Semantic Technologies for Historical Research: A Survey

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

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

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

Abstract. The diversity of sources of information for historical research fill a continuum between individual accounts transmitted for instance in letters but also in poems and songs, and aggregated statistical information as in the case of historical census. Historiography shares this heterogeneity and complexity of source material with other humanities fields. Methods to order this rich material, and by this ordering also to determine the way history is told are as old as history writing and vary among the different branches (or subdisciplines) of historical research.

In this paper we focus on the work of historians, and even more specifically economic and social history.At the crossroad of information and historical sciences, so-called *Historical Informatics* or *History and Computing* emerged as a specific profession during the nineties of the last century. Together with computer scientists historians created a research agenda concentrating around questions how to create, design, enrich, edit, retrieve, analyze and present historical information with help of information technology. There exist a number problems and challenges in this field; some of them are closely related to semantics and meaning of knowledge in general. In this context, Semantic Web technologies can be applied in a number of situations, environments, applications of historical computing and historical information science. However, only a few number of contributions have yet considered these technologies. In this survey we present an overview of the past and present problems, challenges and advances of historical science computing, from out the perspective of Semantic technology.

Keywords: Semantic technology, Digital Humanities, Historical datasets, Historical Computing

1. Introduction

This survey covers current approaches, namely papers, projects, online resources, and tools and technologies, on how to apply semantic technology to historical research. As of now, 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. Hence, the intended audience of the paper is twofold. On the one hand, section 3 offers a general overview of recent advances in eHistory, historical computing and historical information sciences, as well as a landscape of recent problems that can be tackled with technologies provided by the Semantic Web community. In turn, on the other hand, section 4 gives a summary of current Semantic Web technology developments, as well as typical scenarios in which these technologies have provided valuable benefits. Hence, both communities, namely scholars doing research in history and Semantic Web advocates, clearly identify their domains and, at the same time, have the opportunity to gain insights into the other field.

Goal. We claim that some current issues in research done by historians can be successfully addressed applying semantic technologies. These issues find their nature in the inner semantics implicitly contained in historical sources, which can be appropriately identified, formalized and linked using. For example, any reference to a historical person contained in an historical source can be conveniently modeled in RDF, given meaning with a set of appropriate OWL vocabularies, and linked with other entities (for example, events in which that person participated, places that person visited, actions that person did), contributing to a very rich graph of knowledge that machines can easily read and process, and from which scholars can extract meaningful knowledge to their research and facilitate their role as history detectives.

Methodology. We proceed with a twofold analysis. First, we introduce historical computing, identifying current problems. Second, we go through a set of 67 relevant contributions, organized into published papers, research projects, online resources and tools and technologies, and build a catalog on them depending on how they give importance or priority to several areas in the crossroads of the Semantic Web and research by historians: knowledge modeling and ontologies, text processing and mining, search and retrieval, and semantic interoperability.

Contribution. This paper offers the current landscape on advances in faithfully representing historical data with semantic technology. Few contributions cover the full spectrum of what is desired from a complete workflow, going from historical sources to semantic data, which can be easily linked and integrated, and although there exist efforts implementing several parts of the process, much more work is needed to successfully understand, process and represent historical sources using Semantic Web formats. Nevertheless, this paper identifies the most relevant problems found in research by historians, and suggests how these issues can be addressed using Semantic Web technologies.

In section 2 we describe elements of the epistemic frames used in this field together with ideal-typical workflows, and in section 3 we shortly introduce Historical Information Sciences. The aim is to identify problems where Semantic Web technological could offer solutions. A prerequisite to that is to bridge between different terminologies concerning sources, data, data models and structured or unstructured data. In section 4 we start from the Semantic Web technology and apply the concepts developed in section 3 to classify past and current research (as documented in publications but also projects) along dimensions such as. Having all these contributions in a single view will hopefully help to realize how the divers the landscape in the borders of historical computing and the Semantic Web research is and also indicate promising research lines for the future.

2. Historical research and semantic technology

The field of historical research concerns the understanding of our past and it is currently undergoing major changes in its methodology, largely due to the growing interest in and the advent of high-quality digital resources [86]. Historical data encompasses statistical information, texts, images, and objects. Datasets contain both factual and terminological knowledge about events, people and processes with a temporal perspective that makes them valuable resources and interesting objects of study. What makes the humanities scholar work even more interesting for a semantic description, is its context dependence and the variety of possible interpretations. Semantic technology could play an important role when it comes to keeping and comparing different perspectives; to relate different sources - both in terms of historical facts as well as of their perception and interpretation; and to transfer knowledge produced in scholarly discourse to ongoing reflection of society and mankind.

Historical research is also part of other disciplines than history itself. The difference in data handling is, that historians want *to look through the sources* to understand a society of the past, while art history and history of literature are much more focused on the art work itself. This implies also differences in encoding information. In this paper we address history or historiography as a research domain.

Historical sources - the data of historians

Historical sources can be characterized and divided in many ways, but a basic distinction used by historians is between primary and secondary sources. Primary sources can be distinguished roughly into administrative sources and narratives like biographies, or chronicles. Secondary sources are all material that has been written by historians or their predecessors about the past [33]. We will restrict here ourselves to primary sources.

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. A good example is the project *Historic Sample of the Netherlands* (HSN[1]) The source material for the HSN database consists mainly of the certificates of birth[2], marriage[3] and death[4] and of the population registers[5]. 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 analyze with techniques of Natural Language Processing (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 [21].

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. Also here knowledge encoding is interesting, and it requires more sophisticated ontologies than currently available.

One has to be aware of the fact, that historical data (as present in sources) are fundamentally different from hard science data: historical data have not been produced under the controlled conditions of an experiment. So historical research will always have something of the work of a detective: he must be careful not to destroy certain details (read: annoying inconsistencies) because 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, harmonizing expressions of what is later used as variables is needed. Harmonization, the process of making datasources uniformly accessible is closely related to issues of standardization and formalization [**?**]. But, harmonization can have different targets and purposes.

Moreover, there is no common language to label facts: the terminology is time and space bound, which makes formalization difficult and results dependent on different interpretations.

We will discuss in this paper how semantic technologies can support both research practices we depicted archetypically above: the careful, authentic and preserving collection of a variety of evidence (historian as detective) and the ordering of those evidences along concepts, theories and models to form explanations debatable in scientific discourse (historian as scientist).

Structures and data models

In section 3 we use the distinction between structured and non-structured to classify historic data. We would like to underline that this differentiation 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, both administrative and narrative sources can ap-

[1] `http://www.iisg.nl/hsn/index.html`
[2] `http://www.iisg.nl/hsn/data/birth.html`
[3] `http://www.iisg.nl/hsn/data/marriage.html`
[4] `http://www.iisg.nl/hsn/data/death.html`
[5] `http://www.iisg.nl/hsn/data/population.html`

pear in the form of *structured* or *unstructured* data in Computer Scientist jargon.

eHistory and the digital humanities

The identification of un-explored links between semantic technologies and historiography requires a careful explanation of the science history of earlier encounters between computer scientists and historians. We present a short history of the so-called historical information sciences at the beginning of section 3. It is also important to realize that the interdisciplinary collaboration between computer science and humanities researchers paves the way for true integration of semantic technologies in what one could call eHistory.

The challenge of a semantic enrichment of historical sources may bring new facilities, on the one hand, for humanities researchers also outside of history, allowing them to search, retrieve and compare information they need for their everyday work using a variety of dimensions and scopes; and on the other hand, for practitioners, giving them new data sources to develop historical-aware applications for public institutions, private companies and citizens.

What we call through this paper eHistory, is nothing more than a common label for the emergence of new research technologies. Co-evolving with them, new theoretical and methodological frameworks are penetrating existing historical research, widening and shifting boundaries to other academic fields and partly defining new scientific subfields. In this paper we focus on most recent technological trends connected to the emergence of the internet and the web [25], and look in particular into Semantic Web technologies [3]. We would like to point out nevertheless that changes in historical research are closely connected to the emergence of new scientific methods already since centuries. Statistics has influenced many fields including history, and paved the ground for quantitative studies [20]. However, these kind of historical studies are more and more the domain of sociologists, economists and demographers than scientists educated as historians [27][6].

More recently, the invention of computers and the formation of computer sciences have inspired historians from the start. *History computing* or *Humanities computing* were labels used in the pre-Internet area [22]. 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 [86]. Though not all research dreams materialized in the primary envisioned way [93], nowadays historians in particular in the area of economic history and social history aim for world-wide, large scale collaborations. This kind of web based cooperation allows to collect, distribute, annotate and analyze historical information all around the globe [8]. While we concentrate on historical research and more specific branches as social and economic history in it, we also would like to stress that similar solutions emerge also in other humanities fields at the turn to e-humanities or Digital Humanities [5,28]. As historical research overlaps with literary studies, ancient language studies, archeology, 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 or other humanities fields.

3. Historical Information Science

Since this paper focuses on applications of Semantic Web technologies in research by historians, it makes sense to introduce briefly the field of historical information science, describing the state-of-the-art and giving some insights in current challenges.

