# BiographyNet: Methodological issues when NLP supports historical research

Fokkens, A.; ter Braake, S.; Ockeloen, N.; Vossen, P.; Legêne, S.; Schreiber, G.

**Publication date**
2014

**Document Version**
Final published version

**Published in**
Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014

# BiographyNet: Methodological issues when NLP supports historical research

**Antske Fokkens♣, Serge ter Braake♢, Niels Ockeloen♠,**
**Piek Vossen♣, Susan Legêne♢ and Guus Schreiber♠**
♣Faculty of Arts, CLTL ♢Faculty of Arts, History Dept. ♠Faculty of Sciences, Dept. of Computer Science
The Network Institute, VU University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{antske.fokkens,s.ter.braake,niels.ockeloen}@vu.nl
{piek.vossen,s.legene,guus.schreiber}@vu.nl

**Abstract**

When NLP is used to support research in the humanities, new methodological issues come into play. NLP methods may introduce a bias in their analysis that can influence the results of the hypothesis a humanities scholar is testing. This paper addresses this issue in the context of BiographyNet a multi-disciplinary project involving NLP, Linked Data and history. We introduce the project to the NLP community. We argue that it is essential for historians to get insight into the provenance of information, including how information was extracted from text by NLP tools.

**Keywords:** Digital history, provenance modeling, Linked Data

## 1. Introduction

Digital humanities has been a much discussed topic in art faculties all over the world for the past five years. It was among other things 'big news' at the 2009 Modern Language Association (MLA) Annual Convention in Philadelphia (Liu, 2012, p. 8, 20), (Kirschenbaum, 2012, p. 7). Digital humanities deals with a wide variety of topics touching the use of digital materials (big data) for research and the tools to analyze them (Zaagsma, 2013, p. 15), (Kirschenbaum, 2012, p. 4).

This relatively new direction of research often involves expertise from different fields which each come with their methodological baggage and requirements. In digital humanities projects intense communication between researchers of the different fields is imperative to reach satisfying results. (Svensson, 2012, section III) , (Siemens, 2009), (Ter Braake, 2014). While working in an interdisciplinary field, it is important to be aware of the requirements and methodological approaches that are used by other disciplines involved. For instance, Rieder and Röhle (2012, p. 70-76) and Lin (2012, p. 306) point out that digital tools rely on assumptions and should not be considered to be objective. Both developers of such tools as well as scholars using them need to be aware of the fact that such assumptions may introduce a bias while being used for research.

In this paper, we discuss methodological issues that come into play when Natural Language Processing (NLP) is used to support historical research. The discussion is centered around BiographyNet,[1] a multi-disciplinary project bringing together history, Linked Data and NLP. BiographyNet aims at enhancing the research potential of the Biography Portal of the Netherlands,[2] a heterogenous collection of Dutch biographies. The portal contains short biographies and a limited set of metadata of more than 76,000 different people mentioned in various resources. Research in BiographyNet is meant to increase its potential for historical research by transforming the available data into a semantic knowledge base and through the creation of a demonstrator.

The goal of this paper is two-fold. First, we introduce the BiographyNet project to the NLP community. The paper thus particularly focuses on the role of NLP in the project. Second, we reflect on methodological issues that come into play when automatic text analysis is used to support academic research in another discipline. We argue that provenance modeling plays an essential role in historical research. There should not only be an indication of the used sources, but also insight into the process that led to a specific result. Awareness of provenance is both important for the historians using NLP output as well as for NLP researchers developing a system for historians. We will address provenance from both of these angles in this paper.

The rest of this paper is structured as follows. Section 2. provides background information on the data from the Biographical portal. In Section 3., we describe the project from the perspective of historians. The role of Linked Data in this project is outlined in Section 4.. We present examples of how NLP methods may introduce a bias and explain what information on provenance should be provided in Section 5. In Section 6., we describe a first basic system for information retrieval. This section also provides an example of how provenance of a simple NLP pipeline may be modeled. We present our conclusions in Section 7.

## 2. The Biography Portal

The Biography Portal of the Netherlands provides access to over 125,000 entries describing 76,000 people considered (more or less) prominent figures of Dutch history. Each entry contains metadata and around 80% includes biographical text. It is a heterogenous collection made up out of 23 sources. The oldest sources date from the eighteenth century, while other sources are still being updated today. The largest sources are the nineteenth century *Biographisch Woordenboek der Nederlanden* (van der Aa, 1878) and the *Nieuw Nederlandsch Biografisch Woordenboek* (Blok and Molhuysen, 1937, NNBW) from the first half of the twentieth century, together good for over 45,000 biographies.

