CEUR-WS.org/Vol-3152/BD2019_paper_4.pdf

Mining the *Biographical Dictionary of Republican China*, from Print to Network Exploration

Pierre Magistry, Cécile Armand, Christian Henriot

Aix Marseille Univ, CNRS, IrAsia, Marseille, France
pierre.magistry@univ-amu.fr, cecile.armand@gmail.com, christian.r.henriot@gmail.com

Abstract

This article describes preliminary experiments conducted in the context of the ENP-China project, which examines the transformation of elites in modern China. The project is centered on exploiting information from untapped textual sources at a large scale which requires to investigate new methodologies for data-rich history and to rely on Natural Language Processing (NLP). The first experiments presented in this paper were designed on a smaller scale and better-known materials to test and develop adequate tools and methodology - at a humanly manageable scale - before eventually enlarging the corpus and scale of analysis. We focus on the Biographical Dictionary of Republican China (BDRC) edited by H. Boorman and aim at extracting biographical information and transforming a conventional dictionary into a reservoir of data on elites in modern China.

Keywords: China, History, Biography, Data Mining, NLP, Graph Visualization

1 Context

The ENP-China project (ERC advanced grant) proposes a step-change in the study of modern China reliant upon scalable data-rich history to create a new dimension in the study of the transformation of elites in modern China. The key issue that the project wants to address is breaking through existing limits of access to historical information that is embedded in complex sources and its transformation into refined, re-usable and sustainable data for contemporary and future study of modern China. This project is rooted in historical research, but it adopts a highly multidisciplinary approach, including computational linguistics to explore and process large textual corpora such as the mid-19th mid-20th press. The project focuses on elites in urban China as actors whose status, position, and practices were shaped by the power configurations that developed over time and whose actions through institutions and informal/formal networks in turn were a determining factor in redrawing social and political boundaries.

2 Introduction

The Biographical Dictionary of Republican China (BDRC) (Boorman et al., 1967) has served generations of China historians, mostly as a reference work to check out major historical figures of the Republican era. Although the work has fallen into oblivion due to its format (print), obsolete transliteration system (Wade-Giles), and current research practices (digital), it remains a formidable source on elites in Republican China. We take it as a container of historical data on a much wider range of figures than the 588 selected individuals with a view to reconstitute it as a database with updated and enriched data. Our objective is, on one hand, to examine the characteristics of the BDRC population and on the other hand to challenge this sample as representative of the elites in Republican China.

The BDRC consists of 4 volumes of about 500 pages each and an index volume produced separately a decade after initial publication. The first four volumes describe the 588 individuals in 1.1 million tokens.

The purpose of processing the BDRC was threefold:

- to conduct a first experiment on a manageable dataset, bearing in mind the different nature and scale of the press corpora to be mined in the course of the project;
- to enable interactions at the very first stage of the project between computational linguistics, NLP, and historical inquiry ;
- to initiate the production of atomized biographical data through the transformation of text into data and to start designing an appropriate database for the project.

This paper is organized as follows: The next section introduces related works and briefly explains where our work falls in the field of biographical data processing. Section 4 gives an overview of the NLP pipeline used to obtain an annotated version of the corpus as a graph database. Section 5 describes different types of information that can be extracted from the annotated corpus. Finally, Section 6 introduces the tool and method we adopted to visualize the extracted data. As we are presenting a work-in-progress, we will not provide a proper conclusion. However, in sections 7 and 9 there is an attempt to share more insights and our plan on how to scale up.