This paper looks also forward on how Semantic Web technology can be applied to historical datasets, and how these technologies can facilitate, boost and improve research by historians. To build a meaningful discourse, it is necessary to explain that semantic technologies need specific requirements in order to be correctly deployed in history: they need to be applied to historical source datasets in a complex, layered and properly adapted pipeline. That pipeline, hence, is not static at all: its configuration strongly depends on the way and degree the datasets are structured In Computer science and history may use different meanings for the term structure which will lead to confusion and ambiguity. For this reason, we dedicate an entire subsection to cover existing classifications of historical datasets, how their inner structure is somehow related to their type of source, and how it strongly influences further semantic pipelines to be applied. As far as we use the term *structure* we mean the software engineering definition in which the term *structured data* refers to information that is contained in a pre-defined data

[6] `http://www.hist.umn.edu/~ruggles/hist5011/Decline.pptx`

model independent of the content (the structure is an empty shell).

Likewise, the term semantics has to be interpreted from the computational point of view. Computational semantics[7] deal with how to automate the process of constructing and reasoning with meaning representations of natural language expressions, which is one of the most common ways in which historical sources are written (that is, any kind of free speech or free discourse text like letters, biographies, manuscripts or memories). Specifically, Semantic Web technologies use formal languages (such as RDF, RDFS or OWL) to define concepts, persons, places and any kind of entity, so that they can be identified and referred from those texts in a way that machines can easily differentiate the human expressions from the meaningful, linked knowledge behind them.

Section 3 is organized in two subsections. The first subsection gives an overview on historical computing explaining, first, the evolution (life cycle) of data streams in research by historians and, secondly, the diverse types of historical sources and their classifications. The second subsection focuses on underlying problems and challenges when dealing with these sources from a semantic point of view.

3.1. State of the art in Historical Information Science

Ever since the advent of computing historians have been using them in their research or teachings in one way or the other. The first revolution in the 1960ï£¡s 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 [2]. However this was a marginal group within the historical research, in general the usage of computers by humanists could be described as occasional [10]. 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.

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 [86]. Instead of improving the quality of their work and assisting them in this process, software developed for historians often requires attending several summer schools before being able to use them [7]. Currently there are still many challenges and information problems in historical research. These difficulties mainly range from textual, linkage, structuring, interpretation or visualization problems [86].

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..), archeology (impossible without GIS nowadays), and other fields using sources that have been digitized for historical (comparative) research and converted to databases [86]. 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 corpus, visual objects etc. These sources or isolated databases often contain a lot of semantics, but their data models were asynchronously designed, making them difficult to compare (interoperability problem). 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 [14].

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. In this article we look into the benefits of the utilization of semantic web technologies in the field of history, providing novel insights for historians to deal with these problems and use contemporary methods and technologies to gain better understanding of historical data sources.

3.1.1. The life cycle

Like all scientific objects historical objects go through several distinct phases with specific transformations in order to produce an outcome suitable for historical research or specific needs of a researcher. The main object of study in historical information sci-

[7] `http://en.wikipedia.org/wiki/Computational_semantics`

ence is historical information, and the various way to create, design, enrich, edit , retrieve, analyze and present historical information with help of information technology [86]. Accordingly, in line with this perspective, historical information can be laid out as sequential phases of a *historical information life cycle* (see Figure 1. It should be mentioned that these stages, although sequential, do not always have to be passed through in sequence and some can be skipped when necessary. The stages are also quite comparable with the practice in other fields of science.

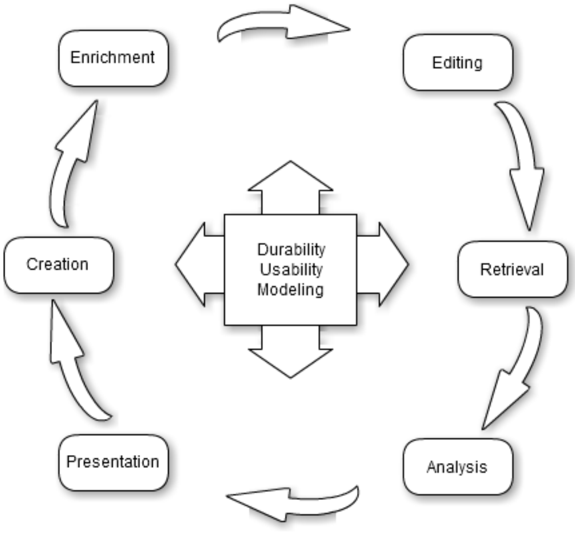


Fig. 1. The life cycle of historical information.

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 projectï£¡s design. In the enrichment stage the most important aspect is to enrich the data which has been created with metadata, describing the historical information in more detail. It is suggested to do this in a standardized way using systems such as Dublin Core[8] for this. Historical context enrichment also involves to link the data which belongs together (nominal record linkage), for example think of persons with the same name, places or events. The editing phase involves activities which aim to enhance the data further and entails the actual encoding of textual information.Examples are: entering data in databases, inserting mark-up tags, annotate original data with extra information, bibliographical references and creating links to related passages. Once the data has been created and processed it is ready to be viewed and used. The retrieval stage mainly involves selection mechanism look-ups such as SQL-queries for traditional databases or Xpath[9] and Xquery[10] for text retrieval. The analysis stage both involves statistical analysis of historical data sets as well as qualitative comparisons and analysis of text. The presentation of historical information is very diverse and can be expressed in different ways such as digital text editions, databases, visualizations, statistical views or even virtual exhibitions.

At the center of the historical information life cycle, three aspects are identified which are central to computing in the humanities in general. Durability ensures the long term use of the data, usability refers to the ease of efficiency, effectiveness and user satisfaction and modeling here denotes to more general modeling of research processes and historical information systems.

3.1.2. *Structured and unstructured historical datasets*

We have presented a general overview on historical computing, putting emphasis on how historical information science faces the cycle of historical information [86], and how the cycle starts with constructing the relevant historical datasets. At the end of this creation phase one may expect to have a set of all data needed for further processes. However, the nature of many of the next steps to be taken thereafter may strongly depend on the way the resulting dataset is structured. Indeed, the way we could attach semantic web technologies to these historical sources (e.g. to extract RDF triples) is strongly dependent on the degree of structure of these sources.

To mark these differences we introduce the dichotomy of *structured* and *unstructured* data. *Structured data* refers, then, to information that is based on sources that have a very clear pre-defined data model like census material published in rows and columns, while *unstructured data* refers to information which has not. A data model is an abstract model that documents and organizes data for communication, and is used as a plan for developing applications.

Since in this paper we only consider digitized sources, we talk about a *structured historical digitized source* (shortly, structured source) when, indeed, such an abstract model for the data contained in the digi-

[8] http://dublincore.org/

[9] http://www.w3.org/TR/xpath/

[10] http://www.w3.org/TR/xquery/

tized source does exist. Well known examples of such a structure are sources encoded as relational databases, XML files, spreadsheet workbooks or RDF triplestores. It is easy to see that all these examples meet a certain abstract model for the data they represent (relational schemes, DTD constraints, tabular formats and RDF triple statements). In case such a data model is lacking we define these as *unstructured historical digitized sources*, so these are the ones that have only a few or no structure at all: commonly, unconstrained corpora.

Some historians [86] working from an information-oriented perspective are coming near to this perspective of structured and unstructured. They proposed to structure historical data depending on their (probably) required further machine processing: 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.

Nevertheless, it has to be said that we donï£¡t consider source classifications as a major issue in the semantic technologies pipeline. Although structure really matters for deciding what 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 (for instance, a secondary source may require an OWL ontology with a vocabulary allowing to describe *historical interpretations*), 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.

Having that said, we propose the source classification system depicted in Figure 2, distinguishing between three levels of inner structure in the digitized historical source: *structured*, *semi-structured* and *unstructured*. Each level of structure can be divided into several *types of structure*. In turn, dotted arrows express how typical workflows tend to transform data contained in unstructured sources, identifying entities, relations and events mentioned in natural (unstructured) language, and modeling them into formal (structured) languages.

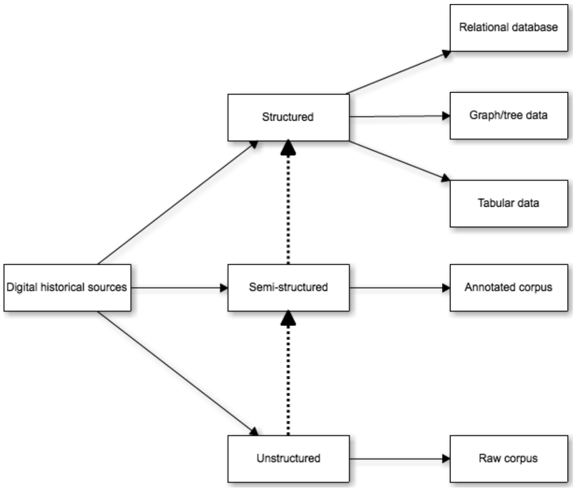


Fig. 2. Classification of sources according to their degree of internal structure.

Structured historical sources can be divided into *relational databases*, *graph/tree data* and *tabular data*. Relational databases are the most fixed, static 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 [86] with some concrete conception of reality in mind, but as a drawback their schemas are the most common obstacle in data integration. On the other hand by converting them into a flat file this problem is easy to overcome. Other solutions are converting the data into a more common data structure with standardized variables such as the Intermediate Date Structure in use by large databases with historical micro-data [1]. More serious are misconceptions of the data itself (bad conceptualization of entities and relationships, inconsistent normalization etc.) that troubles a historian maybe more than other scientists.

Graph/tree data is found in historical samples that come in formats such as XML (trees), RDF (graphs) or JSON. 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), these formats also follow some assumptions to put structure on historical data.