This variety in sources from different time periods and with different themes (e.g. socialists, women, artists) is unique in comparison to other online national biographical dictio-

[1] http://www.biographynet.nl/
[2] http://www.biografischportaal.nl/en

nary projects, which usually limit themselves to one or two sources.[3]

The portal can be searched with a full text search and for basic information such as name, religion, date and place of birth and category. The basic information is provided in an additional layer of metadata created for each individual (Hoekstra, 2013). The searchable metadata are based on metadata from different sources. In case of conflict, information from sources considered more reliable was selected. The completeness of the metadata varies significantly: some sources come with rich metadata, others with very limited metadata. The completeness of metadata for an individual thus depends on the biographical sources he or she is described in. The metadata of some sources provides a relatively complete picture of the individuals that are described. For instance, the metadata from the Parlementary Documentation Center includes all information known about someone's education and carreer path. Other resources provide elaborate information in the text, but minimize metadata to the name and place and date of birth and death. Furthermore, the structure of texts, language use and length of texts varies from one biographical source to another. The possibilities provided by information extraction as well as the challenges involved in this task thus highly differ depending on the source. We will elaborate on these properties in Section 5. First, we will elaborate on digital history in the context of BiographyNet.

## 3. Digital History and BiographyNet

The Biography portal of the Netherlands can already be seen as a digital history project in itself, since it brings together a wide variety of sources and offers opportunities for digital research. BiographyNet tries to fully tap into the research potential of the portal by transforming the available biographical data into a semantic knowledge base, facilitating more complex queries and creating a user friendly interface, which meets the provenance requirements of both the historian and the computer scientist. The development of the BiographyNet demonstrator requires a careful consideration of historical research methods.

### 3.1. Historic Methods

Historic research has some challenges that are generally agreed to be specific for the discipline (Ankersmit, 1983, p. 280), (Mink, 1966, p. 39-40), (Van den Akker, 2012, p. 245-247). First, a conclusion cannot be confirmed or falsified by running a test, but has to become evident from the provided source material and its analysis. Furthermore, historians present visions and syntheses that are based on facts, but do not derive from them directly, i.e. they must *explain* what happened and not only present what happened. Analysis, logic, interpretations, models, theory and quantification lead historians to their view of the past. Because of this creative element in historic research, two historians using the same material can come to different visions (Ricoeur, 2004, p. 242). In theory, the more visions on a certain subject exist, the better for historical scholarship (Ankersmit, 1996, p. 21), because opposing visions may help to get closer to the truth.

One of the main objectives of text mining for historical research is the extrapolation of facts or events. Every historical narrative is built on these basic *building blocks*. There are however, many degrees of how reliable the evidence for a historical event can be. It is for example quite possible to proof a famous historical event like the defeat of Napoleon in 1813 at the battle of Leipzig by referring to a wide variety of sources like pamphlets, chronicles, diaries, newspapers, accounts et cetera. It is more difficult to proof the presence of a lower army officer at the battle of Leipzig, because fewer sources will be available to show that he was there, or maybe the historian has to rely on circumstantial evidence. Events for historical research therefore have a wide variety of "provability". Most academic historical books are based on events that are easy and difficult to prove as well as reconstructions of events.

If we want to visualize what happens when a historian writes a book or article, we can use the metaphor of a *house* (see Figure 1).[4] All the data or facts at disposal can be considered the *building blocks* of the house. The methods used to put those building blocks together, e.g. the logic, analysis, quantification, theory and models, can be seen as the *concrete*. The house is the synthesis, e.g. the history of the fall of Napoleon at Leipzig. Multiple independent building blocks can be put together by the concrete to form a dependent building block. This dependent building block is by nature less stable than the building blocks it was built from since more blocks run the risk of being falsified. This new building block can be used again to form yet another, even more dependent, building block. If either building blocks or concrete are of deficit quality, then the house will fall, or become drafty. If a construction is made well however, a house will not fall if only one block crumbles.

These narrative constructions will endure over time if they are built well. They have their place within a setting of intertextuality (of other works with a similar topic) like a house does in urban space. If the narrative crumbles, it will eventually be demolished and not be used any more by future historians as a reliable source.

