BiographyNet: Managing Provenance at multiple levels and from different perspectives

Niels Ockeloen, Antske Fokkens, Serge ter Braake, Piek Vossen, Victor de Boer, Guus Schreiber, and Susan Legêne

The Network Institute, VU University Amsterdam
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{niels.ockeloen,antske.fokkens,s.ter.braake,piek.vossen
v.de.boer,guus.schreiber,s.legene}@vu.nl
http://wm.cs.vu.nl

Abstract. The BiographyNet project aims at inspiring historians when setting up new research projects. The goal is to create a semantic knowledge base by extracting links between people, historic events, places and time periods from a variety of Dutch biographical dictionaries. A demonstrator will be developed providing visualization and browsing techniques for the knowledge base. In order to establish its credibility as a serious research tool, keeping track of provenance information is crucial. This paper describes a schema that models provenance from different perspectives and at multiple levels within BiographyNet. We will present a concrete model for the BiographyNet demonstrator that uses elements from the Europeana Data Model [6], PROV-DM [17] and P-PLAN [11].

Keywords: eHumanities, Linked Data, PROV-DM, P-PLAN, ORE, EDM

1 Introduction

E-humanities investigates what can be done in humanities with modern techniques which we could not do before, or only could do with a great deal of effort. E-history is a subdomain of e-humanities which offers a way of linking pieces of information and discovering relationships which otherwise would be difficult to trace. It generally aims at improving methods of existing historical research rather than introducing a whole new way of historical research [22]. It creates pathways through information, rather than being the closing factor or end result in historical research [1, 41]. Efforts in e-humanities often concentrate on how to mine 'big data', which we define as data which is very difficult to handle manually for a traditional researcher. More challenging, and in general also more interesting, are projects which aim to go beyond the simple data mining and endeavor to answer difficult research questions like the similarity between and interdependability of two or three texts, tracing and defining the subjective elements and descriptions, or signaling traces of political or cultural influences

from a society during a given period. These new ways of mining historical data lead to new questions on provenance of information. It is imperative for historians to keep a good oversight over the sources which were used to produce a certain output. How reliable are the sources which were used and what do they tell about the significance of the outcome? What differences are found in the information that individual sources provide? When information differs, how are specific points of view distributed over different sources? How can results be manipulated by adjusting queries for a more accurate result? For these reasons, the historian needs to have an aggregated view of the process from query to output and, if necessary, inspect the whole process step by step to learn which additional sources and heuristics were involved.

1.1 Use Case: BiographyNet

The BiographyNet project is an e-history project bringing together researchers from history, computational linguistics and computer science. The project uses data from the Biography Portal of the Netherlands (BP), which contains approximately 125,000 biographies from a variety of Dutch biographical dictionaries, describing around 76,000 individuals. The aim of BiographyNet is to develop a demonstrator which supports the discovery of interrelations between people, events, places and time periods in biographical descriptions. Through a combination of data enrichment, quantitative analysis, visualization and browsing techniques, the demonstrator should provide leads and insights that may be hard to discover using traditional methods. As such, it may inspire historians to investigate more ambitious research questions.

The BP links biographies written by thousands of authors with very different temporal and academic backgrounds. This results in many levels of reliability of the 125,000 entries in this melting pot of Dutch biographies. Provenance information is therefore an important factor. It must however be noted that provenance information on the original sources does not go beyond the information that is provided by the BP such as author, publisher or the book from which a text was taken.

2 Motivation

The demonstrator should help historians do their research. This goal can only be met if the validity of the demonstrator's results can be verified. To this end, information needs to be available on performed operations as well as on used sources. According to Groth et al. [12], "data can only be meaningfully reused if the collection processes are exposed to users. This enables the assessment of the context in which the data was created, its quality and validity, and the appropriate conditions for use". Hence, provenance plays an important role in establishing the demonstrator's credibility.