In tabular data, typically represented by means of comma-separated values (CSV[11]) or Excel workbooks (XLS[12]) (although other tabular data formats[13] are also common), historical information is encoded following a model based on cells, columns and rows, which suits well with datasets containing essentially numerical variables. Census spreadsheets are a good example of structured tabular data.

Semi-structured sources consist solely of *annotated corpora*. Although they appear more often not as sources, but as some intermediate representation between unstructured and structured historical data representations, annotated corpora can be treated as historical sources as well. Typical technologies applied here are markup languages, such as XML, to denote special characteristics of historical texts in specific regions of the corpus.

Unstructured sources are the most common representation for digital historical samples. They use to be digital transcriptions of historical texts. Objects with a high variety of historical nature can be included in this category: letters, books, memoranda, acts, etc.

3.2. Information problems

The weapon of choice of historians was and remains the database, particularly in relational form [2]. This not only enables 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. Although many advances have been made in different field and computers are seen as valuable assets, a vast amount of historians are unknown with or remain unconvinced that semantic technologies may become a new methodological asset [2,31]. Therefore, needless to say historians typically do research using their *own* datasets, resulting in the creation of a vast amount of scattered data and specific technological challenges. By understanding the use and advantages of semantic technologies, practitioners and researchers of historical 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. In historical research, information problems can be divided into four main categories [86]. Namely, information problems of historical sources, information problems of relationship between sources, information problems in historical analysis and information problems of the presentation of sources or analysis. In section 3.2 we will elaborate on these issues, except historical analysis. In section 3.2.1 we will go into the issue of semantic operability and historical sources.

Historical Sources As historians often have different interpretations and no clear research question when starting an investigation, it is neither possible nor desirable to model the data according to certain requirements in advance. The main information problem with historical sources relates to the fact that different sources have been produced throughout different periods in history with different views and motives. These sources often vary in format, structure, are not consistent, often unclear or ambiguous but also incomplete. Historical census data is a great example of these inconsistencies, varying structures and changing levels of detail which hinders comparative social history research both in past and present efforts [26]. Moreover, in historical research the meaning of data cannot exist without interpretations [86]. Textual problems for example include the meaning of words, its relation to other objects, the context or underlying thoughts of the data which are subject to different interpretations. Due to drifting concepts in history, different interpretations could exist with regards to certain data. However as interpretation of data is a subjective matter, this information should be added in a non destructive way, preserving the original source data. Another main issue relates to the data structuring problem of historical data. As historical researchers often deal with various (isolated) sources, they often face the problem of how to integrate these dissimilar sources for their purposes and have to decide on what is an adequate data model for historical data. The main discussion regarding this involves whether to use a source or a goal oriented data

[11] See RFC 4180: http://tools.ietf.org/html/rfc4180

[12] http://en.wikipedia.org/wiki/Microsoft_Excel#File_formats

[13] http://en.wikipedia.org/wiki/Comparison_of_spreadsheets

model for historical data. 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 wthe database is supposed to be shared with other researchers are is to be used in the future [21].

Relationships Between Sources Quite often several sources are used in historical research which makes linking different sources another key problem in historical research. Think of for example micro data of the same person contained in different censuses, parish registers, marriage or death certificates etc. The challenges are here on how to relate the appearance of a person in one source document to the appearance in another document. 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. Other problems related with the spatial and temporal context of historical data are for example, occupational titles which evolve over time or the problem of changing geographical boundaries of a country (compare for example the contemporary geographic position of Poland with the situation in 1930 and in 1900). 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 [86].

Historical Analysis 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 [86]. For example, various statistical techniques are borrowed from the social sciences, multilevel regression or techniques which have been specifically developed for historical research such as *event history analysis*.

Presentation of Sources One of the main problems of comparative historical researchers relates to an adequate presentation of sources. As noted before the presentation of historical information may take different shapes varying from digitized documents, bad and good modeled databases to visualizations and GIS-representations. Currently there is a great need for tools and methods to present changes over time and space. Moreover for historians, different types of presentations are suitable at different stages of a research project.

3.2.1. Semantic Interoperability

So far, we have presented a classification of digital historical sources according to their level of structure and the software they may have been encoded with. Moreover, we have stated that further workflows that researchers may need to apply to these sources strongly depend on such structural types. The same could be said for challenges of semantic interoperability challenges that these workflows may encounter while extracting and transforming information contained in the sources. This section analyzes which specific semantic challenges may occur in these processes, assuming that disparate sources of the same kind (e.g. spreadsheets coming from a dozen sources, XML documents with different data structuring) have to be commonly aligned (that is, being able to uniformly querying them).

Structured digitized sources

Relational databases. Relational databases have their own languages (SQL) and systems (MySQL, Postgres) to represent and store historical data. Semantic issues, especially when trying to merge different entity-relationship schemas from disparate databases, appear often in historical datasets encoded this way. The field of data integration is currently producing results in this respect.

Graph/tree sources. Hierarchical sources may be unable to be merged or compared due the following semantic mismatches.

- *Schema mismatch* occurs when two different sources cannot be compared because semantic differences in their defining schemas. For instance, two XML files conformant to different DTD schemas may define and structure differently the same historical fact, event or person. Or, two RDF datasets may describe these historical facts, events or persons by the means of triples which use different (and not necessarily compatible) vocabularies.
- Accordingly, *constraints on values regarding range or type* may also mismatch across datasets, though 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*. Likewise, categorical values may be represented as numerical in other samples, or have spe-

cial restrictions (for instance, if a person has a job then he or she must be at least 16 years old).

Tabular data. Tabular sources, which are typically represented by means of Excel spreadsheet workbooks (XLS files) or CSV, may present a variety of semantic heterogeneities.

- *Variable mismatch* occurs when using variables which do not perfectly overlap. For example, one column *marital status* and *civil status* may refer to the same concept (i.e. the attribute stating if a person is married or not) along different sources.
- *Value mismatch* may happen even though having two sources using the same variable or concept. For example, two data tables may encode a variable *occupation*, but due to temporal, geographical or cultural differences values for that variable may not be comparable.
- *Problems with hierarchy in variables and values*. This problem is the best showed with tabular data. Since tables offer a very opened model consisting of cells arranged in columns and rows, data may be represented according to some particular logic of hierarchical columns and rows. For example, it is very common to find splitted and spanned columns in census tabular sources, as well as rows with hierarchies that classify territories, occupations or social classes.

Semi-structured sources

Annotated corpora. An annotated corpus is one of the most common examples of semi-structured data. They are usually raw historical texts with annotations on defined text sections, usually implemented with a markup language, like XML. However, semantic variations may be observed across these kind of datasets.

- The *annotation procedure* in which annotations have been generated can be different between samples. We distinguish here two ways of annotating corpora: *manual* annotations, and *automatic* (or semi automatic) annotations. If the corpus has been manually annotated, then divergence in criteria from one human expert to the other may cause semantic mismatches when trying to make the datasets mutually compatible. If the corpus has been annotated by some algorithm, then semantic heterogeneity may have been generated by the use of different algorithms or configurations.
- In a similar way to what happens with structured sources, especially hierarchical sources, the *schemas or vocabularies used for annotation may differ* to one historical sample to the other, producing additional semantic gaps.

Unstructured sources

Raw corpora. To bring structure into an unstructured soured using natural language processing (NLP) techniques seems the best option. Two major concerns appear during this process: one on how to extract semantics (meaning) from historical texts; and another on how to represent the extracted semantics in a structured way.

- *Different NLP pipelines and algorithms* can be chosen to extract relevant information from unstructured historical datasets. Due to the high variety of choices, several concerns regarding semantics arise here: the precision of the technique used (e.g. while detecting persons on a text, which ratio of them are indeed detected with respect to the total in the source), the pipeline being constructed (which strictly depends on the goal to be achieved), the nature of the text being analyzed, and the existence and availability of other related texts for supervised learning.
- Likewise, *different vocabularies and schemas* can be chosen to represent the extracted data. Languages like XML or RDF can be useful for representing data in a more structured way, climbing to a more formal representation for concepts, individuals and relationships between them contained in historical texts. This also concerns which vocabulary or schema is more appropriate for convenient historical data modeling.

4. Semantic technologies for Historical Information Science

So far, we have come to a general description of the state-of-the-art in historical computing, highlighting the life cycle of historical data, describing convenient classifications for sources, and identifying the most relevant issues that history scholars still have to face in the colliding borders of history and computing. As we show in this section, semantic technologies can provide solutions to some of these problems. First we offer a brief description of the relevant semantic technology, and then we describe how these technologies can be (and, in some cases, are currently being) applied to digital historical sources. Since it is intended as an overview on Semantic Web, most Semantic Web advocated may prefer to skip the former and go directly to the latter.

4.1. The Semantic Web and semantic technologies

Semantic technologies do not require much introduction in an article in a Semantic Web journal. Still, we would like to briefly refer to this notion for scoping and reference. Throughout the paper we use *Semantic technology* and *Semantic Web technology* in the most general possible way, hereby referring both to the existing standard technology as well as to novel extensions to come. More concretely, Semantic technologies are based on formal (usually symbolic) representation languages where some meaning is encoded separately from data and content. Without overcommitting to any specific formalisms in the analysis of current work in the historical domain we usually start our analysis with the standardized Semantic Web languages and data models in mind, such as the Resource Description Framework (RDF[14]), the Web Ontology Language (OWL[15]) and the SPARQL Protocol and RDF Query Language (SPARQL[16]).

