Connecting People Across Borders: a Repository for Biographical Data Models

Antske Fokkens and Serge ter Braake

CLTL, Vrije Universiteit Amsterdam / Media Studies, University of Amsterdam
De Boelelaan 1105 1081 HV Amsterdam, the Netherlands / Turfdraagsterpad 9, 1012 XT Amsterdam
antske.fokkens@vu.nl,sergeterbraake@gmail.com

Abstract

This paper proposes a practical approach for sharing knowledge about biographical datamodels circumventing issues with copy-right. We furthermore provide the main observations of a study analyzing the data structures of eight biographical resources, two platforms for biographical information and four biographical data models. We outline an approach for designing a generic model that can be used for linking information from different models despite differences in structure.

Keywords: Biographical data models, RDF

1 Introduction

The biography genre has a long history. Plutarch (45-ca. 120 AD) is often considered the father of the biography. He did not only provide syntheses of people's life, but he also tried to compare them to a similar person in a 'double biography'. Other than length, there is a difference between such full length biographies and biographical entries in biographical dictionaries. Biographies in biographical dictionaries tend to be more factual. They provide a chronicle of the lives of noteworthy people, without necessarily giving much attention to social environment, political circumstances or comparisons to other people. Full length biographies paint a biographical narrative, while short biographical entries in biographical dictionaries or encyclopedia such as Wikipedia mostly provide 'biographical data'. These biographical data can be the building blocks for full length biographies, or they can serve to build group portraits and systematically compare people (see also Harrison (2004)).

Over the past twenty years the amount of online available 'biographical data' has increased rapidly, with the advent of the Internet and large digitization projects. The potential for biographical research, network analysis and group portraits seem to be endless when all of this data can be linked and shared for analysis (Fokkens et al., 2017; Arthur, 2017, e.g.).

Projects aiming at making biographical data available, first need to address the question of how to represent this data. Individual projects have dealt with this issue in different ways. Where some introduced or reused formally defined models, others used basic approaches using comma-separated-values to represent the information most commonly provided by the original resource.

Because many projects did not consider data representation a central issue in their digitization efforts, the number of publications about this part of the process remained limited and, as a consequence, knowledge about existing models and best practice for modeling biographical data is not sufficiently shared. This resulted in two challenges for researchers working with biographical data. First, researchers working on new digitization projects for biographical data are 'reinventing the wheel' and run into the same problems others have dealt with before them. Second, most biographical datasets have their own data representation making it challenging to carry out research across datasets.

A few examples of successful integration of biographical data and standardization of metadata from different sources are the national Australian Dictionary of National Biography,[1] the Biography Portal of the Netherlands,[2] and the transnational Biographie-Portal[3] and the APIS (Austrian Prosopographical Information System) project (Gruber and Wandl-Vogt, 2017).

This paper proposes a practical approach that addresses the problems faced when integrating biographical data from different sources into one repository. We introduce a repository for biographical data models that provides examples and descriptions of existing data models. The repository provides illustrations of data models used in different projects using fictional biographies, accompanied with fictional biographical data. Researchers working with the models can add information about the process, why the model is designed in a particular way and problems and advantages they experienced from their modeling choices. In addition, the samples in the repository are used to design a generic overarching model that can combine data represented in different formats.

The main contributions of this paper are:

1. We compare and classify the design of models for modeling biographical data from fourteen resources
2. We introduce a repository that provides insight into the structure of one of these models
3. We outline our approach for connecting models that use different frameworks, formats and structures

The remainder of this paper is structured as follows. Section 2 discusses related work. The comparative analysis of biographical data models is presented in Section 3. We describe the set-up and current status of the Repository for Biographical Data Models (BDM) and our proposal for designing a generic model for connecting data from various

[1] http://adb.anu.edu.au
[2] http://www.biografischportaal.nl
[3] http://www.biographie-portal.de

projects in Section 4. We conclude in Section 5. Appendix A describes the resources we studied for this paper.

2 Background and Related Work

Even though a handful of publicly available standards exist for biographical data and some initiatives define their models in RDF (Resource Description Framework) and make use of existing vocabularies, most projects have designed their own model. This can be for historical reasons, either by the desire to stay close to the structure of an original (non)-digital source or by the direct research goals that were outlined in early stages of the digitization process. It is however likely that this is at least partially due to lack of knowledge on existing resources. This lack of knowledge is not due to lack of interest, but to the fact that it is non-trivial to obtain this information. Experience in creating structured data often stays project internal: publications on formalizing biographical data are limited, biographical resources are often part of national projects written in a local language or their use is restricted by copyright.

Making use of other people's experience in their digitization and enrichment projects not only saves work, it can also help avoid problems further down the line. It is difficult to foresee exactly what information various researchers interested in a resource may need later on. Investigating data structures that have already been used for various use cases can provide valuable insight into what works and what does not. Following examples from other projects has the additional advantage that it will be easier to make connections between different datasets facilitating, for instance, comparative biographical research across borders.