Provenance needs to be modelled from different perspectives and at multiple levels for BiographyNet. These different perspectives include 1) the perspective

of the information used to produce the results provided by the demonstrator, e.g. which original sources contributed to the outcome, 2) the perspective of the processes involved in creating the results and 3) the perspective of the people that were involved in setting up the pipeline of processes. The various levels include 1) provenance at component level, recording each aspect of the processing steps involved such as tool name, version, etc. and 2) an aggregated view of the provenance information for the interlinked processes as a whole. The latter is targeted at the end user of the system, in this case the historian, while the former is needed by the computer scientist in case the outcome of an aggregated process is pulled into question.

In the next two sections, we address the requirements for provenance modelling specific for BiographyNet. First, we will address the point of view of historians who are primarily interested in the reliability of the system. We will explain how the requirements for historians relate to the categories for provenance on the web defined by Groth et al [12]. Section 4 will outline BiographyNet from the point of view of the system developers whose primary interest is to improve the technology behind the demonstrator. Section 6 will describe the BiographyNet schema devised to allocate the required provenance information as described in the preceding sections.

3 Requirements for Historians

There are two main requirements for the historian regarding provenance when using the demonstrator: A trace back to the text and metadata in the original source, and insight into the processes manipulating and selecting the original data. We will explain the first requirement through a research question on the background of the 71 governors-general of the Dutch Indies between 1610 and 1949. If, for instance, we run a query to find out what the average age of these individuals was at the time of their appointment, provenance information of different granularity should be present: a) an overview of the sources (in our case biographical dictionaries) that were used for the overall outcome and how often each individual source was consulted, b) an overview of potentially relevant data that was excluded from the end result. This is important in case of conflicting data, where one source generally considered more reliable was used rather than another and c) the sources that were used for a specific results (i.e. the age of a specific governor at the time of his appointment).

One can assume that few historians will have the background to completely (or even partly) understand the finer technical details of how data are processed in order to answer a query. Even when a new generation of 'e-historians' is trained, one cannot expect them to be computer scientists. Therefore, provenance of data manipulation should be modelled as simple as possible and focus on aspects that may directly influence the outcome of research questions. First and foremost, it should always be indicated whether information is directly extracted from the metadata or the result of automatic interpretation of text. Complete accuracy in automatic text interpretation cannot be guaranteed. Information

extracted from text should therefore always include a direct link to the original source. Provenance should also indicate the overall performance of the system that interpreted the text; depending on the kind of question, the historian may want to have results that aim for high recall or high precision. Finally, a global description of heuristics used when interpreting data should be provided. While resolving ambiguous location names, for instance, a strategy that always prefers locations in or near the Netherlands is likely to lead to good results within the BiographyNet project. However, if the historian wants to investigate the ties between officials in the former Dutch colonies (where cities with Dutch names can be found), this strategy would bear a direct undesirable influence on the results. The historian should thus be able to check whether the interpretation process used any strategies that may introduce a bias that influences results.

If we translate this to the categories outlined in [12], this leads to the following requirements.[1] The **objects** for which we need to model provenance are texts from several sources, metadata and statements extracted from the text. Texts and metadata are **attributed** to publishers and authors of this data. Extracted information should also indicate the author or publisher of the original text and, in addition, point to the system used to extract the information. There is thus a tight link between the process and the attribution while modelling provenance of automatically extracted text. Attribution plays a significant role in establishing the reliability of information and this includes the reliability of the methods that were used to extract information from text.

Information on the **process** should include detailed indications of the system?s overall performance: i.e. it should indicate the precision and recall of the system for specific categories. Furthermore, the **version**, publication date and person **responsible** for generating the output should be indicated in case the historian wants to replicate their results at a later stage. Finally, provenance should include **justifications** for decisions made in the extraction process, in particular concerning techniques used to disambiguate terms or resolve entities. The historian may need such information to check whether the information extraction used heuristics or forms of **entailment** that may interfere with the outcome of the research question addressed by the demonstrator, as illustrated by the location disambiguation example above.