The purpose of this article is wider, though, than to study particular languages or paradigms. In general we are interested in challenges and potential of Semantic technology, which implies that any extension, variant or alternative of such methods is highly relevant to the work presented in this paper. This not just applies to alternative languages for representing vocabularies, such as SKOS[17] or Provenance (such as ProVO[18]) but also to novel semantic formalisms, such as multi-dimensional, temporal, contextualized, spatial languages (and more more of this kind) that are currently considered as extensions of the standardized formalisms.

4.2. How can semantic technologies help with historical information science challenges?

So far, we have presented the most significant challenges regarding semantics in historical information science, as well as an overall introduction on the Semantic Web and linked data. But, how are semantic technologies supposed to help solving these historical computing semantic issues? This section presents a general view on how Semantic Web facilities like RDF* and OWL can (and actually are) helping historians to solve semantic gaps in their datasets, and describes deeper the relations between historical datasets issues and identified semantic tools.

In section 3 we have discussed five well known problems history scholars face when dealing with digital historical datasets: problems with historical sources, problems with relationships between sources, problems with historical analysis, and problems with historical data presentation and visualization. Problems with data interoperability belong more to the computing domain, but they must be taken into account, for example, in scenarios where disparate sources represented with different data models or formats need to exchange information or have to be queried uniformly. Still, those problem domains are too widely defined to be of help to analyze how semantic technologies can fit each. Therefore, we introduce six more specific problems or challenges we think semantic technologies have been proven to be helpful (see Figure 3). Among them are the *lack of formalized historical domains*. Classification and ontologies do exist, but not for all areas, not in Semantic Web languages and not always agreed upon. The absence of mechanisms for automatic inference, so that new, *implicit historical knowledge* can be derived is another issue. While there exist quite a lot of different data and information sources, not always they are 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. This is why we defined the existence of *non-interoperable historical data sources* as a separate issue. Hereby, we particularly point to the co-existence of incompatible data models, which profoundly hampers exchange of information. It is considered to be a bad practice in historical research to not get your historic data modeling right at the beginning. What is valuable in terms of scientific rigor can become a barrier for the comparison of different historic data models. Comparison requires a shared framework and a flexibility of the models to be able to match them to one another. We marked that with the notion of *non-flexibility of historical data modeling*. At the end, that enforces history scholars to make their data selection and processing dependant of a certain data model that can not be (easily) replaced or altered if needed; this can happen usually in environments with very changeable requirements (like research) or requirements creep [17]. The non-flexibility of data models is related to a non-flexibility of historical data transformations. Digital historical data sources get altered in the life cycle of historical information

[14] http://www.w3.org/TR/rdf-primer/

[15] http://www.w3.org/TR/owl-ref/

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

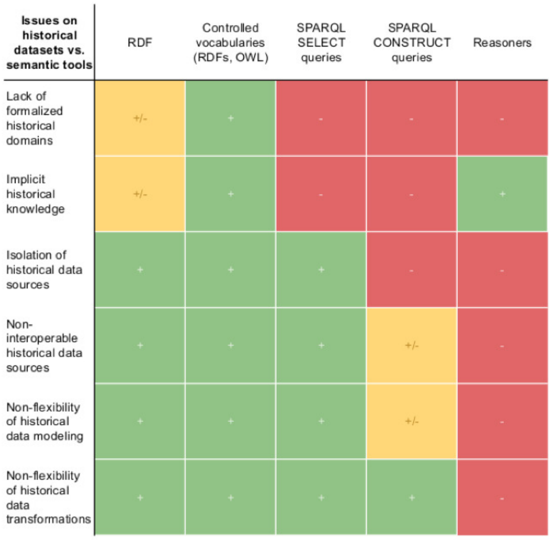
[17] http://www.w3.org/2004/02/skos

[18] http://www.w3.org/2011/prov/

(see Figure 1). But, if update, enrichment, analytic and interpretative operations are not controlled those transformation lead to a variety of historical data representations which can hardly be related to each other any more, nor in terms of provenance nor in terms of relatedness.

Some of these concrete issues are, obviously, subsets of the broader ones analyzed in section 3; for instance, one may say that *isolation of historical data sources* is part of the problem of relationships between historical sources, or *non-interoperable historical datasets* is part of the problem of semantic interoperability.

Figure 3 compiles both general historical data issues identified and semantic technologies, and offers as well an overall perspective on which semantic tools may be useful for historical computing in order to solve issues typically found during the development of projects. Green cells, also marked with a (+) sign, denote semantic technologies which are strongly related to concrete solutions for a historical computing problem. Yellow cells, also marked with a (+/-), identify tools which are related to the described problems, but in a less degree than the ones marked in green (+). Finally, red cells, also marked with a (-) sign, are assigned to technologies which, though in some cases could make sense to apply, are less related as solutions to those problems. We describe in the following some steps how to address those problems. While doing this we also indicate which of the five selected semantic technologies are useful for which problem, underlying the decisions behind Figure 3, about the usefulness of which technology for which problem.



Issues on historical datasets vs. semantic tools	RDF	Controlled vocabularies (RDFs, OWL)	SPARQL SELECT queries	SPARQL CONSTRUCT queries	Reasoners
Lack of formalized historical domains	+/-	+	-	-	-
Implicit historical knowledge	+/-	+	-	-	+
Isolation of historical data sources	+	+	+	-	-
Non-interoperable historical data sources	+	+	+	+/-	-
Non-flexibility of historical data modeling	+	+	+	+/-	-
Non-flexibility of historical data transformations	+	+	+	+	-

Fig. 3. Semantic technologies (columns) supporting historical computing issues (rows).

4.2.1. *Creating controlled vocabularies and ontologies for historical information as a way to address lack of formalization*

The first problem found when trying to model historical datasets into RDF is the lack of controlled vocabularies for describing historical facts. Although some ontologies have been developed for describing events[19] (such as the Simple Event Model [35]), these models are insufficient for the vast amount and variety of historical data that still has to be published in the web of data, especially when key issues for historians like *interpretations* or *evidences* need to be modeled and conveniently linked as well. Historical ontologies and vocabularies have been a reality in recent approaches. OWL ontologies describing classes and properties of some historical concern, such as concepts around the Pearl Harbor attack in 1941 [90], are an exciting modeling exercise for researchers but also a necessary step for better structuring historical information in the web. OWL ontologies and RDF 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 an historical reality depending on the temporal dimension of events described [91]? Moreover, how can differences in meaning and relations between concepts be traced, as time and historical realities change these concepts [36]? These questions, which comprise semantic technologies, knowledge acquisition and knowledge modeling techniques, are not yet completely understood and are a significant challenge in semantic historical information science research. On the other side over the centuries dictionaries, thesaurus, classification systems have been developed which are relevant. Think in terms of classification of historic occupations, or historic dictionaries or lexicons. 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

[19] http://labs.mondeca.com/dataset/lov/

specific tools and computer sciences, often relying on much broader and generic ontologies.

4.2.2. *Inferencing or implicit knowledge discovery by means of ontologies*

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.

Indeed, OWL historical ontologies can be used to facilitate historical knowledge discovery through inference, using OWL reasoners. Assuming that a particular domain is completely formalized (as long as OWL logic capabilities allow users to do so), then it is possible to run a reasoner to highlight derived, inferred facts that were not present in the original model as explicit knowledge (i.e. knowledge that was directly introduced by the user as input), but that were there as implicit knowledge. For instance, if an ontology describes, on the one hand, the fact that a letter was sent from one government diplomatic institution to another, and on the other hand, the fact that government diplomatic institutions have a person responsible of sending and receiving letters, then it may be possible for the reasoner to infer those concrete persons that sent and received, respectively, the mentioned letter. As the knowledge base grows and gets bigger and bigger, implicit knowledge can not be as evident as it was in the beginning, and reasoners may facilitate an enormous work and produce highly valuable pieces of historical knowledge.

4.2.3. *Data integration as a mean to link between isolated data sources*

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 URI 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.

Although this kind of studies can be (and actually are) made, applying semantic technologies and Semantic Web principles like in the example above may provide significant improvements. First, data retrieval is automatically solved by SPARQL: once pointed to the appropriate endpoint(s), triples needed to answer the query are fetched automatically. Second, data comes with warranty of the authority providing the SPARQL endpoint: well-known domain names embedded in entity-defining URIs are the signature of the institution issuing the data. Third, potential semantic heterogeneity problems with the data are solved if appropriate vocabularies are applied. Fourth, SPARQL technology provides handy mechanisms to perform automatically data transformations needed, according to user requirements. Last but not least, results from SPARQL queries can be customized in structure (i.e. which variables are returned) and format (raw text, HTML, JSON, CSV, etc.), which makes easy a direct plug to visualization tools (and in general any tool consuming the requested data).

4.2.4. *Data interoperability and new search & retrieval possibilities*

Data interoperability is one of the major challenges to be achieved in historical databases. Very often, research efforts tend to divide research tasks into several teams, project groups or even research calls, fragmenting the data modeling process across several database schemas, and thus making data definitely impossible to query in an uniform way. In section 3 we have identified a set of semantic interoperability problems that may appear depending on the type of structure in the source historical datasets.

How can these data problems be addressed? A major advantage of linked data is that exposed RDF datasets, when published through controlled vocabularies, are interoperable enough to enable cross-querying without manually solving semantic or schema heterogeneities. That is, they can be uniformly queried as long as they share standard vocabularies to expose common factual knowledge. SPARQL queries can homogeneously query these endpoints as long as the user knows the vocabularies being used in the remote graphs.

