Semantic Technologies for Historical Research: A Survey

Editor(s): Aldo Gangemi, LIPN Université Paris13-CNRS-Sorbonne Cité, France and ISTC-CNR Roma, Italy
Solicited review(s): Monica Wachowicz, University of New Brunswick, Canada; Jack Owens, Idaho State University, USA; Aldo Gangemi, LIPN Université Paris13-CNRS-Sorbonne Cité, France and ISTC-CNR Roma, Italy

Albert Meroño-Peñuela [a,b], Ashkan Ashkpour [c], Marieke van Erp [d], Kees Mandemakers [c], Leen Breure [e], Andrea Scharnhorst [b], Stefan Schlobach [a], and Frank van Harmelen [a]

[a] *Department of Computer Science, VU University Amsterdam, De Boelelaan 1081a, 1081HV Amsterdam, NL*
E-mail: {albert.merono, k.s.schlobach, frank.van.harmelen}@vu.nl
[b] *Data Archiving and Networked Services, Anna van Saksenlaan 10, 2593HT Den Haag, NL*
E-mail: {albert.merono, andrea.scharnhorst}@dans.knaw.nl
[c] *International Institute of Social History, Cruquiusweg 31, 1019AT Amsterdam, NL*
E-mail: {ashkan.ashkpour, kma}@iisg.nl
[d] *Faculty of Arts, VU University Amsterdam, De Boelelaan 1081a, 1081HV Amsterdam, NL*
E-mail: marieke.van.erp@vu.nl
[e] *Universiteit Utrecht, Princetonplein 5, De Uithof, 3584 CC Utrecht, NL*
E-mail: l.breure@uu.nl

Abstract. During the nineties of the last century, historians and computer scientists created together a research agenda around the *life cycle of historical information*. It comprised the tasks of creation, design, enrichment, editing, retrieval, analysis and presentation of historical information with help of information technology. They also identified a number of problems and challenges in this field, some of them closely related to semantics and meaning. In this survey paper we study the joint work of historians and computer scientists in the use of Semantic Web methods and technologies in historical research. We analyse to what extent these contributions help in solving the open problems in the agenda of historians, and we describe open challenges and possible lines of research pushing further a still young, but promising, historical Semantic Web.

Keywords: Semantic Web, Historical research, Digital Humanities

1. Introduction

Historians have a long tradition in using computers for their research [16]. The field of historical research is currently undergoing major changes in its methodology, largely due to the advent and availability of high-quality digital data sources. More recently, the Web has shaken the paradigm of research data publication, particularly since the inception of the Semantic Web [13] and the Linked Data principles [43]. This paper looks forward on how Semantic Web technology has been applied to historical data, and how these technologies can facilitate, boost and improve research by historians. This survey revisits the open problems in historical data and historical research, and analyses current contributions, namely papers, projects, online resources and tools, that apply semantic technologies to solve such problems. We study how successful these solutions have been and propose some challenges for the future.

Historical research is an interesting domain for the Semantic Web. Historical data are extremely context dependent, and always open to a variety of possible interpretations. Availability on the Web of historical research data, which concerns the study and understanding of our past, is growing. The Semantic Web is an evolution of the existing Web (based on the paradigm of the document) into a Semantic Web (based on the paradigm of structured data and meaning). It 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. This survey studies the crossroads of the Semantic Web and history as research domains.

We consider surveying the state of the art in Semantic Web and history a fundamental task for both fields. First, it is necessary as a knowledge organisation task, in order to articulate research and discern contributions. Second, it fosters development of semantic technology and history, both individually and as a unique field, and helps on building research agendas. Other attempts on gathering research efforts on Semantic Web and history exist, but most of them study specific history subfields [75,97,131] or analyse concrete task-oriented tools [37,76] and methodologies [43,50,51]. Moreover, none of them consist in surveys or literature reviews. To the best of our knowledge, this is the first survey reviewing contributions on history and the Semantic Web as generic fields of research.

The elaboration of the study in this paper is not free of obstacles. The first of them is the large amount of research contributions to survey, which had to be filtered to fit strictly the Semantic Web goals and the historical research goals in order to be feasible. By historical research we mean strictly research performed by historians, and talk about history as a research domain. Thus, we exclude other fields of the humanities in which historical research is also performed, such as art history or history of literature. Nevertheless, in the end the number of contributions amount to more than a hundred. Secondly, and even though the corpus of available literature is large, we also encountered difficulties on accessing some of the sources. To solve this, we combined the contributions with the knowledge of domain experts, conducting eight interviews with pioneers in this area. Third, structuring and articulating all this work is an arduous task that requires a lot of schemas, tables and discussions. Finally, the clash of the vocabularies used by two different research communities, usually pointing at similar issues, is problematic. To bridge different jargon we devote some space to cover existing classifications of historical data, especially discussing terms like *structure*, and we map historical data problems in terms of Semantic Web solutions.

The paper delivers four contributions. First, it describes a classification of historical data depending on several factors, merging existing distinctions by historians with structural approaches from computer science. Second, it articulates the research conducted in the emerging field of historical Semantic Web in terms of several *tasks*, and depicts the current landscape on advances in representing historical data with semantic technology. Third, we map the open problems of historical data with the solutions provided by the surveyed research. Finally, we show some open challenges for the future, considering first the not (or only partially) solved problems, and secondly Semantic Web facilities still to be explored in historical research.

While we concentrate on historical research, similar solutions emerge also in other humanities fields at the turn to e-humanities or Digital Humanities [15,95]. As historical research overlaps with literary studies, ancient language studies, archaeology, art history and other humanities fields, these areas of encounter are also predestined candidates for the travel of generic methods developed from a semantic technology perspective for historical research to other humanities fields [64].

The survey is organized as follows. In Section 2 we introduce some background on historical research and the Semantic Web. In Section 3 we study the ecosystem of historical data. We describe the life cycle of historical information, propose a classification for historical sources, and show open problems of historical data. In Section 4 we articulate contributions that apply semantic technologies to historical research. In Section 5 we answer the question on how the contributions presented in Section 4 solve the open problems we describe in Section 3. In Section 6 we show the challenges that are still left to solve. Finally, in Section 7 we discuss our findings and establish some conclusions.

2. Background

2.1. Historical research

The field of historical research concerns the study and the understanding of the past. The field is currently undergoing major changes in its methodology, largely due to the advent of computers and the Web [16].

Computer science has inspired historians from the start. *History and computing* or *Humanities computing* were labels used before the inception of the Web [63].

Many pioneers in computer aided historical analysis have a background both in history and in informatics, and reflected early on about the usefulness of computational and digital techniques for historical research [16]. Ever since the advent of computing, historians have been using it in their research or teachings in one way or the other. The first revolution in the 1960s allowed researchers to harness the potential of computational techniques in order to analyze more data than had ever been possible before, enabling verification and comparisons of their research data but also giving more precision to their findings [3]. However this was a marginal group within the historical research: in general, the usage of computers by humanists could be described as occasional [34]. The emphasis was more on providing historians with the tools to do what they have always done, but now in a more effective and efficient way. Concretely,

- *databases and document management systems* facilitated the transition from historical documents to historical knowledge through text analysis;
- *statistical methods* were used predominantly for testing hypotheses, although with time were more valued as a descriptive or exploratory tool than as an inductive method;
- *image management* aided historians to digitize, enrich, retrieve images and visualize data [16].

Although computing tools are currently embedded in the daily life of most researchers, the use of these tools did not revolutionized all sciences equally. Accordingly, history failed to acknowledge many of the tools computing had come up with [16]. Instead of improving the quality of the work of historians and assisting them in their processes, software developed for historians often requires attending several summer schools [17]. Currently there are still many challenges and information problems in historical research. These difficulties mainly range from textual, linkage, structuring, interpretation, to visualization problems [16].

Despite these challenges, computing in history and in the broader sense the humanities, also brought some significant contributions in certain fields like linguistics (corpus annotations, text mining, historical thesauri etc..), archaeology (impossible without geographic information systems (GIS) nowadays), and other fields using sources that have been digitized for historical (comparative) research and converted to databases [16]. The use of electronic tools and media is incredibly valuable and important for opening up various sources for research which would otherwise remain unused. Open access to research data has always been an issue, especially in the humanities. However, over the past years various efforts have been made in opening up these black boxes and making them available for researchers. These different sources contain rich information from various fields, which are often digital in nature in the form of databases, text corpora or images. These sources, in practice isolated databases, often contain a lot of semantics, but their data models were asynchronously designed, making them difficult to compare. So, while more and more sources are being digitized, more attention has to be given to the development of computational methods to process and analyze all these different types of information [41].

A key issue for historians and other humanities researchers when dealing with historical data for comparative research concerns the lack of consistency and comparability across time and space, due to changing meanings, various interpretations of the same historical situations or processes, changing classifications, etc.

Though not all research dreams materialized in the way initially envisioned [56], the inception of the Web allowed historians to aim for world-wide, large scale collaborations, especially in the area of economic and social history. This kind of web based cooperation allows to collect, distribute, annotate and analyze historical information all around the globe [29].

Changes in historical research are closely connected to the emergence of new scientific methods, and this co-evolution holds for decades and centuries. Statistics has influenced many fields including history, and paved the ground for quantitative studies [57]. However, these kind of historical studies became more and more the domain of sociologists, economists and demographers than scientists educated as historians [91]. Late important changes are consequences of recent technological trends connected to the emergence of the Web [71] and the inception of Semantic Web technologies [5].

2.2. *The Semantic Web*

The advent of the Semantic Web poses new perspectives, challenges and research opportunities for historical research. Envisioned in 2001 by Berners-Lee, Hendler and Lassila [13], the Semantic Web was conceived as an evolution of the existing Web (based on the paradigm of the document) into a Semantic Web (based on the paradigm of structured data and meaning). By that time, most of the contents of the Web were designed for humans to read, but not for computer programs to process meaningfully. Although computer programs could parse the source code of Web pages to extract layout information and text, computers had no mechanism to process the semantics. In other words, 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* [13].

More practically, the Semantic Web can be defined as the collaborative movement and the set of standards that pursue the realization of this vision. The World Wide Web Consortium [14] (W3C) is the leading international standards body, and the Resource Description Framework [135] (RDF) is the basic layer in which the Semantic Web is built on. RDF is a set of W3C specifications designed as a metadata data model. It is used as a conceptual description method: entities of the world are represented with nodes (e.g. *Dante Alighieri* or *The Divine Comedy*), while the relationships between these nodes are represented through edges that connect them (e.g. Dante Alighieri *wrote* The Divine Comedy). These statements about nodes and edges are expressed as *triples*. A triple consists of a subject, a predicate, and an object, and describes a fact in a very similar way as natural language sentences do (e.g. subject: *Dante Alighieri*; predicate: *wrote*; object: *The Divine Comedy*). Subjects and predicates must be URIs (Uniform Resource Identifiers, the strings of characters used to identify and name a web resource like a web page), while objects can be either URIs or literals (like integer numbers or strings) [43]. RDF can be considered a knowledge representation paradigm where facts and the vocabularies used to describe them have the form of a graph. This setting makes RDF very suitable for data publishing and querying on the Web, especially when (a) the dataset does not follow a static schema; and (b) there is an interest of linking the dataset to other datasets.

Efforts on standardization have produced ontologies and vocabularies to describe multiple domains. An ontology is an *explicit specification of a conceptualization* [40] and contains the classes, properties and individuals that characterize a given domain, such as history. In the Semantic Web, the design of ontologies is done using the Web Ontology Language [132] (OWL). OWL consists of several language variants built upon different modalities of Description Logics [9] (DL), a family of formal knowledge representation languages. Such languages allow reasoning, that is, to extract or deduce consequences and new knowledge from the original.