The situation of biographical data models is far from unique and some efforts have been made to address this issue. Franzini et al. (2016) aim to provide an overview of properties of digital editions and RIDE[4] offers a Review journal for digital editions and resources. In the typical case, the data model used in digital humanities projects is determined by structure of the original resource or specific research questions from the early phases of the project. This is only natural, because staying close to the original source minimizes loss of information and current research questions form a concrete set of requirements that can be used for designing the model.

In the remainder of this section, we first provide background information on data structures and clarify who related terminology will be used in the remainder of this paper. We then introduce previous projects that provide a common model for multiple biographical resources.

2.1 Formal Modeling and Linking

Data can be unstructured (such as flat text), semi-structured (e.g. CSV (comma separated values) files containing descriptions in natural language) or fully structured (e.g. a representation in RDF). Note that an RDF representation can also contain unstructured elements (e.g. a literal value that is a text) and that CSV can also be used to provide fully structured information (e.g. only information that is numerical or ontologically defined). In this paper, we only deal with semi-structured and structured data representations.

[4] `https://ride.i-d-e.de`

Comparing data representations is complex, because formats and models are regularly confused. In particular, advantages and disadvantages of using RDF or XML ((eXtensible Markup Language) and JSON (JavaScript Object Notation) are frequently discussed even though XML is a serialization format and RDF is a data model, that can be represented in several formats including XML or JSON. Likewise, XML and JSON can be used to represent data models that are not RDF, specified in e.g. the DTD (Document Type Definition) of the XML. When comparing XML to RDF, people generally mean the possibility of capturing information through its structure when using XML (by embedding elements or placing them in some order), where RDF enforces making all information explicit.[5] Even though we are aware of the fact that XML and RDF operate on a different level and thus cannot be compared, we distinguish between models using RDF and models using non-RDF based XML or non-RDF based JSON or CSV. Unless specified otherwise, the terms XML and JSON will refer to (semi-)structured representations that are not defined in RDF in the remainder of this paper, where we use *RDF* to refer to RDF models regardless of the format they are represented in.

Structured data forms the basis for applying digital models, but structure in itself does not provide the means to connect or compare data from various resources. In order to automate a process of connecting data, its category must be formally defined. In RDF, identifiers are used to refer to entities or their properties. These entities and properties can be formally defined, which also allows us to define correspondences between entities and properties. These correspondences can link data across resources. We therefore aim to work towards a generic model in RDF.

A full discussion of related work on linking data within the digital humanities is beyond the scope of this paper. We therefore limit this overview to projects that directly influenced the approach proposed in this paper. In our proposal, we follow de Boer et al. (2012), who outline a procedure for converting cultural heritage data structured in XML to RDF with a minimum of data loss. Their approach will be explained in detail in Section 4.2. They ultimately map their converted data to a common data model for cultural heritage data: the Europeana Data Model (Doerr et al., 2010, EDM). We propose to follow this example for biographical data, where we keep data representations as close as possible to their original form and then connect them by defining categories occurring in individual models by relating them to a generic model for biographical data.

2.2 Work on Biographical Datamodels

The BiographyNet project applied the procedure outlined by de Boer et al. (2012) to data from the Biography Portal of the Netherlands (BPN) as described in Ockeloen et al. (2013). The BPN forms a collection of biographical dictionaries describing people who are Dutch or lived in the Netherlands. It is one of the projects that already proposed an overarching generic structure for a heterogeneous

[5] See for instance Fokkens et al. (2014) for a more elaborate discussion on this matter.

dataset, resulting in an event-centric model for biographical data (Hoekstra, 2013).
The national Australian Dictionary of National Biography[6] (ADNB), is part of a larger effort of data aggregation, collaboration and cooperation together with the Humanities Network Infrastructure (HuNI) (Arthur, 2017).
The transnational "Biographie-Portal"[7] which combines nine biographical resources from four countries (Germany, Austria, Switzerland and Slovenia) and can be searched on name and occupation. Richer developments for these resources, and in particularly the Austrian Biographical Lexicon (ÖBL) are developed as part of the APIS project (Gruber and Wandl-Vogt, 2017).
A handful of projects have made use of linked data for enrichment and connecting biographical data to external resources. It is used for connecting data in the HuNI and ADNB data aggregation projects. The Deutsche Biographie (DB) also represents information in RDF. However, to our knowledge, neither of these resources represent all their metadata in RDF. The BPN was converted to linked data as part of the BiographyNet project, which also enriched the metadata by processing the biographical text automatically and linking extracted information to external sources (Fokkens et al., 2017). The model that is used to represent this data in RDF including an elaborate schema for representing provenance in a detailed manner can be found in Ockeloen et al. (2013).
To our knowledge, none of the projects discussed above make use of linked data to provide a generic overarching model. The work by Leskinen et al. (2017) comes closest to this idea. They provide a basic structure that can be used for prosopographical research defining name, lifespan and gender. More elaborate information can be defined using externally defined data models such as the Simple Event Model (van Hage et al., 2011, SEM).
The Biographical Data Model Repository proposed in this paper is intended to be complementary to all initiatives mentioned above. It does not provide a platform for aggregating the data itself like BNP, the ADNB or the transnational Biographie-Portal. Its goal is to primarily provide examples of a wide variation of biographical data models. These can be collected across projects with relatively limited effort. To illustrate, the fourteen resources presented here were collected in a couple of weeks. The method we propose for converting and linking data aims to go beyond defining a basic generic model for representing biographical data as developed by Leskinen et al. (2017). We propose a bottom up approach for representing various resources in RDF, which can consequently be mapped on a high or fine-grained level to other sources.

