Graph Technologies for the Analysis of Historical Social Networks Using Heterogeneous Data Sources

Sina Menzel*[1] Mark-Jan Bludau*[2] Elena Leitner*[3]
Marian Dörk[2] Julián Moreno-Schneider[3] Vivien Petras[1]
Georg Rehm[3]

*The authors contributed equally to this work as first authors.

[1] Humboldt-Universität zu Berlin
[2] FH Potsdam – University of Applied Sciences
[3] DFKI – Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Abstract

Over the last decades, cultural heritage institutions have provided extensive machine-readable data, such as bibliographic and archival metadata, full-text collections, and authority records containing multitudes of implicit and explicit statements about the social relations between various types of entities. In this paper, we discuss how approaches to the creation and operation of advanced research infrastructure for historical network analysis (HNA) based on heterogeneous data sources from cultural heritage institutions can be examined and evaluated. Based on our interdisciplinary research, we describe challenges and strategies with a special focus on the issue of data processing, sketch out the advantages of human-centered project design in the form of a preliminary co-design workshop, and present an iterative approach to data visualization.

In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): Graph Technologies in the Humanities - Proceedings 2020, published at http://ceur-ws.org.

1 Introduction

The study of historical events is relevant to many disciplines in the digital humanities, with the analysis of relationships between agents often being crucial for the understanding and explanation of social, political, and cultural phenomena. Given that historical research is heavily dependent on information from the respective time period, the combination of as many historical sources as possible is essential for the reconstruction of historical networks – and this is where the method of historical network analysis (HNA) comes into play. Derived from social network analysis, HNA is characterized by the same dependency on numerous historical sources that ideally support each other (Jansen and Wald, 2007).

One limiting factor in HNA can be a lack of awareness with regard to the availability of suitable research data. At the same time, over the past decades, cultural heritage institutions have produced very large amounts of machine-readable and, in many cases, standardized and well-organized data in the form of bibliographic and archival metadata, full-text collections, and sets of authority or reference records. These datasets contain a plethora of implicit and explicit statements about social relations, which can in turn be exploited for HNA research. However, systematically combining multiple data sources (not to mention extracting and visualizing the complex resulting networks) currently requires extensive knowledge in graph theory as well as time-consuming manual work carried out by the individual researcher. One reason for this is the heterogeneity of the data sources made available by cultural heritage institutions, for example, in terms of data formats.

The research project SoNAR (IDH): Interfaces to Data for Historical Social Network Analysis and Research[1] addresses this issue. We examine and evaluate approaches to the development and operation of HNA-supporting research infrastructure based on heterogeneous cultural heritage data. In this paper, we present a number of preliminary insights related to the process of modeling and transforming heterogeneous data sources, and to the design of user-centered visualization for historical social networks. By sharing our approach and its accompanying challenges, we aim to contribute to the ongoing discussion on the suitability of bibliographic big data for HNA and the development of corresponding research technologies.

2 Related Work

The following section gives an overview of previous research, and discusses projects related to graph modeling and visualization approaches within the

[1] https://sonar.fh-potsdam.de

digital humanities from the perspective of historical network analysis.

2.1 Related Projects

In recent years, open knowledge graphs have frequently been used as an alternative to a document-based approach (Auer and Mann, 2019). Several large-scale initiatives such as EOS,[2] Europeana,[3] and CLARIN[4] provide researchers in the digital humanities with access to cultural data. Meanwhile, the issue of decentralized and heterogeneous bibliographic data sources is being addressed by projects such as Culturegraph (Vorndran, 2018) and DARIAH-DE[5] in the digital humanities, Lynx[6] in the legal domain, and, to a certain extent, ELG[7] in language technology (Rehm et al., 2020). Most of these initiatives connect to infrastructures of cultural heritage institutions, often hosted by libraries or archives.

Even though these initiatives provide, among other things, access to new, previously unidentifiable or implicit information, they do not primarily focus on the extraction of network data. Therefore, HNA researchers are often left to create their own individual graphs after gathering data that is suitable to address their research question(s), in many cases using open source software tools such as Gephi,[8] Palladio,[9] or VennMaker,[10].

Along with the establishment of network analysis as a method in historical research, there has been an increase in joint research projects that are focused on the extraction of historical networks within the social sciences and the humanities.

For example, the project Six Degrees of Francis Bacon[11] applies statistical methods to the base data with the goal of inferring relations that permit the reconstruction and visualization of historical social networks in Early Modern Britain. The project allows for the expansion and curation of the data through collaborative annotation by the users (Warren et al., 2016). The histoGraph[12] project follows a similar approach by offering users an opportunity to collaboratively explore and research historical social networks by

[2] European Open Science Cloud, https://www.eosc-portal.eu

[3] https://www.europeana.eu

[4] Common Language Resources and Technology Infrastructure, https://www.clarin.eu

[5] https://de.dariah.eu

[6] http://www.lynx-project.eu

[7] https://www.european-language-grid.eu

[8] https://gephi.org

[9] https://hdlab.stanford.edu/palladio/

[10] https://www.vennmaker.com

[11] http://www.sixdegreesoffrancisbacon.com

[12] http://histograph.eu

means of extensive multimedia collections, with a special focus on crowdsourced indexation (Novak et al., 2014). In a joint project involving several European research institutions, Issues with Europe – A Network Analysis of the German-Speaking Alpine Conservation Movement (1975-2005)[13] is currently examining the disputes over European alpine transit policy, while the Austrian project APIS – Mapping historical networks has been working on the extraction and visualization of networks from more than 18,000 records in the Austrian Biographical Encyclopedia.[14] Finally, the German project Gesellschaftliche Wissensproduktion in der Aufklärung – Text- und netzwerkanalytische Diskursrekonstruktion considers full texts of more than 300 periodicals published in Halle, Germany between 1688 and 1815, and combines the methods of topic modeling with historical network analysis in order to systematically analyze public discourse during the Age of Enlightenment (Purschwitz, 2018).

These are only a few examples of the ongoing efforts to provide users with direct access to networks in existing data collections. In our project, we are working with data sources that have not been modeled for HNA before. Our generic data approach is closely connected to similar projects, like the North American cooperative SNAC – Social Networks in Archival Context[15] and the French project PIAAF,[16], which both have a strong focus on archival metadata and full texts.