4.2.5. *Flexibility, changeability and mapping of data models*

It is a well known problem that the choice of a particular data model to represent historical data is a critical issue for most historical computing projects. Moreover, the election of some ÔappropriateÕ data model may seem a good design decision at some stage of the project, but new requirements, research directions or stakeholder priorities may convert that data model (that once upon a time was the most suitable one) into a nightmare. Indeed, flexibility of data with respect to the data model used to represent historical facts is something desired to avoid restructuring entire databases again and again.

Applying semantic technologies and linked data principles to historical datasets may have a major advantage regarding historical data workflows: a complete flexibility regarding the historical data modeling process. As explained before, two different approaches regarding historical data modeling have been followed traditionally in historical computing: the *source-oriented* representation, and the *model-oriented* (also known as *goal-oriented*) representation [86]. The source-oriented representation tries to design a database schema which concerns and respects the historical source structure. As opposed to this, model-oriented representation models historical data according with user or application goals, and thus committing to a particular view on the historical reality being encoded. Both have advantages and drawbacks, and several examples follow an hybrid approach between them.

Semantic technologies allow a flexible representation of historical datasets. RDF represents factual knowledge by means of triples, thus using a fine-grained granularity for expressing data. With this in mind, RDF triplestores can act as a middleware representation of further views on the data [23], which can be modeled 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 workflow design, or adopted as early as the user may desire, providing complete flexibility regarding the data schema. This feature also prevents historians of designing their databases over and over again, avoiding spending resources on data migrations.

4.2.6. *Flexibility regarding data transformations (non-destructive updates)*

Another big problem when dealing with historical sources is supporting data transformations under two constraints: (a) without modifying source data (so the originals stay intact); and (b) with the ability to trace all the changes performed on data. Destructive updates are, thus, a major concern while selecting, aggregating and modifying data. On the one hand, modifications to specific kinds of digital files (such as CSV, spreadsheets or XML documents) do not support non-destructive updates: if one wants to keep original data, some versioning system has to be maintained so previous statuses can be retrieved. On the other hand, relational databases can be inefficient if all transformations, edits and manipulations have to be recorded using a certain provenance model.

Non-destructive updates and specific data views are both supported by the `CONSTRUCT` and `SELECT` SPARQL queries. The former allows the construction of RDF triples according to the supplied graph pattern, facilitating data transformations without altering consistency of previous factual knowledge in the knowledge base. The latter, as described before, permits a certain selection of triples (again, according to some graph pattern) and disposing the data they contain according to any desired view-format (for example, columns matching some interesting historical variables). This way, SPARQL can provide equivalent features to the `UPDATE` and `SELECT` SQL queries, plus non-destructive updates and the ability to retrieve previous data transformations and statuses, via the appliance of some provenance model for linked data.

5. Current Historical Semantic Web

In section 3 and section 4 we reviewed concepts and problem definitions as articulated in historical information science and mapped them to approaches developed by the Semantic Web community. The goal of this operation was to identify possible bridge heads for common problem solving processes on a conceptual level. In the preparation for such a shared conceptualization we reviewed a variety of publications, but also projects, datasets, and technologies relevant for the foundation of a long-term collaboration between specialists from both fields. We build the collection starting from some key publications [86], main journals (*Historical Methods*, *Journal of the Association for History and Computing*), and conducted about 8 interviews with pioneers in this area in the Netherlands. we accompanied this by an extensive web search. Still, we are aware that our selection covers only a small part of the relevant literature.

Another specific characteristics of this survey is that the information steams from two at least two different communities: history and computer sciences. That makes the set very heterogenous, but also very special. Eventually our selection is driven by the curiosity to identify relevant contributions to *semantic research questions*. Here, with *semantic* we donÕt point exclusively to an understanding as cultivated in the computer sciences. We identify *semantic research questions* as shared objects of concern both by historians and by computer sciences. How to achieve such a shared understanding has been elaborated in the previous sections. In this final section we present an analysis and a categorization of 67 sources we have found to be relevant for depicting the roadmap for a historical Semantic Web.

The sources are divided into the categories of scientific papers, research projects, online resources (like presentations or online articles), and tools and technologies (like ontologies, demos, applications or programming libraries). Figures 4, 5, 6, 7 and 8 show these sources and a classification based on the general goals they are aiming at. We have used the colors green, yellow and red, and accordingly the signs (+), (+/-) and (-) to express that a particular work is strongly related, more or less related, or weakly related, respectively, to each one of these general goals.

The first identified semantic research question is *knowledge modeling and ontologies*, and under this category we consider contributions that intensively apply technology to model historical knowledge or historical facts, essentially using semantic technologies such as RDF, OWL and SPARQL. In *text processing and mining* we gather all the work that at some stage deals with unstructured text, facing problems like text storage in convenient database systems, or automatic or semi-automatic entity extraction (such events or persons) via NLP techniques. 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 for information retrieval. Finally, under the category of *semantic interoperability* we analyze to what extent contributions consider the problem of data integration and use the Semantic Web to deal with that problem, which often faces data model mismatches, schema incompatibilities and disparate source formats. In the following we describe those generic semantic questions in greater detail, and we review selected papers shown in FIgures 4, 5, 6, 7 and 8 to illustrate how the identified questions are reflected in the literature. At the end there are a couple of sources which can not been disambiguously allocated to one of the four questions. Those we review under a last group called *holistic approaches*.

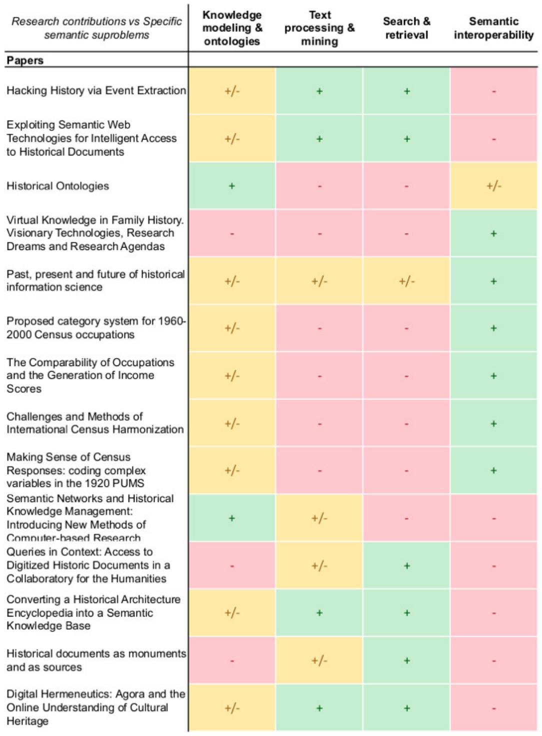
5.1. Knowledge modeling and ontologies

Under this category we have grouped work that contributes to a semantically enabled historical web by the following main streams of research: classifying historical knowledge by means of classification systems; building and publishing historical structured data; and designing historical data models and ontologies.

5.1.1. Classification systems (and census management)

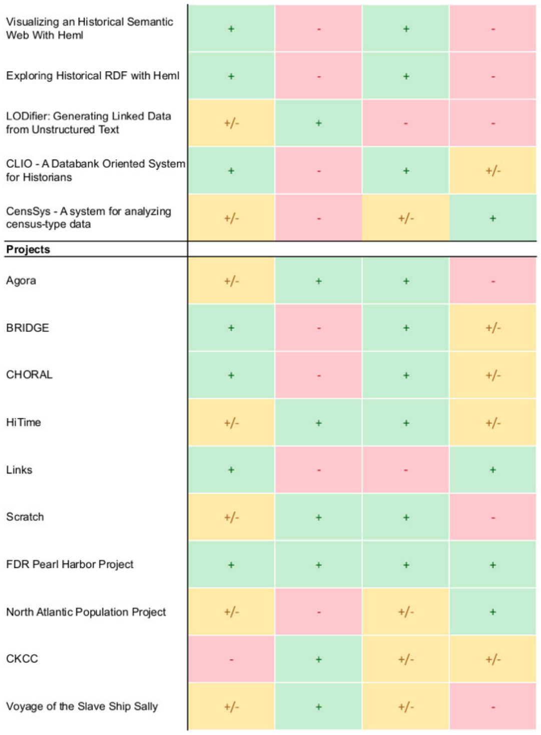
When dealing with vast amounts of historical data, classification systems are a necessity in order to organize and make sense of the data. The main goal of a classification system is therefore to put things into meaningful groups [4]. This entails an allocation of classes which are created according 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. 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 [27].



Research contributions vs Specific semantic suproblems	Knowledge modeling & ontologies	Text processing & mining	Search & retrieval	Semantic interoperability
Papers				
Hacking History via Event Extraction	+/-	+	+	-
Exploiting Semantic Web Technologies for Intelligent Access to Historical Documents	+/-	+	+	-
Historical Ontologies	+	-	-	+/-
Virtual Knowledge in Family History. Visionary Technologies, Research Dreams and Research Agendas	-	-	-	+
Past, present and future of historical information science	+/-	+/-	+/-	+
Proposed category system for 1960-2000 Census occupations	+/-	-	-	+
The Comparability of Occupations and the Generation of Income Scores	+/-	-	-	+
Challenges and Methods of International Census Harmonization	+/-	-	-	+
Making Sense of Census Responses: coding complex variables in the 1920 PUMS	+/-	-	-	+
Semantic Networks and Historical Knowledge Management: Introducing New Methods of Computer-based Research	+	+/-	-	-
Queries in Context: Access to Digitized Historic Documents in a Collaboratory for the Humanities	-	+/-	+	-
Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base	+/-	+	+	-
Historical documents as monuments and as sources	-	+/-	+	-
Digital Hermeneutics: Agora and the Online Understanding of Cultural Heritage	+/-	+	+	-

Fig. 4. Research contributions to historical Semantic Web (rows) related to specific semantic issues (columns) 1/5.