### 3.2. A Narrative from Biographies

In BiographyNet, NLP tools are used to extract facts from biographies, i.e. they should identify *building blocks* in the text. These *building blocks* are stored as Linked Data (see Section 4.) to facilitate tools that allow the historian to identify connections between people and events. By doing so, the historian creates new narratives that go beyond the stories of individuals. Metaphorically speaking, the historian uses the *building blocks* extracted from biographies to create new *houses of different dimensions and shape*.

The quality of narratives depends on the reliability of the facts it is built on. It is essential that the historian can

[3] See for example Germany, the Neue Deutsche Biographie: `http://www.ndb.badw-muenchen.de/ndb_aufgaben.htm` (accessed 10 April 2013); United Kingdom, the Oxford Dictionary of National Biography: `http://www.oxforddnb.com/` (accessed 10 April 2013); Sweden, Svenskt Biografikst Lexicon: `http://www.nad.riksarkivet.se/sbl/Start.aspx?lang=se` (accessed 10 April 2013).

[4] This metaphor expands on a similar metaphor used by Ricoeur (2004, p. 150-151).

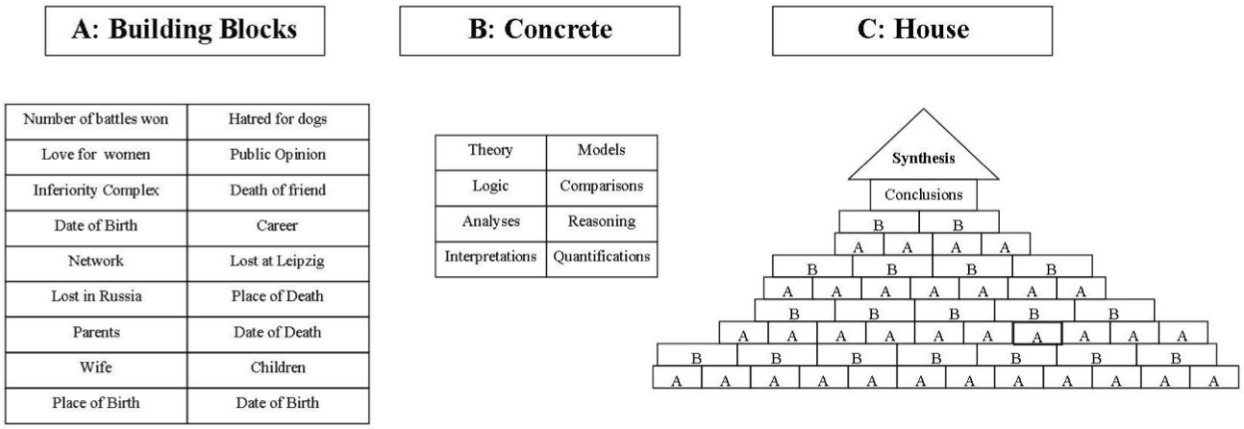


Figure 1: The Historical Narrative

trace the provenance of individual facts. In addition to a pointer to the original document, the historian should know whether information was retrieved from text by NLP tools or whether it was extracted directly from the database. The method that was used and the reliability of the output needs to be presented as well. This information allows historians to choose between alternative methods depending on whether they need high precision or high recall. Ideally, the historian should also be able to gain insight into whether the NLP method is likely to introduce a bias in the interpretation. Providing such information where needed in an understandable method is, however, less trivial. We will address the challenges involved in Section 5. In the next section, we present a brief description of the role of Linked Data in our project.

## 4. Linked Data in BiographyNet

For BiographyNet the information from metadata and extracted from text are converted to Linked Data. Representing data in the Resource Description Framework (RDF) has the advantage that we can use Semantic Web technology. Moreover, we can easily link information from the Biography Portal to information from other related projects, such as Agora (Van den Akker et al., 2011) which attempts to place objects from various museums in a historic context and Verrijkt Koninkrijk ("Enriched Kingdom"),[5] improving the searchability of De Jong (1969 1991)'s 14 volume description of the Netherlands during the Second World War. A detailed description of the conversion process can be found in Ockeloen et al. (2013). We will briefly highlight the main properties here.

One of the main goals was to stay close to the original data structure to prevent information loss due to interpretation and modeling decisions. The following steps were followed in the conversion process.