2.2 Network Visualization

As far as the visualization of data for HNA is concerned, many interfaces have been developed over the years that offer explorative, web-based network visualization tools for historical network analysis. Examples include the above-mentioned Six Degrees of Francis Bacon (Warren et al., 2016) and histoGraph (Novak et al., 2014), as well as Visualizing the Republic of Letters (Chang et al., 2009), Kindred Britain,[17] and Deutsche Biographie.[18]

Graph visualization is an extensive field in itself, which is accompanied by a substantial body of literature on issues such as graph-related algorithms (e. g. Gibson et al., 2012; Jacomy et al., 2014; Behrisch et al., 2016), task taxonomies for graph visualization (e. g. Lee et al., 2006; Ahn et al., 2013; Kerracher et al., 2015), state-of-the-art visualization interaction techniques and

[13] https://www.uibk.ac.at/projects/issues-with-europe/index.html.en

[14] Österreichisches Biographisches Lexikon, https://apis.acdh.oeaw.ac.at

[15] https://snaccooperative.org

[16] Pilote d'interopérabilité pour les autorités archivistiques françaises https://piaaf.demo.logilab.fr

[17] http://kindred.stanford.edu

[18] https://www.deutsche-biographie.de

developments (e. g. van Ham and Perer, 2009; von Landesberger et al., 2011; Pienta et al., 2015), as well as the use of visual facilitators for the construction of graph queries (e. g. Pienta et al., 2017). Nevertheless, existing research and taxonomies mostly address the wider field of graph visualization. More often than not, visualizations and digital practices are not specifically adapted to the requirements of HNA research or established data practices in the humanities, and are ill-suited to address issues such as uncertainty, subjectivity, or observer-dependence (Drucker, 2011).

2.3 Human-Centered Design

A key element in the examination and development of a new research infrastructure designed for human-computer interaction is how well it meets the needs of the people it is intended to assist. This human-centered approach is closely related to Grounded Theory, which generates inductive results by means of sociological methods (Glaser and Strauss, 1967).

Isenberg et al. (2008) adapted Grounded Theory for the evaluation of information visualizations. They suggest iterative evaluation throughout the process of system development using several points of qualitative inquiry to ensure the focus of a system's intended use, including field research to examine potential contexts of human interaction with the system. In keeping with this argument for grounded evaluation, the neuralgic points for evaluation in our project are based on Munzner's nested model for visualization design and validation (Munzner, 2009), which allows for iterative improvement of the prototypes. The stages of evaluation include the assessment of possible use cases, and the investigation of the problems and data of a particular user domain at the top level. In order to better address such issues, it is becoming more and more common to include domain experts in the creation process of digital humanities-related projects. This kind of co-creation is precisely what Chen et al. (2014) attempted to foster with a workshop, wherein the participants were asked to create collages to make sense of a photo archive with the aim of creating collection-sensitive interfaces. Henry and Fekete (2006) used a similar participatory approach in the development of a tool for the exploration of social networks: they invited social science researchers to create paper prototypes, which in turn led to a list of domain requirements for their tool and resulted in a prototype with novel features. A thorough evaluation of such co-creation methods, conducted in a co-design process with social science researchers, found that domain experts in general appreciate their additional empowerment in the process and the domain-customized results based on their specific needs. Nevertheless, regarding their personal involvement and necessary time commitment, some participants did not perceive their personal involvement as beneficial for the facilitation of their own research (Molina León and Breiter, 2020). Besides the use of co-design techniques, there is also a shift from perceiving visualizations as mere tools for humanities-related research towards the acknowledgment of visualization and visualization processes as a methodology and facilitator of cross-disciplinary research in and of itself (Hinrichs et al., 2019). While we have noticed increased attention to the method of HNA, to the best of our knowledge, there has so far been little investigation of the modeling and visualization of (bibliographical) big data for this purpose.

3 Data Sources

The interdisciplinary project SoNAR (IDH), which studies the potential of large heterogeneous data collections for HNA, includes partners from the fields of historiography, information visualization, and artificial intelligence, as well as computer and information science. This variety of disciplines opens different perspectives on the requirements and challenges connected to the use of heterogeneous (meta)data for HNA. What distinguishes our approach is the synchronous operation of all components of the project, so that the design of the data technology, the development of a model research design for HNA, and the development of innovative visualization and interface approaches with the involvement of HNA experts are all intertwined and influence one another.

The project is based on heterogeneous source data from authority files, bibliographic records, and full texts. The data is available in various XML-based formats such as MARC21 (Kruk et al., 2005), EAD (Allison-Bunnell, 2016), and METS/ALTO[19] (Cantara, 2005):

- The Integrated Authority File (GND)[20] represents and describes 8,295,047 entities (people, corporations, conferences, geographical areas, technical terms, and works);
- The German National Library (DNB)[21] provides descriptions of bibliographic resources. The dataset has 19,926,573 records of books, magazines, newspapers, sheet music, music recordings, audio books etc.;
- The German Union Catalogue of Serials (ZDB)[22] describes newspapers, magazines, serial titles, yearbooks, etc. and contains 1,908,334 records;

[19] http://www.loc.gov/standards/alto

[20] https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

[21] https://www.dnb.de/EN/Home/home_node.html

[22] https://zdb-katalog.de/index.xhtml

- The Kalliope Union Catalog (KPE)[23] is a collection of personal papers, manuscripts, and publishers' archives, which consists of 26,752 records;
- The Newspaper Information System (ZeFYS)[24] represents 2,596,641 digitized pages of historical newspapers and full texts;
- The Exile Press[25] represents German-language exile journals between 1933 and 1945 and consists of 5,336 digitized pages.

Since the source data – describing entities (authority files) and resources (bibliographic files) – is encoded in various formats, these formats must first be analyzed in order to enable the design of an appropriate data model and allow their transformation into a uniform, generic format. Full texts are prepared for automatic enrichment (i. e. named entity recognition and linking) and converted to a corresponding format.

4 Data Processing

In this section, we will give an overview of the data transformation and graph modeling process, and outline the challenges that we have encountered along the way.

The technical goal of our project is the integration of the various source datasets into a common research infrastructure. We currently use the graph database Neo4j,[26] which is well suited to the efficient storage and high-performance analysis of large amounts of highly networked information (Efer, 2016; Matschinegg and Nicka, 2018; Wintergrün, 2019). Entities are modeled as nodes and relations as edges with absolute and relational features.

There are a total of 9 entity types extracted from the source data:

1. Person `PerName`;
2. Corporate body `CorpName`;
3. Place or geographic name `GeoName`;
4. Conference or event `MeetName`;
5. Subject heading `TopicTerm`;
6. Work `UniTitle`;
7. Temporal information `ChronTerm`;
8. Information about ISIL[27] `IsilTerm`;
9. Resource `Resource`.

