Linked Data for Digital History

Lessons Learned from Three Case Studies

Victor de Boer, Albert Meroño-Peñuela and Niels Ockeloen

Department of Computer Science, VU University Amsterdam
Amsterdam, the Netherlands
{v.de.boer, niels.ockeloen, a.merono}@vu.nl

1 Introduction

As the notion and praxis of Digital Humanities is gaining ground, humanities researchers are producing more and more digital data. To foster cross-researcher and cross-project collaborations, re-usability of these data is key, allowing for data integration and new types of integrated analysis. In this paper, we focus on the subdomain of digital history, where historical researchers collect data from historical archives for their specific research questions. Currently, these datasets are often not published in public archives and if they are, they are not easily reusable because of lack of standard data representations. To further the digital history agenda, representing and sharing data is key (Cohen et al. 2008, Meroño-Peñuela et al. 2012)

The principles of the Semantic Web (Berners-Lee et al. 2001) and specifically the practice of Linked Data (Bizer et al. 2009) provide key technologies to allow this type of data re-usability and data integration. Linked Data can be used to publish datasets that are the result of historical research; to integrate these datasets and to re-use these integrated datasets, allowing for new types of (historical) analyses.

In this paper, we present a number of case studies which use Linked Data principles for the representation and publication of digital history datasets. The three research projects that we present here are all collaborations between Dutch humanities researchers and computer scientists from VU University Amsterdam. They are in no way exhaustive and other Linked Data for Digital History projects are described (e.g. Warren, 2012, Hyvonen et al. 2012). In the next section, we introduce the principles of Linked Data and present their opportunities. In subsequent sections, we present the three projects, highlighting their motivations, the datasets produced and the new types of analyses and research questions that they support. We close with conclusions and lessons learned from these three cases.

2 Linked Data for Digital History

2.1 Linked Data principles

Shortly after the birth of the Web in 1989, Tim Berners-Lee (Berners-Lee et al. 2001) envisioned the Semantic Web as an evolution of the original World Wide Web, which was originally built on HTML documents. Most of the contents of these documents were designed for humans to read, but not for computer programs to process meaningfully. Computer programs certainly could parse the source code of Web pages to distinguish layout information and text, but before 2001 there were no mechanisms to process their semantics, their meaning. The "semantic" Web enables the sharing of content from databases and other structured data sources that are not directly published on the Web. The Semantic Web "is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee et al. 2001).

This vision is being realized nowadays by a collaborative movement and a set of standards lead by the World Wide Web Consortium (W3C). The Resource Description Framework (RDF) (Cyganiak et al. 2014) is the basic layer on which the Semantic Web is built. In RDF, entities of the world are represented with subjects and objects, while the relationship between the two are represented with predicates that connect them (e.g. "Amsterdam" "is located in" "The Netherlands"; "Amsterdam" is the subject, "The Netherlands" is the object, and "is located in" is the predicate). Hence, RDF is a knowledge representation system in which facts and their properties are expressed as subject-predicate-object sentences called triples. All these connected triples have the form of a graph (e.g. all the other cities located in the Netherlands are also connected to the resource "the Netherlands" through the predicate "is located in"). Finally, all unique subjects, predicates and objects are assigned a Uniform Resource Identifier (URI) that uniquely identifies them on the Web. Once converted and published, RDF data can be queried online through the query language SPARQL[1] (SPARQL Protocol and RDF Query Language). In 2008, the practice of using HTTP, RDF and URIs became a new paradigm on publishing data on the Web, today known as Linked Data. Tim Berners-Lee recommended four principles on a technical note[2], describing how HTTP, RDF and URIs could be used to publish Linked Data on the Web:

1. Use URIs to name (identify) things.
2. Use HTTP URIs so that these things can be looked up (interpreted, "dereferenced").
3. Provide useful information about what a name identifies when it's looked up, using open standards such as RDF, SPARQL, etc.
4. Refer to other things using their HTTP URI-based names when publishing data on the Web.

2.2 Linked Data for Digital History

More and more sources for historical research are being published under the paradigm of Linked Data in the Semantic Web (Meroño-Peñuela et al. 2015). To better understand how the principles of Linked Data can be used for Digital History, we use the life cycle of historical information (see Figure 1) of (Boonstra et al. 2004). In this cycle, historical objects traverse six phases:

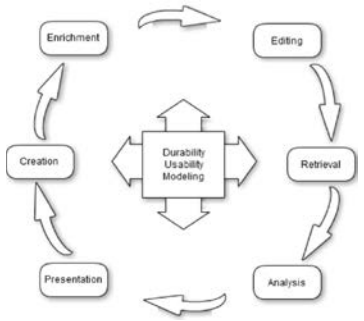


Figure 1: The life cycle of historical information as proposed by Boonstra et al. (2004)

1. **Creation,** where physical creation of digital data takes place. This includes the design of the information structure and the research project. Examples of activities in this phase are the data entry plan, the digitisation of documents, or the choice of a specific database management system.

[1] http://www.w3.org/TR/rdf-sparql-query/

[2] Design Issues: Linked Data (https://www.w3.org/DesignIssues/LinkedData.html)

2. **Enrichment,** in which data created in the previous step is augmented with metadata. These metadata provide additional details to the historical information. This phase also comprises the linkage of data of different datasets that belong together to a historical reality, because these data belong to the same person, place or event.
3. **Editing,** where actual encoding and data entry of historical information is performed. Example activities would be inserting markup tags, or entering data in the fields of database records. All data transformations through algorithmic processes prior to analysis also belong to this phase. These include annotating original data with background information, adding bibliographical references, and linking related quotations.
4. **Retrieval,** where information is selected, looked up, and used (e.g. via SQL or XPath).
5. **Analysis,** which varies depending on the data and goals, from qualitative comparison and assessment of query results, to statistical analysis and model construction.
6. **Presentation,** where historical information is communicated through multiple forms of presentation (e.g. text editions, online databases, virtual exhibitions, visualizations).

Three invariant aspects are central along the development of these phases: *durability*, which ensures the long term deployment of the produced historical information; *usability*, which ensures efficiency, effectiveness and user satisfaction; and *modeling*, which occurs during research processes and within historical information systems.