First, the biographies in the portal are converted from XML (the format we received them in) to RDF using the "XML-RDF" tool[6] for ClioPatria (Wielemaker et al., 2008; Sheth et al., 2008). We then carry out a conversion following De Boer et al. (2012) that allows us to link multiple biographies for the same person to a single resource representing that person. These biographies can contain possibly conflicting views and are therefore treated as seperate sources. We use the PROV-Ontology (Moreau et al., 2012) to model provenance. Because data that resulted from text analysis needs additional provenance information on the process of extracting information in text, we treat the output of text analysis as a new source of information on the individual. This allows us to record the performed NLP processes using PROV-O and relate this provenance information to the resulting "new source" and the original sources on which the processes where performed. Furthermore, we use the P-PLAN ontology (Garijo and Gil, 2013) to model what is supposed to happen at each stage in the NLP pipeline. This information can be helpful for error analysis and debugging. A description of the overall schema that is used to include provenance of what happened and information on the planned processes can be found in Ockeloen et al. (2013). The next section will address provenance modeling in NLP in more detail.

## 5. Provenance modeling of NLP research

NLP research within BiographyNet can be divided in two categories. For now, focus lies on extracting information from text. In future work, NLP will also play a role in comparing style and topics of the different sources in the portal to find out how historic research and interest changes over time (e.g. Herbelot et al. (2012) for philosophical texts). Both directions of research require understandable information on the methods that were used to create the results. We will illustrate this focusing on extracting information from text.

### 5.1. Motivation for Provenance modeling

Historians must be able to verify the validity of the facts they base their conclusions on. Automatic text analysis adds a new dimension to this verification process. We will illustrate how NLP methods may influence the outcome of a historical question through two examples. The first example, an extended description of an example we previously

[5] http://www.niod.nl/en/projects/enriched-kingdom

[6] http://cliopatria.swi-prolog.org/packs/xmlrdf

presented in Ockeloen et al. (2013), shows how heuristics may temper the validity of research. The second example addresses potential problems when machine learning approaches are involved.

Geographical locations are often ambiguous. Many cities existing in Europe are also used to indicate locations in the United States. The biographical portal only describes individuals related to the Netherlands (either because they are Dutch, lived in the Netherlands or played an important role in Dutch history). A simple approach that always assigns the location in or closest to the Netherlands in GeoNames[7] will achieve high accuracy on this corpus. If the historian investigates the origin of civil servants in The Hague, errors introduced by this approach will hardly have an impact on the results. If the investigation addresses the interaction of Dutch officials overseas with their home country on the other hand, this bias would be highly problematic since many place names in former Dutch colonies also existed in the Netherlands. This question thus requires an approach using a more sophisticated approach for disambiguiting locations.

The bias introduced by a heuristic as the one explained above is easy to spot. For other approaches, biases may be less transparent. A machine learning approach to classify biographies according to the topics of the biography (e.g. politics, science, feminism, etc.) could lead to possible biases as well. This could for instance occur when investigating a family that has produced several members active in politics. A potential question the historian might ask would be how many of these family members mentioned in the portal are known for their political involvement. Because the family has a relatively high correlation of people involved with politics, a machine learning approach using simple bag of words as features may have learned to associate the family name with politics. The approach thus increases the likelihood of biographies of this family as being classified as political directly influecing the outcome of this question.

## 5.2. Raising awareness

Historians that use BiographyNet tools must be aware of these potential biases. For each NLP approach, an overview of technologies involved should be given together with an explanation of how they work. The rules and heuristics of rule-based approaches should be made explicit. For machine learning approaches, the training data and features that are used should be indicated.

One of the main challenges lies in explaining how heuristics, rules or features in machine learning may introduce a bias. Historians may expect a bias from heuristics, but the second example of a potential bias is not that easy to spot for a historian. A simple overview of the technology is therefore not sufficient. Accessible explanations of how technologies use features and heuristic and how information may influence an experiment's outcome should also be provided. A complicating matter is that usually the better a tool works the more difficult it is for a lay person to comprehend more than the basics (Lin, 2012, p. 306). Providing such explanations therefore is not a trivial task and will be ongoing work throughout the project.