[23] https://kalliope-verbund.info/en/index.html

[24] http://zefys.staatsbibliothek-berlin.de/index.php?id=start&L=1

[25] https://www.dnb.de/EN/Sammlungen/DEA/Exilpresse/exilpresse_node.html

[26] https://neo4j.com

[27] International Standard Identifier for Libraries and Related Organizations

Six entity types (i. e., person PerName, corporate body CorpName, place or geographic name GeoName, conference or event MeetName, subject heading TopicTerm, and work UniTitle) are taken from the corresponding classes of the authority files. Bibliographic entities are represented as Resource. We added two types to this list: ChronTerm, which describes temporal information encoded in entity types from authority files; and IsilTerm, which is used to identify the libraries related to other entity types. Each entity has general features, such as a unique source identifier, URI, name, link, etc., and specific features, such as age, gender, coordinates, etc. Furthermore, there are also nine relation types that correspond to entity types, such as RelationToPerName, RelationToCorpName, RelationToGeoName. Relations between entities include information about the relation source, relation source type, information about temporal validity, and additional information (Figure 1).

4.1 Data Model

While the relations between entities are explicitly described in authority files, relations between actors such as persons or corporate bodies that are identified or defined in the resource are only implicitly encoded in bibliographic files. Our aim is to automatically infer these implicit relations with the assistance of a set of strict guidelines (e. g. a connection between two persons can be assumed if both are co-authors of a scientific publication), and to make them available as explicitly encoded data. In order to derive corresponding relation types, the role of actors regarding a specific resource (e. g. as author, editor, or addressee) and the resource type (bibliographic files of primary sources of the Kalliope Union Catalog and of secondary sources of the German National Library and the German Union Catalog of Serials) are to be taken into account. Using this approach, we were able to infer additional relations (but marked them as computed), for instance between co-authors, co-publishers, and authors/addressees, to further enrich the data.

In order to prepare full texts for analysis, named entities are automatically recognized, disambiguated, and linked to their associated authority files (e. g. the Integrated Authority File or Wikidata[28]). Next, relations between detected entities are automatically recognized, added to the graph database, and connected with their respective full texts, represented as nodes.

4.2 Challenges and Solutions

Overall, the authority and bibliographic files used by us contain approximately 30 million records that describe entities and resources in detail. As was to be expected, normalization of the data revealed a number of errors and

[28] https://www.wikidata.org

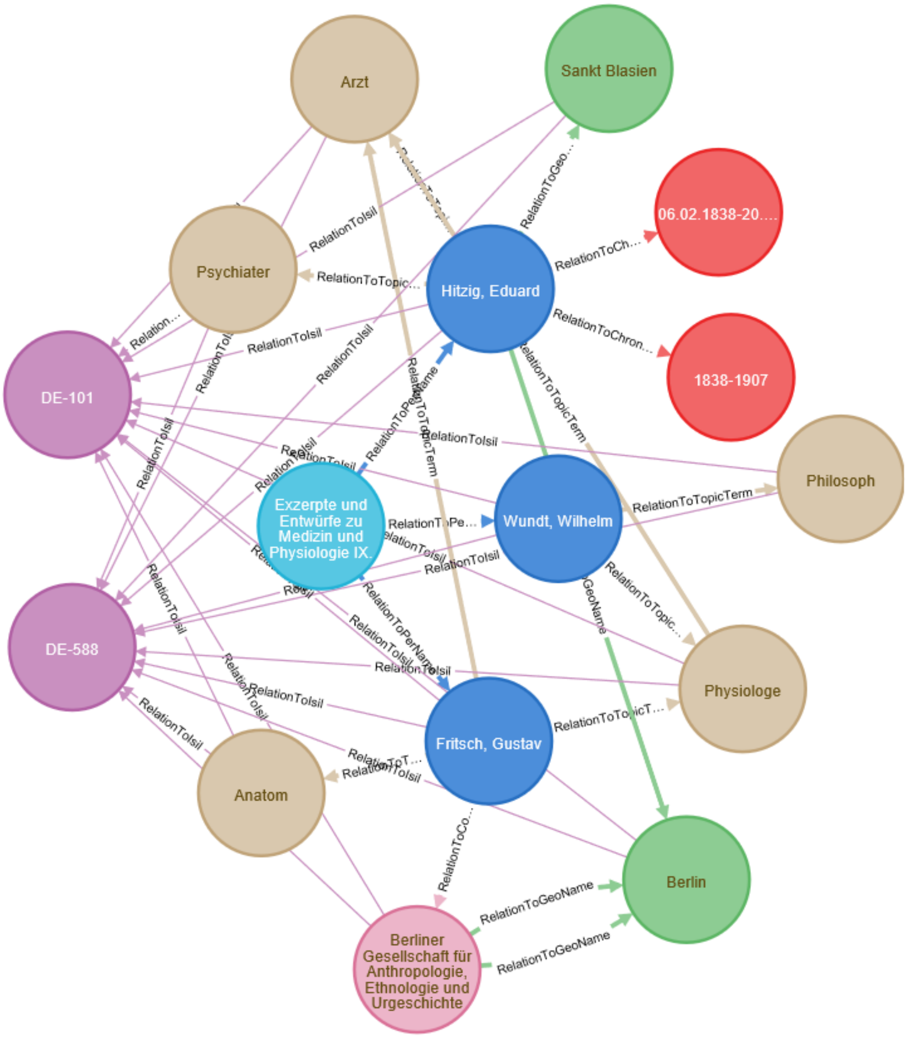

Figure 1: Data modeled in *Neo4j*. Persons are shown in blue, locations in green, subject headings in light brown, works in pink, ISILs in purple, temporal expressions in red, and resources in light blue.

inconsistencies. In this section, we would like to describe some particularly problematic areas in more detail and suggest possible solutions.

We have modeled and transformed data for the graph database in such a way that identifiers are used as coordinates for relations between entities. In the Integrated Authority File, entities with old identifiers were found, so that an appropriate connection of two entities was not possible. The first challenge was to detect old identifiers and replace them with valid ones in order to enable error-free representation. All replacements were written in a log file. However, during a consistency check we also found relations to entities within the source data that were without identifiers. Since such entities could not be clearly assigned to existing entities with identifiers, ambiguous relations of this kind had to be ignored.

Information that was encrypted in internal codes in the Integrated Authority File, the German National Library, and the German Union Catalogue of Serials (in format MARC21) was also checked for codes of general and specific entity types, codes of relation types between an agent and a resource, and country codes. Further examinations were performed on the consistency of entity names, resource titles, and identifiers. Again, all errors or inconsistencies were written in a log file.

Building on the conclusions that we were able to draw from testing Neo4j, we decided to adapt the data model to our needs. In order to simplify searching and filtering according to temporal dimension, time information from the source data was adjusted. First, while retaining the source data, we additionally separated time intervals, noted as "begin" and "end." Second, in order to facilitate more performant visualization and querying of the data, we added a feature to resource descriptions that reflects the year of publication (in addition to the publication date). Thirdly, differing time expressions in MARC21 and EAD were normalized.