3 A comparative analysis

We collected samples from two platforms for sharing biographical data, eight biographical databases and four data models, two of which were specifically designed as part of a digitization/enhancement project related to one of the databases. This total of fourteen resources was collected as part of the preparation for the *Workshop on Biographical Data and Datamodels*.[8] A short description of each project can be found in Appendix A. The models we observed as part of this investigation come from a wide variety of projects. Some projects mainly focus on the digitization process or historical research where designing a model for presenting biographical data emerged as a by-product. Others specifically aimed at developing a formal model for biographical data.
We compare the models on the level of content (what kind of information is provided), the framework (is the model formalized and how) and formatting (how is data represented). In this investigation, we only consider components of the data that are (semi-)structured: raw text is not analyzed in depth.

3.1 General Observations

3.1.1 Content

We first examine what kind of information can be included in the models in a (semi-)structured manner. As expected, all models we examined represent the person's name and lifespan (if known). When looking at richer models, we observe common themes in the kind of information that is provided. Most resources and models address the individual's career, education, family relations and residence. Furthermore, several resources make the reason for including a person explicit by providing information labeled 'kind of person', 'category' or 'claim to fame'.
The main differences lie in the level of granularity of the information provided. Where some only indicate the sector in which a person worked, others provide detailed information about the firm, dates and time lines of the employment. The same can be observed for education.

3.1.2 Framework and Structure

The level of formalization highly differs from one model to another. The least formalized models make use of text fields for providing information. They use words represented as strings to define various categories of information and values are presented as descriptions. In these cases, minor differences can already be observed in the way dates are represented or the same location may appear using a different name. Other models use predefined classes and relations. This particularly holds to a large extent for the models that are defined in RDF. Finally, a handful of models adapted their basic structure from TEI P5, which defines a generic XML structure.
Basic representations in strings have the advantage that unstructured and semi-structured data from the original sources can be represented in its surface form in a simple and straight-forward manner. However, it may be worthwhile to invest in defining models and ontologies: predefined categories have the advantage that identical information is presented in a consistent manner. Formally defining information in RDF facilitates the process of connecting it to external resources.

[6] http://adb.anu.edu.au
[7] http://www.biographie-portal.de
[8] http://www.biographynet.nl/dh-biographical-data-workshop/

		general					categories								
		model	framework or format	event/relation	accessibility	metadata/in-text	lifespan	gender	faith	claim-of-fame/ person-type	education	occupation	residence	personal relations	further specifics
Repositories	AINM	TEI P5	XML	relation	AFR	MD+IT	✓	✓	✓		✓	✓			-
	ANB	TEI P5	XML	event	CRR	MD+IT	✓	✓				✓			-
	BPN	TEI P5	RDF/XML	event	OS/AFR	MD	✓	✓	✓	✓	✓	✓	✓	✓	-
	CBD	own	RDB	relation	OS	MD	✓			✓		✓	✓	✓	9
	CBW	SNAC	CSV/JSON	n.a.	OS	MD	✓	✓		✓					-
	DB	own	RDF/XML	relation	OS/AFR	MD+IT	✓	✓	✓	✓	✓	✓	✓	✓	-
	ODNB	TEI P5	XML	event	CRR	MD+IT	✓	✓	✓		✓	✓	✓	✓	✓
	ÖBL	own	RDB	relation	AFR	MD	✓	✓			✓	✓		✓	

Table 1: Overview of properties of individual biographical databases

3.1.3 Representation

We compared choices of representation for various data models. The most basic form of structuring data is through CSV. Advantages of using CSV are clear: it is an easy to understand format that can be operated well by humans as well as machines. On the other hand, it provides little support for defining more complex relations. Most data entries consist of rows defining the identifier for the person described, name, dates of birth and death and possibly room for a 'claim-to-fame' category and parents. They become less convenient when defining properties of which a person may have more than one during their life: professions, schools attended, residence, children, etc. They also fall short when defining more complex relations, for instance, the start and end date of each profession together with the location of the position. It is therefore not surprising that CSV is mainly used for resources that only represent a relatively modest amount of metadata on the person.