A large number of RDF datasets have been published and interlinked on the Web, using these ontologies and vocabularies and following the Linked Data principles [12]. In the middle of the document-Web and the data-Web, formats and vocabularies for rich structured document markup (such as RDFa [134] or schema.org [94]) are enabling software agents to crawl semantics from web pages, bridging the gap between the Web for humans and the Web for machines. These efforts have evolved the Web into a global data space [43] where data can be queried e.g. using the SPARQL query language (SPARQL Protocol and RDF Query Language) [136]. Although the transition from the document-Web to the database-Web exists in the form of these standards and technologies, the simple idea of the Semantic Web remains largely unrealized [98].

3. Historical data

Since the introduction of computers in the field, historical research has produced high-quality digital resources [16]. Historical datasets encompass texts, images, statistical tables and objects that contain information about events, people and processes throughout history. Converted or born-digital, historical datasets are now analyzed at big scale and published on the Web. Their temporal perspective makes them valuable resources and interesting objects of study.

In this section we describe the ecosystem where historical information lives. First we introduce the life cycle of historical information, which is the framework we use to study how historical data is created, enriched, edited, retrieved, analysed, and presented. Then we propose a classification of historical data depending on several factors. Finally, we revisit the traditional open problems of historical data. Some of these problems have found solutions in current Semantic Web developments we present in Section 4.

3.1. The life cycle

The main object of study in historical research is historical information, and the multiple ways to create, design, enrich, edit, retrieve, analyse and present historical information with help of information technology. It is important to distinguish historical information from raw data in historical sources. These data are selected, edited, described, reorganized and published in some form, before they become part of the historian's body of scientific knowledge. We use the life cycle of historical information proposed by Boonstra et al. [16] to study the workflow of historical information in historical research.

Historical objects go through distinct phases in historical research. In each phase, these objects are transformed in order to produce an outcome meeting specific historical requirements. The phases can be laid out as the workflow of a *historical information life cycle* (see Figure 1). The phases, although sequentially presented, do not always have to be passed through in rigorous order; some can be skipped if necessary. The phases are also quite comparable with the practice in other fields of science.

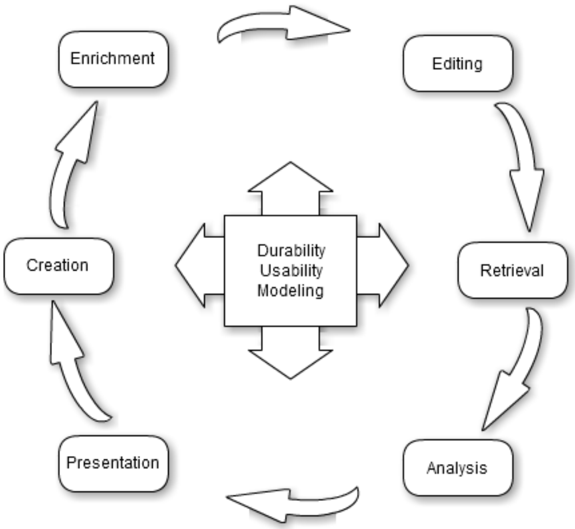

Fig. 1. The life cycle of historical information (Boonstra et al. [16]). The phases in the life cycle are: (1) creation; (2) enrichment; (3) editing; (4) retrieval; (5) analysis; and (6) presentation.

The life cycle of historical information consists of six phases:

1. **Creation.** The first stage of the life cycle is the creation stage. The main aspect of this stage consists of the physical creation of digital data, including the design of the information structure and the research project. Examples of activities in this phase would be the data entry plan, digitisation of documents (through e.g. OCR), or considering the appropriate database software.
2. **Enrichment.** The main goal of this phase is to enrich the data created in the previous step with metadata, describing the historical information in more detail, preferably using standards such as Dublin Core [28], and intelligible to retrieval software. This phase also comprises the linkage of individual data that belongs together in the historical reality, because these data belong to the same person, place or event.
3. **Editing.** Editing includes the actual encoding of textual information, like inserting mark-up tags or entering data in the fields of database records, with the intention of changing or adding historical data of convenience. All data transformations through algorithmic processes prior to analysis also belong to this phase. Editing also extends to annotating original data with background information, bibliographical references and links to related passages.
4. **Retrieval.** In this phase information is retrieved, that is, selected, looked up, and used. The retrieval stage mainly involves selection mechanism look-ups such as SQL-queries for traditional databases or Xpath [137] and Xquery [138] for XML-encoded texts.
5. **Analysis.** Analyzing information means quite different things in historical research. It varies from qualitative comparison and assessment of query results, to advanced statistical analysis of data sets.
6. **Presentation.** Historical information is to be communicated in different circumstances through multiple forms of presentation. It may take very different shapes, varying from electronic text editions, online databases, virtual exhibitions to small-scale visualizations. It can happen frequently in other phases as well.

In the middle of the historical information life cycle, three aspects are identified which are central to history and computing, but also in the humanities in general:

- *Durability* ensures the long term deployment of the produced historical information.
- *Usability* refers to the ease of efficiency, effectiveness and user satisfaction.
- *Modeling* denotes to more general modeling of research processes and historical information systems.

3.2. A classification of historical data

The continuous usage of computing in different areas of historical research has produced digital historical data with different formats, perspectives and goals. To be used in the Semantic Web, these historical data have to be represented semantically, using the current standards (see Section 2.2). In this section we propose a classification of historical data in order to bridge the gap between the data representation tradition in historical research, and the standard modelling paradigms of the Semantic Web [5,43].

3.2.1. Primary and secondary sources

Historical sources can be characterized and divided in many ways. A basic distinction used by historians is between *primary* and *secondary* sources.

Primary sources are original materials created at the time under study [11]. They present information in its original form, neither interpreted, condensed nor evaluated by other writers, and describe original thinking and data [8]. Examples of primary sources are scientific journal articles reporting experimental research results, persons with direct knowledge of a situation, government documents, legal documents (e.g. the Constitution of Canada), original manuscripts, diaries (e.g. the Diary of Anne Frank) and creative work. Primary sources can be distinguished into *administrative* sources and *narrative* sources, like biographies or chronicles. Administrative sources contain records of some administration (census, birth, marriage and death rolls, administrative accounts of taxes and expenses, resolutions minutes of administrative bodies, deeds, contracts, etc.). Typically, historians want to extract the facts in order to gather statistical data. Narrative sources are full text documents containing a description of the past, made by an author being an eyewitness: think of diaries, chronicles, newspaper articles, diplomatic reports, political pamphlets, etc. Historians may be interested in both, factual information and the author's vision and the bias.

Secondary sources are materials that have been written by historians or their predecessors about the past [127]. They describe, interpret, analyze and evaluate the primary sources. Usually, secondary sources gather modified, selected, or rearranged information of primary sources for a specific purpose or audience [8]. Examples of secondary sources are bibliographies, encyclopedias, review articles and literature reviews, or works of criticism and interpretation.

Since historical data have not been produced under the controlled conditions of an experiment, historical research always has something of the work of a detective, and certain details (read: annoying inconsistencies) cannot be destroyed or manipulated. These details may contain relevant information. On the other hand, to be able to extract statistical information and come up with more general statements, some formalization, relating information and harmonizing expressions of what is later used as variables is needed. Harmonization, the process of making data-sources uniformly accessible without altering its original form, is closely related to issues of standardization and formalization [65].

3.2.2. Intended further processing

Some historians [16] propose to structure historical data depending on their required further machine processing. They distinguish between *textual data*, *quantitative data* and *visual data*. Textual data comprises the whole set of unstructured historical sources, such as letters, memoranda or biographies, all in a form of free text. Quantitative data can be seen as historical sources aiming at a quantitative analysis, like church registers, census tables and municipality micro-data. Finally, visual data gathers all kinds of historical evidence not encoded by text or numbers, such as photographs, video footage and sound records.

3.2.3. Source oriented vs. goal oriented

Researchers make the distinction between *source oriented* and *goal oriented* historical data [16]. When dealing with historical data it is important to decide in an early stage whether the data should be modeled according to a source or goal oriented approach. The source oriented approach aims to postpone enforcing any standards or classifications, resemble the underlying source data as close as possible (schema free representation) and hence allow room for multiple interpretations of the data. Another approach is the goal or model oriented approach. Historical data is often plagued with inconsistencies, changing structures and classifications, redundant or erroneous data and so forth. The goal oriented perspective therefore advocates the use of more sound data models to start with. This means restructuring the data according to certain views or goals which are mainly dependent on expert knowledge. Accordingly, this perspective commits to a certain data model in an early stage.

3.2.4. Level of structure

At the end of the creation phase (see Section 3.1) one may expect to have a historical dataset suitable for further processes. However, the nature of the steps to be taken thereafter may strongly depend on the way the resulting dataset is structured. Indeed, attaching Semantic Web technologies to these historical sources (e.g. to extract RDF triples from them, or enrich them semantically) is strongly dependent on their level of structure. We propose the historical data classification shown in Figure 2. We distinguish three levels of inner structure in historical datasets: *structured*, *semi-structured* and *unstructured*. Each level of structure can be divided into several *types of structure*.

Structured data. Structured data refers to sources that have a clearly defined data model. A data model is an abstract model that documents and organizes data for communication, and is used as a plan for developing applications. An example of a structured dataset would be census material published in rows and columns, or a database of historical events. Well known generic examples of such a structure are sources encoded as relational databases, XML files, spreadsheet workbooks or RDF datasets. It is easy to see that all these examples meet a certain abstract model for the data they represent (relational schemas, DTD constraints, tabular formats and RDF triple statements). Structured historical data are usually managed with *relational databases*, *graph/tree representations* and *tabular representations*. Relational databases are the most well-known way of committing to some schema for representing historical objects and their relationships. Because their structure, relational databases are ideal for goal or model-oriented representation of historical data [16] with some concrete conception of reality in mind.

Relational databases. Relational databases have their own languages (SQL) and systems (MySQL, Microsoft SQL Server, PostgreSQL, Microsoft Access, Oracle, etc.) to represent and store historical data. They all follow the relational model [26]. Some issues, especially when trying to integrate data from different databases and modelled with different conceptual schemas, appear often in historical datasets encoded this way.

Graph/tree representations. Relying on graph theory, graph databases offer mechanisms for storage and retrieval of data with less constrained consistency models than traditional relational databases. They provide variable performance and scalability, but high flexibility and complexity support. AllegroGraph, IBM DB2, OpenLink Virtuoso and OWLIM are typical examples. To exchange historical data in graph form, RDF (see Section 2.2) is used. Graph/tree data is found in historical samples that come in formats such as XML (trees), RDF (graphs) or JSON (JavaScript Object Notation). Although they are conceived for modeling data in very disparate models (a tree, a graph and nested dictionaries, respectively) and purposes (e.g. JSON is mainly used for data interchange between web applications and services), these formats also follow some assumptions to put structure on historical data.

Tabular representations. Some historical datasets are encoded in tabular form. Tables consist of an ordered set of rows and columns, the latter typically identified with a name. The intersection of a row and a column is a cell. Depending on the specific encoding (Comma-separated values (CSV), Microsoft Excel spreadsheets, etc.) tables can offer variable features. Tables are used to store all kinds of historical data, especially meso, macro and microdata about individuals, registries, or historical population censuses.

Semi-structured data. Semi-structured data appear more often as an intermediate representation between unstructured and structured historical data than as raw historical data. Typical technologies applied here are markup languages, such as XML, to denote special characteristics of historical texts in specific regions of the corpus. *Annotated corpora* are the most important example of semi-structured data. They usually consist of raw historical texts with annotations on well-defined text regions, usually implemented with a markup language, like XML.

Unstructured data. In case a data model such as we described for structured data does not exist, we talk about unstructured data. In unstructured data there is scarce or no structure at all. The typical example is unconstrained, raw corpora encoded in plain text files. Unstructured sources are the most common representation of historical data, typically transcriptions of historical texts. Objects with a high variety of historical nature can be included in this category: letters, books, memoranda, acts, etc.