As shown in (Meroño-Peñuela et al. 2015), Linked Data is being currently used in Digital History to address four important open problems in this field:

- **Historical sources**. The first set of open problems in historical research happens in phase 1 of the historical data life cycle, in which the historical data are created. The creation of the dataset might present the first issues: what is the word that is written on this thirteenth-century manuscript? What is its meaning? Provenance is also important: even if the source is clearly identified and its meaning deciphered, the historian needs to know more: To what events does it relate? Why was it put there? Who was the author? How has it survived? And also important: how can all these be modeled into a coherent data model? A shown in (Meroño-Peñuela et al. 2015), Linked Data and the Semantic Web have provided two useful resource sets to address these issues: historical ontologies, and

Linked Data vocabularies. These semantically rich models “aid historians to describe, at least, the baseline historical entities and relations in historical domains: events are combinations of persons, places and moments in time when something historically relevant happened”. The modelling and publication of historical Linked Data is supported by these basic terminologies, that can be furtherly refined.

- **Relationships between sources**. As typically historical researchers work with multiple isolated data sources at a time, they face the problem of how to integrate information in these dissimilar sources for their purposes (second phase in Figure 1, enrichment). For instance: is this Lars Erikson, from this register, the same man as the Lars Eriksson from this other register? Obvious linkage problems are name disambiguation, management of changing names, and how to standardization of spelling variations. Additional problems appear when linking historical data with their spatial and temporal context. As described in Section 2, one of the goals of Linked Data consists of solving this linkage problem. Previous research has shown that the technical infrastructure for providing a global knowledge base to support the discovery and linkage of relationships between sources in historical research by using Linked Data is already set up and running (Meroño-Peñuela et al. 2015). Unfortunately, the number of linked historical datasets is still too small to provide a useful historical linkage platform for every case, and “little background knowledge can help today these historians in solving e.g. errors or inconsistencies” in name disambiguation and record linkage. More Linked Historical Data of high quality would dramatically increase the usefulness of Linked Data for historians in this respect.

3 Dutch Ships and Sailors

3.1 Motivation

The Dutch Ships and Sailors project (de Boer et al. 2015)[3], was a digital history project that ran from 2013-2014 as a collaboration between maritime historians and computer scientists. The goal of this project was to publish a number of Dutch maritime historical data sources in an

[3] This section is based on work previously published in de Boer et al. 2015 and de Boer et al. 2014. It reproduces a number of figures and paragraphs.

integrated linked data cloud. As much of Dutch history is found on the water, being central to economic, social and cultural life, it is one of the best historically documented sectors of human activity in the Netherlands. In the past few decades, much of the data in the preserved historical source material has been digitized. Even though these datasets are related through common entities such as places, ships and ship types, persons or historical events, up until now, these data sources can be considered data silos with no explicit references outside of the individual datasets. During a workshop on maritime historical databases, professional and amateur historians agreed that the integration of these data sources could significantly improve historical research and open new possibilities for explorations and analyses[4]. Initially, four of such maritime databases were connected using Linked Data principles (de Boer, 2014). In a follow-up project, two more datasets were added to the data cloud (Entjes, 2015). The resulting data cloud of Dutch Ships and Sailors is already a useful data source that may be used for research that transcends the possibilities of the original, unconnected datasets, but its significance lies also in the open nature of the linked data cloud. By design, new sets may be incorporated into the cloud using the same linked data.

3.2 The DSS data cloud

We list the six different datasets that currently make up the DSS data cloud below. Within a pilot study, over 25 Dutch maritime historical datasets were identified as potential candidates for the DSS data cloud and gathered on a web page with information about availability[5]. Out of these 25, four were selected on the basis of availability and interestingness. Two of these datasets (GZMVOC, MDB) were not digitally published before, but were the result of running research by project participants. Two other datasets (DAS, VOCOPV) were central to Dutch maritime historical research and describe Dutch voyages in the 17th Century. The last two datasets were added in a follow-up project focussing on economic voyages in the 19th Century (Entjes, 2015).

- **GZMVOC:** The “Generale Zeemonsterrollen VOC” (GZMVOC) (En: “General sea muster rolls VOC”) is a dataset describing the crews of all ships of the Dutch East India Company (VOC) from 1691--1791. The data was gathered by a Dutch social historian

[4] Workshop Dobberende informatie: van zeevarenden naar databases (7 September, 2012). See http://dutchshipsandsailors.nl/wp-content/uploads/2013/12/2012-07-12-Workshop-Maritime-Historical-Datasets-Dobberende-Informatie.pdf

[5] http://dutchshipsandsailors.nl/?page_id=11

Matthias van Rossum in the course of his research on labor situations for European and Asiatic crews on Dutch East India Company (VOC) ships. The data consists of the size of the captain and crew as well as its composition (number of European and Asiatic sailors, soldiers and passengers), geographical location of the counting as well as data on the name and type of ship. References to the Dutch Asiatic Shipping (DAS) records were also present..

- **MDB:** The "Noordelijke Monsterollen Databases" (En: "Northern muster rolls databases") is a dataset describing mustering information found in mustering archives in the three northern Dutch provinces (Groningen, Friesland, Drenthe) in the period 1803--1937. The original Noordelijke Monsterollen Databases (MDB) was provided as a SQL dump file by the original maker of the data, historian Jurjen Leinenga.
- **VOCOPV:** The original dataset "VOC Opvarenden" (van Velzen, 2010) is the result of a manual digitization of the personnel data of the VOC in the 18th Century. The original data consists of three separate parts (En: `voyagers', `salary books' and `beneficiaries') and was downloaded from DANS Easy website[6].
- **DAS:** The Dutch Asiatic Shipping (DAS) dataset contains data regarding outward and homeward voyages of more than 4,700 ships that sailed under the flag of the VOC between 1595 and 1795. The dataset is a conversion of a previously digitized DAS dataset hosted at Huygens ING[7]. Between 1595 and 1795 the Dutch East India Company (VOC) and its predecessors before 1602 equipped more than 4,700 ships to sail from the shores of the Netherlands bound for Asia.
- **Elbing:** The Elbing dataset contains toll registry information from voyages starting in 1585 until 1700 (Lindblad et al., 1995). In the creation of this dataset, all shipping that had goods headed towards the Netherlands regardless of captain or ship nationality, and all shipping by Dutch captains regardless of the destination of cargo, have been included.
- **Archangel:** The data about Archangel contains entries of voyages to Archangel and other European ports from 1594 until 1724. This data was originally gathered by Piet de Buck (1931 - 1999). Its sources are cargo contracts and other notarial acts from the Amsterdam

[6] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:33602
[7] http://resources.huygens.knaw.nl/das/index_html_en

city archives. It contains information about the ship, captain, the freight brokers and the ship's intended route and cargo pricing. The dataset contains roughly 4700 acts.