We also decided to change gender-specific names of professions. These are represented in the Integrated Authority File as two different entities with their own identifiers, male and female. Conceptually, however, what we are dealing with is a single entity with two versions, so these versions must be merged in the graph database and represented as a node. One challenge is to adequately display all information from the two versions without making the search more difficult. In this case, we are currently still looking for a suitable solution.

5 Co-Design Workshop

In accordance with the principles of grounded evaluation (Isenberg et al., 2008), we aim to closely integrate domain experts into the data modeling and visualization process. The presence of HNA experts in our project team means that all internal decisions that are made take the domain perspective into account. Additionally, the inclusion of external domain experts is another integral part of our research design. Conducting studies with researchers from various fields allows us to iteratively improve the project's outcome.

At the beginning of the project, it was important to us to stimulate discussions on the potential of bibliographic (meta)data for HNA, and on the requirements for the visualization of historical networks. Following the approach proposed by Chen et al. (2014) and Henry and Fekete (2006), we organized a co-design workshop that included domain experts in order to help identify key aspects and gain new insights into historical network research and visualization.

5.1 Procedure

The workshop consisted of ten participants, including four historical/social network practitioners as domain experts, two project-internal information visualization designers/engineers, two members of our project-internal evaluation team, one member of our team of data scientists (responsible for the data transformation), and an external participant who had a background in design and previous experience with the co-design format. The interdisciplinary composition of the group was intended to enrich the discussion by offering a multitude of perspectives on the topic of HNA through the lens of HNA experts, with fresh insights being provided by participants from other (project-relevant) fields. Since we aim to develop an infrastructure for HNA that can be used by researchers from all disciplines working with this method, the participation of experts from fields other than history was especially welcome.

The workshop was scheduled for three hours in total. As suggested by Fekete and Plaisant (2002), we started off with a brief presentation of various recent developments in the field of network visualization, including some of the more novel and experimental approaches.

We started the process of conceptualizing network visualizations with a short, hands-on visualization exercise, during which the participants were asked to visualize a very small social network (ten nodes) based on a data matrix we provided. After this warm-up, we gave a short introduction to the goals of our project and the data we are using. The participants were then asked to create a collage depicting possible approaches to HNA research with our specific data and project in mind (see Figure 2). For the collages, we supplied a variety of materials (e. g. construction paper, pencils and markers, sticky notes). While Chen et al. (2014) provided visual material from

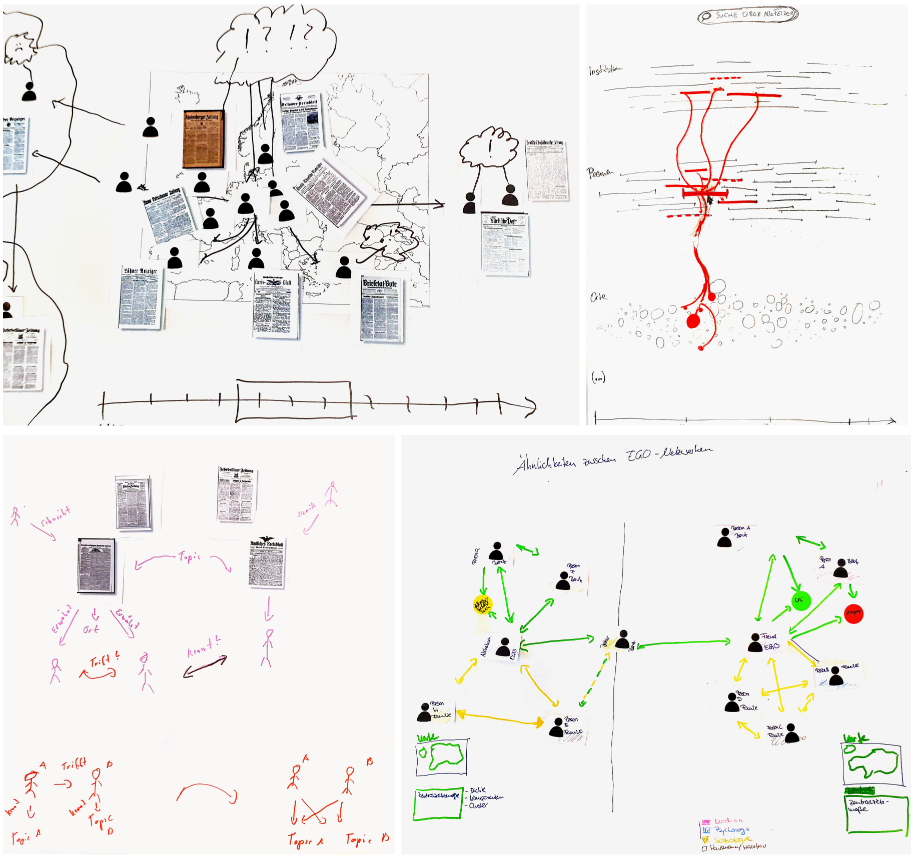

Figure 2: Selected collages created in the interdisciplinary co-design workshop

their photographic collection, our data is more abstract and less visual. To compensate for this, we printed out and distributed further visual material including an empty map, various icons (e. g. as representations of network nodes), and a small number of scans from our full-text data sources. We then introduced several questions to help initiate the creative process (e. g. "How would you like to move through the data?" and "What role do data dimensions such as time, space, or semantic relationships play?"), but encouraged the participants to feel free to disregard them. The task we had in mind was not to create wireframe sketches for a concrete user interface, but to envision desired functionalities as well as general approaches and entrance points to HNA research and our data.

After about 30 minutes, each of the collages was discussed. First, the participants not involved in making a collage were asked to interpret and speculate about what they were seeing. Afterwards, the creators of the collages were asked to give explanations and discuss their approach with the group. In this step, the almost inevitable misinterpretations were meant to foster further discussion and novel ideas. In the final step, each participant was asked to give a closing statement recapitulating the most important insights from the process and the most prominent topics or themes in the discussion.

For further analysis and documentation, the entire workshop was audio-recorded and photographed. The audio recordings were transcribed and encoded in a tool for qualitative data analysis. This allowed us to assess various qualitative aspects of the workshop discussions at a later date. As pointed out above, the goal of our workshop was not to create functional wireframes or concrete interaction principles, but to stimulate discussion, foster sensibility towards the domain and data, and highlight important domain-specific research aspects and challenges. The following section will discuss some of the most relevant insights concerning our visualization process.