The use of the terms *structured* and *unstructured* in computer science to describe datasets is very different from the use of those notions in history, where administrative sources are often labeled as *structured* and the textual secondary sources as *unstructured*. Also narrative sources have internal structures, which can be made explicit. From the 19th century onwards historians have made scholarly source editions, which contain structured and annotated information. Nowadays the printed source editions are replaced and supplemented by databases and XML-based digital editions. So, *structured* or *unstructured* are relative notions: administrative sources usually have an obvious structured layout, while narrative sources have a latent, at first sight *hidden* structure, which is made explicit as soon as they appear in a scholarly source edition. So,

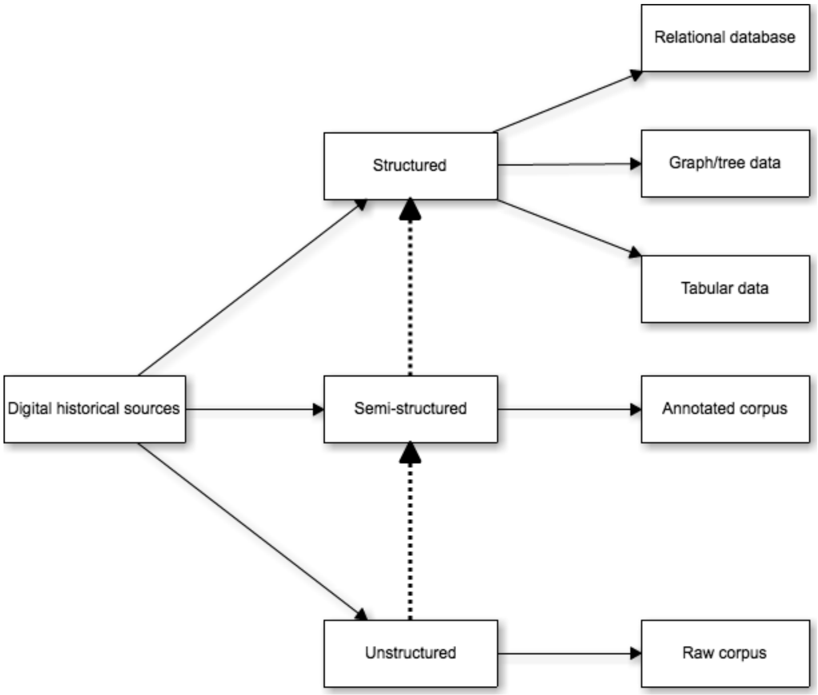

Fig. 2. Classification of historical data according to their level of structure. Dotted arrows indicate the direction of usual transformations in workflows that identify historical entities (and their relations), from unstructured to structured representations.

both administrative and narrative sources can appear in the form of *structured* or *unstructured* data in computer science jargon.

3.2.5. Discussion

Although structure really matters for deciding what specific computing technique or semantic model has to be applied to the sources, being those sources administrative or narrative, deliberate or inadvertent, does not really matter if their inner structure is clearly identified. Their belonging to one type of another may have an influence at some point, but in general the procedure to extract RDF triples from the sources strongly relies on the type of source we have regarding their structure. The goal is a faithful representation of the source in Semantic Web formats: a source-close representation allowing to model data as-is, meeting the same requirements of faithfulness than critical source editions (which is the standard for historians). It is critical for semantic representations to consider *context* and *source structure* as critical editions do, because they may be relevant for interpretation of the data. A digitized, semantically-enabled historical source should ideally preserve context and structure and support goal-oriented extraction of data, in order to construct historical facts in the framework of a certain research. By means of dataset interlinking and appropriate design and usage of ontologies and vocabularies, *context* and *source structure* should be able to be preserved using semantic technologies. To this end, ontologies can be contextualized to conciliate a party's subjective view of a domain [18].

3.3. Open problems

The classification proposed in Section 3.2.4 is not strict and admits hybrid examples. For instance, annotated digital text sources can be provided both as XML files or stored in a relational database (e.g. for statistical analysis). Some authors classify sources that combine primary and secondary sources like these as tertiary sources [72].

Although many advances have been made in different fields and computers are seen as valuable assets, a high percentage of historians are unfamiliar with or remain unconvinced that semantic technologies may become a new methodological asset [3,103]. The reason is that the weapon of choice of historians was and remains mostly the database, particularly in relational form [3]. This not only enabled historians to retain some of the integrity of the original data sources, but also paved way for rapid advances on issues such as classifications and record linkage. Therefore, historians typically do research using their *own* datasets, resulting in the creation of a vast amount of scattered data and specific technological challenges. In this section we revisit the traditional open problems of historical research derived from this tradition. In Section 5, after presenting the state of the art in the application of semantic technologies to historical research in Section 4, we point to the specific open problems described here that are eventually solved with such technologies.

Historical data problems can be divided into four main categories: information problems of historical sources, information problems of relationship between sources, information problems in historical analysis, and information problems of the presentation of sources [16].

3.3.1. Historical sources

The first set of open problems in historical research happens in phase 1 of the historical data life cycle (see Section 3.1). This is when the historical data are created.

Manually encoded or OCR-scanned, the creation of the dataset reveals the first barriers. Some characters, words or entire phrases in the original material may be lost or impossible to read or recognise by the human or the computer. Moreover, different techniques may extract historical entities differently. An example would be: what is the word that is written on this thirteenth-century manuscript?

The next question usually is: what does it mean? Background knowledge is provided by libraries in the offline world. But the computer aiding tools also need to have means to help the historian, using the Web as channel and semantics as meaning.

Related to background knowledge is the provenance of the data. Even if the source is clearly identified and its meaning deciphered, the historian needs to know more. To which issue does it relate? Why was it put there? Why was the text written? Who was the author? Who was supposed to read the manuscript? Why has it survived?

Another main issue relates to the structuring problem of historical data [87]. How can historical objects be encoded in a database? Researchers have to decide on what is an adequate data model for their datasets. As historians often have no clear research question when starting an investigation, it is neither possible nor desirable to model the data according to certain requirements in advance. Moreover, different sources have been produced throughout different periods in history with different views and motives. Historical census data is a good example, having varying structures and changing levels of detail which hinders comparative social history research both in past and present efforts [87].

The main discussion regarding this involves whether to use a source or a goal oriented data model for historical data (see Section 3.2.3). Researchers in favor of the source oriented approach claim that a commitment to a certain data model suitable for analysis should be postponed to the final stages of a project, in order to maintain flexibility and build on the data in a non destructive manner. This is especially the case when the database is supposed to be shared with other researchers or used in the future [62].

3.3.2. Relationships between sources

As historical researchers deal with various isolated sources, they face the problem of how to integrate these dissimilar sources for their purposes. This typically happens in phase 2 (enrichment) of the life cycle of historical information (see Section 3.1). An example would be: is this Lars Erikson, from this register, the same man as the Lars Eriksson from this other register?

Quite often several sources are used in historical research, which makes linking different sources another key problem. Micro data of the same person contained in different censuses, parish registers, marriage or death certificates are a good example. Obvious linkage problems are how to disambiguate between persons with the same name, how to manage changing names (e.g. in case of marriage of a woman) and how to standardize spelling variations in the names. In databases, several issues affect data comparability. *Schema mismatch* occurs when two different databases cannot be compared because of semantic differences in the concepts of their defining schemas. For instance, two XML files conformant to different DTD schemas may define and structure differently the same historical entity. Additionally, *value mismatch* occurs when the allowed values for columns or variables in two databases are different. It may also happen across datasets despite being schema or vocabulary-compatible. For instance, an attribute may encode the variable *social class* with categories A, B, C, while other dataset may do so with categories *high, medium, low*.

Other problems relate to how to link historical data with their spatial and temporal context. For example, some historical facts may need to be linked with occupational titles that evolve over time [45] or with countries with changing geographical boundaries (compare for example the contemporary geographic position of countries in Europe with the situation in 1930 and in 1900; or the fact that the city of Rotterdam suffered nine major changes in its composition between 1886 and 1941 [1]). As historical research often deals with changes in time and space, historians require tools which enable them to deal with these aspects. Accordingly several techniques have been developed for historical research, but the applicability of these has yet to be determined [16].

3.3.3. Historical analysis

Historical analysis is a fundamental part of the life cycle (see phase 5 in Section 3.1). It usually implies data transformations that aid historians in guiding their research. It also builds the bridge between their hypotheses and historical evidence.

The first issue in analysis is the massive treatment of historical data processed in previous stages to satisfy historical requirements, or to support a specific historical interpretation. An example would be: from this huge amount of digital records, is it possible to discern patterns that add to our knowledge of history? Various statistical techniques are borrowed from the social sciences to this end, like multilevel regression, and other techniques have been specifically developed for historical research, such as *event history analysis*. However, addressing historical data analysis in a broad sense remains essentially unsolved.

In historical research the meaning of data cannot exist without interpretations [16]. Due to drifting concepts in history, different interpretations could exist with regards to certain data [140]. However as interpretation of data is a subjective matter, this information should be added in a non destructive way, preserving the original source data.

3.3.4. Presentation

Presentation is the final phase of the historical information life cycle (see Section 3.1). Its goal is to use visualizations to aid the study and comprehension of historical data. An example problem of such phase would be: how do you put time-varying historical information on a historical map?

Presentation of historical data must be adequate. Different types of presentations are suitable at different stages of a research project. Presentation may take different shapes, varying from digitized documents, poorly and well modelled databases, or visualizations and representations on Geographic Information Systems (GIS). Currently there is a great need for tools and methods to present changes over time and space.

4. Findings

In this section we review the current state of the art in the application of semantic technologies to historical research, describing relevant contributions towards a historical Semantic Web.

The contributions are classified in the categories of *scientific papers*, *research projects*, *online resources* (presentations, online articles), and *tools, ontologies and lexical resources* (ontologies, demos, applications or programming libraries). Tables 1, 2, 3 and 4 show these contributions. Additionally, we map each contribution to one or more specific *tasks*. These tasks are shared areas of concern for both historical research and the Semantic Web. We identify the following four tasks: *knowledge modelling*, *text processing and mining*, *search and retrieval*, and *semantic interoperability*. We develop the work on each task in the following sections. We group contributions that cover the full spectrum of tasks under a final section in *longitudinal approaches*.

4.1. Knowledge modelling

Under this category we study research that has been conducted to model historical knowledge or historical facts using standard Semantic Web representations (see Section 2.2). We group contributions to a semantically enabled historical web by the following emphasis of research: historical ontologies, and linking historical data.

4.1.1. Historical ontologies

Data models are necessary for giving structure to any historical data, since they are the abstract models that document and organise data properly for communication. Ontologies encode such models in the Semantic Web [13] (see Section 2.2), and attention has been given to the need of historical ontologies [51]. In historical research, ontologies are the providers of metadata and background knowledge in phases 2 (enrichment) and 3 (editing) of the historical information life cycle (see Section 3.1). Semantic Wikis [88,101] are a great resource for historians to collaboratively build such ontologies.