One of the main benefits of Linked Data is that it allows data integration on a loose, ad-hoc basis. There is no need for a rigorous data schema definition that the individual datasets should adhere to. This allows for the use of heterogeneous data models to represent the individual datasets, while interoperability is still achieved through linking of instances, classes and properties. This light-weight integration also allows for later updating of datasets and easy integration of new datasets. For each of the six datasets, RDF representations were created following the methodology as described in (de Boer et al. 2012). For more details on this methodology and the different tools used in it we point the reader to (de Boer et al. 2014) and (Entjes 2015).

The end result consists of a number of datasets modeled as RDF *named graphs* that each have their own structure and model. Next to this, each of the datasets is accompanied by an RDF schema file that lists the classes and properties used in that graph. To achieve interoperability, we then construct a number of additional named graphs containing links:

- We establish links at a schema level: these link classes and properties from one dataset to other datasets and to a generic (DSS) schema. For example a "Master" in DAS or a sailor in VOCOPV are both defined as being a subclass of "Person" as defined at the DSS level.
- We link shiptypes, rank types and geographical entities to common vocabularies available in the Linked Data Cloud. These include DBPedia, the Getty Art and Architecture Thesaurus and GeoNames[8].
- We link individual entities between two datasets. For example, references from GZMVOC to DAS identifiers are formalized as RDF relations.
- We also link to non-Linked Data resources, more specifically to digital historical newspaper articles from the Dutch Royal Library (KB)[9].

[8] http://dbpedia.org/, http://www.getty.edu/research/tools/vocabularies/aat/ and http://geonames.org respectively

[9] http://kranten.delpher.nl

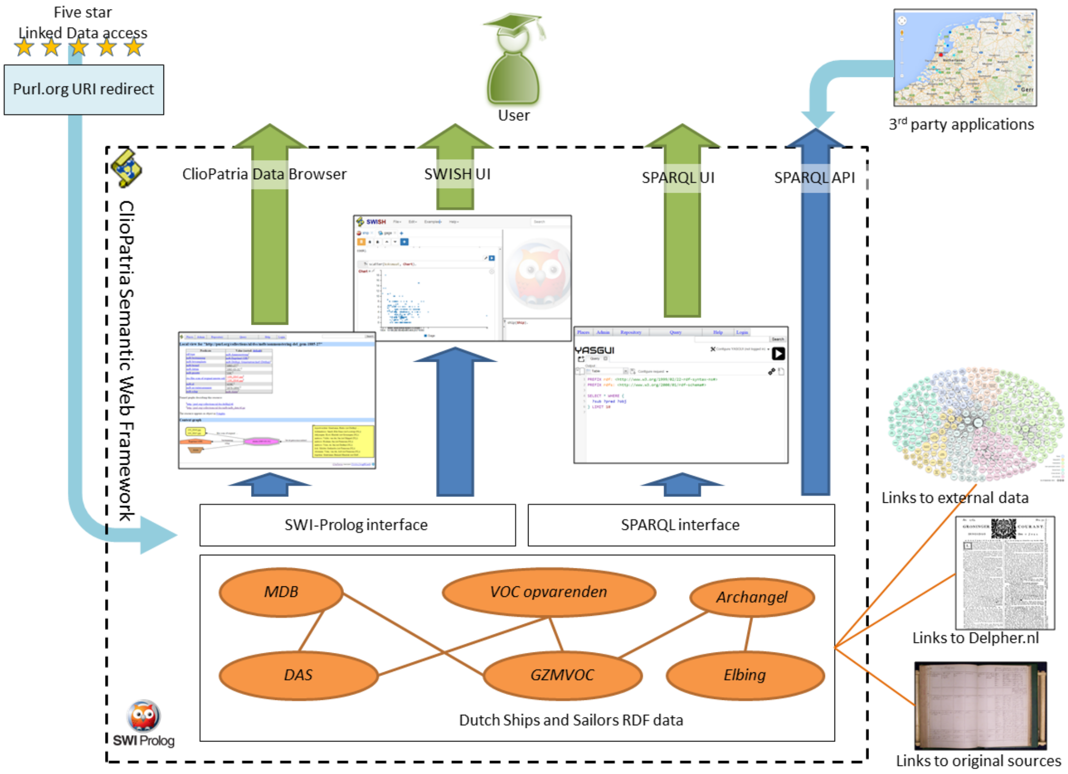


Figure 2: Overview of DSS data cloud and how it is made accessible.

3.3 Accessing the Data cloud

Figure 2 shows how the different datasets are connected and how they are made accessible. The different datasets and the links between them are separate *named graphs* that are loaded into a live ClioPatria semantic server[10]. At present there are two versions running live: a stable version at http://dutchshipsandsailors.nl/data and a development server at http://dutchshipsandsailors.nl/data. The ClioPatria framework includes a triple store and offers a number of APIs and User Interfaces: For live querying, there is a standard SPARQL API available and both servers also provide a live query environment through which (human) end users can query the data. The SPARQL API can be used by external applications. An example is an application that visualizes geographic data[11]. ClioPatria furthermore offers a web interface that allows for full-text search through the entire data cloud and a graphical browser to explore the knowledge graph. Finally, there is also the possibility to query the data via a SWISH web

[10] http://cliopatria.swi-prolog.org

[11] http://entjes.nl/jeroen/thesis

interface allowing arbitrary SWI-Prolog commands. Through a collaborative environment, these queries and visualizations can be shared among researchers. Next to this, all URIs are dereferencable as per the rules of Linked Data (Section 2).

3.3 Provenance

Provenance plays an important role in historical research and specifically in archival research. In the DSS cloud we model the provenance on the *named graph* level using the PROV-O[12] datamodel . Each named graph is a separate set of triples that come from one source. This can be either an original data source, or the result of an enrichment or linking process. Provenance triples describe for each named graph a) the process from which it originates b) (software) actors involved in those processes and c) datasets used as input. We also annotate a number of automatically generated linksets with metadata about the level of confidence of the links in those graphs. This provenance data can be queried alongside the data itself, allowing users to consider only trusted parts of the data and exclude other parts.

3.4 Queries and research questions

In collaboration with the historical researchers, a number of example use cases and queries have been developed. Editable example SPARQL queries are presented at the development server. We here present some examples of interesting types of analyses possible on this data cloud:

- **Cross-dataset search and comparison**: Due to the mapping of dataset-specific properties to more generic ones, it becomes possible to search resources across the different datasets. For example, we can search for all persons with the name “Veldman”, results of which includes captains from DAS as well as a cook from the MDB dataset. It also allows for comparisons between datasets.
- **Exploiting Geographic background knowledge:** The links from different resources to GeoNames entities allows us to use that background knowledge for new types of analyses and visualisations. This can be simple plots of ship or person locations on a map, but also more complex like the example shown in figure 3, where birthplaces of sailors are visualised per Dutch province over multiple years.

