STOLE: A Reference Ontology for Historical Research Documents

Laura Pandolfo

DIBRIS, Università di Genova, Via Opera Pia, 13 – 16145 Genova – Italy
laura.pandolfo@edu.unige.it

Abstract. Historical documents are a relevant part of cultural heritage. It is well-established that this domain is very complex: data are often heterogeneous, semantically rich, and highly interlinked. For this reason, searching and linking them with related contents represents a challenging task. The use of Semantic Web technologies can provide innovative methods for more effective search and retrieval operations. In this paper we present STOLE, a reference ontology that provides a vocabulary of terms and relations that model the history of Italian public administration' domain. STOLE can be considered a first step towards the development of a knowledge exchange standard on this specific domain.

1 Context and Motivation

Historical texts and documents are considered an important component of cultural heritage. Among other sources, historical newspapers and journals archived in libraries represent a valuable source of information for historians.

In the last years, several portals and digital libraries in the cultural heritage field have been enhanced with Semantic Web (SW) technologies in order to implement methods for more effective search and retrieval operations. In fact, they can offer effective solutions about design and implementation of user-friendly ways to access and query content and meta-data – see [11] for a survey in the historical research domain. The use of SW technologies can improve the productivity of research in historical domain, since they help to identify implicit and explicit knowledge included in the documents, e.g., reference to a historical person contained in a historical source can be discovered and related to other entities, for example events in which that person has been involved in.

One of the main elements of the SW infrastructure are ontologies. They provide a shared and clear representation for a specific domain and they may play a major role in supporting knowledge extraction, knowledge discovery, and data integration processes. An ontology is usually defined as a formal specification of domain knowledge conceptualization. A *Knowledge Base* (KB) is composed of two elements, i.e., a *Terminological Box* (TBox) and an *Assertional Box* (ABox). The TBox represents the intensional knowledge of the KB and it is made up of classes of data and relations among them. In the following, we use the terms TBox and ontology interchangeably to denote the conceptual part of a KB. The

ABox contains extensional knowledge, which is specific to the individuals of the domain.

In this paper we present STOLE[1], a reference ontology that provides a vocabulary of terms and relations with which it is possible to model the history of Italian public administration' domain. STOLE represents a first step towards the development of a knowledge exchange standard on this specific domain. To the best of our knowledge, this is the first time that ontological modeling has been undertaken for this specific domain.

The paper is organized as follows: in Section 2 we describe the specific case study that motivated this research. In Section 3 we present the STOLE ontology and its design process. This ontology is built on an extended version of STOLE [1]. We conclude the paper in Section 4 by identifying stimulating directions and challenging issues for continued and future research in this application domain.

2 Case Study

In the last years the way of doing historical research has deeply changed. From traditional research in archives and libraries, which necessarily required the presence on site and had precise limits and constraints, the digitization of historical documents allows new perspectives and ways to conduct research. The use of SW technologies has upset times, costs, methods of historical research, and is also changing the so-called culture of the document.

The heritage of the history of public administration represents a fundamental element to understand the history of the Italian institutions as well as the history of the country in general. One of the main sources used in this field of research is represented by historical text and documents, including journals and newspapers of the age. Recently, several web sites concerning specific domain databases on Italian institutions, e.g., the Bibliography of Italian Parliament [2] and Internet Archive [3], offer to the scholars the opportunity to easily access the sources, which are sometimes totally unknown.

The development of the STOLE ontology responds to the needs of some researchers of Department of History of the University of Sassari which, since the 1980s, have been involved in a project designed to collect and digitalize historical journals regarding the origin and the evolution of institutions, customs, usages, and rules in the Italian public administration. As a result, the ARAP[4] digital archive of the University of Sassari was created and it actually collects a large amount of information about some of the most relevant journal articles published between 1848 and 1946 concerning the legislative history of public administration in Italy. The main goal of ARAP was to offer to the scientific community

[1] STOLE is the acronym for the Italian "STOria LEgislativa della pubblica amministrazione italiana", that means "Legislative History of Italian Public Administration".

[2] `http://bpr.camera.it`

[3] `https://archive.org`

[4] ARAP is the acronym for the Italian "Archivio di Riviste sull'Amministrazione Pubblica", that means "Archive of Journals on Italian Public Administration".

interested in those documents, such as historians, lawyers, political scientists, and sociologists, a repository of important sources, otherwise hardly accessible.