In order to address the aspects of **trust** and **accountability** as outlined above, it must be crystal clear which information comes directly from original sources, and which information is the result of the processing or interpretation of these sources. Hence, the schema for BiographyNet should accommodate for this. The distinction should be marked prominently, because automatic processes add a dimension to reliability that not all historians will be familiar with. One of the main challenges therefore is that technical processes should be explained in terms that are understable to researchers who generally do not have a technical background. Strong collaboration between the historians and system designers is thus required when designing this part of provenance modelling throughout the project. At this level, an indication of responsibility is necessary so that

[1] Concepts that are addressed in [12] will be marked in **bold font**.

historians can contact the persons who designed the interpretation pipeline in case of an unexpected outcome or if questions arise on the made assumptions or used heuristics.

4 Requirements for computer scientists

Researchers working on demonstrators are mainly interested in provenance because it helps to make experiments replicable and it supports research to improve existing technologies. We use the term **replication** to refer to the process of following the exact same procedure as in the original work and thereby obtain the exact same output. This is different from reproduction where the same question is answered using different means (e.g. a new implementation or evaluation set). The validity of research results increases when they can be reproduced, whereas replication only verifies that an outcome was valid under specific conditions [8]. Within our setup, replication matters for two reasons. First, we need to be able to create the exact same dataset for historians if they want to compare new results to previous results. Second, when results cannot be reproduced, it is almost impossible to find the cause without being able to replicate the original results [18].

It is well known that both replicating and reproducing results is challenging when computer programs are involved. This especially holds if the code is not available [19, 18] but even if code is present [21, 10]. Fokkens et al. [10] define five categories that may influence results in pipelines that involve Natural Language Processing (NLP). They are preprocessing (e.g. tokenization, cleaning up data), experimental setup (e.g. splitting folds for 10-fold cross validation, evaluation set), versioning (e.g. version of resources such as WordNet [9], or tools such as Mallet [15] for machine learning), system output (e.g. the exact features for specific tokens, intermediate output of the system in a pipeline) and system variation (e.g. treatment of ties, thresholds). This information must be explicit in order to replicate results.

Information on influential factors immediately contributes to the second use of provenance for computer scientists: improving existing technologies. Individual tools and datasets interact in different ways with each other. Systematic testing of influential parameters, exchanging tools for subtasks and combining the output of different tools can lead to significant improvement in performance. The interaction between performance of subtasks and overall performance of the system is not always straightforward. The output of the sentence splitter, for instance, influences the output of the parser. However, even if the output of the parser of the utterance as a whole is incorrect, we may still obtain the grammatical relations we need to identify the participant of an event.

The **object** for which we need to model provenance thus is the data at various stages of the provenance pipeline. This data is **attributed** to a specific tool that has taken data from the previous stage and possibly one or more external resources as input. Again, attribution is tightly linked to the **process**. Modeling the process is the most complex aspect of modelling provenance for

the NLP pipeline. It requires registering detailed information on all tools and data sets involved including preprocessing steps, steps to generate features and the process of creating training data for machine learning. For all tools and resources, the **version** should be indicated. A detail in implementation or a small step or setting can make a significant difference in the results. It should therefore be registered who is **responsible** so that differences can be traced when third parties do not manage to reproduce results. Finally, documentation should clearly describe the decisions made in the setup which both serves as a **justification** of the approach and a way to indicate any form of **entailment** that may be required by the historian.

5 Retrieving information from text

One of the main challenges of building a demonstrator lies in creating tools that can automatically interpret text and extract information from it. The design of the system that is responsible for automatic text interpretation is work in progress. We will therefore provide a description of what this process is likely to look like based on the work carried out so far as well as systems used in related work. The main purpose of this section is to provide an indication of the different steps involved in automatic text interpretation.