[12] http://www.w3.org/TR/prov-o/

- **Use of background knowledge:** Through the link with other background data, we can automatically analyze the data in new ways. For example, the ship type hierarchy from AAT can be used to analyze features of specific ship types. One of the example queries lists persons that embarked on coastal ships (which has a number of subtypes such as “kof” or “tjalk”).

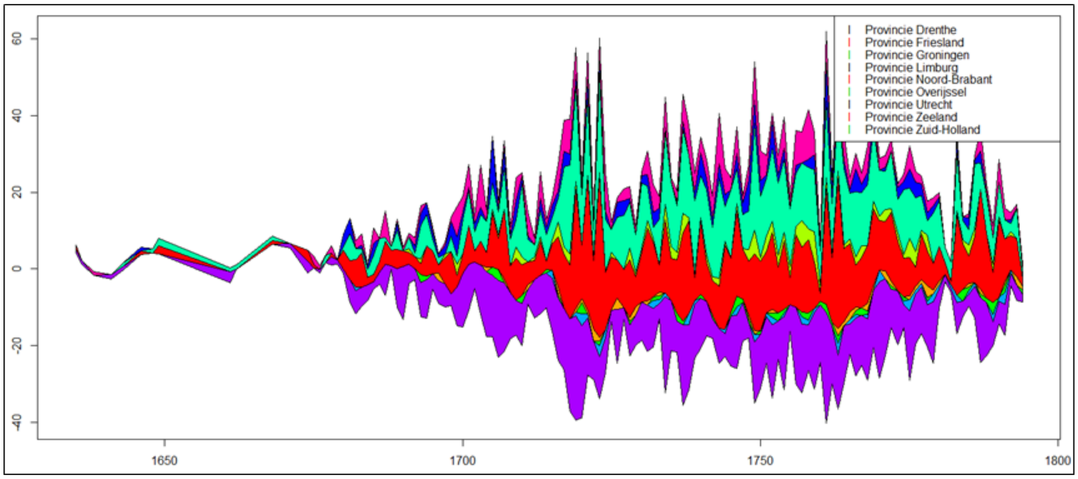


Figure 3: Birthplaces of sailors per Dutch province 1600 - 1800.

4 CEDAR

4.1 Motivation

Historical censuses are a key data source for social history research. Through History, censuses regularly and systematically collected loads of information about members of a population, recording valuable society snapshots. By studying these snapshots, social historians extract knowledge about the experiences of ordinary people in the past, one of the fundamental angles to better understand History.

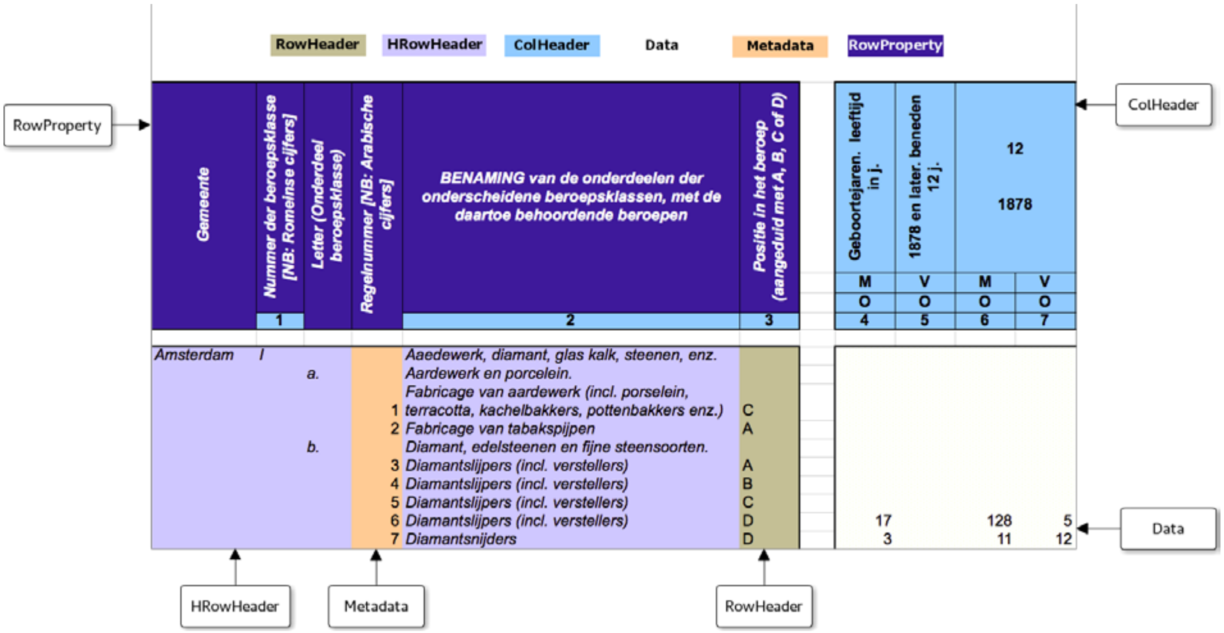
In the Netherlands, traditional censuses were conducted from 1795 until 1971 in 17 different occasions, once every 10 years. In these, the government counted the whole population of the country, door to door, and aggregated their characteristics in three different collections: the demographic census (age, gender, marital status, location, belief); the occupation census (occupation, occupation segment, position within the occupation); and the housing census (ships, wagons, private houses, government buildings, occupation status). The unprecedented level of detail of this dataset, the loss of the original survey data it is based on, and the fact that the whole population was counted (instead of sampled) make it a unique and invaluable source to social historians.

However, accessing data in this collection is extremely cumbersome. The recorded observations are scattered across 2,288 census tables. The digitization of these tables as computer spreadsheets certainly improved their access with respect to original printed materials. Nevertheless, many challenges beyond digitization remain. Social historians are still forced to work with these tables using visual inspection and data munging, which keeps the costs of longitudinal analyses too high, from days at best to months at worse. The challenges implied by such longitudinal analyses come mostly by changes introduced in each of the census editions, due to varying information needs: changes in the survey questions; changes in the variable design; changes in the classification systems behind the possible answers; and changes in the organization and layout of the census results. This heterogeneity makes comparisons across time very difficult, and is often solved by means of methods and techniques of census harmonization (Esteve and Sobek 2003).