Resources that do aim to define more complex relations either represent their data in RDF, which can be represented in e.g. XML, turtle or LD-JSON, or they use some other XML format or JSON structure. XML and JSON both provide straightforward means to define multiple entries of the same categories (e.g. a list in JSON or sequence of XML elements) as well as the means to define more elaborate relations. It is possible to provide formal definitions of what constitutes well-formed XML of a given data structure, including the elements, attributes and values that are permitted. However, XML itself does not offer the means to formally define the meaning of these elements, attributes and values. To summarize, RDF models provide, in principle, the richest formal definitions and are most (explicitly) expressive, followed by (non-RDF defined) XML structures, JSON and finally CSV. The order of complexity of the model, the effort involved in defining them properly and possibly the order of the gentlest learning curve for people starting to work with them, is the inverse: CSV is the simplest, followed by JSON, XML and RDF.

3.2 Data Sample Analyses

We compared samples of fourteen biographical data resources outlined in Appendix A[9] paying attention to the level of formalization, the overall structure (relation-based, event-based or both) of the model as well as the categories provided for most entries or, for the four datamodels, which categories they specifically formalize. We also indicate the availability of the data itself for the eight databases.

3.2.1 Databases

Table 1 provides an overview of the properties of the databases. The left side of the table indicates general properties. The first column indicates the generic model that was used as a basis for the model employed by the database: three projects invented their own model from scratch, CBW makes use of representations developed as part of SNAC and all others have taken TEI P5 as a basis. The second column indicates whether the database makes use of the framework RDF and otherwise, which representation format is used. Both databases that have RDF representations also represent information in plain XML. ABD and CBDP are relational databases that can be queried using SQL. CBW uses CSV and JSON for data representations. The third column indicates whether the structure of the representation is event-centric or mainly relational. The model used for CBW is not rich enough to make this distinction. Two databases are copyright restricted (CRR), two databases can be made available for research purposes (AFR), two are open source (OS) and two are partially open source and can partially be made available for research (OS/AFR), as indicated in column five. The sixth column indicates whether the database only provides structured data as metadata (MD) or whether it also provides structured data tagged in the biographical text (+IT).
The right side of the table indicates which categories of information are provided as specifically structured data. It should be noted that lack of a checkbox does not necessar-

[9]The abbreviations used in our comparison are introduced in the Appendix as well.

ily mean that the information is not present in the resource. The information can standardly provided in the biographical text or it can be provided in a semi-structured manner, rather than being part of the structured dataset. The last column indicates the extent to which alternative categories are provided in a structured way. The ÖBL has at least 36 additional relations defined, CBDP has 9 additional information fields and ODNB mainly provides relatively fine-grained subcategories.

3.2.2 Platforms and Data models

What information is formally represented in the two platforms and four models is presented in Table 2. The information provided by APIS and BiographyNet (BNET) correspond to that included in the respective databases they are related to (ABD and BNP). For reasons of space, we omitted categories that are only provided by one of these two resources.

APIS provides the same 36+ relations that are indicated for the ABD. The other resources can provide richer structured information due to their ability to be combined with other models. BNET, BCRM and DFKI are defined in RDF for this exact reason. SNAC and EIBIO do not represent their data in RDF, but do make use of external links to connect information from various sources.

	framework/format	event/relation	lifespan	gender	education	occupation	personal relations	external links/extensions
APIS	RDB	rel.	✓	✓	✓	✓	✓	✓
BNET	RDF	event	✓	✓	✓	✓	✓	✓
BCRM	RDF	event	✓	✓		✓	✓	✓
DFKI	RDF	event	✓		✓	✓		✓
SNAC	JSON	rel.	✓		✓			✓
EIBIO	CSV	rel.	✓					✓

Table 2: Overview of properties defined in models and platforms

3.2.3 Summarizing the analysis

Overall, we observe that all resources provide ways for specifying a person's life span in a structured way. Almost all resources provide means to specify a person's occupation or gender, CBW being the only exception when it comes down to education and ABD and CBDP being the only two sources that do not seem to have a field to specify gender. The other categories, faith, person-type/claim-to-fame, education, residence and personal relations each occur in four to eight resources. The division between event-based and relational based structures is about 50-50. Notably resources that make use of RDF seem to have a preference for event-centric structures. A probable reason for this will be outlined in Section 4.2, where we describe the process of connecting data, including a conversion step to representations in RDF.

4 The BDM Repository

As a practical approach to address the two main drawbacks of developing models independently outlined in Sections 1 and 2, we initiated a repository of biographical data models (the BDM repository). We first describe the process of collecting models in the BDM repository and then outline the process we intent to follow to connect the models collected in this repository.

4.1 Collecting Data

The Biographical Data Model (BDM) Repository is a place for collecting and connecting biographical data models. The BDM Repository serves three purposes: First, researchers faced with the task of representing biographical data can find various examples of models used by other projects in one place. Second, the repository forms a natural environment for comparing data models and recording advantages and disadvantages of various representations. Third, the repository will support the process of representing models in RDF (for those that are not represented in RDF already) and defining correspondences between models. These correspondence definitions can be used to link data from various models, which in turn, enables a wide range of comparative research.