We find a first category of such models in the form of (typically XML-encoded) taxonomies for histori-

Paper title	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
Hacking History via Event Extraction [97]	○	✓	✓	
Exploiting Semantic Web Technologies for Intelligent Access to Historical Documents [50]	○	✓	✓	
Historical Ontologies [51]	✓			○
Virtual Knowledge in Family History. Visionary Technologies, Dreams and Research Agendas [56]				✓
Past, present and future of historical information science [16]	○	○	○	✓
Proposed category system for 1960-2000 Census occupations [68]	○			✓
The Comparability of Occupations and the Generation of Income Scores [102]	○			✓
Challenges and Methods of International Census Harmonization [31]	○			✓
Making Sense of Census Responses: coding complex variables in the 1920 PUMS [39]	○			✓
Semantic Networks and Historical Knowledge Management: Introducing New Methods of Computer-based Research [54]	✓	○		
Queries in Context: Access to Digitized Historic Documents in a Collaboratory of the Humanities [126]		○	✓	
Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base [141]	○	✓	✓	
Historical documents as monuments and as sources [27]		○	✓	
Digital Hermeneutics: Agora and the Online Understanding of Cultural Heritage [129]	○	✓	✓	
Visualizing an Historical Semantic Web with Heml [89]	✓		✓	
Exploring Historical RDF with Heml [90]	✓		✓	
LODifier: Generating Linked Data from Unstructured Text [7]	○	✓		
CLIO - A Databank Oriented System for Historians [125]	✓	○	✓	○
CensSys - A system for analyzing census-type data [73]	○		○	✓
A discursive analysis of itineraries in an historical and regional corpus of travels: syntax, semantics, and pragmatics in a unified type theoretical framework [69]	○	✓	○	
A Comparison of Knowledge Extraction Tools for the Semantic Web [37]	✓	✓		

Table 1

Reviewed papers. The ✓and ○ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

Project name	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
Agora [2]	○	✓	✓	
BRIDGE [19]	✓		✓	○
CHoral - access to oral history [24]	✓		✓	○
Historical Timeline Mining and Extraction (HiTiME) [46]	○	✓	✓	○
LINKing System for historical family reconstruction (LINKS) [59]	✓			✓
SCRipt Analysis Tools for the Cultural Heritage (SCRATCH) [96]	○	✓	✓	
FDR Pearl Harbor Project [85]	✓	✓	✓	✓
North Atlantic Population Project (NAPP) [70]	○		○	✓
Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic (CKCC) [25]		✓	○	○
Voyage of the Slave Ship Sally [93]	○	✓	○	
Multilingual Access to Large Spoken Archives (NSF-ITR/MALACH) [61]	○	○	✓	✓
H-BOT [42]	○	✓	✓	
Clergy of the Church of England Database (CCEd) [21]	✓		✓	○
Armadillo: Historical Data Mining [6]	✓	✓	✓	✓
Historical Event Markup and Linking (HEML) [44]	✓		✓	
SAILS [92]	✓		✓	✓
CLARIN-Verrijkt Koninkrijk [131]	✓	✓	✓	✓
Historical International Standard Classification of Occupations (HISCO) [45]	✓			✓
Historical Sample of the Netherlands (HSN) [48]	○		✓	✓
CEDAR [22]	✓	○	✓	✓
Linking History in Place [58]	○		✓	✓

Table 2

Reviewed projects. The ✓and ○ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

cal research. A taxonomy is a collection of controlled vocabulary terms organized into a hierarchical structure, in general with less expressivity than an ontology. The first important example of such knowledge organization is the CLIO system, a databank oriented system for historians [125] appeared in 1980. CLIO included a tag/content representation for historical data that could be structured in complex hierarchies, supporting the recoding of material with doubtful semantics. CLIO remained as *the* system for organizing historical knowledge until the inception of the Web.

More recently, the Semantic Web for Family History [82] exposes a set of genealogy markup languages based on XML to semantically tag genealogical information on sources containing that kind of historical data. In the context of the Text Encoding Initiative [119] (TEI) there is an important discussion about building the bridge between XML (taxonomies) and OWL (ontologies) in historical data. SIG: Ontologies [78] contains a full log on contributions on how to use ontologies with TEI formats; namely, how TEI-XML encoded documents can refer to historical concepts and properties that have been previously formalized in an external OWL ontology.

The Historical Event Markup and Linking Project [44,90] (HEML) was probably the first project with the goal of creating a Semantic Web of history. Started in 2001, it explored the use of W3C markup technologies to encode and visualize historical events on the Web. Although in the beginning XML was the selected lan-

Resource name	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
Semantic Web approaches in Digital History: an Introduction [75]	✓			✓
Fawcett: A Toolkit to Begin an Historical Semantic Web [76]	✓	✓	✓	✓
Spatial cyberinfrastructures, ontologies, and the humanities [77]	✓	✓	✓	✓
SIG Ontologies [78]	✓			✓
CultureSampo - Finnish Culture on the Semantic Web 2.0: Thematic Perspectives for the End-user [79]	✓	✓	✓	✓
Text Mining for Historical Documents: Topics and Papers [80]	○	✓	○	
RDF vocabularies for historic placenames and relations between them [81]	✓			✓
The Semantic Web for Family History [82]	✓		✓	✓
Data portal for Social Sciences. Open data with SPARQL endpoint [83]	✓		✓	✓

Table 3

Online resources. The ✓and ○ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

guage to provide tagging and markup for describing historical events, the project later experimented with RDF to model and visualize them [89]. This transition was also happening in the whole historical ontologies community, as researchers better understood RDF and its differences with XML.

The modelling and representation of events, often defined as *persons* doing an *activity* in a certain *place* and *time*, has received a lot of attention in the development of historical ontologies, and most practical results show that the concept of the event is at the core of historical knowledge modelling. Van Hage et al. [130] design the Simple Event Model [110] (SEM), intended to model events in the domains of history, cultural heritage, multimedia and geography. Similarly, the Event Ontology [33], inspired in the musical domain, models the representation of events as combinations of persons, places, moments in time, and factors. Finally, LODE: An ontology for Linking Open Descriptions of Events [60] is especially intended for the publication of historical events as Linked Data. Interestingly, these ontologies have a great overlap in their conceptual modelling of events even coming from different domains. On the other hand, some studies point out specific modelling needs for different historical domains, stressing that historical ontologies should reflect how a particular time frame influences the definitions of concepts [51].

Another big focus in historical ontologies is given to geographical modelling. Owens et al. [84] describe a geographically-integrated history, and stress the importance of dynamics and semantics in Geographic Information Systems (GIS). They set an agenda for historical GIS systems that includes important semantic modelling tasks involving ontologies and geography for historical analysis. Moot et al. [69] depict the interesting crossroad between text analysis, historical semantics and geography in a work that structures geographical knowledge from a historical corpus of itineraries. Vocabularies for historical place names are under discussion [81]. Although not intended for historical research, the GeoNames ontology [38] is the reference for geographical modelling in the Semantic Web.

Since entities like places, persons or events change their over history and time, there is work raising the importance of a change-aware modelling in ontologies [35,64,66]. In historical research and the Semantic Web this is especially true for geographical names, places and regions [49], but also for demographical, social and economical indicators, such as occupations [45].

4.1.2. Linking historical data

By understanding the use and advantages of semantic technologies, practitioners and researchers of his-

Tool, ontology, lexical resource	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
NLP2RDF [105]	○	✓		
SIMILE/Timeline [106]	○		✓	
Gapminder [107]	✓		✓	✓
TokenX [108]	○	✓		
TAPoR [109]	✓	✓		
SEM event model [110]	✓			○
OpenCYC [111]	✓			✓
XCES [112]		✓		✓
Dublin Core [28]	✓			○
GATE [113]		✓		
WordNet [114]	○	✓		
Framenet [115]	✓	✓		
SUMO [116]	✓			○
MILO [117]	✓			○
AskSam [118]		✓		
TEI (Text Encoding Initiative) [119]		✓		✓
SGML [120]		✓		✓
TACT [121]		✓	○	
Wordcruncher [122]		✓		
Atlas.ti [123]		✓	○	
NLTK [124]		✓		
FRED [36]	✓	✓		
WAHSP and BILAND [139]	○	✓		
The Event Ontology [33]	✓			○
LODE [60]	✓			○
Semantic MediaWiki (SMW) [101]	✓		✓	✓
Europeana Data Model (EDM) [32]	✓			○

Table 4

Tools, ontologies, and lexical resources. The ✓and ○ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

torical data can not only connect their own data sources but moreover, also disseminate their data into the Semantic Web and integrate it with other data sources which were previously not possible or cumbersome. The approaches reviewed in this section match the historical data problem of the *relationships between sources* (see Section 3.3.2). In most cases, the use of semantic technologies solves it.

If one side of knowledge modelling stresses the importance of ontologies and formalization of the semantics of historical domains, the other side pursues the usage of such ontologies to interlink related historical data on the Web. Some researchers in history have centered their interest in how semantics can help relating and linking historical sources and entities: *historical, semantic networks are a computer-based method for working with historical data. Objects (e.g., people, places, events) can be entered into a database and connected to each other relationally. Both qualitative and quantitative research could profit from such an approach* [54]. Linking historical datasets appropriately is an old and very well known problem in historical research [16]. The landscape on current projects linking historical data (typically extracted from unstructured sources) shows a tendency on publishing more and more historical Linked Data in RDF.

There is a wide variety of project types looking for that structure, though not doing so solely (or explicitly) in RDF. For instance, the Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic [25] (CKCC project) studies the epistolary network for circulation of knowledge in Europe in the 17th century, extracting all entities and links from the correspondence of scientific scholars of that time. The LINKing System for historical family reconstruction (LINKS) project [59] reconstructs the links between individuals of historical families across several registries. The CCed [21] project follows a similar approach with clerical careers from the Church of England Database. While these projects mine the historical sources for important historical personalities and their relationships, other approaches, such the SAILS [92] project, dive into more concrete historical events and links various World War I naval registries together. The common goal in these initiatives is to produce a *semantic network of historical data containing objects like people, places and events connected to each other*, which clearly matches the intended purpose of historical ontologies (see Section 4.1.1), but also the general mission of the Semantic Web [13] and Linked Data [43].

Many other projects expose their domain specific historical datasets using RDF. These datasets facilitate their linkage to others using existing ontologies (see Section 4.1.1), achieving shared goals with the old task of historical record linkage. For instance, the Agora project [2] aims at formally describing museum collections and linking their objects with historical context using the SEM [110] (Simple Event Model). Historical events are found elsewhere in historical data. The FDR Pearl Harbor project links events, persons, dates, and correspondence found on government letters and memoranda on the surroundings of the Pearl Harbor attack on 1941 between the US and Japanese governments. All these entities are represented in RDF to model a graph of historical knowledge about that particular event. From a more socio-historical point of view, the Verrijkt Koninkrijk [131] project links RDF concepts found on a structured version of De Jong's studies on *pillarization* of the Dutch society after the World War II. More focused on media, the Poli Media project [86] mines the minutes of the general state debates to link historical entities to the archives of historical newspapers, radio bulletins and television programs. The goal is to create a unified historical search environment, facilitating a cross-media analysis [55].

Some general purpose tools facilitate the creation of historical Linked Data. The Fawcett toolkit [76] and the Armadillo project [6] are good examples. The latter exports RDF from any unstructured historical source, producing an RDF graph of historical knowledge that encodes the historical entities and their relationships expressed in that source. Other tools like Open Refine [74] or TabLinker [104] are tailored to produce such Linked Data from structured sources like tables (see Section 3.2).

4.2. Text processing and mining

In *text processing and mining* we revise work that deals with processing unstructured text. Textual resources play an important role in history research. We especially survey work on automatically extracting historical entities (such events or persons) via Natural Language Processing (NLP) techniques. The purpose of NLP is to enable computers to derive meaning from human, natural, or unstructured language input (see Section 3.2).

Structuring historical information from textual resources for further analysis is the bottom line of many research projects. The interesting differences come usually from the various source materials these projects mine. The general public-aimed Agora project [2] enriches museum collections with historical knowledge in order to help users place museum objects in their historical contexts. To this end, Agora employs information extraction techniques from statistical natural language processing to extract named entities (actors, locations, times, event names) from textual resources such as Wikipedia and collection catalogues which are used to populate SEM [110] (see Section 4.1.1) instances. From the object descriptions, also relevant historical entities are extracted which can be linked to the events. To formalize this workflow, Segers et al. [97] present a prototype extraction pipeline for extracting events and their properties from text using off-the-shelf natural language processing tools such as named entity recognition and pattern-based approaches. The main problem they encounter is that the notion of events is still ill-defined in NLP research, and as such tools are not yet readily available.