We start by identifying linguistic information in text, where we distinguish two processes: named entity recognition and concept identification. Named entity recognizers identify names of persons, organizations and locations. Some also identify dates. We will use an off-the-shelf named entity recognizer for Dutch, for instance LingPipe[2]. Concept identification involves linking words in text to a set of concepts of interest. We will use revisions of tools described in [20] and [5]. Their approach is based on McCarthy et al's [16] observation that words tend to have a predominant sense within a specific genre or domain. The approach involves two steps. Concepts of interest are first identified in the corpus whereafter an executing step is performed in which these concepts are labeled in the text. We will briefly describe the two steps below.

- First, candidate terms are identified in the text. In a basic system, these may be verbs and nouns co-occurring in a sentence. We thus start by running a sentence identifier, tokenizer and part-of-speech tagger and lemmatizer over the entire corpus.
- Next, we link all these terms to WordNet entries and create hypernym chains. This process results in an overview of the hypernym chains identified in the text. For each hypernym, the set of hyponyms occurring in the text is given. We manually select a set of hypernyms from this overview. This set of hypernyms constitutes our concepts of interest. As soon as we have created a set of concepts of interest, we can tag these concepts in the text. First, we create a corpus by running a tokenizer, part-of-speech tagger and lemmatizer over the text. For each lemma in the corpus, we check whether

[2] `http://alias-i.com/lingpipe/web/demo-ne.html`

one of its senses is a hyponym of one of the concepts of interest. In this case, we associate the lemma to this concept of interest. Lemmas are thus only linked with selected concepts of interest and the senses that are related to these concepts constitute their predominant sense within our domain.

Together, named entity recognition and concept identification provide a corpus in which persons, organizations, times, locations and concepts are labelled.

Consequently, we can apply two strategies to extract useful information from text: a rule based approach and an machine learning approach (ML). We can define basic mapping rules that directly map the resulting labels within this corpus to usable metadata. If for instance, we encounter a person name identified by the named entity recognizer in close proximity of a profession tagged by our concept identifier, we assign this profession to the person.

The ML strategy uses existing metadata to discover similar information in biographies for which that metadata is missing using named entities and concepts as features. The biographies obtained from the BP are accompanied by metadata that includes information on the subject of the biography. The completeness of this metadata varies significantly from source to source. Biographies with rich metadata can be used to learn to identify information in text and hence find this information in biographies with poorer metadata. We have created a corpus in which information from metadata is tagged in the original text of the biography. This corpus can be used as a training set for machine learning to discover information in texts that is missing in the metadata. For example, we found that the metadata field 'religion' was available for only 6 out of the 71 governor-generals in our use case. However, using ML we found this information in the text for 20 governors.

Together these strategies form the core of our system for text interpretation. It should be noted that the descriptions provided above illustrate a basic system that is currently under development. Throughout the project, we will incrementally improve the system by adding more linguistic information.

6 The BiographyNet schema

Having outlined the main concerns and requirements for the BiographyNet demonstrator, the following section describes the schema devised to manage the data used and produced for the demonstrator. It describes how data from both original sources and enrichments is stored, how provenance information is handled for involved processes and how this ties into the formulated requirements. An impression of the schema can be found at: `http://www.biographynet.nl/schema/`. The following subsections are best read with the schema alongside. The mentioned concepts and relations can then be traced and followed in the schema. Description of the various parts of the schema generally takes place from left to right. Please note that this impression includes the various aspects described in this section in order to provide a general overview of the schema for BiographyNet. It does not include every aspect of the biographical data and provenance information in order to maintain overview. Information on individual Activities,

Entities etc. such as start times, version numbers etc. is left out and qualified relations are only modelled if needed to illustrate the ideas behind the schema.