The journal articles included in ARAP have a remarkably value for the wealth of information they contain. In fact, starting the research from the journals could make a positive and significant contribution to the current knowledge. For example, through the study of these documents it is possible to establish connections between authors, institutions, persons and historical events mentioned in an article. Specifically, the link between an author and the people cited can reveal a lot, such as the political and cultural reference of the author. Clearly, the relationship between an institution and an historical event can offer support to understand the evolution of public administration. The use of certain concepts or the recurring of names in relation to some historical events can be an indicator of a trend.

In this context, STOLE represent the core element of the ARAP digital archive and its main goal is to to model historical concepts and gain insights into this specific field in order to support historians in their research tasks.

3 The STOLE Ontology

The main steps of the STOLE ontology design process concerned the identification both of the key concepts for this specific domain, and the proper language for the TBox implementation. Moreover, we populated the ontology, i.e., filling the ABox with semantic annotations. A team of domain experts was involved during the whole ontology development process. In particular, they contributed at the early stage in order to define the key issues related to the application domain.

In the first phase, we detected the main categories of data expressed in the considered historical documents. The results of this process enabled us to detect three categories of elements: 1) Data concerning the authors of the articles, e.g., name, surname and biography; 2) Data concerning the journal and the article, e.g., article title, journal name, date and topics raised in the article; 3) Data concerning some relevant facts and people cited in the article, e.g, people, historical events, institutions. In this specific domain, historical analysis is based on these categories of information and focused on the interrelations among these data. For instance, the relation between an author and the people mentioned in an article could provide valuable information to historians, e.g., if an author has often referred to Giuseppe Mazzini then it could easily be interpreted that this author was favourable to the republic.

During the second phase, the TBox of the STOLE ontology has been designed building on some existing standards and meta-data vocabularies, such as Bibliographic Ontology (BIBO) [5], Bio Vocabulary (BIO) [6], Dublin Core (DC) [7], Friend

[5] `http://bibliontology.com`
[6] `http://vocab.org/bio/0.1/.html`
[7] `http://dublincore.org`

of A Friend (FOAF) [8], and Ontology of the Chamber of Deputies (OCD) [9]. In details, BIBO has been used to describe information about documents, for instance `bibo:volume`, `bibo:issue`, `bibo:pageStart` and `bibo:pageEnd`, denoting values of volume, issue, page start and page end of an article, respectively. Both BIO and FOAF have been used in order to describe information about people. In the case of FOAF, we reused terms such as `foaf:person`, `foaf:firstName`, `foaf:surname`, `foaf:gender`. Other information about people, their relationships and the events in their lives have been described using BIO concepts, such as `bio:birth`, `bio:death`, `bio:event`, `bio:place`, `bio:biography`. From DC we reused concepts related to the structure and characteristics of a document, e.g, `dc:title`, `dc:isPartOf`, `dc:publisher`, `dc:description`. Considering OCD, we noticed that it is a relevant source for STOLE ontology since, for example, most part of the authors of articles related to the history of the public administration topics during their lives were also involved in government activities. In particular, OCD provided us valuable concepts in order to give detailed information about people involved in political offices. Finally, the usage of these core ontologies allows both extensibility and interoperability of STOLE with other resources and applications.

In dealing with the modeling language, we decide to use OWL2 DL [8] since it allows to properly model the knowledge for our application domain by means of constructs like cardinality restrictions and other role constraints, e.g., functional properties. The TBox is composed of 268 axioms, 19 classes, 34 object properties, and 33 data properties.

In the following, we describe main classes of the STOLE ontology.

`Article` represents our library, namely the collection of historical journal articles. Every instance of this class has data properties such as `articleTitle`, `articleDate`, `pageStart`, and `pageEnd`.

`Institution` is used to represent the public institutions cited in the articles. This class contains four subclasses, namely `Central`, `Local`, `PoliticalInstitution` and `EconomicInstitution`. Notice that `Central` and `Local` are disjoint classes. The same holds for `PoliticalInstitution` and `EconomicInstitution`. In this way, for example, an institution can be both a local institution and a political institution at the same time.

`Jurisprudence` is a subclass of `Article` and contains a series of verdicts which are entirely written in the articles. Every individual of this subclass has the following data properties: `verdictDate`, `verdictTitle`, and `byCourt`.