Textual encoding of the media have also been the source to extract historical knowledge in several projects. The Bridge project [19] aims at bringing more cohesion into Dutch television archives by finding relevant links between the official archives maintained at the Netherlands Institute for Sound and

Vision and other information sources such as program guides and broadcasting organizations websites. It is thus focused on improving access to television archives for media professionals. In order to do so, relevant entities are extracted from archives by using statistical NLP techniques. Furthermore, they will detect interesting events in television archives by detecting redundant stories, utilising the structure of the archive to identify links between different entities [20]. The Poli Media project [86] mines the text of minutes of the general state debates to extract and link historical entities from the archives of historical newspapers, radio bulletins and television programs.

The Historical Timeline Mining and Extraction (HiTiME) project [46] is aimed at detecting and structuring biographical events. To this end they analyze biographies of persons from the Dutch union history to create timelines that tell the life story of these persons, and social networks of the persons they interacted with. Van de Camp and Van den Bosch [128] describe an approach to build networks of historical persons by mining biographies for person names and relationships between persons. They use standard named entity recognition tools and utilise the inherent structure of biographies (the topic of the biography is a particular person, and any persons mentioned in this biography should have something to do with this person) to detect interpersonal relations.

Many ehumanities and ehistory projects are exploring document summary techniques or document enrichment techniques from NLP to aid search in their archives. One of these techniques is topic modelling, which can be used to add topic indicators to a document, which may help cluster search results or create more fine grained indexes of archive records. Wittek and Ravenek [142] explore the state of the art in topic modelling techniques to index 19,000 letters of correspondences between 16th and 17th century Dutch scientists.

Other high-level text analysis methods, such as frequency-based corpus analysis to compare e.g. work from different authors or investigation of other stylometry characteristics, are also popular in the ehumanities domain [30]. These methods are not domain-dependent and fit more easily into the ehumanities researcher search-based toolbox.

The spectrum of tools to extract knowledge from unstructured historical data is wide. Important contributions are essentially domain-independent [7], thus not particularly focused on historical text processing. Gangemi [37] presents a recent and complete comparison of generic knowledge extraction tools for the Semantic Web, which will aid historical researchers working in the phases 2 (enrichment) and 3 (editing) of the historical information life cycle (see Section 3.1).

4.3. Search and retrieval

In *search and retrieval* we include systems that exploit semantic formalisms as a new way of indexing, querying and accessing historical data, instead of relying on the traditional text-based or keyword-based algorithms. This task matches the phase 4 (retrieval) of the life cycle of historical information (see Section 3.1).

It is not a coincidence that a high number of contributions that aim at extraction of structured entities from historical data also point at some desired system able to improve search and retrieval of such entities. Indeed, by means of constructing a semantic graph of historical knowledge, search and retrieval of that knowledge, as well as indexing systems that give exact pointers to the source in which particular historical entities are mentioned, can be easily built and improved. The Agora [2] (museum collections), BRIDGE [19] (historical TV metadata), CHOoral [24] (historical audio metadata), Historical Timeline Mining and Extraction (HiTiME) [46] (biographical events), Verrijkt Koninkrijk [131] (Dutch post-war social clusters concepts) and FDR Pearl Harbor [85] (historical events around Pearl Harbor attack on 1941) projects are all good examples of this tendency. Once the knowledge is successfully extracted from the historical sources and formalized appropriately, entities structured this way can be used for a graph-based search and retrieval, for instance through SPARQL queries (see Section 2.2), although most systems use specific access methods [50]. Other projects, like the H-BOT [42] project, use a natural language interface instead of a query system for retrieval of such historical structured knowledge.

Indexing of historical contents is another way of improving search and retrieval of historical data. Indexing and historical data storage systems have a long tradition in historical research [16]. CLIO [125] is a traditional example of such a system, nowadays indexing is performed by XML annotation-oriented approaches, such as described by Robertson [90]. These initiatives should consider the emerging RDFa, microformats and microdata technologies (see Section 2.2) to study the ways they fit in the vast domain of historical text annotation systems.

4.4. Semantic interoperability

In this section we analyze to what extent contributions consider the problem of data integration and use the Semantic Web to deal with it. The specific problems encountered are data model mismatching, schema incompatibilities and disparate source formats. Semantic interoperability has much to do with data integration, namely, how to commonly query and uniformly represent data that come from multiple sources (i.e. fitting several, probably non-compatible data models).

Semantic heterogeneity of historical sources is especially present on social history projects. The North Atlantic population project [70] (publication of micro-data of several Atlantic countries) has this problem of data harmonization, in which heterogeneity of sources requires an intense work on resolving data model inconsistency between datasets.

The source material for the Historical Sample of the Netherlands [48] (HSN) database consists mainly of the certificates of birth, marriage and death, and of the population registers. From those sources the life courses of about 78.000 people born in the Netherlands during the period 1812-1922 have been reconstructed. Stored in a database and downloadable as files, this information forms a unique tool for research in Dutch history and in the fields of sociology and demography. As in the case of the HSN this type of sources is usually stored in archives, and, for the majority from a more remote past, not yet machine readable and not easy to analyse with NLP techniques. There is one major pitfall in linking this kind of data: extracting data about persons, events, institutions, locations is one thing, but linking to their different instantiations (for instance different name spellings, or persons with the same name) and keeping good documentation is the real challenge [62].

The CEDAR project [22], located in the crossroads of the Semantic Web, statistical analysis and social history, exposes the Dutch historical census data in the Semantic Web. Censuses are a great source of non-biased socio-historical information, but they present complex problems in both internal (i.e. between the time series) and external (i.e. other datasets) interlinking [67].

The work developed by Sieber et al. [100] provides a deep analysis of how semantic heterogeneity can be addressed exclusively with semantic technologies, and describes how to achieve success in environments with very disparate data models. In the history-related domain of geographic information systems (GIS), already discussed in Section 4.1.1, Manso and Wachowicz [4] provide an extensive review on current issues in interoperability.

4.4.1. Classification systems

Multiple publications in classification systems [31, 39,68,102] are especially aimed at solving interoperability problems in historical data. Classification systems provide a standard mechanism to compare such data, but their specific implementation and effectiveness depends on the orientation towards source or goals of the historical data (see Section 3.2.3) created in phase 1 of the historical data life cycle (see Section 3.1).

When dealing with vast amounts of historical data, classification systems are a necessity in order to organize and make sense of the data. The main goal of a classification system is therefore to put things into meaningful groups [10]. This entails an allocation of classes which are created according to certain relations or similarities. The main issue with historical classification systems is that they are not consistent over time, making comparative historical studies problematic. Historical census data is a typical example of this problem [22,65]. Census data is the only historical data on population characteristics which are not strongly distorted and yields an extremely valuable source of information for researchers [91].

However, major changes in the classification and coding of the different censuses, have hindered comparative historical research in both past and present efforts [87]. Researchers are forced to create their own classifications systems in order to answer their research question; however, this process often results in disparate systems, which are not comparable, contain a lot of expert knowledge, different interpretations of the data and could not be easily (re)used by other researchers. The fact that many of the modelling techniques are destructive in nature (we cannot go back to the source) makes it even more cumbersome to comprehend these sources. In order to deal with the changing classifications and vast differences at both national and international level, we need to connect the gaps between the datasets and conform to certain *standard* classification systems.

Currently several significant efforts have been made in this direction. The Integrated Public Use Micro-data Series (IPUMS) project [52] for example faces the problem of bridging 8 different occupational classification systems and a total of 3200 different categories, containing the richest source of quantitative information on the American population. The North Atlantic Population Project [70] (NAPP) project provides a machine-readable database of nine censuses from several countries. The main focus of the NAPP project is to harmonize these data sets and link individuals across different censuses for longitudinal and comparative analysis. Their linking strategy involves the use of variables which do not change over time. In this process records are only checked if there is an exact match for some variables, such as race and state of birth. Other variables like age and name variables are permitted to have some variations. Another significant historical classification system is the Historical International Standard Classification of Occupations [45] (HISCO). As occupations are one of the most problematic variables in historical research, HISCO aims to overcome the problem of changing occupational terminologies over time and space. It encodes historical occupations gathered from different historical sources coming from different time periods, countries and languages, and classifies tens of thousands of occupational titles, linking these to short descriptions and images.

4.5. Transversal approaches

Finally, there are few but key contributions we have classified as being *transversal*, because they cover a wide spectrum of the list of overlapping tasks between the Semantic Web and historical research. They also influence almost every phase in the historical information life cycle (see Section 3.1).

The CLIO system [125], a databank oriented system for historians, is the first of such contributions. CLIO was, for decades, *the* system for creating, enriching, organizing and retrieving historical knowledge from historical data in the pre-Web era. Although not using Semantic Web technologies (see Section 2.2), it had a strong emphasis on semantics as key for structuring historical knowledge.

In the Linked Data universe, the Agora project [2] is one of such transversal contributions. It generates historical RDF of events extracted using NLP techniques from unstructured texts, uses it for enhanced search and retrieval, improves semantic heterogeneity and gives context by linking to other datasets. Similarly, the Verrijkt Koninkrijk [131] and Multilingual Access to Large Spoken Archives (NSF-ITR/MALACH) [61] projects perform these tasks in their particular domains (see Section 4.1.2). The FDR Pearl Harbor project [85] also contributes on this line, but additionally opening the very promising field of historical knowledge inference through the formalization and usage of historical OWL ontologies. All these are good examples on how historical data get much richer when their semantics are explicitly expressed and they are interlinked through standard vocabularies and ontologies.

Regarding tools, the Armadillo architecture of Semantic Web Services [6] and the Fawcett toolkit [76] contain the generic plot behind all these contributions, and cover the whole pipeline of semantic historical data management. The latter extracts RDF event-oriented triples from unstructured texts, and additionally allows historians to install a full semantic toolbox with widgets to experiment with their data. Open Refine [74], in combination with its RDF-export plugin, allows the extraction, transformation, modelling and publishing of historical Linked Data when the sources come in tabular format.

Additionally, the theoretical study by Boonstra et al. [16] envisages possibilities on how the Semantic Web can enhance research by historians. It constitutes, besides, a major work on the evolution of historical computing, ehistory and historical information science, and gives a deep intuition on how computer science can help to solve ancient problems in historical research.

5. Solving historical problems

In this section we point to the open historical data problems revisited in Section 3.3 which are addressed or solved by the Semantic Web contributions reviewed in Section 4 as tasks. The mapping between the open problems and the tasks is shown in Table 5.

The first interesting result is that some of the problems identified in *historical sources* (Section 3.3.1) are mostly solved by the approaches we review in historical ontologies (Section 4.1.1). Concretely, our perception is that the structuring of historical data and the development of historical data models have been a success due to the creation of standard vocabularies and ontologies. These ontologies 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 large number of projects exposing historical Linked Data on the Web using these ontologies (see Section 4.1.2) prove their usefulness and success. There is space, though, for improvement. Although it is commonly agreed that cur-

Open historical data problems	Historical ontologies (4.1.1)	Linking historical data (4.1.2)	Text processing and mining (4.2)	Search and retrieval (4.3)	Classification systems (4.4.1)	Transversal approaches (4.5)
Historical sources (3.3.1)	✓	○				
Relationships between sources (3.3.2)		✓			✓	
Historical analysis (3.3.3)						○
Presentation of sources (3.3.4)						○

Table 5

Mapping between the open problems of historical data (see Section 3.3) and the surveyed contributions in historical Semantic Web (see Section 4). The sign ✓indicates that the problem is directly addressed in the Semantic Web task. The sign ○ indicates that the problem is indirectly or partially addressed in the Semantic Web task.

rent historical ontologies model the core semantics of historical research, authors also agree that they are still scarce and need further development [51,84].