The first challenge this repository faces is that many biographical data collections are copy-righted. From the collections described above, only two are completely open source and two are partially open source. Samples from the other resources cannot be made openly available to everyone. To circumvent this problem, we wrote a handful of biographies of fictional characters and make the texts and metadata we (partially) invented available under the Creative Commons License. The idea is that the repository will ultimately include representations of these non-copyrighted texts in all biographical data models we are aware of. This allows us to illustrate the structure of the models without sharing their copy-righted content. It has the additional advantage that it becomes easier to compare information between models, since different samples provide the same information.

The BDM repository currently provides samples for all 21 dictionaries included in the Biographical Portal of the Netherlands. They are illustrated by the biography of Mary Morstan, protagonist in one of the Sherlock Holmes books and later wife of dr. Watson. The biographies are written in English, but otherwise follow the conventions of the original resources (concerning abbreviation and semi-structure in text). The information provided on Morstan currently covers the categories included in the BPN models and will be extended accordingly as models with structure for additional information are added. The latest version of the BDM repository can be found on github.[10]

[10] https://github.com/cltl/BiographicalDataModels

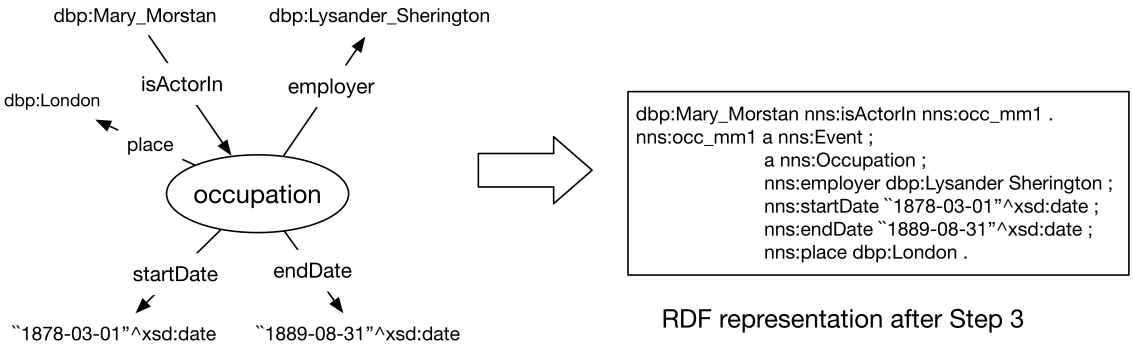

Figure 1: Illustration of conversion of event-centric data representation to RDF

4.2 Connecting Biographical Data

Once multiple data models have been included in the BDM repository, we can investigate how to connect them. We plan to achieve this by representing all models in RDF. Once individual models have been formally defined, we can define correspondence between them. In this section, we outline this process.

4.2.1 From CSV or XML to RDF

The first step is to provide RDF representations for models that have not been defined in RDF so far. When converting from one representation format to another, there is always a risk of loss in information. This particularly applies when the data is converted to a standardized model. We avoid this by following the procedure outlined in de Boer et al. (2012) for converting XML to RDF and adapting a similar approach for converting CSV and JSON files. The procedure consists of the following steps (adapted from de Boer et al. (2012), page 735): 1) XML/CSV/JSON ingestion. 2) Crude conversion to RDF. 3) RDF restructuring. 4) Design metadata mapping scheme. 5) Align vocabularies with external sources. 6) Publish as Linked Data.

In the first step, the original structure is interpreted. Then a direct conversion to RDF maintaining the full original structure takes place. As also explained by de Boer et al. (2012), data in XML can be complex: elements can be nested deeply within other elements, they may be grouped in a specific manner or ordered by the structure. Some of these structural properties are meaningful (e.g. elements within a group are connected by some implicit link, or the order of elements indicates their order in time), but many do not express information that needs to be maintained in the RDF structure. If the original XML (or JSON) is complex, the resulting RDF structure is likely to be messy. The third step addresses this by restructuring the RDF so that structures containing implicit information are translated to flatter (non-embedded) representations that make this information explicit and idiosyncratic complexities are removed. The first three steps ideally result in an RDF representation that is as simple as possible, but still provides all information from the original data.

In the fourth step, researchers explore which categories and relations expressed in the generated RDF correspond to definitions and classes defined in other vocabularies. Based on this exploration, correspondences between the resulting RDF and existing models and vocabularies can be defined. In the fifth step, these correspondences are used to link the generated RDF to external sources after which it is possible to publish the model as linked data. The BDM repository aims to help researchers carry out the first four steps. Since the repository only provides mock-up samples of data, the actual alignment of the resource and publication as linked data is out of scope. In the next subsection, we will explain how correspondences may be defined between a relational based and event-centric model.

4.2.2 Conversions and Linking

Figure 1 provides an illustration of the conversion of an event-centric representation to RDF. We illustrate the representation of the event after Step 3, before the step mapping it to other resources. The namespace `nns:` stands for a new namespace for the dataset. Conversion to RDF is relatively straight-forward: a unique identifier is assigned to the event, this is typed as an occupation and all other information can be defined directly as properties of the event. In the next step, these relations can be mapped to other existing models. We can use the Simple Event Model (van Hage et al., 2011) for instance to define the location, the begin time and end time. Categories that commonly occur in biographical data, such as occupations, should ideally also be defined by the same vocabulary across resources.

