

# An Ontology for Historical Research Documents

Giovanni Adorni<sup>1</sup>, Marco Maratea<sup>1</sup>, Laura Pandolfo<sup>1</sup>, and Luca Pulina<sup>2</sup>(✉)

<sup>1</sup> DIBRIS, Università di Genova, Via Opera Pia, 13, 16145 Genova, Italy  
{adorni,marco.maratea}@unige.it, laura.pandolfo@edu.unige.it

<sup>2</sup> POLCOMING, Università di Sassari, Viale Mancini N. 5, 07100 Sassari, Italy  
lpulina@uniss.it

**Abstract.** In this paper we present the conceptual layer of STOLE, our ontology-based digital archive aiming at helping historical researchers to organize data, extract information and derive new knowledge from historical documents.

## 1 Context and Motivation

Historical documents are considered a rich and valuable source of information related to, e.g., events and people used by researchers and scholars to investigate history. In the last decades, the digitization of historical documents has been mainly focused on developing applications that enable users to access, retrieve and query information in a highly efficient way [1]. In fact, historical documents are characterized by being syntactically and semantically heterogeneous, semantically rich, multilingual, and highly interlinked. They are usually produced in a distributed, open fashion by organizations like museums, libraries, and archives, using their own established standards and best practices [2].

It is well-established that ontologies can offer a clear conceptual representation and they provide a valuable support to knowledge extraction, knowledge discovering, and data integration. More, they can offer effective solutions about design and implementation of user-friendly ways to access and query content and meta-data – see [3] for a survey in the historical research domain.

In this paper we describe the STOLE<sup>1</sup> ontology, that represents the conceptual layer of our ontology-based digital archive. The main goal of the STOLE ontology is to clearly model historical concepts and, at the same time, to gain insights into this specific field, e.g., supporting historians to find out some unexplored but useful aspects about a particular event or person. STOLE collects information about some of the most relevant journal articles published between 1848 and 1946 concerning the legislative history of public administration in Italy. These documents are regarded as an estimable source of information for historical research since through the study of these texts it is possible to trace the course of Italian history.

---

<sup>1</sup> STOLE is the acronym for the Italian “SToria LEgislativa della pubblica amministrazione italiana”, that means “Legislative History of Italian Public Administration”.

The rest of the paper is organized as follows: in Sect. 2 we provide some explanations about the main steps and the key decisions which marked the ontology design process, and we describe the STOLE ontology in detail. In Sect. 3 we briefly describe the architecture of our ontology-based digital archive in order to provide an overall view of the system. Finally, we conclude the paper in Sect. 4 with some final remarks and future work.

## 2 The STOLE Ontology

### 2.1 Design At-a-Glance

Narrative and statistical documents represent the main sources used by historians to conduct their research. Typically, historians want to extract the facts from these documents in order to gather information for reconstructing specific historical events.

Our design process derived from the needs of the researchers of Department of History of the University of Sassari which, since the 1980s, have been involved in a project designed to collect, digitalize and catalogue historical journals concerning genesis and evolution of the Italian public administrations and institutions. The research was conducted on a wide selection of magazines owned by the following Italian institutions' libraries: Central Archives of the State, Chamber of Deputies, Supreme Court of Cassation, University of Bologna and University of Sassari. As a result, the ARAP<sup>2</sup> archive of the University of Sassari was created and it actually collects a large amount of narrative sources. Currently, historians can access to several websites which allow them to flip between pages of relevant documents, however effective analysis tools are rarely provided by these applications. Semantic web technologies can address some specific issues in historical domain, since they allow to identify implicitly and explicitly knowledge included in the documents. For example, reference to a historical person contained in a historical source can be discovered and related to other entities, e.g., events in which that person participated, providing a rich representation from which historians can extract meaningful knowledge to their research [3].

In the following, we can summarize the main phases of the creation process:

1. Identification of key concepts.
2. Identification of the proper language and Tbox implementation.
3. Ontology population, i.e., filling the Abox with semantic annotations.