As part of the problems in historical sources, provision of background historical knowledge has been successful only partially. The infrastructure (Linked Data cloud, SPARQL endpoints on historical data) is set up and running. But the amount of historical data available is still too low to give good support to any historian creating historical datasets in the beginning of the life cycle (see Section 3.1). Consequently, little background knowledge can help today these historians in solving e.g. errors or inconsistencies at that phase. Similarly, the generic infrastructure for provenance publishing and retrieval in the Semantic Web is very mature and extensively used in other domains [133], but scarce or non existing in the historical domain although being identified as a very important requirement (see Section 3.3.1). The provision of such provenance on historical datasets needs to be guaranteed in projects using semantic technologies to publish historical data.

Solutions to the problem of *relationships between sources* are probably the greatest achievement of the application of semantic technologies to historical research. The large number of projects linking historical data we survey in Section 4.1.2 proves that the Semantic Web delivers working solutions to the problem of connecting isolated historical data sources. The usage of developed ontologies and vocabularies has been key to this end. Additionally, the existence of classification systems (Section 4.4.1) helps on data comparability in the Semantic Web. Because we see that the body of historical knowledge in the Semantic Web is still small, we expect the problem of finding related links between historical entities and datasets to grow in the future, although the Semantic Web has generic solutions for this [99].

The problems in *historical analysis* and *presentation of sources* (see Sections 3.3.3 and 3.3.4) are only partially addressed in approaches we have classified as transversal. These works cover a wide spectrum of the life cycle of (semantic) historical data, including analysis and presentation (phases 5 and 6, Section 3.1). Consequently, they deal with some analyses and visualizations. However, there is a lack of contributions tackling directly the problem, or considering explicitly historical research requirements with respect to analysis and visualization. The transversal tools are hence very generic, and they could be inappropriate for some historians. Therefore, it is very important to distinguish what analysis requirements are specific to historical research, and which ones are domain-independent. Our hypothesis is that these problems overlap only partially with the goals of the Semantic Web (i.e. representing and linking meaning on the Web). However, historians could benefit from analysis and visualization tools for historical semantic data, not as specific as project-oriented, but not as generic as domain-independent.

In Table 5 all open problems have Semantic Web tasks providing solutions, but not all tasks are mapped to some historical open problem. Concretely, the tasks of *text processing and mining* (Section 4.2) and *search and retrieval* (Section 4.3) do not seem to solve any of the identified problems. Why do we find contributions on these areas? First, although not being identified by historians as primary problems, they constitute secondary problems that need to be solved when representing and linking semantic historical data. These problems are not exclusively historical, but they needed to be reimplemented in the Semantic Web realm (e.g. natural language processing for extracting historical RDF triples, SPARQL to query historical semantic data on the Web). Secondly, the goals these tasks aim at were quite well solved in historical research before the inception of semantic technologies (e.g. manual input of historical data, SQL queries in historical relational databases), and thus historians did not consider them into the primary problem space.

6. Open challenges

The use of semantic technologies has contributed significantly to solving the open problems of historical data (see Section 5). However, there is a lot of room for improvement. The open problems are being addressed as shown, but they are far from being solved until they get additional attention. The scarce amount of historical data on the Semantic Web is a good example. Other problems, some more specific, some more generic, could be also tackled with semantic solutions. In this section we explore some aspects of the Semantic Web that have not been used yet or could be furtherly exploited in historical research.

6.1. Semantics of time, change, language, uncertainty and interpretation

Classifications and ontologies in history do exist, but not for all areas, not in Semantic Web languages and not always agreed upon. Although several historical ontologies have been developed (see Section 4.1.1), these models are insufficient for the vast amount and variety of historical data that still has to be published in the Semantic Web, especially when key issues for historians like *interpretations* or *evidences* need to be modelled and conveniently linked. Historical ontologies and vocabularies have been a reality in recent approaches. Ontologies describing classes and properties of some historical concern, such as concepts around the Pearl Harbor attack in 1941 [50], are an exciting modelling exercise for researchers but also a necessary step for better structuring historical information in the Web. Ontologies and vocabularies offer a way of controlling the predicates, classes, properties and terms that the community uses as a standard for describing factual and terminological knowledge about history. Designing good ontologies for historical domains is also an area with plenty of challenges: how can ontologies comprise the many conceptions of history depending on the temporal dimension of events described [51]? Moreover, how can differences in meaning and relations between concepts be traced, as time and historical realities change these concepts [140]? To what extent these meaning differences relate to the complexity of the language (e.g. Latin, Middle languages) and uncertainty (e.g. fuzzy dates and locations)? These questions, which comprise semantic technologies, knowledge acquisition and knowledge modelling techniques, are not yet completely understood and are a significant challenge in semantic historical research. On the other side, over the centuries, dictionaries, thesaurus, classification systems have been developed. How to mount those specifically grown ordering principles to the Web in a way that makes them explorable and linkable to other ontologies is one interesting challenge which requires a close collaboration between historians, knowing and designing those specific tools, and computer scientists, often relying on much broader and generic ontologies.

6.2. Reasoning

From the point of view of Linked Data, ontologies and vocabularies are designed in order to control the terms in which datasets may express data, as well as the data model in which these data are represented. However, in a more Semantic Web perspective, one may expect these ontologies and vocabularies to facilitate new knowledge discovery; that is, to make explicit some implicit fact that was not trivial to deduce for the human eye, especially in big knowledge bases.

Reasoning is one of the key mechanisms of the Semantic Web still to be used in historical research. The absence of specific methods and tools for automatic historical inference, so that new, *implicit historical knowledge* can be derived, is another issue. We claim that reasoning could be fundamental for historical analysis 3.3.3 and tasks in the phase 5 (analysis) of the historical information life cycle (see Section 3.1).

Historical ontologies can be used to facilitate historical knowledge discovery using reasoners. Assuming that a particular domain is completely formalized as historical ontologies, then it is possible to run a reasoner on these ontologies to produce derived, implicit rules and facts that were not present in the original model as explicit knowledge (i.e. specifically encoded in the ontology), but that were there as underlying knowledge. For instance, if an ontology describes, on the one hand, the fact that a letter was sent from one government to another, and on the other hand, the fact that governments have a person responsible of sending and receiving letters, then it may be possible for the reasoner to infer what concrete persons sent and received what letters. As the knowledge base grows, implicit knowledge is not evident anymore and reasoners can facilitate an enormous work and produce high-value pieces of historical knowledge.

Since historians have different interpretations and no clear research question when starting an investigation, abductive reasoning (i.e. given the conclusions and a rule, try to select possible premises that support the conclusion) may be more convenient than deductive reasoning (i.e. deduce true conclusions given a premise and a rule) in historical research [23,47]. These would revert the order of some phases of the life cycle of historical information (see Section 3.1), generating a more bottom-up, data-based generation of hypotheses supporting evidence. The impact of abductive reasoning in historical research and its relationship with the life cycle needs further study and clarification.

The introduction of any kind of reasoning in the life cycle needs to be done with the goal of supporting, not replacing, the task of the historian, who must keep control of the implementation of the different phases.

6.3. Linking more historical data

We show in Section 4.1.2 that great efforts are being devoted to publish historical Linked Data. However, the amount of structured historical knowledge available on the Web is still insufficient to aid tasks that need high amounts and different kinds of historical background knowledge. While many different data and information sources exist, they are not always interlinked. This isolation of historical data sources hampers that they can be found, but it also inhibits how they can be further processed and connected.

One of the big claims of linked data is that, by linking datasets, relations established between nodes of these datasets highly enrich the information contained in them. That way, browsing datasets is not an isolated task anymore: by allowing users (and machines) to explore entities through their predicate links, data get new meanings, uncountable contexts and useful perspectives for historians.

For example, consider a scenario with three different SPARQL endpoints exposing RDF triples of a census with occupational data, a historical register of labour strikes, and a generic classification system for occupations (in the context of one particular country, for instance). Suppose that: the occupational census of the data exposes triples with countings on occupations (for example, how many men and women worked in a particular occupation in a concrete city), the historical register of labour strikes contains countings on how many people participated in labour strikes (number of women and men, per occupation and city), and the generic classification system harmonizes names of the occupations between both previous datasets (for example, gives a common number for representing occupation names that may vary between census occupations and labour strike occupations). Then, it is clear that several SPARQL queries can be constructed to give very meaningful and interesting linked data to the historian. For instance, such a query may return, given a city and an occupation code, which ratio of men and women followed a particular well-known labour strike. Another SPARQL query may return an ordered list of historical labour strikes by relevance, according to several indicators (strike successfulness ratio, total number of workers on strike, density of people on strike depending on the location, etc.). It is obvious that the possibilities increase if we think of more related historical sources to link, like datasets describing historical weather or historical geographical names and areas.

6.4. Flexibility of data models

It is considered to be a bad practice in historical research not doing the historical data modelling at phase 1 of the historical information life cycle (see Section 3.1). The choice of a particular data model to represent historical data is a critical issue for most historical computing projects. The election of some appropriate data model may seem a good design decision at some stage of the project. However, new requirements, research directions or stakeholder priorities may convert that data model into an obstacle more than an aid. Flexibility of data with respect to the data model used to represent historical domains is desired to avoid restructuring entire databases. Comparison in historical research requires flexibility of the models to be able to match them to one another. At the end, that enforces historians to make their data selection and processing dependent of a certain data model that can not be easily replaced or altered if needed. This happens usually in environments with changeable and creep requirements [53].

Applying semantic technologies and Linked Data principles to historical data may have a major advantage regarding historical data models, providing flexibility at the historical data modelling phase. Two different approaches regarding historical data modelling have been followed traditionally in historical computing: the *source-oriented* representation, and the *model-oriented* (also known as *goal-oriented*) representation (see Section 3.2.3). Do semantic technologies allow a flexible representation of historical data? Can the Semantic Web found a new standard on a source-and-goal, hybrid approach? RDF databases (see Section 2.2) can store the middleware representation of further views on the data [65]. These views can be modelled as close to any particular historical interpretation as needed. This way, the decision of what data model suits better the historical source can be postponed until the very end of the life cycle (see Section 3.1), or adopted as early as necessary.

Moreover, additional questions arise when considering the traditional perspectives on data modelling: the conceptual, logical and physical data models. These perspectives help in detaching data management technology, like relational databases or RDF triplestores, from conceptual schemas (i.e. the semantics of a domain). While conceptual data models are currently shared on the Web as e.g. historical ontologies (see Section 4.1.1), the flexibility of the whole modelling stack towards semantic changes needs to be better understood.

6.5. Non-destructive data transformations

The non-flexibility of data models (see Section 6.4) is related to the non-flexibility of historical data transformations. Historical data are modified in the life cycle of historical information (see Section 3.1). But if update, enrichment, analytic and interpretative operations are not controlled, these transformations lead to different historical data representations which can hardly be related to each other any more, nor in terms of provenance nor in terms of relatedness.

Another issue is supporting data transformations under two constraints: (a) without modifying source data (so the originals stay intact); and (b) with tracking of changes. Consequently, destructive updates are a major concern when selecting, aggregating and modifying historical data. On the one hand, modifications to specific encodings (CSV, spreadsheets, XML) do not support non-destructive updates, and version control systems are necessary to retrieve previous states. On the other hand, relational databases can be inefficient when querying all recorded transformations, edits and manipulations.

Non-destructive updates are well supported by current Semantic Web technology like SPARQL (see Section 2.2). SPARQL `CONSTRUCT` allows the construction of RDF triples according to the supplied graph pattern, facilitating data transformations without altering consistency of previous states in the knowledge base. SPARQL `SELECT` selects, according to some graph pattern, the desired data in the RDF graph, and disposes them according to any desired view format (for example, columns matching some interesting historical variables). However, standardization on how semantic technologies cover data transformations in all phases of the historical data life cycle (see Section 3.1) is needed.