The goal of CEDAR is to publish the Dutch historical census data on the Web in a way they can be openly and easily harmonized, integrated, shared and enriched with other datasets. To achieve this, in CEDAR we use Semantic Web standards and formats to represent the harmonized Dutch historical census data and link them to other Linked Data on the Web; but also to connect them to the original historical sources they are derived from, and to expose their conversion, curation and harmonization processes.

4.2 The CEDAR data model

We publish the Dutch historical censuses dataset as a 5-star Linked Data dataset (Meroño-Peñuela et al. 2016). The Dutch historical censuses are multidimensional data. In essence, this means that any census observation can be defined as a value or measure affected by a set of variables or dimensions. For example, in Figure 4 a valid observation is “128 single (O, ongetrouwd in Dutch) males (M) aged 12 and born in 1878 worked in Amsterdam as diamond polishers (*diamantslijpers*) of the lowest work category (D) in 1889”.



Gemeente	Nummer der beroepsklasse [NB: Romeinse cijfers]	Letter (Onderdeel beroepsklasse)	Regelnummer [NB: Arabische cijfers]	BENAMING van de onderdeelen der onderscheidene beroepsklassen, met de daartoe behoordende beroepen	Positie in het beroep (aangeduid met A, B, C of D)	Geboortejaren. leeftijd in j.	1878 en later. beneden 12 j.	12 1878	
						M	V	M	V
						O	O	O	O
	1			2	3	4	5	6	7
Amsterdam	I			Aaedewerk, diamant, glas kalk, steenen, enz.					
		a.		Aardewerk en porcelein.					
			1	Fabricage van aardewerk (incl. porselein, terracotta, kachelbakkers, pottenbakkers enz.)	C				
			2	Fabricage van tabakspijpen	A				
		b.		Diamant, edelsteenen en fijne steensoorten.					
			3	Diamantslijpers (incl. verstellers)	A				
			4	Diamantslijpers (incl. verstellers)	B				
			5	Diamantslijpers (incl. verstellers)	C				
			6	Diamantslijpers (incl. verstellers)	D	17		128	5
			7	Diamantsnijders	D	3		11	12

Figure 4: One of the census tables of the Dutch historical censuses, displaying part of the occupation census of 1889, province of Noord-Holland.

To represent this information, it is natural to choose RDF Data Cube (QB) (Cyganiak and Reynolds 2014) as our goal data model to express all census data in RDF, since QB provides a means "to publish multi-dimensional data, such as statistics, on the web in such a way that they can be linked to related datasets and concepts" (Cyganiak and Reynolds 2014). In QB, any data point is an observations, primarily composed of a measure ("128" in the previous example of Figure 4) and a set of dimensions (e.g. "marital status", "sex", "age", "year of birth", "municipality", "occupation", "job position", "year") and values (e.g. "single", "male", "12", "1878", "Amsterdam", "diamond polishers", "D", "1889") qualifying that measure. Dimensions can be arbitrarily combined to refer to unique observations in the cube. The same information can be represented as RDF Data Cube as follows:

```
cedar:example-observation a qb:Observation;
                  maritalstatus:maritalStatus maritalstatus:single ;
                  sdmx-dimension:sex sdmx-code:sex-M ;
                  sdmx-dimension:age “12”^^xml:integer ;
                  cedar:yearOfBirth “1878”^^xml:integer ;
                  sdmx-dimension:refArea gg:Amsterdam ;
                  cedar:occupation hisco:88030 ;
                  cedar:occupationPosition cedar:job-D ;
                  cedar:population “128”^^xml:integer .
```

This model allows for linking together harmonized observations that share some common dimensions. For example, one entire census edition can be constructed by retrieving all observations that share the same year. Likewise, longitudinal analyses can be performed by gathering all observations that share, for instance, the same location or the same occupation, but not necessarily the same year. Together with the harmonized and linked census data, which we call the release data, we link all observations to the harmonization technique we applied to integrate them (we call this the rule data) and to the original source table information they were derived from (we call this the raw data). Besides this internal linkage, we also link relevant census data to other Linked Data on the Web (Meroño-Peñuela et al. 2016). Figure 5 shows other datasets linked to the Dutch historical censuses and providing additional context and richness, like the Dutch Ships and Sailors (see Section 3), *gemeente geschiedenis* (an RDF dataset with information on historical borders and shapes of Dutch municipalities), DBPedia (a machine readable version of Wikipedia) and HISCO, the Historical International Standard Classification of Occupations.

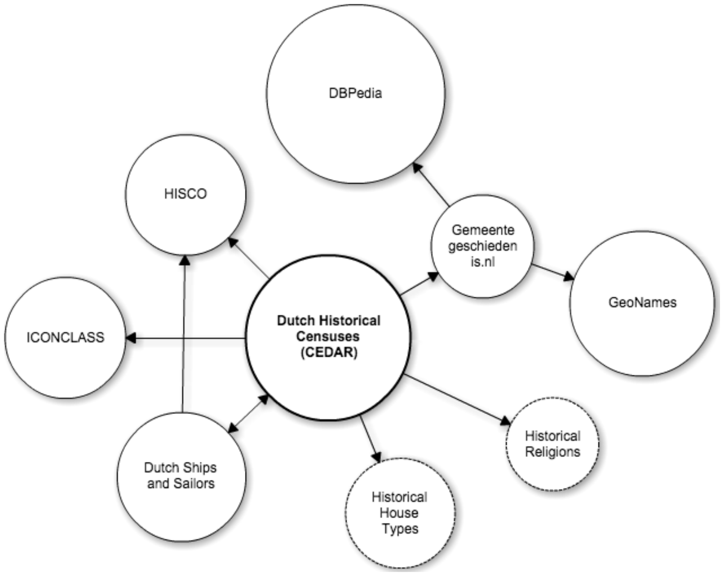


Figure 5: Other datasets represented as Linked Data that CEDAR links to/is linked from.