`Law` is also a subclass of `Article`, and it contains a set of principles, rules, and regulations set up by a government or other authority which are entirely written in the articles. This subclass has data properties such as `lawDate` and `lawTitle`.

`Event` denotes relevant events. It contains five subclasses modeling different kinds of events: `Birth` and `Death` are subclasses related to a person's life;

[8] `http://www.foaf-project.org`
[9] `http://data.camera.it/data`

`BeginPublication` and `EndPublication` represent the publication period of a journal; `HistoricalEvent` contains the most relevant events that have marked the Italian history.

`Journal` denotes the collection of historical journals. This class has data properties such as `journalArticle`, `publisher`, and `issn`.

`Person` is the class representing people involved in the Italian legislative and public administration history. This class contains one subclass, `Author`, that includes the contributors of the articles. Every instance of this class has some data properties as `firstName`, `surname`, and `biography`.

`Place` represents cities and countries related to people and events.

`Subject` is a class representing topics tackled in the historical journals.

In conclusion, there are other examples of ontologies designed for the cultural heritage domain, e.g., CIDOC CRM [5], however these are rather generic to be applied in our context where a detailed modeling of the singular domain is needed. Despite its specific nature, STOLE could be used in several fields of research, e.g., admnistrative law, political science, history of institutions.

4 Current Work and Open Problems

Currently, we are extending and improving in many ways the implementation of this work. First, we are dealing with a key issue for historians, namely how to disambiguate individuals and how to manage changing names, e.g., different people with the same name or institutions that changed name retaining the same functions. This point still represents an open challenge in this application domain– see [11].

Another open problem relates the ontology population process. For the current version, STOLE has been populated leveraging a set of annotated historical documents comprised into the ARAP archive. Semantic annotations were provided by a team of domain experts and individuals were added to the ontology by means of a JAVA program built on top of the OWL APIs [9]. This activity requires specialized expertise, it is time consuming and resource-intensive. Given these reasons, we are studying to find solutions for its automatization on the basis of some recent contributions – see, e.g.,[7, 10, 13, 6]. Most approaches for automatic or semi-automatic ontology population process from texts are based on the following techniques: Natural Language Processing [4], Machine Learning [2], and Information Extraction [12]. Once the ontology will be fully populated, we are planning to perform an experimental analysis on the STOLE ontology involving state of the art DL reasoners on both classification and query answering tasks.

Acknowledgments I would to thank the anonymous reviewers for their valuable suggestions, which were helpful in improving the final version of the paper. Moreover, I would also like to thank my supervisors, Giovanni Adorni and Luca Pulina, for their support, and Salvatore Mura and Prof. Francesco Soddu for the valuable discussions about the application domain.

References

1. Adorni G., Maratea M., Pandolfo L. Pulina L. An Ontology for Historical Research Documents. Web reasoning and Rule Systems. Springer (2015) 11–18
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
3. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - A Crystallization Point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web **7**(3) (2009) 154–165
4. Dale, R., Moisl, H., Somers, H.L.: Handbook of Natural Language Processing. CRC (2000)
5. Doerr, M.: The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Mag. **24** (3) (2003) 75–92
6. Faria, C., Serra I., Girardi, R.: A Domain-Independent Process for Automatic Ontology Population from Text. Journal of Science of Computer Programming **95** (2014) 26–43
7. Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E.: Semantically Enhanced Information Retrieval: An Ontology-based Approach. Web Semantics: Science, Services and Agents on the World Wide Web **9**(4) (2011) 434–452
8. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: Owl 2: The next step for owl. Web Semantics: Science, Services and Agents on the World Wide Web **6**(4) (2008) 309–322
9. Horridge, M., Bechhofer, S.: The OWL API: A Java Api for OWL Ontologies. Semantic Web **2**(1) (2011) 11–21
10. Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N.K., Alpaslan, F.N.: An Ontology-based Retrieval System Using Semantic Indexing. Information Systems **37**(4) (2012) 294–305
11. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F.: Semantic Technologies for Historical Research: A survey. Semantic Web Journal (2014)
12. Moens, M.F.: Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer Netherlands (2006)
13. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based Semantic Similarity: A New Feature-based Approach. Expert Systems with Applications **39**(9) (2012) 7718–7728