3 Related Works

A substantial amount of works in Digital Humanities focuses on processing biographical materials. It covers theoretical issues regarding the definition of data structures (Beretta, 2017; Thierry and Sprugnoli, 2017; Fokkens and Ter Braake, 2017), and more practical aspects of the processing methods as well as the design of query interfaces. The present paper does not address the theoretical aspects about ontological definition of what constitute biographical data. We did not adopt a specific data schema from the start and chose to proceed in the opposite direction, which could be described as *bottom-up*. We ran experiments with information extraction tools to study what kind of data can be retrieved automatically.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Due to the relatively small scale of our experiments and some copyright issues to be resolved, we cannot provide an online interface to the full-texts through a web portal similar to those presented in other works such as (Reinert et al., 2015; Raghallaigh and Cleircín, 2015). We still plan to release most of our productions under a permissive licence when possible (see Section 8 for details). Regarding the language technogolies used in this work, our NLP processing pipeline for data extraction, which is described in the next section is very similar to the one used in (Dib et al., 2015). The main originality of our approach is to rely on a graph-based exploration at different stages of the data processing. It ranges from clean, manually curated and augmented data, to noisy output of the NLP pipeline.

4 NLP Pipeline

In this section, we will describe the set of tools that we relied on to process the texts and extract information. When processing documents in English, our goal is not to design better models and NLP techniques for every step of the analysis, but rather to provide a relevant and complete tool chain for our scenario. We tried first to assess what could be achieved with off-the-shelf tools such as CoreNLP (Manning et al., 2014). The main issue we encountered in fact was at the tokenization step, which forced us to adapt the tokenizer to our corpus. Other modules were left unchanged except from some simple and systematic post-processing of the Named Entity Recognition (NER).

41 The Big Picture

Figure 1 illustrates the whole processing of the BDRC dictionary. Starting from an OCRized version of the text (keeping volume and page number information), the first step was to split the four volumes into a biography entry basis, to obtain one document per individual. Each document was then segmented into paragraphs and sentences. Sentences are our basic unit of information and undergo NLP analysis with tokenization, part-of-speech tagging, syntactic (dependency) parsing and NER. All these steps were performed with pretrained modules from the CoreNLP toolkit and will progressively be replaced by more state-of-the-art tools. The resulting annotation is stored in a graph-database for seamless querying and information extraction. We always keep the trace of the document for each information stored in this database. Visualization of the NER annotation layer is also provided using FLAT[1], after a conversion of the document into the FoLiA format (van Gompel and Reynaert, 2013).

42 Tokenization

It may seem surprising to the NLP community, where much of the effort focuses on tasks of higher abstraction such as parsing, NER or entities linking, but some of the most damaging and obvious processing errors occurred as early as the tokenization step. The default English tokenization in CoreNLP is simply unable to handle the transliteration of Chinese names from that period, which use the Wade-Giles romanization system. An example of such names is *Ch'en Kung-po*, which is split into 6 tokens (Ch / ' / en / Kung / - /po) where we would expect two (Ch'en / Kung-po). This kind of tokenization mistakes results in a cascade of errors in the subsequent modules and failures in the syntactic parsing and NER. This concerns the majority of person names and an important part of locations and organizations. It is however fairly straightforward to derive a Wade-Giles-aware tokenization module for the CoreNLP toolkit and avoid such mistakes.

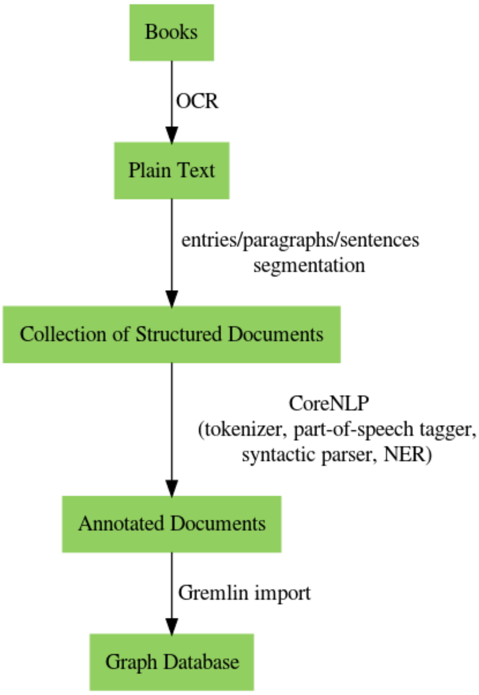

Books
OCR
Plain Text
entries/paragraphs/sentences segmentation
Collection of Structured Documents
CoreNLP (tokenizer, part-of-speech tagger, syntactic parser, NER)
Annotated Documents
Gremlin import
Graph Database

Figure 1: overview of our processing pipeline

43 Annotation and Storage

Other modules from CoreNLP were used without modification to provide annotation layers on the tokens . The output is illustrated on Figure 3 and 4.