4.3 New types of research questions

We distinguish three different types of new research questions enabled by research in CEDAR: (i) those enabled by the availability, for the first time, of easy and reliable longitudinal analysis on the Dutch historical censuses dataset; and (ii) those enabled by links connecting the Dutch historical censuses to other databases; and (iii) those enabled by the combination of symbolic and statistical reasoning. Research questions enabled by longitudinal analysis allow fast and reliable query answering of questions that previously would have taken an immense investment on time and resources. For example, now the number of empty houses (i.e. houses without inhabitants spread across the country and over time) can be used as an indicator of public investment in different regions, migratory behaviour, and labour distributions. The existence of explicit and resolvable links between the census and other datasets enable cross-domain querying on the Web: how prominent were occupations related to ship construction in the regions where ship captains were born and raised? What is the relationship of the wealth increase of the Dutch municipalities and their changing shapes and borders? What other datasets (possibly censuses) on the Web describe the same statistical variables as the Dutch historical censuses in other regions of the world? Finally, the combination of statistical and symbolic reasoning can lead to the unveiling of new socio-historical analysis perspectives: do historical occupations with similar semantics also display similarity in their longitudinal distributions? What is the relationship between two statistically dependent variables of the census, and their semantic similarity?

4.4 Links between DSS and CEDAR

Some of the facts expressed in the Dutch Ships and Sailors (DSS) and the Dutch historical censuses (CEDAR) datasets share location and time. Therefore, historians might be interested in exploring the connections between these two datasets, in order to answer some of the research questions described in the previous section. This is our motivation to use Linked Data to link some meaningful resources between the DSS and CEDAR datasets.

The mutual connection of both datasets are historical occupations. In DSS, registers describe labour roles of individuals; in CEDAR, one of the ways of grouping people together was by considering their occupations. Therefore, we use the RDF/SKOS version of the Historical International Standard Classification of Occupations[13] (HISCO) to create links from both datasets to HISCO. We have shown in the example observation of Section 4.2 how a group of 128 young Amsterdammers can be linked to the HISCO code 88030, which describes their occupation (diamond polishers). We use the same technique in DSS. For example, the following statements[14]:

[13] http://historyofwork.iisg.nl/

[14] Other links are available at https://github.com/biktorrr/dss/blob/master/rdf/alignments/datathon_links.ttl

```
vocopv:rank-Advocaat skos:exactMatch hisco:12110.
vocopv:rank-Barbier skos:exactMatch hisco:57030.
vocopv:rank-Beeldhouwer skos:exactMatch hisco:16120.
```

Indicate how different ranks in the DSS dataset link to the HISCO resource identifiers for lawyers, barbers, and sculptors, respectively. The CEDAR observations of the occupational censuses link to these HISCO resource identifiers too. Consequently, groups of census individuals that share the same occupation with persons described in DSS (and further other characteristics, such as place and time) can be retrieved by exploring the incoming links of that occupation.

5 BiographyNet

5.1 Motivation

The BiographyNet project[15] is an e-history project bringing together researchers from history, computational linguistics and computer science[16]. The project is funded by the Netherlands eScience Center[17] and 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 that supports the discovery of interrelations between people, events, places and time periods in biographical descriptions.

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.

[15] http://www.biographynet.nl
[16] The sections on BiographyNet are based on work previously published in Ockeloen et al. 2013.
[17] https://www.esciencecenter.nl

It is imperative for historians to keep a good oversight over the sources that were used to produce a certain output. How reliable are the sources that 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.

5.2 The BiographyNet dataset

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. In the original collection, the available biographical data is not linked to any other sources. To facilitate linking to external sources and in order to reason over the data, the BiographyNet demonstrator will be based on Linked Data (Bizer et al. 2009) principles. Therefore, the collection of XML files is converted to RDF[18], using a conversion process as described in de Boer et al. (2012).

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 facilitate multiple biographies for the same person, ORE 'proxies' (Lagoze and van de Sompel 2007, Lagoze et al. 2008) are used in a way similar to how they are used in the Europeana Data Model (EDM) (Doerr et al. 2010), making the BiographyNet schema compatible with EDM. The RDF version of the biographical data is available through a public SPARQL endpoint[19].

[18] https://www.w3.org/TR/rdf11-concepts/

[19] http://data.biographynet.nl/sparql/

5.3 Retrieving information from text

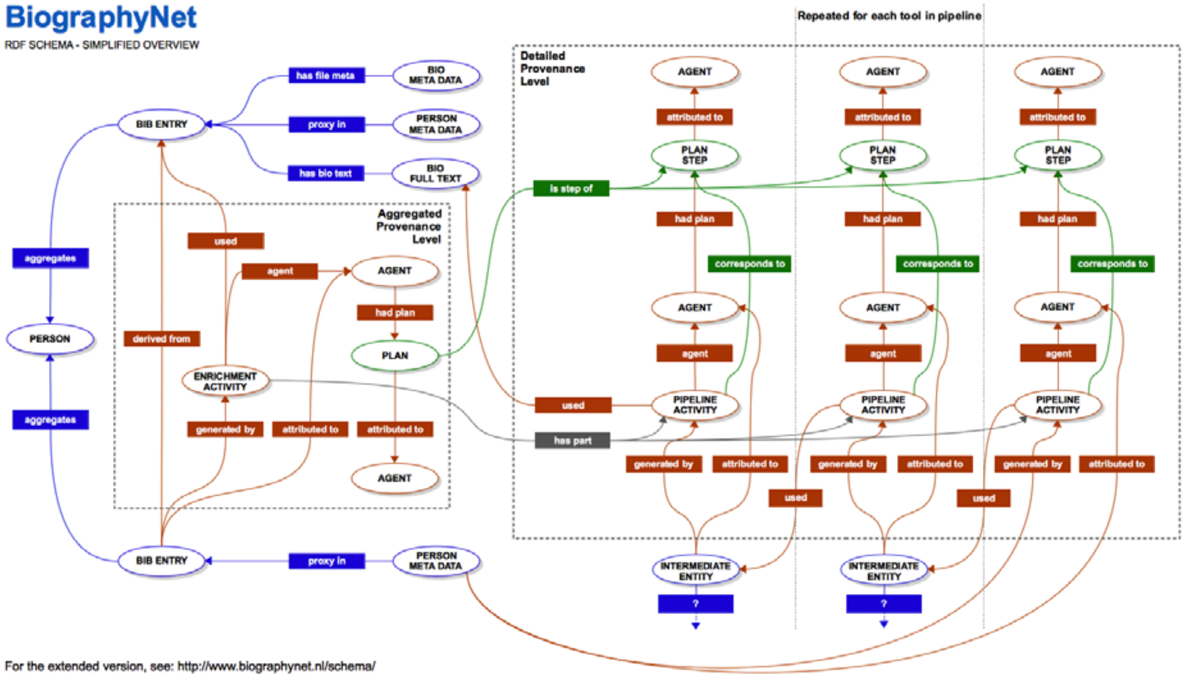
One of the main challenges of building a demonstrator lies in creating tools that can automatically interpret text and extract information from it. In particular, our analyses attempt to identify who did what, when and where according to the text. This is achieved in two main steps using open source tools, such as tools from the IXA pipeline (Agerri et al. 2014). In the first step, an NLP pipeline developed as part of BiographyNet and NewsReader[20] identifies time expressions, semantic roles, named entities and events. The pipeline also applies word sense disambiguation, maps semantic roles to FrameNet, disambiguates named entities by linking them to DBpedia where possible and establishes intra-document event coreference. In the next step, we convert the output of the pipeline to RDF representations in SEM (Van Hage et al. 2011) using GAF(Fokkens et al. 2013) to relate events back to their original mention in text. The enrichments resulting from the NLP process are used to form additional sets of metadata for the person described in the processed biographical description. The NLP processing pipeline has been established and tested. More information on the NLP aspects can be found in Fokkens et al. (2014).