Representing a relational based structure in RDF requires more effort for relations that are temporary bound or tied to a specific location. Figure 2 provides an illustration. In principle, the relation itself can easily be translated into RDF by assigning a URI to the relation and specifying its meaning. However, we then need to decide how to specify the duration and location of the employment. The problem of making statements about a triple in RDF is well-known and several solutions have been proposed for solving this challenge. Van Atteveldt et al. (2007) provide an in depth analysis of proposals. We illustrate two commonly used approaches in Figure 2.

On the left-hand side, the statement about Mary's employment is taken as a unit that can receive its own identifier. This approach is used for defining context (Carroll et al., 2005; MacGregor and Ko, 2003, e.g.). In our example, we use a named graph for assigning an identifier to the relation. Information about time and place are then linked to the identifier of the named graph. The advantage of this approach is that it remains close to the original data structure. Following a solution originally designed to define contexts

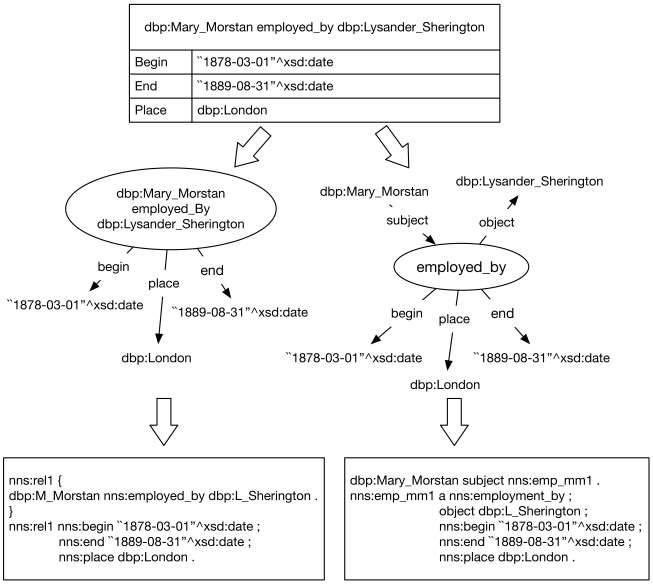

Figure 2: Illustration of conversion of relational data representation to RDF

also intuitively makes sense: the specific relation applied in a given time period and in a given place. On the other hand, we also want to define the context in which the information about time and place is provided: what is the original source of this information? How was it integrated in this database and by whom? What conversions and other operations were applied to this data? Modeling provenance is essential for research in the digital humanities (Ockeloen et al., 2013, among others). We can place the information in the left box of Figure 2 as well and then define provenance information for this new named graph, but (potentially extensive) use of nested named graphs does not improve the usability of our data structure.

The solution on the right-hand side is called *reification*. In this case, a new node is introduced that splits the predicate `employed_by` into two relations: one with the subject of the original triple and one with the object. Properties associated with the relation can then be linked to this new node. This solution changes the original structure making the relation between, in this example, the employer and employee less direct: they are now connected to the same node rather than each other. It also increases the number of relations. On the other hand, it avoids introducing an additional layer of nested named graphs. An additional advantage is that reification of relations that involve a state or event result in event-centric structures (compare the representation on the right-hand side of Figure 2 to the one in Figure 1). Reification thus facilitates the process of defining correspondences between information from these relational based representations to information represented in event-centric models. We will therefore adopt this solution once we start connecting information from various models.

5 Conclusion

Many projects that involve digitizing or enriching biographical data develop their own data model. In addition to the inefficiency of not making use of knowledge acquired in by other resources, this has led to differences between models making it harder to make connections between various resources. We illustrated some of these differences through an analysis of fourteen resources collected as part of the *Workshop on Biographical Datamodels* held in Krakow, July 2016.

The problem of models being developed independently is partially due to the difficulties involved in finding detailed information on data representations used in various projects. In this paper, we have taken a first step in addressing the problem. We propose a practical approach in the form of a biographical data model repository where detailed examples of different models can be collected. The samples will make use of biographical texts of fictional characters and invented data written under the create commons license avoiding issues with copyright.

Once a number of resources have been collected, the repository can furthermore be used to start and define connections between models by mapping them to a generic biographical representation. We outlined a general procedure that starts by converting resources to linked data representations (if they are not provided in RDF already) and consequently linking them to a generic model. We illustrated the process of converting event-centric and relationally structured resources to RDF. We showed that relational resources can be converted to event-centric representations in RDF when applying reification.

As of the moment of submission, the repository illustrates all 23 biographical dictionaries included in the Biography Portal of the Netherlands. In the near future, we plan to add illustrations of the other thirteen resources we collected, as well as encourage researchers involved in other projects with biographical data to add illustrations of their models to the repository. The repository is available on github.[11]

[11] `https://github.com/cltl/BiographicalDataModels`

6 Acknowledgements

This work was supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award to the UvA and VU Universities and NWO VENI grant 275-89-029 awarded to Antske Fokkens. We furthermore would like to thank researchers involved in the individual projects for providing samples of their data as well as the participants of the BDM workshop in Krakow for their input during discussions. We thank the audience of BD2017 and anonymous reviewers for their useful and detailed feedback. All remaining errors are our own.