5.2 Results

We noticed two different types of statement. On a more abstract level, the participants expressed various information needs that commonly arise in the process of their research. In some cases, however, the conversation and the collages yielded very concrete ideas regarding possible features of an HNA infrastructure that would address these needs. As mentioned before, the latter were not regarded as direct assignments to be fulfilled in the visualization process, but rather as indicators for the participant's general receptiveness towards various properties of the user interface of an HNA infrastructure. Table 1 and 2 summarize the main aspects of the workshop discussions in the form of needs and features.

Need	**Number of Mentions**	**Persons**
New perspectives	30	7
Uncertainty	25	7
Data potential	27	4
Graph density	26	4
Entry points	17	4
Data explanations	16	4

Table 1: List of the most frequently expressed needs with the count of their mentions during the workshop and the number of persons (n=10) referring to them

The most pressing topic in the discussions was the envisioned user approaches and use cases the infrastructure is expected to support. Seven of the ten participants expressed the hope that the visualizations could generate *new perspectives*, thereby creating forms of access to the data that would hardly be available based on non-machine-supported cognitive work. In this

Feature	Number of Mentions	Pers.
Timeline	22	4
Tie metrics	18	4
Other filters	16	3
Export and citation	13	4
Location filter	8	3
Source linking	6	4

Table 2: Most frequently desired features with the count of their mentions during the workshop and the number of persons (n=10) referring to them

context, one participant explicitly emphasized the potential of visualizations to raise new questions:

> What kind of relationships you are looking for in the data is something you often notice in the very moment that you look at the pile for the first time.[29]

Since the participants were aware of the fact that we are confronted with a very large amount of data, which can hardly be presented in its entirety (see Section 3), a discussion of possible entry points emerged. There was consensus on the importance of filter options, most importantly time filters:

> Without timelines, the visualizations are of no use to me – neither for analysis, nor for the presentation of results.

In addition to timelines, other filters (e. g. node type and node source) were considered a prerequisite for data exploration. Three participants also mentioned the importance of location filters (e. g. through a map view).

Participants with more HNA experience explicitly stressed the essential role of a multi-layered approach. The capacity to display the evolution of relations (e. g. through time and location) was described as the distinctive feature of HNA when compared to the non-historical analysis of social networks. The sole option of static display was considered insufficient.

Along with possible entry points, another important topic raised in the discussion was data complexity, with introductions and explanations regarding the underlying data being identified as particularly crucial. Some participants suggested addressing this issue with the help of concrete use cases that could give potential users a more specific idea of the possibilities afforded by the HNA infrastructure.

[29] All quotes translated from German into English.

About half of the participants cited the ability to quantify network characteristics as graph metrics during the research process as one of their main motivations for using HNA methods. This includes indicators such as the clustering coefficient, closeness centrality, degree distribution, degree centrality, and betweenness centrality. Four participants also considered density within a selected sample of nodes to be a relevant indicator for a given dataset's potential for network analysis. After the first cluster of possible approaches had been discussed, one participant highlighted the added value of graph metrics when it comes to the identification of anomalies in the data:

> What all these things are actually about is that we are looking for patterns!

Some participants also stressed the potential of tie metrics to accommodate a variety of relation types, and expressed the desire to have the weight of edge properties visualized:

> It is of course a big difference whether you are a family member [...] or whether you are a correspondence partner or whether you met at a congress during a coffee break. These are all relationships, but of course they have different weights in their interpretation. This is, for example, something we would like to see in the visualization.

This statement is representative of another central topic discussed in the workshop, namely the visual marking of missing or uncertain information in the data which can, for example, be the result of inconsistencies in the metadata fields (see Section 4.2). The design expert considered this to be a major desideratum:

> I think this is not done enough in current visualizations to show uncertainties of data.

With regard to the scientific standards of HNA research, a final major issue was the export and citation of the visualizations. This, of course, requires unambiguous and persistent provenance links to the source of each data point, as well as timestamps of the corresponding data import.

Many of the results of our co-design workshop match the challenges in information visualization discussed in the pertinent literature. In the following section, we will draw on these results to describe our prototyping approach and process.

6 Visual Prototyping

The dataset of our project is comprised of a number of elements that go well beyond what can be perceptually or cognitively grasped at a glance. When it comes to encoding, for example, the sheer amount of nodes and relations poses technological as well as visual challenges (Fekete and Plaisant, 2002; Shneiderman, 2008). While some potential users of our technology might have a very specific research question in mind, others might be inclined towards a more serendipitous approach (Thudt et al., 2012), or may wish to use such an infrastructure in order to formulate research questions. Our aim is to provide access points for a broad variety of motivations and research questions, including ones that we cannot as of yet anticipate. Therefore, the conceptualization of a visual representation as an access point to our data in the form of a data exploration interface can be described by a wide and diverse range of challenges and difficulties:

- How can tens of millions of nodes and hundreds of millions of edges be visualized?
- What are possible and meaningful entrance points to the data?
- How can we deal with uncertainty, missing data, and varying data sources?
- How can we deal with multiple data dimensions?
- How can we provide a technology that is complex and open enough for a broad range of undefined research questions, but simple enough for casual use?
- How can we be transparent with regard to the algorithms used?
- How can users move between overviews, detail views, and egocentric views?

Even though our workshop, our conversations with domain experts, and existing task taxonomies (e. g. Lee et al., 2006; Kerracher et al., 2015; Ahn et al., 2013) have already yielded a multitude of potential tasks, needs, and requirements that should be addressed in our graph technology, we see the prototyping process as a form of research through design (Zimmerman et al., 2007) that is not only capable of confirming these requirements, but also of unveiling new ones. Moreover, in contrast to the above-mentioned task taxonomies, we are engaging with humanities-related data and research questions – a field, where traditional visualization approaches are often deemed incompatible with the nature of the objects of inquiry Drucker (2011). Along with the data modeling process and the co-creation approaches described above, our visualization process can thus be described as a form of rapid, experimental, and iterative prototyping process and data exploration.

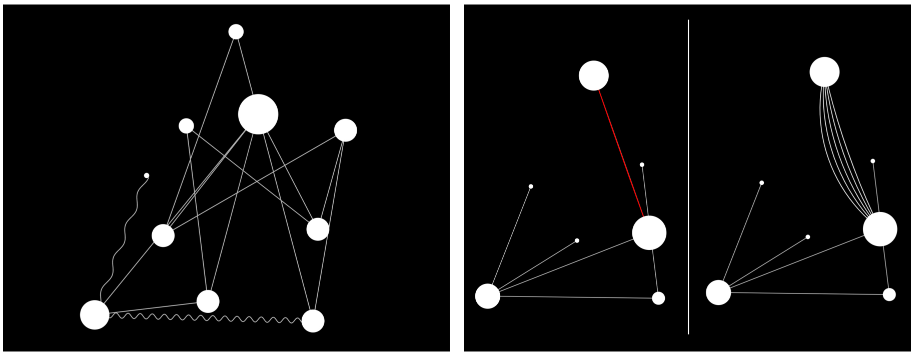

Figure 3: Two small design studies. Left: visualizing levels of uncertainty in edges between nodes by using waves and varying levels of frequency. Right: concept for handling of multiple edges between two nodes. In the initial view, multiple edges are combined into one (marked as the red line) to reduce the overall complexity of a graph. A click fans out the individual edges on demand, visually transitioning from one line to multiple arcs.