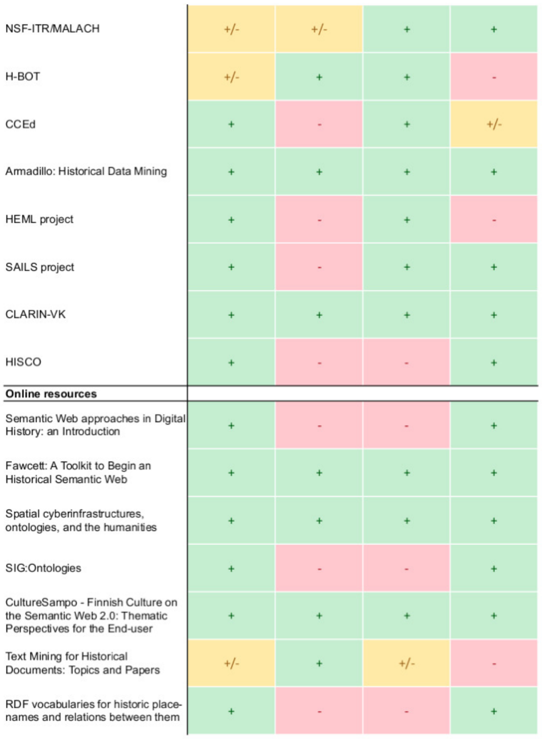
Visualizing an Historical Semantic Web With Heml	+	-	+	-
Exploring Historical RDF with Heml	+	-	+	-
LODifier: Generating Linked Data from Unstructured Text	+/-	+	-	-
CLIO - A Databank Oriented System for Historians	+	-	+	+/-
CensSys - A system for analyzing census-type data	+/-	-	+/-	+
Projects				
Agora	+/-	+	+	-
BRIDGE	+	-	+	+/-
CHORAL	+	-	+	+/-
HiTime	+/-	+	+	+/-
Links	+	-	-	+
Scratch	+/-	+	+	-
FDR Pearl Harbor Project	+	+	+	+
North Atlantic Population Project	+/-	-	+/-	+
CKCC	-	+	+/-	+/-
Voyage of the Slave Ship Sally	+/-	+	+/-	-

Fig. 5. Research contributions to historical Semantic Web (rows) related to specific semantic issues (columns) 2/5.

However, major changes in the classification and coding of the different censuses, have hindered comparative historical research in both past and present efforts [26]. 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 modeling 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 IPUMS International project 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 (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 (theoretically) 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 Code of Occupations (HISCO). As occupation is one of the most problematic variables in historical research (SOURCES), HISCO aims to overcome the problem of changing occupational terminologies over time and space. This system combines various kinds of information on tasks and duties in historical setting, by classifying tens of thousands of occupational titles and linking these to short descriptions and images of the content of the work[20]. Naturally there are many other clas-

[20] http://hisco.antenna.nl/

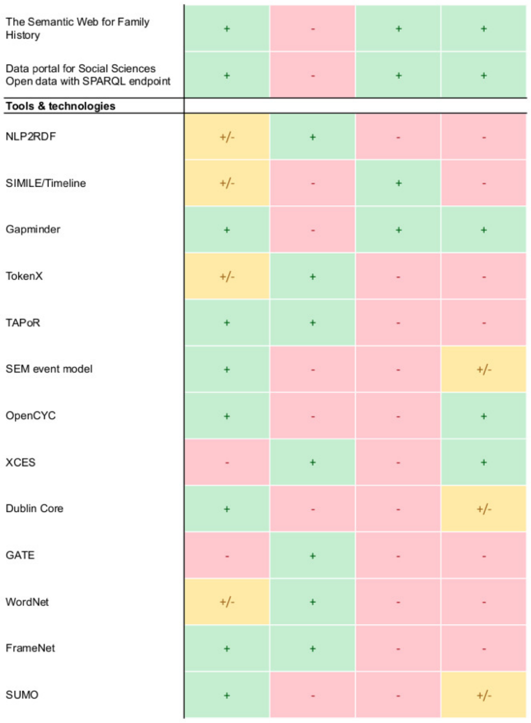


NSF-ITR/MALACH	+/-	+/-	+	+
H-BOT	+/-	+	+	-
CCEd	+	-	+	+/-
Armadillo: Historical Data Mining	+	+	+	+
HEML project	+	-	+	-
SAILS project	+	-	+	+
CLARIN-VK	+	+	+	+
HISCO	+	-	-	+
Online resources				
Semantic Web approaches in Digital History: an Introduction	+	-	-	+
Fawcett: A Toolkit to Begin an Historical Semantic Web	+	+	+	+
Spatial cyberinfrastructures, ontologies, and the humanities	+	+	+	+
SIG:Ontologies	+	-	-	+
CultureSampo - Finnish Culture on the Semantic Web 2.0: Thematic Perspectives for the End-user	+	+	+	+
Text Mining for Historical Documents: Topics and Papers	+/-	+	+/-	-
RDF vocabularies for historic place-names and relations between them	+	-	-	+

Fig. 6. Research contributions to historical Semantic Web (rows) related to specific semantic issues (columns) 3/5.

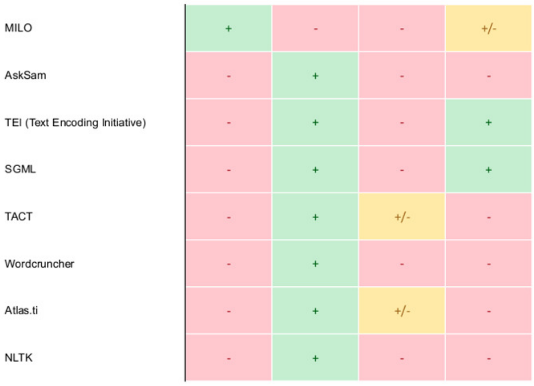
sification systems which aim to put historical data into meaningful groups. However the common characteristic of all these historical classification systems is that they contain loads of semantics which offer a great stimulus for the field of historical information science to look for novel semantic technologies, such as RDF, in order to deal with old problems with contemporary techniques.

There are two main imperatives when applying any classification or model on historical data. When dealing with historical data it is important to decide in an early stage whether the data should be modeled according to a source and 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. In recent years special attention has been given to modeling techniques such as RDF-based representations. By opting for a Linked Data model such as RDF, we can avoid an early commitment to a certain data model and build a DB in a non destructive manner. 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



The Semantic Web for Family History	+	-	+	+
Data portal for Social Sciences Open data with SPARQL endpoint	+	-	+	+
Tools & technologies				
NLP2RDF	+/-	+	-	-
SIMILE/Timeline	+/-	-	+	-
Gapminder	+	-	+	+
TokenX	+/-	+	-	-
TAPoR	+	+	-	-
SEM event model	+	-	-	+/-
OpenCYC	+	-	-	+
XCES	-	+	-	+
Dublin Core	+	-	-	+/-
GATE	-	+	-	-
WordNet	+/-	+	-	-
FrameNet	+	+	-	-
SUMO	+	-	-	+/-

Fig. 7. Research contributions to historical Semantic Web (rows) related to specific semantic issues (columns) 4/5.



MILO	+	-	-	+/-
AskSam	-	+	-	-
TEI (Text Encoding Initiative)	-	+	-	+
SGML	-	+	-	+
TACT	-	+	+/-	-
Wordcruncher	-	+	-	-
Atlas.ti	-	+	+/-	-
NLTK	-	+	-	-

Fig. 8. Research contributions to historical Semantic Web (rows) related to specific semantic issues (columns) 5/5.

goals which are mainly dependent on expert knowledge. Accordingly, this perspective uses more sound models to start with and commits to a sound model in an early stage.

5.1.2. Building and publishing historical structured data

Although not being pure initiatives to publish historical datasets in Semantic Web formats, there are so much contributions extracting and modeling historical information that cannot be described here. However, some researchers in history have centered their interest in how semantics can help relating and linking historical sources and entities, despite of the underlying technology: *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* [92]. Indeed, the landscape on current projects exposing structured historical information extracted from unstructured sources shows a tendency on having more and more contributions exposing data in RDF.

There is a huge variety of nature in projects looking for that structure, though not doing so solely (or explicitly) in RDF. For instance, the CKCC project in the Netherlands tries to formalize an epistolary network for circulation of knowledge in Europe in the 17th century, extracting such knowledge from the correspondence of scientific scholars of that time. On a similar line, the CCed project does so with clerical careers from the Church of England Database. While these projects mine the historical sources for important historical personalities, other approaches such the SAILS project dives into more concrete historical events and links World War I naval registries, although ship's personnel is also analyzed and linked as a primary goal. As the reader may have deducted, the common goal seems to produce that *semantic network of historical data containing objects like people, places and events connected to each other*, which overlaps with the general purpose of the Web of Data (WoD).