7 References

Paul Arthur. 2017. Integrating biographical data in large-scale research resources: Current and future direction. In Á. Z. Bernád, C. Gruber, and M. Kaiser, editors, *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards fr eine europische Biographik*, pages 193–224. New Academic Press, Vienna.

Peter K Bol, Robert M Hartwell, Michael A Fuller, et al. 2004. China biographical database project (cbdb).

Alison Booth. 1999. The lessons of the medusa: Anna jameson and collective biographies of women. *Victorian Studies*, 42(2):257–288.

Jeremy J Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*, pages 613–622. ACM.

Victor de Boer, Jan Wielemaker, Judith van Gent, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. 2012. Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In *ESWC, volume 7295 of Lecture Notes in Computer Science*, pages 733–747, Berlin and Heidelberg. Springer.

Thierry Declerck and Rachele Sprugnoli. 2018. Considerations about uniqueness and unalterability for the encoding of biographical data in ontologies. In *Proceedings of the second Conference of Biographies in a Digital World BD2017*.

Österreichische Akademie der Wissenschaften. 2013. Österreichisches biographisches lexikon 1815–1950. online edition. *Online Publikation: http://www. biographien. ac. at/oebl*.

Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini, and Herbert van de Sompel. 2010. The europeana data model (edm). In *World Library and Information Congress: 76th IFLA general conference and assembly*, pages 10–15.

Bernhard Ebneth and Matthias Reinert. 2017. Potentiale der deutschen biographie als historisch-biographisches informationssystem. In Á. Z. Bernád, C. Gruber, and M. Kaiser, editors, *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards fr eine europische Biographik*, pages 283–295. New Academic Press, Vienna.

Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. Naf and gaf: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. 2017. Biographynet: Extracting relations between people and events. In Á. Z. Bernád, C. Gruber, and M. Kaiser, editors, *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards fr eine europische Biographik*, pages 193–224. New Academic Press, Vienna.

Greta Franzini, Melissa Terras, and Simon Mahony. 2016. 9. a catalogue of digital editions. *Digital Scholarly Editing*, page 161.

Christine Gruber and Eveline Wandl-Vogt. 2017. Mapping historical networks: Building the new Austrian Prosopographical Biographical Information System (APIS). In Á. Z. Bernád, C. Gruber, and M. Kaiser, editors, *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europische Biographik*, pages 271–282. New Academic Press, Vienna.

Daniele Guido, Marten Düring, and Lars Wieneke. 2016. European integration biographies reference database (eibio). In *DH Benelux*.

Brian Harrison. 2004. The dictionary man in: M. bostridge ed. In *Lives for sale. Biographers tales*, pages 76–85.

Rik Hoekstra. 2013. Historische representativiteit in context. over het biografisch portaal als onderzoeksinstrument.

John Kendall. 2014. American national biography. *Reference Reviews*, 28(2):7–10.

Hans-Ulrich Krieger and Thierry Declerck. 2015. An owl ontology for biographical knowledge. representing time-dependent factual knowledge. In Serge ter Braake, Antske Fokkens, Ronald Sluijter, Thierry Declerck, and Eveline Wandl-Vogr, editors, *Biographical Data in a Digital World. Proceedings of the First Conference on Biographical Data in a Digital World. Amsterdam, The Netherlands, April 9, 2015*, pages 101–110.

Katalin Lejtovicz and Amelie Dorn. 2017. Connecting people digitally-a semantic web based approach to linking heterogeneous data sets. In *Proceedings of the Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP 2017*, pages 1–8.

Petri Leskinen, Jouni Tuominen, Erkki Heino, and Eero Hyvönen. 2017. An ontology and data infrastructure for publishing and using biographical linked data. In *Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). CEUR Workshop Proceedings (October 2017)*.

Tom J Lynch. 2014. Social networks and archival context project: A case study of emerging cyberinfrastructure. *DHQ: Digital Humanities Quarterly*, 8(3).

Robert M MacGregor and In-Young Ko. 2003. Representing contextualized data using semantic web tools. In *PSSS*.

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

Legêne. 2013. Biographynet: Managing provenance at multiple levels and from different perspectives. In *Proceedings of the Workshop on Linked Science (LISC2013) at ISWC (2013)*.

Brian Ó Raghallaigh and Gearóid Ó Cleircín. 2015. Ainm.ie: Breathing new life into a canonical collection of irish-language biographies. In Serge ter Braake, Antske Fokkens, Ronald Sluijter, Thierry Declerck, and Eveline Wandl-Vogt, editors, *Biographical Data in a Digital World. Proceedings of the First Conference on Biographical Data in a Digital World. Amsterdam, The Netherlands, April 9, 2015*, pages 20–23.