Compared to the potentially shortest path to a finished 'tool,' our method resembles a curiosity-driven 'sandcastling' (Hinrichs et al., 2019). We understand experimental approaches and detours in the visualization process itself as a methodology of knowledge production. By following this route, visualizations are not necessarily created with the goal of implementing them in a final prototype or concept. Rather, they become a method for exploring the data or individual facets of the data, a tool for investigating the basic challenges of data or their encoding, or a visual facilitator for encouraging cross-disciplinary communication and the development of novel and thought-provoking approaches (Hinrichs et al., 2019).

From the beginning, the entire project has been conducted in an interdisciplinary and concurrent mode, without any delays between its individual steps; data processing, case study development, visualization, and evaluation all occur alongside one another. Initially, the data was neither processed for visualization, nor was it accessible via some form of API, which only allowed us to work with small subsets of selected data. While this made it difficult to anticipate all of the facets and challenges associated with handling the full extent of the data, working with data subsets early on gave us the ability to exert iterative influence on the data processing and the data model.

Instead of trying to combine all potential features and ideas into a single prototype, our approach focuses on small, separate problems and ideas through a multitude of rough prototypes. Many of our design studies or prototypes have been developed in close collaboration with our own HNA specialists, and/or draw extensively on input from our workshop or other external sources of expertise, whereas others are more experimental in nature,

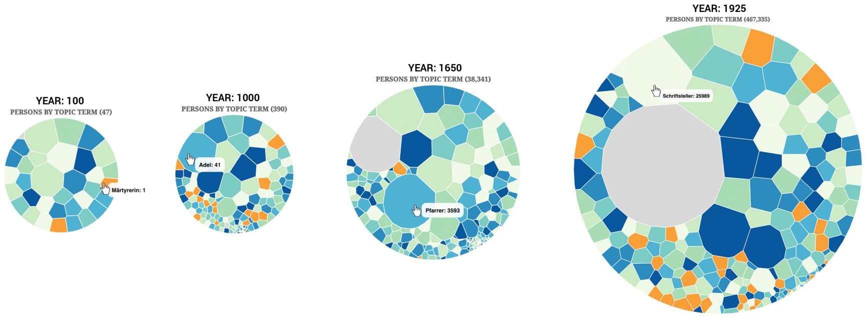

Figure 4: Prototype overviews of a specific data facet (in this case, topic terms related to persons), based on a selected year. A Voronoi map displays distributions of topic terms connected to persons alive in a selected year. Orange represents female-gendered terms.

and are often the product of spontaneous impulses. For the most part, the following examples were designed with the data visualization library D3.js (Bostock et al., 2011), which permitted the development of customized visualizations.

Figure 3, for instance, shows two small design studies from the beginning of the project, without using real data: the one to the left is a visualization of levels of relation uncertainty, while the one to the right represents the testing of an interaction concept with the goal of reducing complexity by merging multiple edges and allowing users to fan them out on demand.

As an example of the influence exerted by visualization on the data model, an early prototype which clusters persons in a small subset of the data based on related topic terms – in most cases occupational titles – revealed that these titles are frequently gendered[30] in our base data (GND), which means that men and women are often not related to the same topic term, even though they practice the same profession. This unexpected differentiation in the data is highly relevant when it comes to search queries and visualization, since it is quite possible that some researchers do not differentiate by gender, and only use the male form that was traditionally considered to be generic. One effect of this differentiation in the data can be seen in another interactive prototype (see Figure 4), where it is possible to select a specific year in the data with a slider, visualizing top topic terms related to persons who were alive in the selected timespan (female-gendered topic terms are colored in orange). The goal of this prototype was to explore the potential of overviews to reveal aspects of the data that might, at a later point, act as entry points

[30] Many German occupational titles are gendered and exist in a male and a female form, as with the English 'actor' and 'actress.'

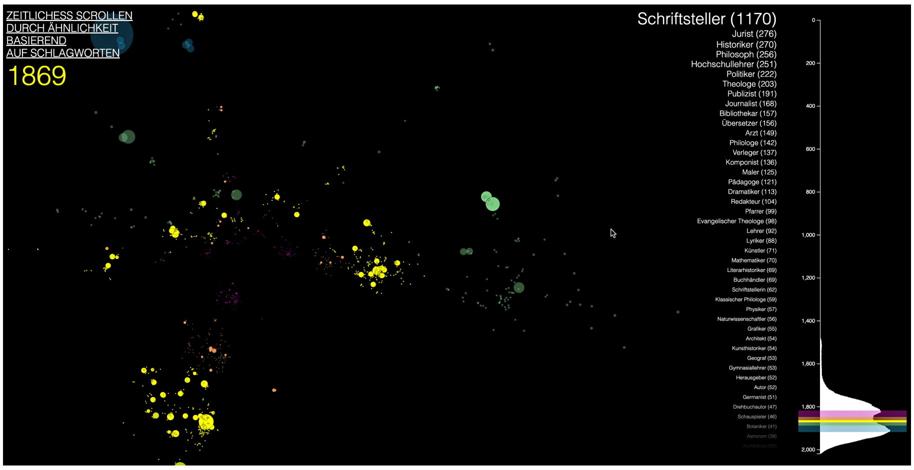

Figure 5: Experimental prototype that enables scrolling through time by means of a UMAP projection of a small subset of our data, which arranges persons on the basis of similarity across topic terms. Color and the sagittal (z) axis are used to encode temporal closeness of a node in relation to a selected year (in this example, nodes that lie inside the selected year 1869 are colored in yellow).

for specific search interests.

Another experimental prototype (see Figure 5) of a small subset of our data also focuses on topic terms and the temporality of the data; an aspect frequently highlighted as important by some of our HNA experts in the workshop. Here, the dimensionality reduction technique UMAP (McInnes et al., 2020) was used to map persons with similar topic term relations in close proximity to each other, effectively forming clusters for certain occupational domains (e. g. authors). A timeline on the right displays the general distribution of all nodes, while a list next to it contains all connected topic terms, ordered by occurrence. Scrolling enables users to move through the temporal dimension of the network, creating the impression of a time tunnel. Nodes belonging to a selected year are displayed in yellow. Temporally close nodes in the past appear more distant from the viewer and are marked in red tones, while those that lie in the future are colored in green and blue tones, and appear to be closer. One insight gained with the help of this prototype was that our data model and processing approach once again needed to be adjusted to make the data more accessible for use in visualizations, especially with regard to temporal filtering.