After a first run of the pipeline, we performed a manual editing of Named Entities at the type level. We ordered the recognized forms by their number of occurrences and established the list of systematic corrections, which could be a change of entity type (e.g. for an organization recognized as a location) or an erasure. This produced a list of corrections to be applied on the whole corpus.

We stored this information in JanusGraph[2], an Apache TinkerPop[3] compatible graph database which allows us to easily perform graph traversals to match and extract patterns. By opposition to similar tools which are specifically crafted for linguistic inquiries such as *Grew* (Guillaume et al., 2012)[4] and are often limited to specific levels of linguistic analysis (such as graph matching in dependency trees), using a more generic graph-database enables us to merge all the layers of analysis, from document structure to NER

[1] https://github.com/proycon/flat

[2] https://janusgraph.org/

[3] https://tinkerpop.apache.org/

[4] http://grew.fr/

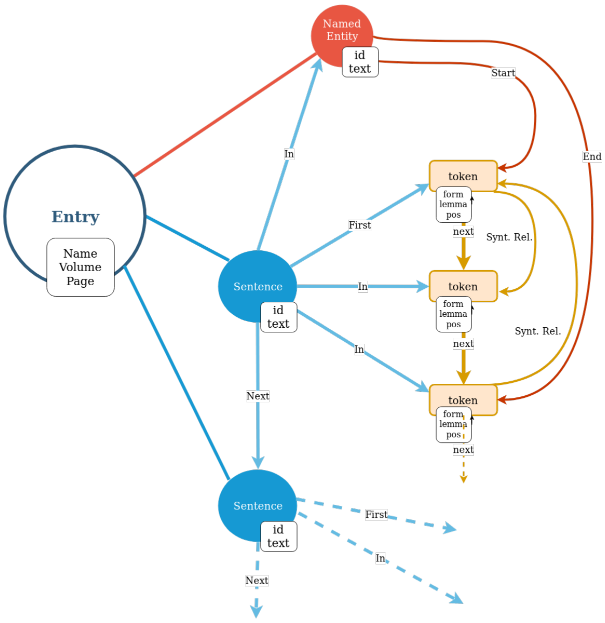


Figure 2: The very beginning of an entry in the graph-database

annotations as well as including syntactic annotation. It is then possible to design graph traversals which combine all the information available. Future additions to annotation layers will be easy and it is possible to write convenient Domain Specific Language (DSL) on top of the *Gremlin* layer[5] to ease the design of complex queries. The data model of the graph-database is shown on Figure 2

5 Information Extraction

Once the entries have been annotated by our NLP pipeline, it becomes possible to design patterns (on strings with labels or on graphs of annotations) to extract the information we are looking for.

51 Information types

Every piece of information is located in a specific sentence, related to a dictionary entry (hence, a specific person). We targeted information of various kinds whereby some could be directly recorded in a biographical database such as the birth date, while others required more subtle mappings such as positions:

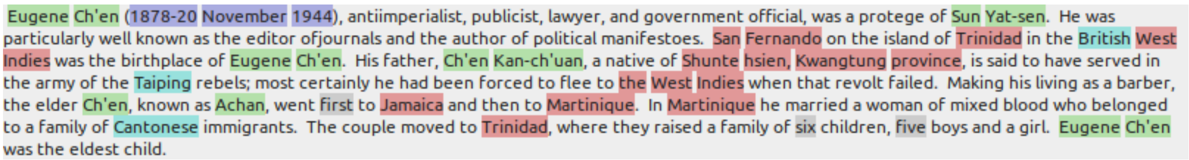
Birth and death From the very first paragraph of each biography (see figure 3), valuable information can be extracted regarding birth and death. Dates could be retrieved thanks to a very simple regular expression (although in a few cases, OCR errors caused the extraction to fail and required manual correction). Many persons described in the dictionary were still alive at the time of writing. In such case the date and place of death had to be retrieved from other sources. The birth places were extracted automatically, but as they were expressed (sometimes inconsistently) in an old transliteration for Chinese place names, they had to be normalized manually to be usable in a GIS. This step involved a large part of manual checking and additions to the original source, but as it concerned only basic information for each person, it was a sustainable workload.

Educational Background Next we wanted to extract information regarding the education background. This information was spread over a larger part of each entry and typically used a free form of writing. For this step and the following one, we first considered relying on FrameNet (Fillmore et al., 2003), but using a concordancer we observed that the vocabulary used in the BDRC was consistent enough to enable relying on the selection of *trigger words* and syntactic patterns to extract relevant phrases and words. Such *triggers* included verbs like “to graduate”, “to study”, “to attend” and nouns like “student”, “university”, “chü-jen”, “hsueh-t’ang”…This kind of patterns is illustrated on figure 4 and described in more details in the next section.

This method yielded a large quantity $(1,610$ entries) of matches, which we could present in tabular format. It was still tractable to go through all the extractions manually to check them and provide normalized forms of fields like the institution where the education was received, the degree(s) obtained and the field of study. Additionally, we had to enrich it with the location (educational institution) and their name in Chinese. With this cleaned dataset, we could produce visualizations using graphs and GIS (see Section 6). Although it may appear that a lot of manual work was involved, it was not so much to correct the extraction itself, than to “update” the extracted data in a standardized format for proper identification or location. It was an unavoidable step that produced new dictionaries to be used in the exploration of the press corpora.

Positions To track the positions occupied by all the individuals, we adopted a similar methodology, changing only the set of trigger words and syntactic patterns. However, we then had to face a much larger set of potential data points (more than $5,000$). Individuals in the BDRC have a single birth and death. They may have enrolled in a couple of institutions during their studies (if we focus on higher education). But the number of positions each may have occupied is much larger and is often listed in succession without date next to each occurrence. Thus it became very difficult to double-check all the data and provide normalized forms for jobs and institutions. We had to turn to data visualization tools before we were able to clean the data. To replace manual normalization, we enriched our data using WordNet (Fellbaum, 1998) to provide links between similar positions. A graph was then constructed from all our instances, it is composed of 14 thousands nodes and 29 thousands edges.

[5] `https://github.com/mpollmeier/gremlin-scala#build-a-custom-dsl-on-top-of-gremlin-scala`



Eugene Ch'en (1878-20 November 1944), antiimperialist, publicist, lawyer, and government official, was a protege of Sun Yat-sen. He was particularly well known as the editor ofjournals and the author of political manifestoes. San Fernando on the island of Trinidad in the British West Indies was the birthplace of Eugene Ch'en. His father, Ch'en Kan-ch'uan, a native of Shunte hsien, Kwangtung province, is said to have served in the army of the Taiping rebels; most certainly he had been forced to flee to the West Indies when that revolt failed. Making his living as a barber, the elder Ch'en, known as Achan, went first to Jamaica and then to Martinique. In Martinique he married a woman of mixed blood who belonged to a family of Cantonese immigrants. The couple moved to Trinidad, where they raised a family of six children, five boys and a girl. Eugene Ch'en was the eldest child.

Figure 3: First paragraph of the biography of Eugene Ch'en, with NER annotations rendered by FLAT The begining of each entry includes similar information regarding basic birth and death informations.