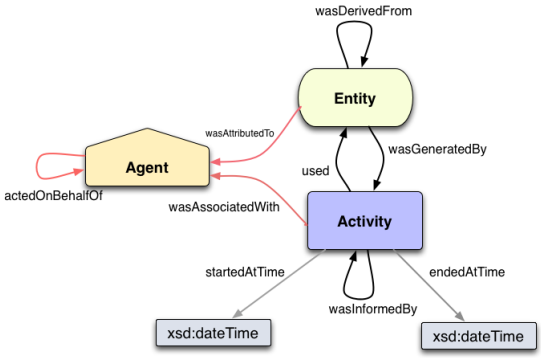


Figure 2: The PROV-DM

Our provenance model should include an overview of known biases as well as observed biases in error analyses. It should furthermore provide a clear indication of all methods, resources and data that was used. Modeling such information also supports NLP research. If we model each tool and resource used, including their version, performance on a particular evaluation task and who carried out the process, we may help to avoid problems reproducing results in the future.

## 5.3. Modeling provenance with PROV-O

As mentioned in Section 4., we use the PROV-DM (Moreau et al., 2012) for provenance modeling in BiographyNet. In Ockeloen et al. (2013), we provide a detailed motivation for this choice. We will provide a brief overview of the most important reasons here as well as a basic description of the structure of the PROV-DM. An example of modeling provenance of a pipeline used in BiographyNet will be given in Section 6.

The PROV-DM is a W3C recommendation for modeling provenance making it widely used for this purpose. Provenance concepts and relations defined in the PROV-DM can be represented in RDF. This makes it straightforward to integrate them in our schema.

The PROV-DM distinguishes three provenance *views*: the data flow view, the process flow view and the responsibility view. The data flow view represents what happens with Entities, the process flow view allows us to model Activities and the reponsibility view associates Entities and Activities with the Agents responsible for them. Figure 2 illustrates the relations between the main classes of PROV-O as provided at `http://www.w3.org/TR/2013/REC-prov-o-20130430/`.[8]

In the PROV-DM, Provenance can be represented on an overall level or more fine-grained level. We can provide a direct pointer to the source text (a provenance Entity) and author of the text (a provenance Agent) of a specific statement. The historian can use this information to go back to the source for verification. It can also provide a more fine grained overview of all the steps that were involved in

[7] `http://www.geonames.org/`

[8] Accessed February 17 2014.

extracting this statement from the text. This includes information on the methods that were used, the resources and data that were involved as well as the people responsible for designing and running the methods.
Accessible explanations of methods cannot be modeled using the PROV-DM. We can however link methods, resources and data present in our provenance model to such explanations.

## 6. A basic system for information extraction

In this section, we describe a basic system that can identify information that falls into metadata categories in the text. We provide a simplified illustration of how we model provenance for this system. The current system uses purely token-based machine learning and yields poor results. We therefore briefly reflect on how the system with its current performance can be useful for historians and present the next steps that will be taken to improve the system.

### 6.1. System description

As explained in 2., the sources in the Biographical Portal are quite diverse as far as their structure, average content and completeness of metadata are concerned. Where some sources provide rich metadata and relatively little text, others almost exclusively provide information expressed in natural language. Several individuals have biographies in more than one resource. Because of this, we can match rich metadata from one source to the text of resources that do not provide much metadata. Text with rich metadata can thus be used to train a supervised machine learning system for identifying metadata in text.
The task of filling out gaps in the metadata of biographies by identifying this information in text is similar to filling infoboxes in Wikipedia. In both cases, we have texts and a manually created set of facts that are likely to be mentioned in these texts which can be used for supervised machine learning. Our approach is therefore inspired by information extraction methods aiming at filling Wikipedia infoboxes, such as Kylin (Wu and Weld, 2007), TextRunner (Yates et al., 2007) and Open IE (the second generation) (Etzioni et al., 2011). Like these approaches, we first identify relevant sentences and then attempt to extract relevant information from these sentences.
We use the following procedure in our baseline system. First, we identify where information from the metadata is mentioned in biographical text. In the next step, we create a corpus for each category of information found in the metadata. The relevant categories are date and place of birth and death, education, occupation, religion and parents. In this corpus, sentences are labeled indicating whether they contain relevant information or not. We use the Mallet (McCallum, 2002) document classifier to identify relevant sentences treating every sentence as an individual document.
In the second stage, we create a corpus of sentences in which tokens containing relevant information are labeled. We train Mallet's conditional random field sequence labeller on this corpus. The sentences marked relevant by our classifier are passed through the sequence labeller to identify the information we are looking for. This first version of our system only uses sentence number, the name of the source and words in the sentence as features. The scripts used for creating the corpora as well as the development data are available at `https://github.com/antske/BiographyNet`.