7. Discussion and conclusions

In this paper we present a general overview of semantic technologies applied to historical research. We describe a general approach to historical research and the Semantic Web, and motivate why the combination of the two is an interesting field of research. We introduce core elements of historical research, such as the life cycle of historical information, several classifications for historical data, and the open problems shared by historians and computer scientists. Then, we overview contributions to the young historical Semantic Web in form of papers, projects and tools, articulating the work into several tasks and trends within these tasks. We provide a mapping to see to what extent the work on these tasks is helping to solve the open problems of historical data and historical research. Finally, we dig out a list of interesting open challenges for the future, like working out the semantics of critical aspects for historians, such as interpretation and time, and encouraging reasoning in the historical Semantic Web.

It is interesting to observe the sparsity in Tables 1, 2, 3 and 4. There is a significant difference in the number of empty spaces (i.e. specificity of the contributions) between Tables 1 and 4 (papers and tools, ontologies), and Tables 2 and 3 (projects and online resources). While the former set has essentially lots of *holes*, the latter has lots of *complete lines*. The reason for this is probably the specificity researchers think research papers and useful tools need. Usually written by computer scientists, papers and tools need to be grounded and tackle a very concrete problem to be worth written or implemented. On the other hand, projects (Table 2) are written in a very generic way covering all tasks, with probably intensive participation of historians and clear aims to solve the whole pipeline. In practice, though, these goals are materialized in very concrete research contributions. This leads us to think that Semantic Web solutions need very specific requirements in order to be correctly deployed in history. They need to be applied to historical data in a complex, layered and properly adapted pipeline. Good practices and standards, and their relationship with the life cycle of historical information, are still needed for the field to continue evolving.

We show how the Semantic Web and history communities understand the need for representing inner semantics implicitly contained in historical sources, and how these semantics can be conveniently identified, formalized and linked. With the appropriate pipelines, algorithms can extract entities from digital historical sources and transform these occurrences into RDF triples, according to some historical ontology or vocabulary. These entities can be linked between them and with other historical Linked Data, contributing to an open, world wide, online persistent graph of historical knowledge: an historical Semantic Web. The work presented in this survey contributes in one phase or another in this graph-building pipeline. We leave to the reader if this historical Semantic Web building pipeline is, in fact, the Semantic Web version of the life cycle of historical information.

The challenge of the realisation of a historical Semantic Web meeting as many requirements as possible may bring new facilities for a number of stakeholders. On the one hand, humanities researchers, also outside history, will be able to integrate the historical Semantic Web to their own information life cycle. They will be able to search, retrieve and compare historical knowledge and use it for the construction of their narratives, still the final outcome of historical research. On the other hand, practitioners will be able to search new data sources to develop history-aware applications for public institutions, private companies and citizens.

Acknowledgements

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, under the auspices of the CEDAR project (see `http://ehumanities.nl`), the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404), and by NWO's CATCH programme via the Agora project, grant 640.004.801.

This work has been supported by the Dutch national programmes COMMIT.

The authors want to thank all contributing colleagues, especially the interviewees Onno Boonstra, Peter Doorn, Jan Kok, Henk Laloli and Dirk Roorda.

References

[1] Ad van der Meer and Onno Boonstra. *Repertorium van Nederlandse Gemeenten, 1812-2006, waaraan toegevoegd de Amsterdamse code*. DANS Data Guide 2, The Hague, 2006.

[2] Agora Project. `http://agora.cs.vu.nl`.

[3] Ian Anderson. History and computing. *Making History*, 2008. `http://www.history.ac.uk/makinghistory/resources/articles/history_and_computing.html`.

[4] Miguel Ángel Manso and Monica Wachowicz. GIS Design: A Review of Current Issues in Interoperability. *Geography Compass*, 3(3):1105–1124, 2009.

[5] Grigoris Antoniou and Frank van Harmelen. *A Semantic Web Primer (Cooperative Information Systems)*. The MIT Press, Cambrdige, Massachusetts, April 2004.

[6] Armadillo: Historical Data Mining Project. `http://www.hrionline.ac.uk/armadillo/armadillo.html`.

[7] Isabelle Augenstein, Sebastian Padó, and Sebastian Rudolph. LODifier: Generating Linked Data from Unstructured Text. In Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Proceedings*, volume 7295 of *LNCS*, pages 210–224, Berlin, Heidelberg, 2012. Springer-Verlag.

[8] James Cook University Australia. Primary, secondary and tertiary sources. `http://libguides.jcu.edu.au/primary`.

[9] Franz Baader, Ian Horrocks, and Ulrike Sattler. Description Logics as Ontology Languages for the Semantic Web. In Dieter Hutter and Werner Stephahn, editor, *Mechanizing Mathematical Reasoning*, volume 2605 of *LNCS*, pages 228–248, Berlin, Heidelberg, 2005. Springer-Verlag.

[10] C Beghtol. Classification Theory. *Encyclopedia of Library and Information Science*, 2010:1045–60, 2010.

[11] Jules R. Benjamin. *A Student's Guide to History*. Bedfors/St. Martin's, Boston, 2004.

[12] Tim Berners-Lee. Linked Data – Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`.

[13] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

[14] Tim Berners-Lee and Jeffrey Jaffe. The World Wide Web Consortium (W3C). `http://www.w3.org/`.

[15] David M. Berry, editor. *Understanding Digital Humanities*. Palgrave Macmillian, New York, 2012.

[16] Onno Boonstra, Leen Breure, and Peter Doorn. *Past, present and future of historical information science*. NIWI-KNAW, Amsterdam, 1st edition, 2004.

[17] B. Bos and G. Welling. The significance of user-interfaces for historical software. *Proceedings of the Eight International*

Conference of the Association for History and Computing, pages 223–236, 1995.

[18] Paolo Bouquet, Fausto Giunchiglia, Frank van Harmelen, Luciano Serafini, and Heiner Stuckenschmidt. Contextualizing Ontologies. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 1(4):325–343, 2004.

[19] BRIDGE Project. `http://ilps.science.uva.nl/node/735`.

[20] Marc Bron, Bouke Huurnink, and Maarten de Rijke. Linking Archives Using Document Enrichment and Term Selection. In *Research and Advanced Technology for Digital Libraries. 15th international conference on Theory and practice of digital libraries, proceedings.*, volume 6966 of *LNCS*, pages 360–371, Berlin, Heidelberg, 2011. Springer-Verlag.

[21] CCEd Project. `http://www.theclergydatabase.org.uk/publications/jeh_article.html`.

[22] CEDAR Project. `http://www.cedar-project.nl/`.

[23] Eugene Charniak and Drew McDermott. *Introduction to Artificial Intelligence*. Addison-Wesley series in computer science. Addison-Wesley, Boston, 1985.

[24] CHORAL Project. `http://hmi.ewi.utwente.nl/choral/`.

[25] Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic (CKCC Project). `http://ckcc.huygens.knaw.nl/`.

[26] Edgar F. Codd. Derivability, Redundancy, and Consistency of Relations Stored in Large Data Banks. Technical Report RJ599, IBM Research, San Jose, California, 1969.

[27] Panos Constantopoulos, Martin Doerr, Maria Theodoridou, and Manolis Tzobanakis. Historical documents as monuments and as sources. In Bernard Frischer, Jane Webb Crawford, and David Koller, editors, *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA). Proceedings of the 37th International Conference*, volume S2079 of *BAR International Series*, Oxford, 2009. Archaeopress.

[28] The Dublin Core Metadata Initiative (DCMI). `http://www.dublincore.org/`.

[29] Stefan Dormans and Jan Kok. An alternative approach to large historical databases. Exploring best practices with collaboratories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 43(3):97–107, 2010.

[30] The eHumanities Group. `http://ehumanities.nl/`.

[31] Albert Esteve and Matthew Sobek. Challenges and Methods of International Census Harmonization. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 36(2):37–41, 2003.

[32] Europeana Data Model (EDM) Documentation. `http://pro.europeana.eu/edm-documentation`.

[33] The Event Ontology. `http://purl.org/NET/c4dm/event.owl#`.

[34] Mary Feeney and Seasmus Ross. Information Technology in Humanities Scholarship British Achievements, Prospects, and Barriers. *Historical Social Research*, 19(1):3–59, 1994.

[35] Giorgos Flouris, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, and Grigoris Antoniou. Ontology change: classification and survey. *The Knowledge Engineering Review*, 23(2):117–152, 2008.

[36] FRED. `http://wit.istc.cnr.it/stlab-tools/fred`.

[37] Aldo Gangemi. A Comparison of Knowledge Extraction Tools for the Semantic Web. In *The Semantic Web: Semantics and Big Data. 10th International Conference, ESWC 2013, Proceedings*, volume 7882 of *LNCS*, Berlin, Heidelberg, 2013. Springer-Verlag.

[38] The GeoNames Ontology. `http://www.geonames.org/ontology/documentation.html`.

[39] Ronald Goeken, Marjorie Bryer, and Cassandra Lucas. Making Sense of Census Responses Coding Complex Variables in the 1920 PUMS. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 32(3):37–41, 1999.

[40] Thomas R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43:907–928, 1993.

[41] Bernhard Haslhofer, Rainer Simon, Robert Sanderson, and Herbert van de Sompel. The Open Annotation Collaboration (OAC) Model. *Computing Research Repository, CoRR*, abs/1106.5178, 2011.

[42] H-BOT Project. `http://chnm.gmu.edu/tools/h-bot/`.

[43] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan and Claypool, 1st edition, 2011.

[44] HEML Project. `http://heml.mta.ca/heml-cocoon/description`.

[45] HISCO Project. `http://historyofwork.iisg.nl/`.

[46] HiTime Project. `http://ilk.uvt.nl/hitime/`.

[47] Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul A. Martin. Interpretation as Abduction. *Artificial Intelligence*, 63(1-2):69–142, 1993.

[48] Historic Sample of the Netherlands (HSN). `http://www.iisg.nl/hsn//`.

[49] Eero Hyvönen, Jouni Tuominen, Tomi Kauppinen, and Jari Väätäinen. Representing and Utilizing Changing Historical Places as an Ontology Time Series. In Ramesh Jain and Amit Sheth, editors, *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications*, Semantic Web and Beyond: Computing for Human Experience, pages 1–25. Springer-Verlag, Berlin, Heidelberg, 2011.

[50] Nancy Ide and David Woolner. Exploiting semantic web technologies for intelligent access to historical documents. *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, pages 2177–2180, 2004.

[51] Nancy Ide and David Woolner. *Historical Ontologies*, chapter Words and Intelligence II: Essays in Honor of Yorick Wilks, pages 137–152. Springer-Verlag, Berlin, Heidelberg, 2007.

[52] Integrated Public Use Microdata Series (IPUMS). `http://www.ipums.org/`.

[53] C. Jones. Strategies for managing requirements creep. *Computer*, 29(6):92–94, June 1996.

[54] Maximilian Kalus. Semantic networks and historical knowledge management: Introducing new methods of computer-based research. *Ann Arbor, MI: MPublishing, University of Michigan Library*, 2007.

[55] Max Kemman and Martijn Kleppe. PoliMedia - Improving Analyses of Radio, TV and Newspaper Coverage of Political Debates. In T. Aalberg and E. Al, editors, *15th international conference on Theory and practice of digital libraries, TPDL 2013, Proceedings.*, volume 8092 of *LNCS*, pages 401–404, Berlin, Heidelberg, 2013. Springer-Verlag.

[56] Jan Kok and Paul Wouters. Virtual Knowledge in Family History: Visionary Technologies, Research Dreams, and Re-

search Agendas. In Paul Wouters, Anne Beaulieu, Andrea Scharnhorst, and Sally Wyatt, editors, *Virtual Knowledge. Experimenting in the Humanities and the Social Sciences*, pages 219–244. MIT Press, Cambridge, Massachusetts, 2013.

[57] Thomas Kuczynski, editor. *Wirschaftsgeschichte und Mathematik. Beiträge zur Anwendung mathematischer, insbesondere statistischer Methoden in der wirtschafts- und sozialhistorischen Forschung*. Akademie-Verlag, Berlin, 1985.

