne
Murai Hajime
Murphy Orla
Murr Sandra
Murray-John Patrick David
Murrieta-Flores Patricia
Musgrave Simon
Mylonas Elli
Nagasaki Kiyonori
Nainwani Pinkey
Nanni Federico
Navarrete Trilce
Neovesky Anna
Nerbonne John
Neuber Frederike
Neuefeind Claes
Newton Greg T
Nieves Angel
Noordegraaf Julia
Nowak Krzysztof
Núnez Alexandra
Nurmikko-Fuller Terhi Maija
Nyhan Julianne
O'Connor Alexander
O'Donnell Daniel Paul
Ocampo Gutiérrez de Velasco Marat
Ochab Jeremi K.
Ohya Kazushi
Olsen Mark
Ore Espen S.
Orekhov Boris V.
Organisciak Peter
Orlowska Anna Paulina
Ortega Elika
Otis Jessica
Overbeck Maximilian
Padilla Thomas George
Page Kevin
Pagé-Perron Émilie
Pairet Laure
Palkó Gábor
Papadopoulos Konstantinos
Paquette-Bigras Ève
Paris Britt
Pawłowski Adam Tomasz
Peaker Alicia Rose
Peña Ernesto
Peña-Pimentel Miriam
Perez Isasi Santiago
Pernes Stefan
Peroni Silvio
Petersen Andrew
Pierazzo Elena
Pimenta Ricardo Medeiros
Piotrowski Michael
Poibeau Thierry
Polyck-O'Neill Julia Geneviève
Powell Daniel James
Preiser-Kapeller Johannes
Pretnar Ajda
Priani Ernesto
Priego Ernesto

Puren M.P.
Puschmann Cornelius
Radzikowska Milena
Ramos Adela María
Ray Murray Padmini
Rebora Simone
Reeve Jonathan Pearce
Rehberger Dean
Rehm Georg
Reiter Nils
Renault Arthur
Ribeiro Cláudia
Ricaurte Paola
Ricciardi Emiliano
Richards-Rissetto Heather
Riddell Allen Beye
Ridge Mia
Ridolfo Jim
Riondet Charles
Risam Roopika
Robertson Stephen Murray
Robey David
Robinson Peter
Robles-Gómez Antonio
Rochat Yannick
Rockwell Geoffrey
Rodighiero Dario
Rodríguez-Roche Sulema
Roe Glenn H
Roeder Torsten
Rogel Rosario
Rojas Castro Antonio
Romanello Matteo
Romary Laurent
Romero-López Dolores
Rosenblum Brian
Rosner Lisa
Rosselli Del Turco Roberto
Rotari Gabriela
Roueché Charlotte
Routsis Vasileios
Röwenstrunk Daniel
Rudman Joseph
Ruiz Fabo Pablo J
Rumyantsev Maxim
Rusinek Sinai
Rybicki Jan
Sahle Patrick
Saklofske Jon
Salvatori Enrica
Sanz Amelia
Saum-Pascual Alex
Sayers Jentery
Scharnhorst Andrea
Scheuermann Leif
Schich Maximilian
Schlarb Sven
Schlesinger Claus-Michael
Schl r Daniel
Schmidt Sara A.
Schmunk Stefan
Schöch Christof
Scholger Walter
Schommer Christoph
Schulz Sarah
Senier Siobhan
Senseney Megan Finn
Serantes Arantxa
Severo Marta
Sharpe CelesteTường Vy
Shaw Ryan Benjamin
Shep Sydney
Shepard David Lawrence
Shepherd Ammon
Sherratt Tim
Shibutani Ayako
Shimoda Masahiro
Shrout Anelise Hanson
Siders Anne R
Siemens Raymond George
Siemens Lynne
Silva Andrea
Sinclair Stéfan
Smithies James Dakin
Snyder Lisa M.
Song Yuting
Sostaric Petra
Spadini Elena
Spence Paul Joseph
Sperberg-McQueen Michael
Spiro Lisa
Sprugnoli Rachele
Stadler Peter
Stalnaker Rommie L
Stertzer Jennifer Elizabeth
Stewart Elizabeth Eleanor Rose
Steyn Zacharias Jacobus
Stokes Peter Anthony
Strötgen Jannik
Stutzmann Dominique
Subotic Ivan
Suire Cyrille
Sula Chris Alen
Swafford Joanna Elizabeth
Swanstrom Elizabeth Anne
Szabo Victoria
Takseva Tatjana
Tambassi Timothy
Tammaro Anna Maria
Tanasescu (MARGENTO) Chris
Teich Elke
Ter Braake Serge

Terras Melissa
Theibault John Christopher
Thomas Lindsay
Thompson Jeff
Thomson Christopher
Tilton Lauren
Tonelli Sara
Tonnellier Gaelle
Tonra Justin Emmet
Tournier Charlotte
Tracy Daniel G.
Travis Charles Bartlett
Tropea Rachel
Tsui Lik Hang
Tuffery Christophe
Tupman Charlotte
Turton Alexander Robert
Valverde Mateos Ana
van den Herik H. J.
van Eijnatten Joris
van Erp Marieke
Van Keer Ellen
Van Kranenburg Peter
Van Zundert Joris Job
Venecek John T.
Viana Vander
Viglianti Raffaele
Visconti Amanda
Vogeler Georg
Volkmann Armin
Volodin Andrei
von Waldenfels Ruprecht
Walkowiak Tomasz
Walkowski Niels-Oliver
Walsh John
Walsh Brandon
Walter Katherine L.
Warwick Claire
Webb Sharon
Weber Andreas
Weidman Robert William
Weidman Sean Gregory
Weigl David M.
Wernimont Jacqueline D
Wevers Melvin
Widner Michael Lee
Wieneke Lars
Wieringa Jeri
Wiesner Susan L.
Wilkens Matthew
Williams Patrick
Williams Helene C.
Wilms Lotte
Winder William
Wintergrün Dirk
Wisnicki Adrian S.
Wittern Christian
Wolff Mark
Worthey Glen
Wrisley David Joseph
Wulfman Clifford Edward
Würsch Marcel
Wuttke Ulrike
Yamada Taizo
Yang Bin
Yeates Stuart Andrew
Yin Xin
Youngman Paul
Zafrin Vika
Zeng Marcia Lei
Zhang Jinman
Zöllner-Weber Amélie
Zwarich Natasha

Digital Humanities 2018

dh2018.adho.org