6.1 Foundations of the BiographyNet schema

The collection of biographies is made available to the BiographyNet project as a collection of XML files. Each XML file contains a 'Biographical Description', which in turn contains three different types of data; A 'File Description' that contains the metadata on the original source, a 'Person Description' that contains limited metadata on the depicted person, and the actual biographical description. Currently, the available biographical data is not linked to any other sources. To be more flexible when it comes to linking to external sources in the near future and in order to reason over the data, the BiographyNet demonstrator will be based on Linked Data [2] principles. Therefore, the collection of XML files is converted to RDF [4]. How this conversion was done in detail is out of scope for this paper, but a similar conversion process is described in [3]. When data needs to be converted, it is advisable to stay as close as reasonably possible to the original schema, in this case defined by the structure of the XML files. Any altering of the schema involves interpretation, and as interpretation can change over time, such a process has the potential for information loss. For this reason we started out with a schema for BiographyNet that closely follows the structure of the original XML files; it contains a resource that represents a 'Biographical Description' (BioDes) that has connections with resources that represent a 'File Description' (FileDes), a 'Person Description' (PersonDes) and a resource for 'Biographical Parts' (BioParts). In the illustration, these are the blue outlined ovals, starting with the second leftmost.

Within the provided collection, multiple biographical descriptions are often available for the same person, originating from different sources. While these are represented as separate XML files in the provided collection, they need to coexist within the created Linked Data corpus. To this end, the BioDes objects are tied together using a resource representing the depicted person. This is the leftmost blue outlined oval. However, this means that -through the BioDes objects- a person can have multiple PersonDes objects containing possibly conflicting sets of metadata. In order to make the semantics of this more clear, we used parts of the Open Archives Initiative's 'Object Re-use & Exchange' ontology (OAI-ORE) [13, 14] in a way similar to how the Europeana Data Model (EDM) [7] uses concepts from that ontology. By defining the PersonDes objects as a subclass of the ore:Proxy class, defining the depicted person as an edm:ProvidedCHO (Cultural Heritage Object) and incorporating the associated predicate relations, the model becomes compatible with the Europeana data model while still staying true to the original data structure. The depicted person can now be viewed as a 'Cultural Heritage Object', of which multiple sets of metadata are made available through proxies, indicating that these sets of metadata represent different 'views' of that person.

This solution also allows for adding a new BioDes object for a person that 'aggregates' multiple other sources (BioDes objects) through the ore:Aggregates

and edm:AggregatedCHO predicates. Besides the original biographical descriptions and an aggregated version of them, the model can also be used to accommodate enrichments. In that sense, an enrichment is a 'new' biographical description which was *derived from* original sources. A FileDes object will not be available for the enrichment, as the enrichment itself does not *directly* come from an original source, i.e. a biographical dictionary. Similarly, a BioDes object for an enrichment will most likely not contain a BioParts object, as it represents a set of metadata resulting from the enrichment process, but does not contain actual biographical texts. By modelling the *was derived from* relation, the enrichment can be traced back to the biographical description it was derived from and its original source, a hard requirement formulated in section 3.

6.2 Extending the schema with Provenance

PROV-DM [17] is the logical candidate for modelling provenance, since W3C[3] made it a recommendation promoting its widespread use. Furthermore, PROV concepts can be modelled in RDF making it suitable for use in the BiographyNet schema. Besides relations such as *was derived from*, the PROV ontology can be used to model Entities, Agents and Activities that played a role during the enrichment process and the creation of the pipeline of processes itself, including their mutual relations. Additionally, concepts from the new P-PLAN [11] are integrated in the BiographyNet schema to specify plans made for the actual activities involved in the enrichment process. Specifying planning information is useful in that it provides a way of verifying to what extend actions were performed according to plan. Hence, integrating this information makes it easier to identify errors in individual processes of the aggregated enrichment process. It also makes replication of results more feasible, as the plans provide a description of what the input and output of activities should look like. As such, the combined use of these ontologies ties into the requirements of the historian to be able to trace which original sources were used to obtain a result and to gather additional information on possible heuristics and biases. It also ties into the requirement of the computer scientist to be able to replicate results. In order to fulfill the requirements of the historian and computer scientist to have both an aggregated view on provenance (i.e. which original sources contributed to an enrichment) and a detailed view (i.e. specified information for all processing steps involved), these two levels are modelled separately in the schema. In the illustration, the aggregated level is represented by the orange outlined ovals (and the green one for the plan) between the two blue biographical structures. The detailed view is made up by the remainder of the schema. Clearly visible in the schema is how these activities and plans are parts and steps of the aggregated enrichment activity and its associated plan. These two views are described in more detail in the subsections below.