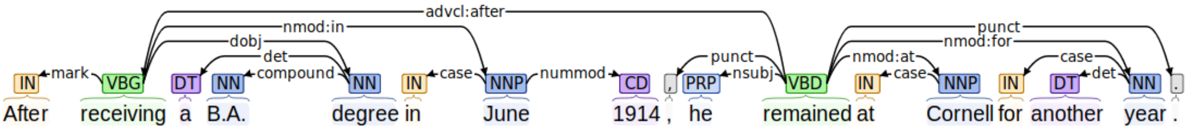


Figure 4: example of syntactic analysis for a sentence containing the trigger word *degree*. The word itself trigger the processing. the *dobj* relation between *degree* and *receiving* confirms it is a relevant match. The *compound* relation allows to enlarge the target and capture the expression *B.A. degree*. *June 1914* and *Cornell* are identified as DATE and ORGANIZATION respectively and are also extracted.

52 Technical details

The string matching to retrieve birth and death information was simply based on regular expressions and Named Entities (Location) annotations. The methods we implemented for education and positions on the other hand deserves a more detailed explanation.

Using a concordancer on the annotated text, we skimmed through examples of trigger words in context. Based on these observations, we were able to describe syntactic patterns to confirm a match around occurrences of the selected trigger words. Because the whole linguistic annotation is merged in a single graph database, we were able to extract sub-parts of the sentences based on traversals on the syntactic graph and NER annotations for each confirmed match. For example, the word *degree* would trigger a possible match in the sentence *"After receiving a B.A. degree in June 1914, he remained at Cornell for another year."* Its morphosyntactic analysis is illustrated on Figure 4. Starting from this occurrence of the word *degree*, we can check that it is in direct object position (*dobj* of *receiving*) and capture the compound *B.A. degree*. With the addition of the NER layer, it is possible to retrieve the DATE span *June 1914* and the ORG *Cornell*. We thus proceeded in three steps:

- finding all occurrences of a trigger word
- checking its syntactic position to confirm a match
- extracting information by following the links in the graph-database.

Extracted information was presented in a tabular format. The table produced could then be used for manual edition or to build graphs for visualization.

6 Graph-based Exploration

We used graph visualization tools to obtain different views of the data we extracted. Both manually cleaned and noisy datasets can benefit from such an approach. It is important to emphasize that there is no single graph which can properly and fully represent a dataset. The way the graph is built needs to be related to some question about the data. In this work we simply wanted to get a better understanding of what data was collected from the biographical dictionary. In order to do so, we relied on Padagraph[6]an online graph exploration tool we introduce in the next section. We will then illustrate its use on the two cases of (cleaned) education data and (noisy) positions data.

61 Padagraph

Padagraph is a web-based tool designed to allow for collaborative graph edition and visualization. It has been successfully adapted for lexicography in the RLF project (Polguère, 2014), and for a bibliographic search engine in the Istex project[7]. We are now exploring its potential on historical data.

Given tabular description of nodes and edges of a graph, which can be CSV files hosted anywhere or an online spreadsheet editing service such as EtherCalc, Padagraph builds the corresponding graph and provides various graph layouts and clustering algorithms. Nodes and edges can receive properties (by adding columns in the tables) which are displayed when a node is selected. We use properties to add quotes of the full text and urls pointing to the source document.

Beside the availability of 3D layout and the possibility to edit the data online in a collaborative way in real time (two features that are not used here), what distinguishes more specifically Padagraph from other similar tools is the possibility to work on very large graphs by displaying only sub-graphs based on searches and expand queries. Random Walk techniques are used to retrieve nodes in the neighborhood of a starting point without being limited to direct neighbors. This enables the user to visualize the local structure of a dataset around a first search, and to iteratively expand the displayed sub-graph around the nodes of interest.

It may be difficult to make sense of the visualization of our full 14k nodes positions graph (which could be done with

[6] `https://padagraph.io/`
[7] `https://www.istex.fr/cillex/`

other offline tools such as Gephi or Cytoscape), but with Padagraph one can explore the graph through an interactive process starting from a specific person or position.