5.4 Provenance

Enrichments resulting from the NLP pipeline are added to the Linked Data corpus. In order not to compromise the original data, the enrichments are added as a new biographical view for the person in question, with detailed provenance information that expresses how that new biographical view is derived from another. Provenance information is stored on various levels using PROV (Moreau et al. 2012) and P-PLAN (Garijo and Gil 2012). 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. (2012), “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 as a

[20] http://newsreader-project.eu

scientific tool for historians: In case results are pulled into question, it must be possible to trace back which specific tools in which specific process produced the possibly faulty results.



For the extended version, see: http://www.biographynet.nl/schema/

Figure 6: Simplified view of the BiographyNet data schema.

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. These provenance levels are illustrated in figure 6 which shows the representation of a Person, a biographical description object for that person, an enrichment added as new biographical description object for that person, the aggregated provenance information between the two and the detailed provenance level describing all processing steps involved in creating the enrichment. Details on the RDF data schema, its requirements and the use of provenance are described in Ockeloen et al. (2013).

5.3 Providing interesting leads

The aim of BiographyNet is not to create a ‘paper machine’ for historical research. The demonstrator provides additional information mined from the full biographical texts using Natural Language Processing (NLP), but leaves the judgement of that information up to the historian. Rather than reaching conclusions or making decisions on its own, the BiographyNet demonstrator aims to provide as much information to the historian as possible, enabling the historian to make the most optional analysis of all this information. The BiographyNet project is still ongoing, and one of the current challenges is how to display all available information to the historian in the most optimal way. Through a combination of data enrichment, quantitative analysis, visualization and browsing techniques, the demonstrator should provide interesting leads and insights that may be hard to discover using traditional methods. As such, it should inspire historians to conduct new research based on the leads found with the demonstrator. Furthermore, the use of Linked Data presents the opportunity to provide additional context by linking to information on similar resources from other datasets.

6 Discussion and Conclusion

In this paper, we have presented the results of three Digital History research projects which make use Linked Data to represent the heterogeneous information that results from these projects.

The CEDAR project shows how a long time series (1795-1971) of government collected statistics can be integrated in a coherent and accessible way. The use of Linked Data vocabularies such as RDF Data Cube provides a uniform conceptual model to express statistical observations, including those of historical nature. To represent historical concepts, these observations are linked to well known classification systems in social history, such as the Historical International Standard Classification of Occupations (HISCO) and the Amsterdamse Code (AC), to represent harmonized occupations and municipalities throughout History, respectively. As a Linked Data dataset, the Dutch census has a great potential to enrich other datasets with demographics, labour and housing data of a key historical period.

The Dutch Ships and Sailors project shows how different datasets around one theme can be brought together in a single linked data cloud. The individual datasets are separately modeled in collaboration with the historical researchers, resulting in standalone datasets. Through links to a common metadata schema as well as direct cross-dataset links we can still achieve a level of integration. This forms a data cloud, which can be integrally explored and queried, opening up new ways of analyzing the data.

The BiographyNet project shows how existing Linked Data vocabularies can be re-used to create a dataset that allows new enrichments to be added to the data, while not compromising the integrity of the original data sources. Re-using existing vocabularies for tasks as provenance and object modelling not only saved time while creating the BiographyNet schema, but also resulted in better compatibility of the data with other sources, especially datasets from Europeana and provenance data in general. Furthermore the use of Linked Data allowed us to be gradually expand the data corpus, which originally consisted of mostly full text, with more and more meta data resulting from Natural Language Processing.

After having looked at the individual projects, we can also identify some common benefits and challenges for using Linked Data for Digital History:

- **Provenance:** Provenance is key to historical research. This holds especially for research that uses digital sources, where datasets are often copied and moved between servers. Linked Data provides standard solutions such as the PROV-O vocabulary for recording provenance of individual statements or collections of statements. This provenance can also connect datasources across (web) locations, and represent the origin of digital historical sources. Additionally, in the CEDAR, BiographyNet and DSS projects, we have seen how provenance is recorded and how it can be used by historical researchers; if a system or tool used by historians manipulates data in any way, provenance information is vital to provide insight on these operations and to establish its credibility. Historical research stands or falls with the traceability of used information; hence, so called ‘black box’ operations are out of the question.

- **Re-use and reusability:** We have seen that by adhering to the standards and principles of RDF and Linked Open Data, we can actually reuse vocabularies and datasets available on the Web of Data. The CEDAR case reuses existing datamodels (RDF Data Cube) and statistical concept schemes (SDMX definitions, HISCO for occupations) to reduce the effort of modeling the data. BiographyNet re-used parts of the Europeana Data Model. The CEDAR, DSS and BiographyNet cases show also how external data sources are used to enrich the data. At the same time, the produced linked data is also easily re-usable in new data integration efforts. One example is the linking of CEDAR and DSS data.
- **Extensibility:** Related to this is the observation that the flexible data models and low level of schema-level commitment in RDF allow for low-effort extension of the data. Rather than having to force new data into an existing rigid data model, one can simply model a new dataset in its own right and establish schema- and instance-level links between the new dataset and other datasets. In the DSS project, this was the case for two datasets, added at a later date. Alignment tools such as Amalgame (Van Ossenbruggen et al. 2011) allow datasets to be connected by establishing links between dataset-specific and shared vocabularies.
- **Data quality:** In all cases, data quality remains an important challenge. Linked Data does in no way guarantee quality of data. Errors in the data can occur because of faulty input data, but also can be introduced because of incorrect data conversion or linking of resources. Data provenance can help increasing trust in data by transparently presenting how specific information was derived.
- **New types of research questions:** In all three projects, we have seen that connecting different datasets generates new possibilities for analyses and new types of research. The links to common-sense background knowledge allows for automatic analyses which can considerably speed up the explorative phases of historical research. In discussions with the historical researchers, we found that they are enthusiastic about these new possibilities but will need some guidance on how these possibilities can be exploited. This calls for further research collaborations between computer and data scientists and historical researchers.