In the first step the domain experts have been involved in order to contribute to the definition of key issues related to the application domain. In particular, we detected the main categories of data expressed in the considered historical documents. The results of this process enabled us to compute a taxonomy composed of the following three elements:

---

<sup>2</sup> ARAP is the acronym for the Italian “Archivio di Riviste sull’Amministrazione Pubblica”, that means “Archive of Journals on Italian Public Administration”.

**Table 1.** Tbox statistics about the STOLE ontology. These data are computed by PROTÉGÉ [5] in the *Metrics* view.

Classes	14
Axioms	440
Object properties	30
Data properties	29

- Data concerning the author of the article, e.g., name, surname and biography.
- Data concerning the journal and the article, e.g., article title, journal name, date and topics raised in the article.
- Data concerning some relevant facts and persons cited in the article, e.g., persons, historical events, institutions.

Historical analysis in this specific domain is based on the above information and focused on the interrelations between these data. For example, the link between an author and the people cited in an article provides valuable information to historians, e.g., if an author has often referred to King Vittorio Emanuele II probably it can easily be interpreted as favorable to the monarchy.

Regarding the second point, the Tbox of the STOLE ontology has been designed building on some existing standards and meta-data vocabularies, such as Dublin Core (<http://dublincore.org>), FOAF (<http://www.foaf-project.org>), the Bio Vocabulary (<http://vocab.org/bio/0.1>), the Bibliographic Ontology (<http://bibliontology.com>), and the Ontology of the Chamber of Deputies (<http://dati.camera.it/data/en>). In particular, the latter is an ontology aiming at modeling the domain of the Chamber during its history. It can be a relevant source since, for example, most part of the authors in our archive were also involved in government activities.

Concerning the modeling language, our choice fall to OWL2 DL [4]. This language allows us to have proper expressivity, and to model the knowledge for our application by means of constructs like cardinality restrictions and other role constraints, e.g., functional properties.

## 2.2 Implementation

In the following, we describe main classes, object properties and data properties of our ontology. Statistics are summarized in Table 1.

**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**.

**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: **sentenceDate**, **sentenceTitle**, 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**.

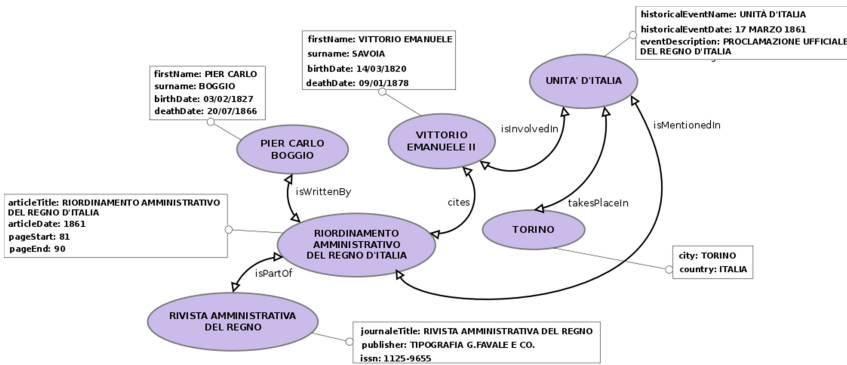
**Event** denotes relevant events. It contains five subclasses modeling different kinds of events: **Birth** and **Death** are subclasses related to a person's life; **BeginPublication** and **EndPublication** represent the publication period of a journal; **HistoricalEvent** contains the most relevant 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.



**Fig. 1.** Example of individuals and their relationships in the STOLE ontology. Ellipses denotes individuals, while information related to their data properties are reported in boxes. Object properties (and their inverse) are denoted by arrows.

Concerning object properties, we describe in the following the ones related to **HistoricalEvent**. The full documentation of the STOLE ontology is available at <http://visionlab.uniss.it/STOLE.DOC>. **HistoricalEvent** has a crucial role in our ontology, and their relationships with other classes are extremely useful for historical research in order to highlight connections between events, people, and articles. In Fig. 1 we show a graphical example of these relations. Looking at the figure, we can see that **Unità d'Italia** is an instance of **HistoricalEvent**, and represents one of the most important historical event in Italian history. This event is in relation to individuals in **Article** by the **isMentionedIn** property: this relationship defines in which articles is mentioned the event **Unità**

d'Italia. The `hasWritten` object property relates an article to its author, and `isPartOf` shows in which journal has been published that specific article. With reference to `HistoricalEvent` class, there are further interesting relationships to emphasize:

- the `takesPlaceIn` object property makes explicit the event's location – in the example depicted in Fig. 1 is the city of Torino;
- `InvolvedIn` connects individuals in `Person` class to a particular event, taking into account all people that played a role in a given historical event.