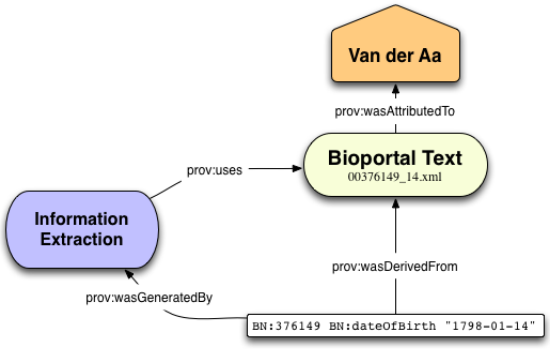


Figure 3: Example: Provenance of extracted birthdate

### 6.2. An example of provenance modeling

Figure 3 provides a simplified illustration of basic provenance information for the date of birth of Thorbecke (an influential Dutch politician from the 19th century) extracted from text. Note that all elements in the representation have their own unique identifier which allows us to define relations between them in RDF.
The extracted information is represented in the white box. Thorbecke is identified by a his BiographyNet identification number, which is unique.[9] The information is extracted from a specific XML file from the Biography portal indicated by the *Information Extraction* activity. There is a direct link between the extracted statement and the original source (*Biographical text 00376149_14.xml*).[10] The biography in question is included in van der Aa (1878). It is therefore attributed to Van der Aa who initiated the biographical dictionary in question in the provenance model.
The information presented in Figure 3 fulfills a number of requirements we defined in Section 3.2. The historian can find the original source, information on who is responsible for the source and that the information was automatically extracted from the text. The provenance information allows the historian to estimate the reliability of the outcome (e.g. Van der Aa is an old source and therefore not the most reliable), to change the query (e.g. to exclude any references to Van der Aa to improve results) and to cite the result obtained by giving a URL in a footnote which can be reproduced and checked for reliability by other colleagues.
Extracting the date of birth from a structured biography is a relatively easy taks and it is unlikely there will be a bias in the way this information was extracted. However, even here correctness of the algorithm cannot be guaranteed and

[9] Note that this unique identifier can be linked to other identifiers for Thorbecke such as his DBpedia entry.

[10] We also have a model for linking the elements of the statement to the specific tokens that refer to them. We use the Grounded Annotation Framework (Fokkens et al., 2013, GAF) to establish this link.

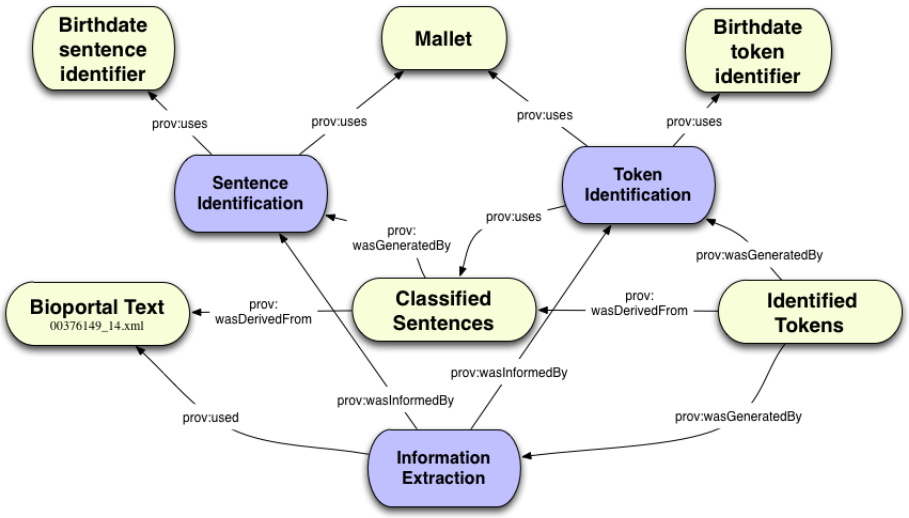


Figure 4: Partial overview of provenance for our basic system

it may be worthwhile to examine the approach in more detail. We can represent the necessary additional information for this examination about the *information extraction* activity in our model. This includes information on the process as a whole, such as the precision and recall found in our evaluation, and a description of insights gained from error analysis. We can also provide information about individual steps taken in our approach for information extraction.