Other projects make explicit use of semantic technologies and expose their datasets using RDF, describing them with well known vocabularies and thus facilitating their linkage to others. For instance, the Agora project aims at formally describing museum collections and linking their objects with historical context, using the SEM (Simple Event Model) for those descriptions. Historical events, however, can be also extracted from other historical sources, like government letters and memoranda. This is the aim of the FDR Pearl Harbor project, that links events, persons, dates, and correspondence on the surroundings of the Pearl Harbor attack on 1941 between the US and Japanese governments: these entities are represented in RDF as well to model a graph of historical knowledge about that particular event. From a more socio demographic point of view, the Verrijkt Koninkrijk project links RDF concepts found on a structured version of De Jong's studies on *pillarization* of Dutch society after the World War II. This set of approaches, which share the RDF entity extraction step from historical sources can be summarized with the Fawcett toolkit and the Armadillo project. The latter exports RDF from any unstructured historical source, producing a particular graph of historical knowledge that encodes the historical entities and their relationships expressed in that source.

5.1.3. Designing historical ontologies and data models

Data models are necessary for giving structure to any historical data, since it is the abstract model that documents and organizes data properly for communication. In this direction, semantic data models are preferably developed in the Web Ontology Language (OWL), but other efforts do so with XML based formats.

Regarding XML we found the Historical Event Markup and Linking Project (HEML[21]), which aims at providing a tag, mark-up specific language for describing historical events. In a more concrete approach, The Semantic Web for Family History[22] 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 (TEI[23]) we find an interesting discussion building the bridge between XML and OWL in historical data: SIG: Ontologies[24] 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 externally pointed OWL ontology.

[21] http://heml.mta.ca/heml-cocoon/description#N10098

[22] http://jay.askren.net/Projects/SemWeb/

[23] http://www.tei-c.org/index.xml

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

In the pure semantic technology world, OWL historical ontologies have begun to be modeled. Though some interesting studies point out specific modeling needs of the historical domain (e.g. how historical ontologies should reflect how a particular time frame influences definitions of concepts [91]), most practical results show how the central concept of events is at the core of historical knowledge modeling. In this line we find the Simple Event Model (SEM) used to model RDF in the Agora project, but also the Event Ontology[25] and efforts on Linking Open Descriptions of Events[26], even though their usage may be subject to particular modeling needs.

5.2. Text processing and mining

Textual resources play an important role in history research, for example primary sources such as letters written by historically important people, but also secondary sources that may describe historical events or persons.

In the Netherlands, currently three projects are underway that are concerned with structuring historical information from textual resources for further analysis: Agora[27], Bridge[28] and HiTime[29]. The CCKC project[30] is also relevant in this context.

The Agora project aims to enrich museum collections with historical context in order to help users place museum objects in their historical contexts. The Agora project is thus explicitly aimed at the general public. 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 instances. From the object descriptions, also relevant historical entities are extracted which can then be linked to the events. In this fashion, it is then possible to relate an object such as the ship model of the Medusa[31] to the battle of Shimonoseki, even though the Battle of Shimonoseki.

The Bridge project aims to bring 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. The Bridge project is focused on improving access to television archives for media professionals. In order to do so, relevant entities are extracted from archives by using statistical natural language processing techniques. Furthermore, they will detect interesting events in television archives by detecting redundant stories.

The HiTime project 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 lifestory of these persons, and social networks of the persons they interacted with.

5.3. Improved search and retrieval

It is not a coincidence that a high number of papers, projects and tools that aim at RDF (or generally structured data) extraction of entities from historical sources also point at some desired system able to improve search and retrieval of those historical 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 (such persons or places) are mentioned, can be easily built and improved, especially comparing to pure textual, keyword-based search systems. The Agora (museum collections), Bridge (historical TV metadata), CHORAL (historical audio metadata), HiTime (biographical events), Verrijkt Koninkrijk (Dutch post-war social clusters concepts) and FDR Pearl Harbor (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 (with RDF, vocabularies and ontologies, but other tools are also used), entities structured this way can be used for a graph-based search and retrieval, for instance through SPARQL queries. Other projects, like the H-BOT[32] project, use a natural language interface instead of a query system for querying such historical structured knowledge (in this case, mined from historical Wikipedia articles).

Indexation of historical contents is another way of improving search and retrieval of historical sources.

[25] http://purl.org/NET/c4dm/event.owl#

[26] http://linkedevents.org/ontology/

[27] http://agora.cs.vu.nl

[28] http://ilps.science.uva.nl/node/735

[29] http://ilk.uvt.nl/hitime

[30] http://ckcc.huygens.knaw.nl/

[31] http://www.rijksmuseum.nl/collectie/NG-MC-485/halfmodel-van-het-schroefstoomschip-medusa

[32] http://chnm.gmu.edu/tools/h-bot/

Indexing and historical data storage systems have a long tradition [86]. Being CLIO [100] a traditional example of such a system, nowadays indexing is performed by XML annotation-oriented approaches, such as HEML (which also includes facilities for visualizing and searching for concrete historical entities through sources) or TEI with ontologies support, with bridges again markup languages and Semantic Web formats. These initiatives should consider the emerging RDFa, microformats and microdata technologies to envisage their fitting in the vast domain of historical text annotation systems.

5.4. Semantic interoperability

Semantic interoperability has much to do with data integration, namely, how to commonly query and uniformly represent data that come from disparate sources (i.e. fitting several, probably non-compatible data models). There is a high number of publications dealing with this problem, especially in classification systems [88,89,94,99]. The HISCO coding system[33] for historical occupations deals directly with such a problem: occupations are encoded very differently if some work has to be performed against a variety of historical sources that come from a different time, region or language.

Semantic heterogeneity of historical sources is especially present on social historical projects, which also have this problem of having very disparate sources: the LINKS project (reconstruction of families), the CEDAR project[34] (exposure of ancient Dutch census data in Semantic Web formats), and the North Atlantic population project (exposure of microdata of several Atlantic countries) share this problem of data harmonization, in which heterogeneity of sources requires an intense work on how to resolve data model inconsistency among datasets. The work developed in *Spatial cyberinfrastructures, ontologies, and the humanities* [29] has also to be mentioned in the more general context of the humanities, but also concerning history: its deep analysis deals with how semantic heterogeneity can be addressed exclusively with semantic technologies, achieving success in environments with very disparate data models.

5.5. Holistic approaches

Finally, there are few, but longitudinal contributions we have classified as being *holistic*, because they cover a complete workflow on semantic issues present in historical datasets, as well as solve some concrete historical problem (see section 3) with the application of some specific semantic technology (see section 4).

The Agora project is one of these contributions, as 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. In a related line of research we can tailor the Fawcett toolkit and the Verrijkt Koninkrijk project; while the former also extracts RDF event-oriented triples from unstructured texts, and additionally allows historian researchers to install a full semantic toolbox with fancy widgets to experiment with their data, the latter is another example on how historical RDF datasets get much more rich when they are linked between them. The FDR Pearl Harbor project also contributes on this line, applying NLP to disparate sources, and additionally opening the very promising field of historical knowledge inference through the formalization and usage of historical OWL ontologies. A good abstract approach, which somehow summarizes and contains the generic plot behind all these contributions, is the Armadillo architecture of Semantic Web Services, which covers all aspects of historical data mining and historical RDF generation in the context of the Semantic Web. The NSF-ITR/MALACH also fits this frame of complete, meaningful workflows, but this time being the perfect candidate for multimedia digital sources (historical videotapes) instead of unstructured texts..

Additionally, there exist some theoretical studies envisaging possibilities on how the Semantic Web can enhance research by historians. The most remarkable one is *Past, present and future of historical information science* [86], a major work on the evolution of historical computing, eHistory and historical information science, which gives a deep intuition on how computer science approaches can help to solve ancient problems in history research.

6. Conclusions

In this paper we have presented a general overview of semantic technologies applied to historical datasets, as well as a survey of recent contributions to this line of

[33] http://hisco.antenna.nl/

[34] http://www.cedar-project.nl/

research. First, we have described a general approach to historical computing, introducing its core elements, such as the historical data life cycle and meaningful classifications for historical sources, putting a special emphasis on how structure is a key characteristic of digital sources to take into account when applying a semantic pipeline. Then, we have overviewed several problematic areas history scholars have to face while working with digital historical datasets. After that, a resume on semantic technologies has allowed us to examine how these technologies can help in solving some of these historical data problems, offering solutions for interlinking datasets, solving semantic interoperability problems, or enabling historical knowledge inferencing. Finally, we have presented a collection of contributions that provide advances in some areas of semantic historical computing.

It is important to realize that current developemnts flagged out as eHumanities or Digital Humanities pave the way for an integration of semantic technologies in what one could call eHistory. While this paper reflects debates between historians and computer scientists we would like to underline that a semantic enrichment of historical sources brings new facilities, on the one hand, for humanities researchers also outside of history, allowing them to search, retrieve and compare information they need for their everyday work using a variety of dimensions and scopes; and on the other hand, for practitioners, giving them new data sources to develop historical-aware applications for public institutions, private companies and citizens.

We claim that semantic technologies are suitable for representing inner semantics implicitly contained in historical sources, which can be appropriately identified, formalized and linked using the cited tools. With the appropriate pipelines, algorithms can extract entities from digital historical sources and transform these occurrences into RDF triples according to some RDFs or OWL provided historical vocabulary of convenience, linking those entities between them and with other external, linked datasets for enrichment, contributing to a open, world wide, online persistent graph of historical linked knowledge. All the work presented in this survey have some relevance in one stage or another in this graph-building pipeline, providing solutions for the previously analyzed problems in historical computing.