[3] `http://www.w3.org/`

6.3 Aggregated provenance information

A prov:wasDerivedFrom relation is made between the BioDes object of the enrichment and the BioDes object of the original source in order to model the information on an enrichment process as a whole. Furthermore, a prov:Activity to represent the aggregated process and its relations to the BioDes objects are specified. That activity has a prov:Agent associated with it. This Agent is the aggregated set of tools used for the enrichment, otherwise known as a 'pipeline'. The desired behavior of the integrated process is described by a prov:Plan object, which has its own provenance information; the plan for the enrichment process is attributed to an Agent, e.g. a computer scientist and can be derived from an earlier version of that plan or another enrichment. This aggregated provenance view allows the end user to identify which enrichments were used to produce a final aggregated view of information. The end user can determine the original sources through the various provenance relations. Furthermore, the end user knows who to contact in case an enrichment process seems to have produced questionable results. The aggregated plan can provide an overview of the input variables used in the underlying processes, as they are referenced through p-plan:isVariableOfPlan relations. This information allows for possible adjustments in order to adjust the output of the overall process.

6.4 Detailed provenance information

The detailed provenance information on individual processes is modeled as a chain of Activities which all have their own input and output Entities, associated Agents and Plan. The Agents are specific tools such as a tokenizer or part-of-speech tagger. The plan describes what the specific tool should do. Each Plan has its own provenance information. These plans are plans in their own right, but are also designated a p-plan:Step to indicate that they are a step of the aggregated plan for the enrichment as a whole. As such, these steps have input and output variables that describe the input and output of the related Activity. These variables correspond to the entities used by and generated by the related activities. An Activity together with its used and generated Entities can be seen as a 'bundle' of objects that together are derived from the Plan for that activity. Each individual Activity is designated as a part of the aggregated enrichment Activity using the Dublin Core 'hasPart' predicate. The order in which the individual Activities are executed can be derived from the prov:used and prov:wasGeneratedBy relations that tie the individual Activities to the Entities representing intermediate results. Besides these intermediate results, other Entities may be used by a specific Activity, e.g. a list of cities for Named Entity Recognition. For both the intermediate results as well as these 'external sources', the data format is unknown. An intermediate result could be a collection of RDF triples, an XML file or plain text file. An external source could be one of those or basically any type of document. In order to cope with this variety, these Entities are represented by a prov:Entity of subclass bgn:IntermidiateResult or

bgn:ExternalSource, that can point to the actual document or serve as Named Graph to contain RDF data.

The aggregated view and detailed view of provenance information are related together by the fact that all Activities in the detailed view are parts of the aggregated Activity, all Plans of the individual Activities are Steps in the aggregated Plan and the biographical description of the 'Source' BioDes object is the actual Entity that is used by the first individual Activity, whereas the Entity produced by the last individual Activity is the resulting set of metadata of the enrichment BioDes object. Any form of pre- or post-processing of input data or results, needed to relate to those objects, needs to be viewed as a separate individual step in the overall plan. For without provenance information on those steps, replicability is not ensured.