Summing up, the example shown in Fig. 1 hints at potentially interesting relationships among elements that can be represented by the STOLE ontology. Our ontology-based application – that will be described in the next section – supports historians in their research, providing to them relations between events, people and documents in an automated way.

Currently, there are no other examples of ontologies that model this particular domain. However, despite its specific nature, STOLE ontology can be used in different application field that relates the history of the Italian administrations and institutions.

Finally, the ontology 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 [6].

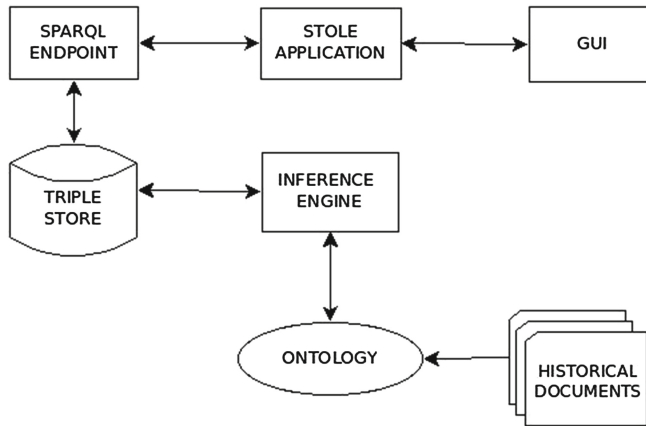


Fig. 2. The architecture of STOLE.

### 3 System Architecture

In Fig. 2 we report the architecture of our ontology-based digital archive built on top of the conceptual layer represented by the STOLE ontology. Looking at the figure, we can see that it is composed of the modules listed in the following:

TIMELINE • TABLE • TILES

45 Items

sorted by: [about](#), [articleDate](#), [partOf](#), and [autlabel](#); then by... • ☐ grouped as sorted

**Atti del Parlamento (1)**

1. <b>Atti del parlamento nazionale</b> Agostino Aliberti, Giuseppe Prato, Paolo Aliberti, and Vincenzo Aliberti <b>Atti del Parlamento</b>	Luigi Zenone Quaglia, Carlo Alberto, and Giovanni Lanza	Instaurazione Parlamento in Piemonte
---	---	--------------------------------------

**Autorità amministrativa (2)**

**1852 (2)**

**Rivista Amministrativa del Regno (2)**

1. <b>Materie generali della responsabilità civile dei comuni</b> Agostino Aliberti, Giuseppe Prato, Paolo Aliberti, and Vincenzo Aliberti <b>Autorità amministrativa</b>	
<b>Materie generali del diritto d'associazione in rapporto all'autorità amministrativa</b> Pier Carlo Boggio	

**Rivista**

45 [Rivista Amministrativa del Regno](#)

---

**Argomento**

1 [Atti del Parlamento](#)

2 [Autorità amministrativa](#)

1 [Azioni giudiziarie dei Comuni](#)

2 [Beni comunali](#)

3 [Bilancio](#)

1 [Brigantaggio](#)

---

**Data pubblicazione**

2 [1850](#)

7 [1851](#)

4 [1852](#)

8 [1853](#)

6 [1854](#)

4 [1856](#)

FIG. 3. Screenshot of the STOLE web GUI.

**Ontology** is the ontology described in Sect. 2.2.

**Inference Engine** aims to accomplish both classification and consistency checking tasks on the STOLE ontology. It interacts with the **Ontology** in order to infer new knowledge to present to the user. Actually, we are using the HermiT reasoner [7].

**Triple Store** and **SPARQL Endpoint** are the modules devoted to store and query the knowledge base, respectively. For these purposes, we are currently using Open Virtuoso<sup>3</sup>.