Matthias Reinert, Maximilian Schrott, Bernhard Ebneth, and Team deutsche biographie.de. 2015. From biographies to data curation - the making of www.deutsche-biographie.de. In Serge ter Braake, Antske Fokkens, Ronald Sluijter, Thierry Declerck, and Eveline Wandl-Vogr, editors, *Biographical Data in a Digital World. Proceedings of the First Conference on Biographical Data in a Digital World. Amsterdam, The Netherlands, April 9, 2015*, pages 13–19.

Wouter Van Atteveldt, Stefan Schlobach, and Frank Van Harmelen. 2007. Media, politics and the semantic web. In *European Semantic Web Conference*, pages 205–219. Springer.

Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics*, 9(2):128–136.

A Appendix: Biographical Databases

This appendix provides a brief description of all resources included in the comparative study (Section 3.2).

A.1 Data collections

AINM.IE (Raghallaigh and Cleircín, 2015, **AINM**) is a collection of biographies describing people who are in some way connected to the Irish language. It contains 1,749 biographies written in Irish of people dating from 1560 until present.[12]

The American National Biography (Kendall, 2014, **ANB**) covers the lives of 19,000 noteworthy American individuals.[13]

The Biographical Portal of the Netherlands (**BNP**) has been introduced in the previous section. It is a collection of 23 different biographical dictionaries of Dutch people.[14]

The China Biographical Database Project (Bol et al., 2004, **CBD**) provides biographical information about approximately 360,000 persons[15] most of whom lived between the 7th and 19th century. It provides detailed information about locations and has comparatively rich information about social structures. It is the only resource in our sample that specifies information about possessions.

The Collective Biographies of Women[16] (**CBW**) provides annotated information on books written in English that contain three or more short biographies describing only women. The collection was originally published as a book (Booth, 1999). The main metadata from this resource is available as CSV and it has been included in SNAC, which will be described below.

The Deutsche Biographie (Reinert et al., 2015, **DB**) (Ebneth and Reinert, 2017) consists of the old and new national German biographical dictionary online.[17] It includes information about 730,000 individuals in German speaking areas covering a timespan from the early Middle Ages until present. The resources also includes approximately 50,000 biographical descriptions.

The Oxford Dictionary of National Biography (Harrison, 2004, **ODNB**) comprises an online version of the old biographical dictionary as well as the new digital born additions.[18] In total, it contains over 60,000 biographies.

The Austrian Biographical Lexicon Online (der Wissenschaften, 2013, **ÖBL**) describes meaningful people born in the Austrian-Hungarian Empire, worked there or lived there and died between 1815 and 1950. It currently contains more than 50,000 biographies.[19]

A.2 Platforms

Our study also included two platforms meant for sharing information. The European Integration Biographies reference database (Guido et al., 2016, **EIBIO**) is a structured repository for information about people. It combines structured data with free text bringing information from external repositories such as VIAF and Wikipedia together that can be queried by an API. The data structure that is used is rather basic (data is shared as a CSV and not enough information is provided to determine whether it is relational or event-centric).

The Social Networks and Archival Context project (Lynch, 2014, SNAC) provides data of people and organizations in their socio-historical context independently from the original resources that provided information about their lives.[20] Data from the CBW is included in this resource which uses JSON as an overall structure.

A.3 Data models

For our analysis we have looked at four data models. **APIS** provides rich structured data for the ÖBL (Gruber and Wandl-Vogt, 2017). Information comes from the original metadata as well as from automated and manual annotations (Lejtovicz and Dorn, 2017). Compared to the other resources, it has a wide range of specifically defined relations between people, organizations and locations.

The BiographyNet project (**BNET**) aims to enhance the possibilities for historical research using the BPN by providing structured information in RDF, extracting information from text and providing access to this information through a demonstrator (Fokkens et al., 2017). Among others, the project resulted in an RDF version of the BPN including an extensive model for representing provenance information (Ockeloen et al., 2013).

[12] `https://www.ainm.ie`

[13] `http://www.anb.org`

[14] `http://www.biografischportaal.nl`

[15] As of April 2015, indicated by the developers

[16] `http://womensbios.lib.virginia.edu`

[17] `http://www.deutsche-biographie.de`

[18] `http://www.oxforddnb.com`

[19] `http://www.biographien.ac.at/oebl`

[20] `http://snaccooperative.org/?redirected=1`

The BioCRM (**BCRM**) is designed for representing biographical information for supporting prosopographical research in the context of the Republic of Letters.[21] It is an extension of CIDOC CRM so that it can easily be used in a variety of digital humanities projects. The model provides the means for defining basic biographical information and is mainly meant to complement or be complemented by other models.

The final model we include in our comparative analysis is the **DFKI** Biography Ontology (Krieger and Declerck, 2015). Contrary to all other resources included here, this model does not provide specific relations for persons, but rather a generic framework that can represent temporarily bound events and states as well as fixed properties of persons. It can be seen as complementary to the other models. The latest status of this ontology and a proposal for moving forward can be found in Declerck and Sprugnoli (2018), this volume.

[21] http://www.republicofletters.net