62 Education Graph

As described in Section 5, the education graph is built after having manually normalized and supplemented the data extracted from the BDRC with other sources. The result of this step is a large table of $1,610$ lines. Each line corresponds to an extraction, it includes the ID of the entry from which it was extracted and the full sentence. It also indicates the original (automatic) analysis output, with the trigger word, the subject and object in each sentence as well as recognized organizations, locations and dates. It is supplemented with curated data including standardized forms for the institution, its country and city, the level of education, the discipline studied and a time span when available.
From this table we can imagine building many graphs. Here we illustrate with a bipartite graph relating individuals to institutions. Among the 1,610 lines of information in the full table, 759 include a standardized institution name. Those relate 474 institutions to 332 individuals. The corresponding graph can be visualized at `http://enpchina.eu/boorman/education/`

63 Positions Graph

When addressing positions, we reached a quantity of extracted information that made it difficult to extensively verify and standardize manually. In this case, we will not be exploring a carefully curated data. Instead we used Padagraph as a tool to explore the output of the automatic analysis. In order to make this exploration more efficient and compensate for the noisy data, we included more information in the graph. Extracted positions are typically noun phrases. In some cases the whole phrase denotes the type of position, in some others it only provides the head noun and the rest of the phrase may be too specific. In other cases, we may want to keep the head noun and its closest adjective or compound. We kept these three level of specificity and created links between position names when they were considered close-synonyms in WordNet. Take for example the sentence *"he became special adviser to the ministry of communications."*. In this case, the node for this sentence in our graph will be linked to *special adviser to the ministry of communications.*, *special adviser* and *adviser*. By exploring the surroundings of this sentence in the graph, we can discover that this is the only mention of a *special adviser to the ministry of communications.*, but that we have another *special adviser*, namely Hu Shih who was *special adviser to the Executive Yuan*. A screenshot zoomed on this part of the graph is presented on figure 5
The initial graph we obtain can be described as a set of $16,058$ nodes connected by $33,460$ edges. Nodes are of the following types:

Sentence $4,419$ extracted snippets

Document 577 documents (entries from the dictionary)

Named Entity $2,919$ types (not including dates)

Date 680

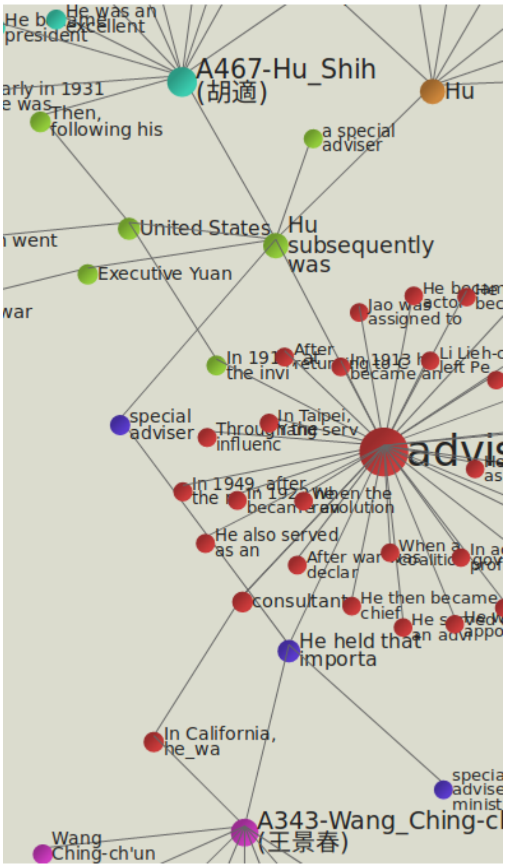


Figure 5: The sentences "Hu subsequently was made a special adviser to the Executive Yuan..." (Hu Shih) and "He held that important post until 1924, when he became special adviser to the ministry of communications." (*he* being Wang Ching-ch'un) are connected through both *special adviser* and *adviser*, the later occurring in many other sentences (in red)

Position $5,076$ nouns (or noun phrases)

Synset $2,387$ synsets from wordnet

A final step of filtering is added to discard nodes with less than two neighbors and nodes with too many neighbors. The resulting graph can be visualized at `http://enpchina.eu/boorman/positions/`.