7 Conclusion

Keeping track of provenance information is essential for the BiographyNet demonstrator to be viewed as a valid research tool for historians. In this paper we described why this is the case, what the requirements are to model provenance from multiple perspectives and which existing ontologies we used to devise a schema for BiographyNet that meets those requirements. We presented a first version of the BiographyNet schema that not only models provenance on what has taken place, but also models plans to compare against. The next step is to proceed with building a first version of the demonstrator. We will then have to evaluate how the schema holds up in practice, and use the output of such evaluation to further improve the schema.

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.

References

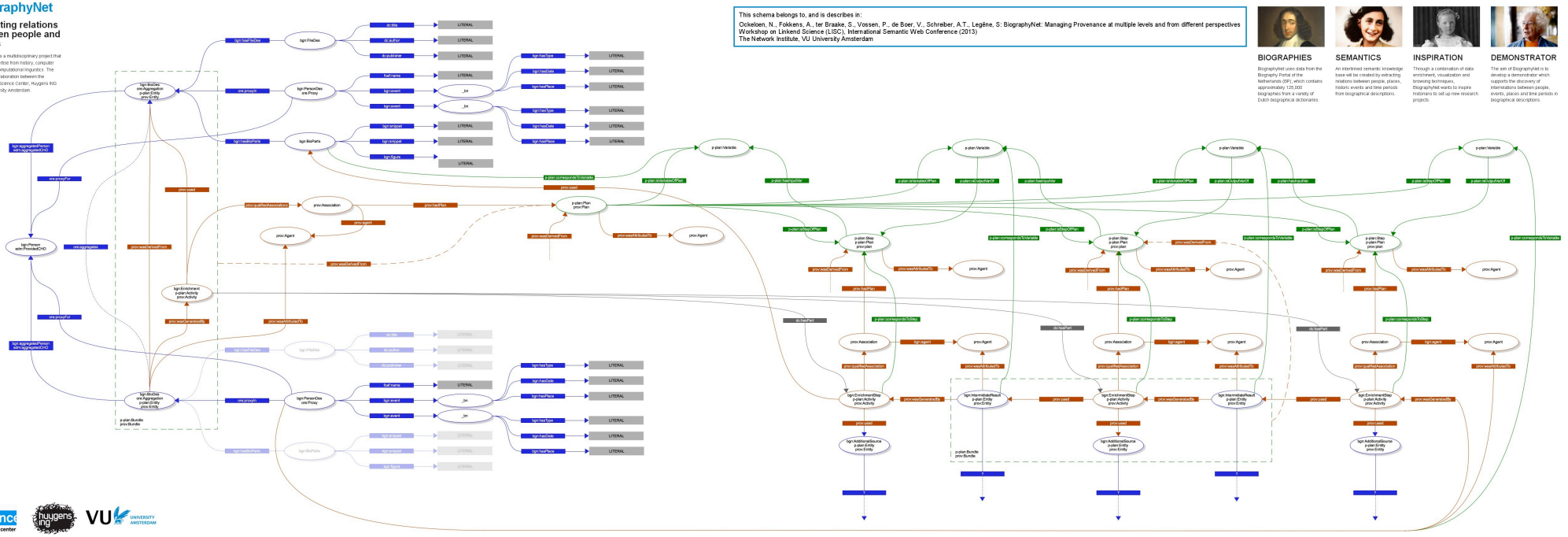
1. Arthur, P.: Exhibiting history. the digital future. Recollections 1(1) (2008)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: ESWC, volume 7295 of Lecture Notes in Computer Science. p. 733?747. Springer Berlin / Heidelberg (2012)
4. Carroll, J.J., Klyne, G.: Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, W3C (Feb 2004), http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