In some cases, as with Figure 6, we developed prototypes out of curiosity for very specific research questions, for example: “Are network communities in the data subset mostly composed of contemporary nodes or do communities stretch over multiple generations?” Here, the prototyping process allowed us to test specific algorithm implementations and design strategies,

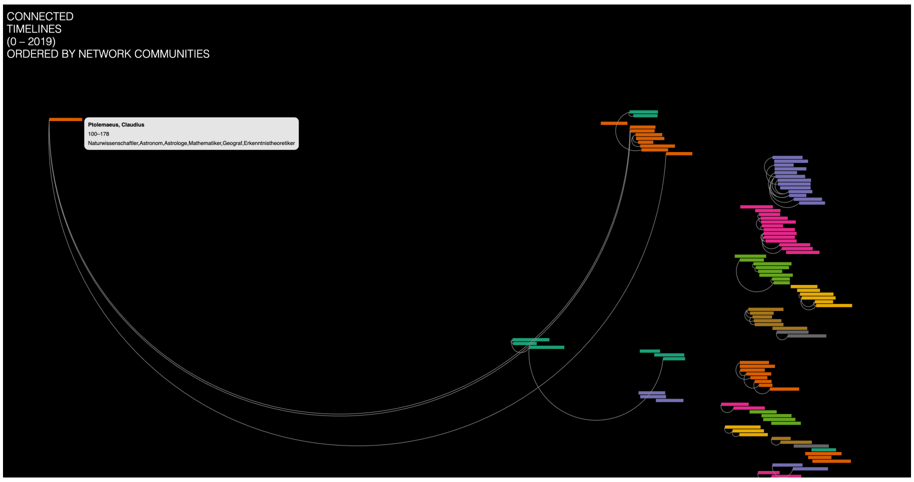
CONNECTED
TIMELINES
(0 – 2019)
ORDERED BY NETWORK COMMUNITIES
Ptolemaeus, Claudius
100–178
Naturwissenschaftler,Astronom,Astrologe,Mathematiker,Geograf,Erkenntnistheoretiker

Figure 6: Prototype overviews of node relations to reveal relations and community clusters over time. First, a community algorithm is applied to the graph data. Then, nodes are ordered and colored based on the community algorithm results, and are placed on a timeline based on their dates of birth and death.

while at the same time being able to obtain deeper insights into the data.

While our research is still in progress, the experiences mentioned above illustrate the benefits of staying curious and open to experimentation throughout the analysis and visualization process. Even though many ideas and concepts are inspired by existing research in the field and, of course, the expertise of our domain specialists, we see additional value in experimenting with the data and generating a multitude of visual representations, even if this means knowingly taking detours. It is precisely these more experimental pathways that can lead to new ideas for tools, or generate fresh insights into the data. The prototypes are non-incremental steps towards a final concept, iteratively informed by feedback from our domain experts and other potential future operators.

7 Conclusion and Future Projects

The converging of multiple heterogeneous data sources containing millions of nodes and edges for a graph-based research infrastructure that enables historical social network analysis creates a plethora of multidisciplinary challenges:

- Difficulties associated with the merging of heterogeneous data sources
- Performance of a system regarding the given scope and further scaling of the data
- Creation of domain-customized interfaces, which are open and flexible with regard to unforeseen research questions

- Integration of domain knowledge into the process
- Visualization of millions of data points to provide explorable access points in addition to search interfaces

We address these challenges by focusing on tight, interdisciplinary collaboration and constant evaluation during the whole research and development process. Our approach brings together historical network specialists, data visualization researchers, data scientists, and experts on the evaluation of information infrastructure, an important example being the initial co-design workshop with additional external HNA practitioners and other domain experts. Building on the contextual data gathered during the co-design workshop, we will continue to follow a human-centered approach towards data modeling and visualization design.

In our next step, we aim to take a closer look at the individual processes behind historical network research in one-on-one interviews with domain experts concerning their approaches to HNA research. Our plans for the future also include the merging of multiple visualization concepts into one prototype, which will join global overviews of our data with local views of specific individual networks inside it. Furthermore, we will make use of our data and our interface to provide exemplary use cases on a variety of historical topics in collaboration with our HNA experts.

Finally, we are experimenting with linked data as an alternative to Neo4j. Here, the source data would be modeled in the form of subject–predicate–object expressions and stored in GraphDB.[31] This approach would simplify the integration of Linked Open Data datasets (Wikidata, DBpedia,[32] GeoNames,[33] etc.), and would provide more sophisticated inference possibilities. In preliminary comparisons of the two approaches, GraphDB also shows better performance, but employing it would mean that the source data must be remodeled in order to display relation features such as relation type, relation source type, and temporal validity.

In this paper, we have described the process of examining the potential of remodeling and merging (bibliographic) big data from cultural heritage institutions into one single gathering point optimized for the use in historical network analysis. It is our hope that by providing insights into emerging challenges and outlining possible solutions, we can encourage additional research and scholarly exchange in and with similar HNA-related projects.

[31] http://graphdb.ontotext.com

[32] https://wiki.dbpedia.org

[33] https://www.geonames.org

Acknowledgements

We would like to thank the participants of our co-design workshop and our project partners Heiner Fangerau, Katrin Getschmann, Thorsten Halling, Hans-Jörg Lieder, Gerhard Müller, Clemens Neudecker, David Zellhöfer, and Josefine Zinck. This research is part of the research project SoNAR (IDH) and is funded by the DFG – German Research Foundation (project no. 414792379).

References

Ahn, J.-w., Plaisant, C., and Shneiderman, B. (2013). A Task Taxonomy for Network Evolution Analysis. *IEEE transactions on visualization and computer graphics*, 20(3):365–376, DOI: `10.1109/TVCG.2013.238`.

Allison-Bunnell, J. (2016). Review of Encoded Archival Description Tag Library – Version EAD3. *Journal of Western Archives*, 7(1):1–3, DOI: `10.26077/af62-2a86`.

Auer, S. and Mann, S. (2019). Towards an Open Research Knowledge Graph. *The Serials Librarian*, 76(1-4):35–41, DOI: `10.1080/0361526X.2019.1540272`.