Figure 4 provides a partial representation of the provenance of the information extraction activity. The *information extraction* activity consists of the *sentence identification* activity and *token identification* activity. These activities are linked to the software that was used to carry them out. We can link the software to descriptions on how they work for historians, but also provide valuable information to computational linguists who may want to experiment with the software or who want to reproduce results. We include information about the version of the software and the data that the classifier and token labeller were trained on. There is, however, always a chance that some information is missing. We therefore also indicate who was responsible for setting up the system and, where possible, who was responsiple for implementing parts of the software. This way, fellow researchers know who to contact in case of questions (this information in not present in the simplified representation in Figure 4).

To come back to our metaphor in Section 3. of a historical narrative as a house built with building blocks with a varying level of reliability: the extracted fact functions as a *building block*, while all the provenance information shows how strong the *building block* is and what influence this has on the strength of the house as a whole.

### 6.3. Evaluation and outlook

The first evaluation of the system aims at creating a portrait of the 71 governors of the Dutch Indies. We looked both at the identification of correct sentences and identification of the correct information within the sentence.

We achieve decent results for sentence identification, but recall for identifying the correct information in the sentence is extremely low (from nothing for education to around 30% for the easiest category of date of birth). Precision, on the other hand, is high (between approximately 60%-100%).

The current results are not of high enough quality yet to truly support the ultimate goal of this project. We do not extract enough information from the text to form a database in RDF that is truly richer than that what can be derived from the metadata.

The first step in our approach can be of use to historians, because it gets decent results in identifying relevant sentences. This can help historians to quickly spot specific information they are looking for in text. However, it can only be used reliably when the historian either knows when all relevant information is found (e.g. unique events such as the birth of an individual) or where recall is not relevant (e.g. looking for examples of an occupation).

Yates et al. (2007) and Etzioni et al. (2011) show that results for information extraction can be improved drastically when linguistic information is used in a clever way. Examination of a comparable development corpus indicates that this probably also applies to the task of extracting information from biographies. Furthermore, basic information about an individual tends to be presented in a consistent manner, at least within a biographical dictionary. Integrating linguistic information in our system and making use of the stable parts of structures in biographies are the next steps to improve our results.

## 7. Conclusion

This paper introduced BiographyNet, a project where NLP and Semantic Web technology are used to support historical research. In this paper, we particularly focused on the role of NLP in the project and addressed methodological issues that come into play when NLP is used to support historical research.

We argued that provenance modeling plays an essential role in historical research and that interdisciplinary research

adds an additional dimension to provenance modeling both historians and computational linguists need to bea aware of. Historians need to be aware that, in addition to verifying reliability of sources as is common in their field, they also need to take the reliability of NLP methods into account when working with automatically extracted information. Looking at the overall performance of the method (e.g. precision and recall of retrieved information) is not always enough, because methods can introduce a bias that directly influences the results.

Computational linguists must be aware of this potential bias, so that they can provide historians with the information they need to be able to interpret their results. Furthermore, methods that introduce a strong bias may require the implementation of alternative approaches. Even if the accuracy of an alternative approach is less high, it may still lead to more reliable results for the historian if it avoids a bias that is highly relevant for the historian's hypothesis.

We provided a basic illustration of what provenance modeling should look like for the basic system we use for information extraction. We introduced the data provided by the Biographical Portal and explained the diversity of data and metadata in the corpus. We furthermore explained why our linguistically naive system performs poorly, addressed the question of how such results may be useful for historians and indicated how we attempt to improve our results in future work.

## Acknowledgements

This work was supported by the BiographyNet project (Nr. 660.011.308), funded by the Netherlands eScience Center (`http://esciencecenter.nl/`). Partners in this project are the Netherlands eScience Center, the Huygens/ING Institute of the Royal Dutch Academy of Sciences and VU University Amsterdam.

## 8. References

Ankersmit, F. (1983). *Narrative logic. A semantic analysis of the historian's language*. Nijhoff, The Hague, Netherlands.

Ankersmit, F. (1996). *De spiegel van het verleden. Exploraties I: geschiedtheorie*. Kok, Kampen, The Netherlands.

Blok, P. J. and Molhuysen, P. C. (1937). *Nieuw Nederlandsch Biografisch Woordenboek*. Sijthoff, Leiden. (1911-1937).

De Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., and Schreiber, G. (2012). Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In *Proceedings of the 9th Extended Semantic Web Conference (ESWC 2012)*, Heraklion, Greece.

De Jong, L. (1969-1991). *Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog*. SDU-uitgeverij, Den Haag, the Netherlands.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam, M. (2011). Open information extraction: The second generation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume One*, pages 3–10. AAAI Press.

Fokkens, A., van Erp, M., Vossen, P., Tonelli, S., van Hage, W. R., Serafini, L., Sprugnoli, R., and Hoeksema, J. (2013). GAF: A Grounded Annotation Framework for events. In *Proceedings of the first Workshop on Events: Definition, Dectection, Coreference and Representation*, Atlanta, USA.

Garijo, D. and Gil, Y. (2013). The P-PLAN ontology.

Herbelot, A., Redecker, E. V., and Müller, J. (2012). Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities EACL 2012*, Avignon, France.

Hoekstra, R. (2013). Historische representativiteit in context. over het biografisch portaal als onderzoeksinstrument. http://www.digitale-geschiedenis.nl/abstracts/historische-representativiteit-context-over-het-biografisch-portaal-als.

Kirschenbaum, M. (2012). What is digital humanities and what is it doing in english departments? In Gold, M. K., editor, *Debates in the Digital Humanities*, pages 3–11. University of Minneapolis Press.

Lin, Y.-W. (2012). Transdisciplinarity and digital humanities: Lessons learned from developing text-mining tools for textual analysis. In Berry, D. M., editor, *Understanding Digital Humanities*, pages 295–314. Palgrave Macmillan.

Liu, A. (2012). The state of the digital humanities. a report and a critique. *Arts and Humanities in Higher Education*, II(1-2):8–41.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mink, L. O. (1966). The autonomy of historical understanding. *History and Theory*, 5:24–47.

Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., and Tilmes, C. (2012). PROV-DM: The PROV Data Model. Technical report.

Ockeloen, N., Fokkens, A., ter Braake, S., Vossen, P., de Boer, V., Schreiber, G., and Legêne, S. (2013). Biographynet: Managing provenance at multiple levels and from different perspectives. In *Proceedings of the Workshop on Linked Science (LISC2013) at ISWC*.

Ricoeur, P. (2004). *Memory,history,forgetting.Translated by Kathleen Blamey and David Pellauer*. University of Chicago Press, Chicago, United States and London, United Kingdom.

Rieder, B. and Röhle, T. (2012). Digital methods: Five challenges. In Berry, D. M., editor, *Understanding Digital Humanities*, pages 67–84. Palgrave Macmillan.

Sheth, A. P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T. W., and Thirunarayan, K., editors. (2008). *The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings*, volume 5318 of *Lecture Notes in Computer Science*. Springer.

Siemens, L. (2009). It's a team if you use „reply all". *Literary and Linguistic Computing*, 24(2):225–233.

Svensson, P. (2012). Envisioning the digital humanities. *Digital Humanities Quarterly*, 6(1).

Ter Braake, S. (2014). Mixed teams in every aspect of the mission: Collaboration in digital humanities. http://sergeterbraake.blogspot.nl/2014/02/mixed-teams-in-every-aspect-of-mission_24.html.

Van den Akker, C., Legêne, S., Van Erp, M., Aroyo, L., Segers, R., Van der Meij, L., Van Ossenbruggen, J., Schreiber, G., Wielinga, B., Oomen, J., and Jacobs, G. (2011). Digital hermeneutics: Agora and the online understanding of cultural heritage. In *Proceedings of the 3rd International Conference on Web Science (WebScience11)*, Koblenz, Germany.

Van den Akker, C. (2012). Narrativist philosophy and the autonomy of history. *Journal of the Philosophy of History*, 6(2):236–257.

van der Aa, A. J., editor. (1878). *Biographisch Woordenboek der Nederlanden*. Brederode, Haarlem, The Netherlands.

Wielemaker, J., Hildebrand, M., van Ossenbruggen, J., and Schreiber, G. (2008). Thesaurus-based search in large heterogeneous collections. In Sheth et al. (Sheth et al., 2008), pages 695–708.

Wu, F. and Weld, D. S. (2007). Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 41–50. ACM.

Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., and Soderland, S. (2007). Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 25–26. Association for Computational Linguistics.

Zaagsma, G. (2013). On digital history. *BMGN/LCHR*, 4:3–29.