5. Cybulska, A., Vossen, P.: Using semantic relations to solve event coreference in text. In: Mititelu, V., Popescu, O., (Eds.), V.P. (eds.) Proceedings of the Workshop Semantic relations-II. pp. 60–67. Istanbul, Turkey (2012)
6. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The europeana data model (edm). In: World Library and Information Congress: 76th IFLA General Conference and Assembly. Gothenburg, Sweden (2010)
7. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The Europeana Data Model (EDM). In: World Library and Information Congress: 76th IFLA general conference and assembly. pp. 10–15 (2010)
8. Drummond, C.: Replicability is not reproducibility: nor is it good science. In: Workshop on Evaluation Methods for Machine Learning IV (2009)
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)
10. Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N.: Offspring from reproduction problems: What replication failure teaches us. In: Proceedings of the 51st ACL. Sofia, Bulgaria (2013)
11. Garijo, D., Gil, Y.: The p-plan ontology (2013), http://www.opmw.org/model/p-plan/
12. Groth, P., Gil, Y., Cheney, J., Miles, S.: Requirements for provenance on the web. International Journal of Digital Curation 7(1) (2012)
13. Lagoze, C., van de Sompel, H.: Open archives initiative object re-use & exchange (2007), http://www.openarchives.org/ore/documents/ore-jcdl2007.pdf
14. Lagoze, C., Van de Sompel, H., Nelson, M.L., Warner, S., Sanderson, R., Johnston, P.: Object re-use & exchange: A resource-centric approach. Tech. rep. (2008)
15. McCallum, A.K.: MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
16. McCarthy, D., Koeling, R., Weeds, J., Carroll, J.: Finding predominant word senses in untagged text. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. p. 279. Association for Computational Linguistics (2004)
17. 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., Tilmes, C.: PROV-DM: The PROV Data Model. Tech. rep. (2012), http://www.w3.org/TR/prov-dm/
18. Neylon, C., Aerts, J., Brown, C.T., Coles, S.J., Hatton, L., Lemire, D., Millman, K.J., Murray-Rust, P., Perez, F., Saunders, N., Shah, N., Smith, A., Varoquaux, G., Willighagen, E.: Changing computational research. the challenges ahead. Source Code for Biology and Medicine 7(2) (2012)
19. Pedersen, T.: Empiricism is not a matter of faith. Computational Linguistics 34(3), 465–470 (2008)
20. P.Vossen, Bosma, W., Rigau, G., Agirre, E., Soria, A., Aliprandi, C., de Jonge, J., Hielkema, F., Monachini, M., Bartolini, R., Frontini, F.: Kyotocore: integrated system for knowledge mining from text (2011)
21. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Experiment databases. Machine Learning 87(2), 127–158 (2012)
22. Zaagsma, G.: Doing history in the digital age: history as a hybrid practice (2013), http://gerbenzaagsma.org/blog/16-03-2013/doing-history-digital-age-history-hybrid-practice

BiographyNet

Extracting relations between people and events

BiographyNet is a multidisciplinary project that combines expertise from history, computer science and computational linguistics. The project is a collaboration between the Netherlands eScience Center, Huygens ING and VU University Amsterdam.



This schema belongs to, and is describes in:

Ockeloen, N., Fokkens, A., ter Braake, S., Vossen, P., de Boer, V., Schreiber, A.T., Legêne, S: BiographyNet: Managing Provenance at multiple levels and from different perspectives
Workshop on Linkend Science (LISC), International Semantic Web Conference (2013)
The Network Institute, VU University Amsterdam

BIOGRAPHIES

BiographyNet uses data from the Biography Portal of the Netherlands (BP), which contains approximately 125.000 biographies from a variety of Dutch biographical dictionaries.

SEMANTICS

An interlinked semantic knowledge base will be created by extracting relations between people, places, historic events and time periods from biographical descriptions.

INSPIRATION

Through a combination of data enrichment, visualization and browsing techniques, BiographyNet wants to inspire historians to set up new research projects.

DEMONSTRATOR

The aim of BiographyNet is to develop a demonstrator which supports the discovery of interrelations between people, events, places and time periods in biographical descriptions.

netherlands eScience center

huygens ing

VU UNIVERSITY AMSTERDAM