Behrisch, M., Bach, B., Henry Riche, N., et al. (2016). Matrix Reordering Methods for Table and Network Visualization. In *Computer Graphics Forum*, volume 35, pages 693–716. Wiley, DOI: `10.1111/cgf.12935`.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 Data-Driven Documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, DOI: `10.1109/TVCG.2011.185`.

Cantara, L. (2005). METS: The Metadata Encoding and Transmission Standard. *Cataloging & classification quarterly*, 40(3-4):237–253.

Chang, D., Ge, Y., Song, S., Coleman, N., Christensen, J., and Heer, J. (2009). Visualizing the Republic of Letters. `https://web.stanford.edu/group/toolingup/rplviz/papers/Vis_RofL_2009`.

Chen, K.-I., Dörk, M., and Dade-Robertson, M. (2014). Exploring the Promises and Potentials of Visual Archive Interfaces. In *iConference 2014 Proceedings*, pages 735 – 741. DOI: `10.9776/14348`.

Drucker, J. (2011). Humanities Approaches to Graphical Display. *DHQ: Digital Humanities Quarterly*, 5(1):1–21.

Efer, T. (2016). *Graphdatenbanken für die textorientierten e-Humanities*. PhD thesis, Universität Leipzig, `https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa-219122`.

Fekete, J. and Plaisant, C. (2002). Interactive Information Visualization of a Million Items. In *IEEE Symposium on Information Visualization, INFOVIS 2002.*, pages 117–124.

Gibson, H., Faith, J., and Vickers, P. (2012). A Survey of Two-Dimensional Graph Layout Techniques for Information Visualisation. *Information Visualization*, 12(3-4):324–357.

Glaser, B. and Strauss, A. (1967). *The Discovery of Grounded Theory*. Weidenfield & Nicolson, London.

Henry, N. and Fekete, J.-D. (2006). Matrixexplorer: a Dual-Representation System to Explore Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):677–684.

Hinrichs, U., Forlini, S., and Moynihan, B. (2019). In Defense of Sandcastles: Research Thinking through Visualization in Digital Humanities. *Digital Scholarship in the Humanities*, 34(1):i80–i99.

Isenberg, P., Zuk, T., Collins, C., and Carpendale, S. (2008). Grounded Evaluation of Information Visualizations. In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization*, pages 1–8. DOI: `10.1145/1377966.1377974`.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). 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`.

Jansen, D. and Wald, A. (2007). Netzwerktheorien. In Benz, A., Lütz, S., Schimank, U., and Simonis, G., editors, *Handbuch Governance: Theoretische Grundlagen und empirische Anwendungsfelder*, pages 188–199. VS Verlag für Sozialwissenschaften, Wiesbaden, DOI: `10.1007/978-3-531-90407-8_14`.

Kerracher, N., Kennedy, J., and Chalmers, K. (2015). A Task Taxonomy for Temporal Graph Visualisation. *IEEE transactions on visualization and computer graphics*, 21(10):1160–1172.

Kruk, S. R., Synak, M., and Zimmermann, K. (2005). MarcOnt – Integration Ontology for Bibliographic Description Formats. In *International Conference on Dublin Core and Metadata Applications*, pages 231–234.

Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., et al. (2006). Task Taxonomy for Graph Visualization. In *Proceedings of the 2006 Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pages 1–5. DOI: 10.1145/1168149.1168168.

Matschinegg, I. and Nicka, I. (2018). REALonline Enhanced. Die neuen Funktionalitäten und Features der Forschungsbilddatenbank des IMAREAL. *MEMO*, 2:10–32, DOI: 10.25536/20180202.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 1802.03426, https://arxiv.org/abs/1802.03426.

Molina León, G. and Breiter, A. (2020). Co-creating Visualizations: A First Evaluation with Social Science Researchers. *Computer Graphics Forum*, 39(3):291–302, DOI: 10.1111/cgf.13981.

Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, DOI: 10.1109/TVCG.2009.111.

Novak, J., Micheel, I., Melenhorst, M., Wieneke, L., et al. (2014). HistoGraph – A Visualization Tool for Collaborative Analysis of Networks from Historical Social Multimedia Collections. In *18th International Conference on Information Visualisation*, pages 241–250. DOI: 10.1109/IV.2014.47.

Pienta, R., Abello, J., Kahng, M., and Chau, D. H. (2015). Scalable Graph Exploration and Visualization: Sensemaking Challenges and Opportunities. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 271–278. DOI: 10.1109/35021BIGCOMP.2015.7072812.

Pienta, R., Hohman, F., Tamersoy, A., Endert, A., et al. (2017). Visual Graph Query Construction and Refinement. In *SIGMOD '17: Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1587–1590. DOI: 10.1145/3035918.3056418.

Purschwitz, A. (2018). Netzwerke des Wissens – Thematische und personelle Relationen innerhalb der halleschen Zeitungen und Zeitschriften der

Aufklärungsepoche (1688–1818). *Journal of Historical Network Research*, 2(1):109–142, DOI: 10.25517/jhnr.v2i1.47.

Rehm, G., Berger, M., Elsholz, E., Hegele, S., et al. (2020). European Language Grid: An Overview. In Calzolari, N., Béchet, F., Blache, P., Cieri, C., et al., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3359–3373. https://www.aclweb.org/anthology/2020.lrec-1.413/.

Shneiderman, B. (2008). Extreme Visualization: Squeezing a Billion Records into a Million Pixels. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 3–12. DOI: 10.1145/1376616.1376618.

Thudt, A., Hinrichs, U., and Carpendale, S. (2012). The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization. In *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1461–1470. DOI: 10.1145/2207676.2208607.

van Ham, F. and Perer, A. (2009). ”Search, Show Context, Expand on Demand”: Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):953–960, DOI: 10.1109/TVCG.2009.108.

von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., et al. (2011). Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. *Computer Graphics Forum*, 30(6):1719–1749, DOI: 10.1111/j.1467-8659.2011.01898.x.

Vorndran, A. (2018). Hervorholen, was in unseren Daten steckt! Mehrwerte durch Analysen großer Bibliotheksdatenbestände. *o-bib. Das offene Bibliotheksjournal*, 5(4):166–180, DOI: 10.5282/o-bib/2018H4S166-180.

Warren, C. N., Shore, D., Otis, J., Wang, L., Finegold, M., and Shalizi, C. (2016). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, 10(3), DOI: 10.17613/M6B020.

Wintergrün, D. (2019). *Netzwerkanalysen und semantische Datenmodellierung als heuristische Instrumente für die historische Forschung*. PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, https://nbn-resolving.org/urn:nbn:de:bvb:29-opus4-111899.

Zimmerman, J., Forlizzi, J., and Evenson, S. (2007). CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502. DOI: 10.1145/1240624.1240704.