In the case of positions, the usage we can have of Padagraph resembles less typical Network Analysis cases and more a search engine which discloses the output of the NLP pipeline, including its imperfections. As mentioned in (Fokkens et al., 2018; Fokkens-Zwirello et al., 2014), relying on NLP to address research questions in Humanities requires to let the researcher have a clear idea about the limitations and the behaviour of the NLP tools. Scores typically provided in most NLP publications are far from giving the full picture of what can be expected from the tools.

> One of the most important aspects of the evaluation is that it should raise awareness to the end user about what NLP analysis can do and what it

> cannot do. As we will argue, the standard precision and recall evaluations are not sufficient to provide the necessary insights to historians using the output of our automatic analyses. It is also important to provide insight into the kind of errors made by analyses, so that end users are aware of potential biases introduced by the tools.

Our use of graphical exploration of the raw output from the NLP pipeline without hiding errors cannot replace proper evaluations, but we hope that providing a convenient interface to navigate through the graph, which can be regarded as "the machine's point of view about our data" can be a way to *raise awareness* among the historians involved in our project.

To summarize the use of graphs in our work, we can distinguish between three main cases. Firstly, the results of the NLP pipeline are stored in a graph database to allow for pattern matching (Figure 2 and 4), these graphs are drawn here as illustrations but are usually not visualized. Secondly we can visualize the graph structure of curated data, to provide a new way to explore a specific dataset. This requires some pre-defined hypothesis or broad questioning to design the graph. Our "Education Graph" is of this kind, as it lets us explore our data following the communities of persons and institutions defined by the relation '*having received education in a specific institution*'. Thirdly, we use graph exploration to expose a larger but more noisy dataset about *positions*. This enables the researcher to have a more global picture of the data contained in the corpus, but may require multiple iterations of filtering and graph re-building to provide a more efficient workflow. Making such iteration easier is one of our main concern for future work.

7 Feedback from historians

In assessing the performance of the tools and techniques implemented in this experiment, one needs to distinguish between issues that relate to the tools/techniques and those that relate to the very nature of the processed document, especially the naming of targeted entities, all labelled in English or in an obsolete transliteration system. The quality and precision of information tends to decrease with the complexity of the searched topics (e.g. birth vs. positions). Yet the main issue is the heterogeneity of the retrieved terms (which is source-dependent) and the overload of information that needs to be curated to be fully usable in a database. Conversely, the manual retrieval of the same information would have required a full month of work based on a conservative assessment, versus a few minutes with NLP tools, and would still have required the same amount of formatting and converting named entities into current standards. The main pitfall identified in implementing NLP tools was the production of a number of "false positives", namely incorrect attribution of properties (education, position) to individuals.

8 Availability of the Results

We will release the outputs of our work under a permissive licence to allow for further research. The BDRC is still covered by copyrights so it seems difficult, prior to an agreement with the publisher, to provide the output of the full-text analysis, but we will provide the extracted and enriched synthetic datasets. The graph visualization is available on our project website, and is linked to the original documents on the Internet Archive[8]. It can thus be used as an enriched index which can serve as digital entry point to benefit from the BDRC. We will, however, create internally X-Boorman, a digitally enriched edition of the BDRC pending further discussions with the publisher.

9 Perspectives

In this work, we focused on a single source. We remained at a small scale compared to the scope of our project, but this experiment will be replicable, with adjustments, to full range of other biographical works such as Who's whos, directories, etc. Yet we already saw the limits of what we can address or correct manually. This work was a test-case to explore the methodology and possible interactions between the disciplines involved in our project.