Acknowledgements

COMMIT, Comp. Humanities / KNAW, CEDAR project. Special thanks: D. Roorda, O. Boonstra.

References

[1] George Alter, Kees Mandemakers, and Myron P. Gutmann. Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Historical Social Research*, Vol. 34(No. 3 = No. 129):78–114, 2009.

[2] Ian Anderson. History and computing. *Making History*, 2008.

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

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

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

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

[7] 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.

[8] Stefan Dormans and Jan Kok. An alternative approach to large historical databases. Exploring best practices with collaboratories. *Historical methods*, 43(3):97–107, 2010.

[9] Albert Esteve and Matthew Sobek. Challenges and Methods of International Census Harmonization. *Historical Methods*, 36(2):37–41, 2003.

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

[11] Fox. Nine step process of historical research. 1969.

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

[13] Good and Scates. Methods of research. 1972.

[14] Bernhard Haslhofer, Rainer Simon, Robert Sanderson, and Herbert Van de Sompel. The open annotation collaboration (oac) model. *CoRR*, abs/1106.5178, 2011.

[15] 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.

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

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

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

[19] Jan Kok and Paul Wouters. Virtual Knowledge in Family History: Visionary Technologies, Research Dreams, and Research Agendas. In Paul Wouters, Anne Beaulieu, Andrea

Scharnhorst, and Sally Wyatt, editors, *Virtual Knowledge. Experimenting in the Humanities and the Social Sciences*, page xxx. MIT Press, Cambridge, Mass., 2013.

[20] Thomas Kuczynski, editor. *Wirschaftsgeschichte und Mathematik*. Akademie-Verlag, Berlin, 1985.

[21] Kees Mandemakers and Lisa Dillon. Best Practices with Large Databases on Historical Populations. *Historical Methods*, 37(1):34–38, 2004.

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

[23] Albert Meroño Peñuela, Ashkan Ashkpour, Laurens Rietveld, Stefan Schlobach, and Rinke Hoekstra. Data Harmonization: a Structured Approach - With a Case-study in Historic Census Data. 2012.

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

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

[26] Bart Van De Putte and Andrew Miles. A Social Classification Scheme for Historical Occupational Data. *Historical Methods*, 38(2):61–92, 2005.

[27] Steven Ruggles and Russell R Menard. The Minnesota Historical Census Projects. *Historical Methods*, 28(1):6–10, 1995.

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

[29] 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*, 2011.

[30] Matthew Sobek. The comparability of occupations and the generation of income scores. *Historical Methods*, 1, 1995.

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

[32] M Thaller. Automation on Parnassus. CLIO - A databank oriented system for historians. *Historical Social Research*, 15, 1980.

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

[34] D. B. Van Dalen. Understanding educational research. *New York: McGraw-Hill*, 1979.

[35] Willem Robert van Hage, VŐronique MalaisŐ, Roxane H Segers, Laura Hollink, and Guus Schreiber. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2), 2011.

[36] Shenghui Wang, Stefan Schlobach, and Michel C. A. Klein. Concept drift and how to identify it. *J. Web Sem.*, 9(3):247–265, 2011.

Surveyed work

[37] Agora project. `http://agora.cs.vu.nl`.

[38] Armadillo: Historical data mining project. `http://www.hrionline.ac.uk/armadillo/armadillo.html`.

[39] Asksam. `https://www.asksam.com/`.

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

[41] Bridge project. `http://ilps.science.uva.nl/node/735`.

[42] Cced project. `http://www.theclergydatabase.org.uk/publications/jeh_article.html`.

[43] Choral project. `http://hmi.ewi.utwente.nl/choral/`.

[44] Ckcc project. `http://ckcc.huygens.knaw.nl/`.

[45] Clarin-vk project. `http://verrijktkoninkrijk.nl/`.

[46] Culturesampo - finnish culture on the semantic web 2.0: Thematic perspectives for the end-user. `http://www.museumsandtheweb.com/mw2009/papers/hyvonen/hyvonen.html`.

[47] Data portal for social sciences open data with sparql endpoint. `http://www.rechercheisidore.fr/`.

[48] Dublin core. `http://dublincore.org/`.

[49] Fawcett: A toolkit to begin an historical semantic web. `http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/175/217`.

[50] Fdr pearl harbor project.

[51] Framenet. `https://framenet.icsi.berkeley.edu/fndrupal/`.

[52] Gapminder. `http://www.gapminder.org/`.

[53] Gate. `http://gate.ac.uk/`.

[54] H-bot project. `http://chnm.gmu.edu/tools/h-bot/`.

[55] Heml project. `http://heml.mta.ca/heml-cocoon/description`.

[56] Hisco project. `http://hisco.antenna.nl/`.

[57] Hitime project. `http://ilk.uvt.nl/hitime/`.

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

[59] Milo. `http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/KBs/Mid-level-ontology.kif`.

[60] Nlp2rdf. `http://nlp2rdf.org/`.

[61] Nltk. `http://www.nltk.org/`.

[62] North atlantic population project. `http://www.nappdata.org/napp/`.

[63] Nsf-itr/malach project. `http://malach.umiacs.umd.edu/`.

[64] Opencyc. `http://www.opencyc.org/`.

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

[66] Sails project. `http://sailsproject.cerch.kcl.ac.uk/2010/07/about-the-sails-project/`.

[67] Scratch project. `http://www.ai.rug.nl/alice/nwo-catch-scratch/index_english.html`.

[68] Sem event model. `http://www.cs.vu.nl/~guus/papers/Hage11b.pdf`.

[69] Semantic web approaches in digital history: an introduction. `http://www.slideshare.net/mpasin/semantic-web-approaches-in-digital-\history-an-introduciton`.

[70] The semantic web for family history. `http://jay.askren.net/Projects/SemWeb/`.

[71] Sgml. `http://en.wikipedia.org/wiki/Standard_Generalized_Markup_Language`.

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

[73] Simile/timeline. `http://www.simile-widgets.org/timeline/`.

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

[75] Sumo. `http://sigmakee.cvs.sourceforge.net/viewvc/sigmakee/KBs/Merge.kif`.

[76] Tact. `http://projects.chass.utoronto.ca/tact/`.

[77] Tapor. `http://portal.tapor.ca/portal/portal`.

[78] Tei (text encoding initiative). `http://www.tei-c.org/index.xm`.

[79] Text mining for historical documents: Topics and papers. `http://www.coli.uni-saarland.de/courses/tm-hist/readings.html`.

[80] Tokenx. `http://tokenx.unl.edu/`.

[81] Voyage of the slave ship sally project. `http://www.stg.brown.edu/projects/sally/`.

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

[83] Wordnet. `http://wordnet.princeton.edu/`.

[84] Xces. `http://www.xces.org/`.

[85] Isabelle Augenstein, Sebastian Padó, and Sebastian Rudolph. Lodifier: Generating linked data from unstructured text. In Axel Polleres Oscar Corcho Valentina Presutti Elena Simperl, Philipp Cimiano, editor, *Proceedings of the 9th Extended Semantic Web Conference*, volume 7295 of *LNCS*, pages 210–224. Springer, Mai 2012.

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

[87] Panos Constantopoulos, Martin Doerr, Maria Theodoridou, and Manolis Tzobanakis. Historical documents as monuments and as sources. -, 2009.

[88] Albert Esteve and Matthew Sobek. Challenges and Methods of International Census Harmonization. *Historical Methods*, 36(2):37–41, 2003.

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

[90] 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.

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

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

[93] Jan Kok and Paul Wouters. Virtual Knowledge in Family History: Visionary Technologies, Research Dreams, and Research Agendas. In Paul Wouters, Anne Beaulieu, Andrea Scharnhorst, and Sally Wyatt, editors, *Virtual Knowledge. Experimenting in the Humanities and the Social Sciences*, page xxx. MIT Press, Cambridge, Mass., 2013.

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

[95] Jan Oldervoll. Censsys Ñ a system for analyzing census-type data. *Computer Applications in the Historical Sciences: Selected Contributions to the Cologne Computer Conference, pages = 17–22, year = 1989*.

[96] B. G. Robertson. Visualizing an historical semantic web with Heml. *Proceedings of the 15th international conference on World Wide Web*, pages 1051–1052, 2006.

[97] Bruce Robertson. Exploring Historical RDF with Heml. *Digital Humanities Quarterly*, (1). `http://www.digitalhumanities.org/dhq/vol/003/1/000026.html`, volume = 3.

[98] 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. `http://doi.acm.org/10.1145/1999676.1999705`.

[99] Matthew Sobek. The comparability of occupations and the generation of income scores. *Historical Methods*, 1, 1995.

[100] Manfred Thaller. Clio - a databank oriented system for historians. *Historical Social Research, Historische Sozialforschung 15*, 1980.

[101] 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. *Aufsatz in Buch*, 2007.

[102] Chiel van den Akker et al. Digital hermeneutics: Agora and the online understanding of cultural heritage. *WebSci 2011 Proceedings*, 2011.

[103] Rene Witte, Ralf Krestel, Thomas Kappler, and Peter C. Lockemann. Converting a historical architecture encyclopedia into a semantic knowledge base. *IEEE Intelligent Systems*, 25:58–67, 2010.