Acknowledgements

Part of the work on which this paper is based has been supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences. For further information, see http://ehumanities.nl. This work has been supported by the Dutch national program COMMIT. Part of this work was supported by CLARIN-NL through the project CLARIN-Dutch Ships and Sailors. This work was also supported by the BiographyNet project (Nr. 660.011.308), funded by the Netherlands eScience Center (http://esciencecenter.nl/).The authors would like to thank those that collaborated in the research projects described above, in particular the historical researchers.

Bibliography

R. Agerri, J. Bermudez, G. Rigau: Ixa pipeline: "Efficient and ready to use multilingual NLP tools". In: Proceedings of the 9th LREC 2014

C. Bizer, T. Heath, & T. Berners-Lee, . "Linked data-the story so far".Semantic Services, Interoperability and Web Applications: Emerging Concepts, 205-227, 2009

O. Boonstra, L. Breure, P. Doorn. "Past, present and future of historical information science". NIWI-KNAW, Amsterdam, 1st edition, 2004.

D. Cohen et al. Interchange: "The promise of digital history." Special issue, Journal of American History, 95, no.2, 2008.

R. Cyganiak, D. Reynolds. "The RDF Data Cube Vocabulary". World Wide Web Consortium W3C Recommendation, 16 January 2014. https://www.w3.org/TR/vocab-data-cube/

R. Cyganiak, D. Wood, M. Lanthaler. "RDF 1.1 Concepts and Abstract Syntax". World Wide Web Consortium W3C Recommendation, 25 February 2014.

V. de Boer, J. Leinenga, M. van Rossum and R. Hoekstra. "Dutch Ships and Sailors Linked Data Cloud". In Proceedings of the International Semantic Web Conference (ISWC 2014), 19-23 October, Riva del Garda, Italy, 2014 (p. 229-244)

V. de Boer, J. Wielemaker, J. van Gent, M. Hildebrand, A. Isaac, J. R. van Ossenbruggen, and G. Schreiber. "Supporting Linked Data Production For Cultural Heritage Institutes: The Amsterdam Museum Case Study". Proceedings of the 9th international conference on The

Semantic Web: research and applications (ESWC'12), Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti (Eds.). Springer-Verlag, Berlin, Heidelberg, 733-747

V. de Boer, M. van Rossum, J. Leinenga and R. Hoekstra. “The Dutch Ships and Sailors Project”. DHCommons Journal volume 1, july 2015 .

M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, H. van de Sompel. “The Europeana Data Model (EDM)”. In: World Library and Information Congress: 76th IFLA general conference and assembly, pp. 10–15, 2010.

J.A. Entjes “Linking Maritime Datasets to Dutch Ships and Sailors Cloud - Case studies on Archangelvaart and Elbing”. MSc. Thesis Information Sciences. Vrije Universiteit Amsterdam, 2015

A. Esteve and M. Sobek. “Challenges and Methods of International Census Harmonization”. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 36(2):37–41, 2003.

A. Fokkens, M. van Erp, P. Vossen, S. Tonelli, W.R. van Hage, L. Serafini, R. Sprugnoli, J. Hoeksema: Gaf: “A grounded annotation framework for events”. In: NAACL HLT 2013

A. Fokkens,S. ter Braake, N. Ockeloen,P. Vossen, S. Legˆene, G. ,Schreiber.: “Biographynet: Methodological issues when NLP supports historical research”. In: Proceedings of the 9th LREC. pp. 3728–3735, 2014

D. Garijo and Y. Gil. "Augmenting prov with plans in p-plan: scientific processes as linked data." CEUR Workshop Proceedings, 2012.

P. Groth, Y. Gil, J. Cheney, S. Miles: “Requirements for provenance on the web”. International Journal of Digital Curation 7(1), 2012

E. Hyvönen, J.Tuominen, E. Mäkelä, J. Dutruit, K. Apajalahti, E. Heino, P. Leskinen, and E. Ikkala. “Second World War on the Semantic Web: The WarSampo Project and Semantic Portal”. In Proceedings of 14th International Semantic Web Conference, 2015

C. Lagoze, H. Van de Sompel, M.L. Nelson, S. Warner, R. Sanderson, P. Johnston. “Object re-use & exchange: A resource-centric approach”. arXiv preprint arXiv:0804.2273 (2008).

C. Lagoze, H. van de Sompel. “Open archives initiative object re-use & exchange”. Online presentation, 2007. http://www.openarchives.org/ore/documents/ore-jcdl2007.pdf

F. Manola, E. Miller. “RDF Primer”. World Wide Web Consortium W3C Recommendation, 10 February 2004. https://www.w3.org/TR/2004/REC-rdf-primer-20040210/

A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, F. van Harmelen. “Semantic Technologies for Historical Research: A Survey”. Semantic Web — Interoperability, Usability, Applicability,

A. Meroño-Peñuela, A. Ashkpour, C. Guéret, S. Schlobach. “CEDAR: The Dutch Historical Censuses as Linked Open Data”. Semantic Web — Interoperability, Usability, Applicability (in press). IOS Press, 2016.

A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, and R. Hoekstra. “Linked humanities data: The next frontier? a case-study in historical census data”. Proceedings of the 2nd International Workshop on Linked Science 2012, 951, 2012.

L. Moreau, P. Missier, K. Belhajjame, R. B’Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, C. Tilmes: PROV-DM: The PROV Data Model. Tech. rep., W3C (2012).

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

W.R. van Hage, V. Malaise, R. Segers, L. Hollink, G. Schreiber: “Design and use of the simple event model (SEM)”. Journal of Web Semantics, 2011

J. van Ossenbruggen, M. Hildebrand and V. de Boer. "Interactive vocabulary alignment." Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, 2011. 296-307.

A.J.M. van Velzen and F.S. Gaastra. “Thematische collectie: VOC opvarenden; voc sea voyagers”. urn:nbn:nl:ui:13-v73-sq8, 2010

R. Warren. “Creating specialized ontologies using Wikipedia: The Muninn Experience”. Proceedings of Wikipedia Academy: Research and Free Knowledge.(WPAC2012), Berlin, Germany, 2012.