To process our whole corpus will require numerous modifications in our work flow, as this corpus will include a century of major periodicals, including newspapers, and a wide range of other materials both in English and Chinese. Putting aside the fact that the "Chinese" language in our corpus is quite different from Modern Standard Mandarin (a difference which deserves a study in itself), the change of scale and the diversity of domains to address is very likely to stifle our *trigger words* approach. We will have to turn to other solutions such as FrameNet or vector-space semantic. As a result, we expect the output of the pipeline to contain more noise. We plan to define a work flow involving a step of subcorpus selection, where an historian can define which documents to work on, and from there to provide a more automatic way to reach the graph exploration step. From this perspective, we can consider the BDRC as a consistent subcorpus of our global collection of documents. We will also connect this work to a manually curated database whose content will help us to improve the recall of the NER and allow us to perform entity linking to create connections between the database and the sources.

10 Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788476).

11 References

Francesco Beretta. 2017. L'interopérabilité des données historiques et la question du modèle : l'ontologie du projet SyMoGIH. In Brigitte Juanals et Jean-Luc Minel, editor, *Enjeux numériques pour les médiations scientifiques et culturelles du passé* , Notions et méthodes. Presses universitaires de Paris Nanterre.

H.L. Boorman, R.C. Howard, D. Howard, J.K.H. Cheng, and J. Krompart. 1967. *Biographical Dictionary of Republican China*. Columbia University Press.

[8] `https://archive.org/details/biographicaldict01boor`

Firas Dib, Simon Lindberg, and Pierre Nugues. 2015. Extraction of career profiles from wikipedia. In *proceedings of the First Conference on Biographical Data in a Digital World, Amsterdam, The Netherlands*, pages 33–38.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250, 09.

Antske Fokkens and Serge Ter Braake. 2017. Connecting people across borders: a repository for biographical data models. In *proceedings of the Second Conference on Biographical Data in a Digital World 2017*, pages 83–92.

Antske Fokkens, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. 2018. Biographynet: Extracting relations between people and events. *Computing Research Repository*, abs/1801.07073.

A.S. Fokkens-Zwirello, S. ter Braake, C.J. Ockeloen, P.T.J.M. Vossen, S. Legêne, and A.T. Schreiber. 2014. Biographynet: Methodological issues when nlp supports historical research. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *LREC 2014, Ninth International Conference on Language Resources and Evaluation*, pages 3728–3735. European Language Resources Association (ELRA).

Bruno Guillaume, Guillaume Bonfante, Paul Masson, Mathieu Morey, and Guy Perrier. 2012. Grew : un outil de réécriture de graphes pour le TAL. In Georges Antoniadis, Hervé Blanchon, and Gilles Sérasset, editors, *12ième Conférence annuelle sur le Traitement Automatique des Langues (TALN'12)*, pages 1–2, Grenoble, France, June. ATALA.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Alain Polguère. 2014. From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4):396–418.

Brian Ó Raghallaigh and Gearóid Ó Cleircín. 2015. Ainm. ie: Breathing new life into a canonical collection of irish-language biographies. In *proceedings of the First Conference on Biographical Data in a Digital World, Amsterdam, The Netherlands*, pages 20–23.

Matthias Reinert, Maximilian Schrott, Bernhard Ebneth, and Malte Rehbein. 2015. From biographies to data curation-the making of www. deutsche-biographie. de. In *proceedings of the First Conference on Biographical Data in a Digital World, Amsterdam, The Netherlands*, pages 13–19.

Declerck Thierry and Rachele Sprugnoli. 2017. Considerations about uniqueness and unalterability for the encoding of biographical data in ontologies. In *proceedings of the Second Conference on Biographical Data in a Digital World 2017*, pages 76–82.

Maarten van Gompel and Martin Reynaert. 2013. Folia: A practical xml format for linguistic annotation a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, Dec.