**STOLE Application** is the module in which we implemented all functionalities related to query the **SPARQL Endpoint** and to process the answer in order to be presented to the user by means of the GUI.

**GUI** is devoted to the user-system interaction. This module is implemented on top of the SIMILE Exhibit API [8], a set of JavaScript files that allows to easily create rich interactive web pages including maps, timelines, and galleries with very detailed client-side filtering. This kind of representations, e.g. timeline of historical events, are widely used in the historical research field. Exhibit allows to display the result of SPARQL queries in JSON format. Figure 3 shows a screen-shot of the STOLE GUI.

## 4 Conclusions and Future Work

In this paper we described the development and implementation of an ontology for historical research documents, and we presented a general architecture overview about the related ontology-based archive.

Currently, we are dealing with a key issue for our domain experts, namely the management of changing names of, e.g., institutions that changed name retaining

<sup>3</sup> <http://www.openlinksw.com/>.

the same functions, across time and space. This particular point still represents an open challenge in this application domain – see [3].

Furthermore, we are developing a data integration layer in order to exploit information coming from relevant external sources, i.e. DBpedia [9] and the Ontology of the Chamber of Deputies, and integrate them in the STOLE ontology.

Concerning data navigation and visualization, we also intend to offer different ways to browse the STOLE resources, e.g., using interactive maps and LodLive<sup>4</sup> data graph representation.

We are also designing a Graphical User Interface to support the ontology population stage, in order to improve this process. More, concerning the ontology population, we are studying solutions for its automatization on the basis of some recent contributions – see, e.g., [10–12]. Finally, 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.** The authors wish to thank the anonymous reviewers for their valuable comments and suggestions to improve the paper. The authors would also like to thank Dott. Salvatore Mura and Prof. Francesco Soddu for the valuable discussions about the application domain. This work has been partially supported by MIUR.

## References

1. Kruk, S.R., Westerki, A., Kruk, E.: Architecture of semantic digital libraries. In: McDaniel, B., Krik, S.R. (eds.) *Semantic Digital Libraries*, pp. 77–85. Springer, Berlin (2009)
2. Ahonen, E., Hyvonen, E.: Publishing historical texts on the semantic web-a case study. In: 2009 IEEE International Conference on Semantic Computing, ICSC 2009, pp. 167–173. IEEE (2009)
3. 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* [under review] (2012). <http://www.semantic-web-journal.net/sites/default/files/swj301.pdf>
4. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: Owl 2: the next step for owl. *Web Seman. Sci. Serv. Agents World Wide Web* **6**(4), 309–322 (2008)
5. Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubézy, M., Eriksson, H., Noy, N.F., Tu, S.W.: The evolution of protégé: an environment for knowledge-based systems development. *Int. J. Hum. Comput. Stud.* **58**(1), 89–123 (2003)
6. Horridge, M., Bechhofer, S.: The owl api: a java api for owl ontologies. *Seman. Web* **2**(1), 11–21 (2011)
7. Shearer, R., Motik, B., Horrocks, I.: Hermit: A highly-efficient owl reasoner. In: *OWLED*, vol. 432 (2008)

---

<sup>4</sup> <http://lodlive.it>.

8. Huynh, D.F., Karger, D.R., Miller, R.C.: Exhibit: lightweight structured data publishing. In: *Proceedings of the 16th International Conference on World Wide Web*, pp. 737–746. ACM (2007)
9. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. *Web Seman. Sci. Serv. Agents World Wide Web* **7**(3), 154–165 (2009)
10. Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E.: Semantically enhanced information retrieval: an ontology-based approach. *Web Seman. Sci. Serv. Agents World Wide Web* **9**(4), 434–452 (2011)
11. Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N.K., Alpaslan, F.N.: An ontology-based retrieval system using semantic indexing. *Inf. Syst.* **37**(4), 294–305 (2012)
12. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: a new feature-based approach. *Expert Syst. Appl.* **39**(9), 7718–7728 (2012)



Web Reasoning and Rule Systems

9th International Conference, RR 2015, Berlin,  
Germany, August 4-5, 2015, Proceedings.

ten Cate, B.; Mileo, A. (Eds.)

2015, XVII, 131 p. 13 illus., Softcover

ISBN: 978-3-319-22001-7