[58] Linking History in Place. `http://cv.vic.gov.au/linking-history/`.

[59] Links Project. `http://www.iisg.nl/hsn/news/links-project.php`.

[60] LODE: An ontology for Linking Open Descriptions of Events. `http://linkedevents.org/ontology/`.

[61] NSF-ITR/MALACH Project. `http://malach.umiacs.umd.edu/`.

[62] Kees Mandemakers and Lisa Dillon. Best Practices with Large Databases on Historical Populations. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 37(1):34–38, 2004.

[63] Willard McCatry. Humanities Computing. In Miriam Drake, editor, *Encyclopedia of Library and Information Science*, pages 1124–35. Taylor and Francis, New York, 2nd edition, 2003.

[64] Albert Meroño-Peñuela. Semantic Web for the Humanities. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data. 10th International Conference, ESWC 2013, Proceedings*, volume 7882 of *LNCS*, pages 645–649, Berlin, Heidelberg, 2013. Springer-Verlag.

[65] Albert Meroño-Peñuela, Ashkan Ashkpour, Laurens Rietveld, Rinke Hoekstra, and Stefan Schlobach. Linked Humanities Data: The Next Frontier? A Case-study in Historical Census Data. In *Proceedings of the 2nd International Workshop on Linked Science (LISC2012). International Semantic Web Conference (ISWC)*, volume 951. CEUR Workshop Proceedings, 2012.

[66] Albert Meroño-Peñuela, Christophe Guéret, Rinke Hoekstra, and Stefan Schlobach. Detecting and Reporting Extensional Concept Drift in Statistical Linked Data. In *Proceedings of the 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC*. CEUR Workshop Proceedings, 2013.

[67] Albert Meroño-Peñuela, Rinke Hoekstra, Andrea Scharnhorst, Christophe Guéret, and Ashkan Ashkpour. Longitudinal queries over linked census data. In *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Satellite Events*, pages 306–307, 2013.

[68] Peter B. Meyer and Anastasiya M. Osborne. Proposed category system for 1960-2000 census occupations. *U.S. Bureau of Labor Statistics*, 2005.

[69] Richard Moot, Laurent Prévot, and Christian Retoré. A discursive analysis of itineraries in an historical and regional corpus of travels: syntax, semantics, and pragmatics in a unified type theoretical framework. In *Constraints in Discourse 2011*, pages 14–16, 2011.

[70] North Atlantic Population Project. `http://www.nappdata.org/napp/`.

[71] Michael Nentwich. *Cyberscience: Research in the Age of the Internet*. Austrian Academy of Sciences Press, Vienna, 2003.

[72] University of Maryland. Primary, Secondary and Tertiary Sources. `http://www.lib.umd.edu/ues/guides/primary-sources`.

[73] Jan Oldervoll. CENSSYS: A System for Analyzing Census-Type Data. *Computer Applications in the Historical Sciences: Selected Contributions to the Cologne Computer Conference*, pages 17–22, 1989.

[74] OpenRefine. `https://github.com/OpenRefine/OpenRefine`.

[75] Semantic Web approaches in Digital History: an Introduction. `http://www.slideshare.net/mpasin/presentations/`.

[76] Fawcett: A Toolkit to Begin an Historical Semantic Web. `http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/175/217`.

[77] Spatial cyberinfrastructures, ontologies, and the humanities. `http://www.pnas.org/content/108/14/5504.full`.

[78] SIG:Ontologies. `http://wiki.tei-c.org/index.php/SIG:Ontologies`.

[79] CultureSampo - Finnish Culture on the Semantic Web 2.0: Thematic Perspectives for the End-user. `http://www.museumsandtheweb.com/mw2009/papers/hyvonen/hyvonen.html`.

[80] Text Mining for Historical Documents: Topics and Papers. `http://www.coli.uni-saarland.de/courses/tm-hist/readings.html`.

[81] RDF vocabularies for historic place-names and relations between them. `http://groups.google.com/group/caa-semantic-sig/browse_thread/thread/ae1db7fa31a1b5a0?pli=1`.

[82] The Semantic Web for Family History. `http://jay.askren.net/Projects/SemWeb/`.

[83] Data portal for Social Sciences Open data with SPARQL endpoint. `http://www.rechercheisidore.fr/`.

[84] J. B. Owens, May Yuan, Monica Wachowicz, Vitit Kantabutra, Emery A. Coppola Jr., Daniel P. Ames, and Aldo Gangemi. Visualizing Historical Narratives: Geographically-Integrated History and Dynamics GIS. In *National Endowment for the Humanities workshop. Visualizing the Past: Tools and Techniques for Understanding Historical Processes*, 2009.

[85] FDR Pearl Harbor Project. `http://www.fdrlibrary.marist.edu/`.

[86] Polimedia Project. `http://www.polimedia.nl/`.

[87] Bart van de Putte and Andrew Miles. A Social Classification Scheme for Historical Occupational Data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 38(2):61–92, 2005.

[88] Thomas Riechert, Ulf Morgenstern, Sören Auer, Sebastian Tramp, and Michael Martin. Knowledge Engineering for Historians on the Example of the Catalogus Professorum Lipsiensis. In *The Semantic Web – ISWC 2010. 9th International Semantic Web Conference, Proceedings*, volume 6496 of *LNCS*, pages 225–240, Berlin, Heidelberg, 2010. Springer-Verlag.

[89] B. G. Robertson. Visualizing an historical semantic web with Heml. In *WWW'06. The 15th International World Wide Web Conference 2006, Proceedings*, pages 1051–1052, New York, NY, USA, 2006. ACM Press.

[90] Bruce Robertson. Exploring Historical RDF with Heml. *Digital Humanities Quarterly*, 3(1), 2009. `http://www.digitalhumanities.org/dhq/`

vol/003/1/000026.html.

[91] Steven Ruggles and Russell R Menard. The Minnesota Historical Census Projects. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 28(1):6–10, 1995.

[92] SAILS Project. http://sailsproject.cerch.kcl.ac.uk/2010/07/about-the-sails-project/.

[93] Voyage of the Slave Ship Sally Project. http://www.stg.brown.edu/projects/sally/.

[94] Schema.org. http://schema.org/.

[95] Susan Schreibman, Ray Siemens, and John Unsworth, editors. *A Companion to Digital Humanities*. Blackwell Publishing Inc, Malden, MA, 2004.

[96] Scratch Project. http://www.ai.rug.nl/alice/nwo-catch-scratch/index_english.html.

[97] Roxane Segers, Marieke van Erp, Lourens van der Meij, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga, Johan Oomen, and Geertje Jacobs. Hacking History via Event Extraction. In *Proceedings of the sixth international conference on Knowledge capture*, K-CAP '11, pages 161–162, New York, NY, USA, 2011. ACM Press.

[98] Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.

[99] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176, 2013.

[100] Renee E. Sieber, Christopher C. Wellen, and Yuan Jin. Spatial cyberinfrastructures, ontologies, and the humanities. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14):5504–5509, 2011.

[101] Semantic MediaWiki (SMW). http://semantic-mediawiki.org/.

[102] Matthew Sobek. The Comparability of Occupations and the Generation of Income Scores. *Historical Methods: A Journal of Quantitative and Interdisciplinary History. Special Issue: The Minnesota Historical Census Project*, 28(1):47–51, 1995.

[103] W.A Speck. History and computing: some reflections on the past decade. *History and Computing*, 6(1):28–32, 1994.

[104] TabLinker. https://github.com/Data2Semantics/TabLinker/.

[105] NLP2RDF. http://nlp2rdf.org/.

[106] SIMILE/Timeline. http://www.simile-widgets.org/timeline/.

[107] Gapminder. http://www.gapminder.org/.

[108] TokenX. http://tokenx.unl.edu/.

[109] TAPoR. http://portal.tapor.ca/portal/portal.

[110] SEM event model. http://www.cs.vu.nl/~guus/papers/Hage11b.pdf.

[111] OpenCYC. http://www.opencyc.org/.

[112] XCES. http://www.xces.org/.

[113] GATE. http://gate.ac.uk/.

[114] WordNet. http://wordnet.princeton.edu/.

[115] FrameNet. https://framenet.icsi.berkeley.edu/fndrupal/.

[116] SUMO. http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/KBs/Merge.kif.

[117] MILO. http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/KBs/Mid-level-ontology.kif.

[118] AskSam. https://www.asksam.com/.

[119] TEI (Text Encoding Initiative). http://www.tei-c.org/index.xm.

[120] Standard Generalized Markup Language (SGML). http://en.wikipedia.org/wiki/Standard_Generalized_Markup_Language.

[121] TACT. http://projects.chass.utoronto.ca/tact/.

[122] Wordcruncher. http://www.wordcruncher.com/wordcruncher/default.htm.

[123] Atlas.ti. http://www.atlasti.com/index.html.

[124] Natural Language Toolkig (NLTK). http://www.nltk.org/.

[125] M Thaller. Automation on Parnassus. CLIO - A databank oriented system for historians. *Historical Social Research / Historische Sozialforschung*, 15:40–65, 1980.

[126] U. Thiel, H. Brocks, A. Dirsch-Weigand, A. Everts, I. Frommholz, and A. Stein. Queries in Context: Access to Digitized Historic Documents in a Collaboratory for the Humanities. In Matthias Hemmje, Claudia Niederée, and Thomas Risse, editors, *From Integrated Publication and Information Systems to Information and Knowledge Environments*, volume 3379 of *LNCS*, pages 117–127, Berlin, Heidelberg, 2005. Springer-Verlag.

[127] John Tosh. *The Pursuit of History: Aims, Methods, and New Directions in the Study of History*. Pearson Education: Harlow 2010, 5th edition, 2010.

[128] Matje van de Camp and Antal van den Bosch. A link to the past: Constructing historical social networks. In A. Balahur, E. Boldrini, A. Montoyo, and P. Martinez-Barco, editors, *Proceedings of the ACL-HLT Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA-2011)*, pages 61–69, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[129] Chiel van den Akker, Susan Legêne, Marieke Van Erp, Lora Aroyo, Roxane Segers, Lourens Van der Meij, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga, Johan Oomen, and Geertje Jacobs. Digital Hermeneutics: Agora and the Online Understanding of Cultural Heritage. In *Proceedings of the 3rd International Conference on Web Science (WebSci 2011)*, pages 1–7, 2011.

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

[131] CLARIN-VK Project. http://verrijktkoninkrijk.nl/.

[132] The World Wide Web Consortium (W3C). OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features/.

[133] The World Wide Web Consortium (W3C). PROV-O: The PROV Ontology. http://www.w3.org/TR/prov-o/.

[134] The World Wide Web Consortium (W3C). RDFa: Rich Structured Data Markup for Web Documents. http://www.w3.org/TR/rdfa-primer/.

[135] The World Wide Web Consortium (W3C). Resource Description Framework (RDF). http://www.w3.org/RDF/.

[136] The World Wide Web Consortium (W3C). SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/.

[137] The World Wide Web Consortium (W3C). XML Path Lan-

guage (XPath). http://www.w3.org/TR/xpath/.

[138] The World Wide Web Consortium (W3C). XQuery 1.0: An XML Query Language (Second Edition). http://www.w3.org/TR/xquery/.

[139] WAHSP and BILAND. http://www.wahsp.nl/.

[140] Shenghui Wang, Stefan Schlobach, and Michel C. A. Klein. Concept drift and how to identify it. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):247–265, 2011.

[141] Rene Witte, Ralf Krestel, Thomas Kappler, and Peter C. Lockemann. Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base. *IEEE Intelligent Systems*, 25(1):58–67, 2010.

[142] Peter Wittek and Walter Ravenek. Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling. In B. Maegaard, editor, *Supporting Digital Humanities 2011: Answering the unaskable*, Copenhagen, Denmark, 2011.