An Ontology-Based Archive for Historical Research

Giovanni Adorni[1], Marco Maratea[1], Laura Pandolfo[1], and Luca Pulina[2]

[1] DIBRIS, Università di Genova, Via Opera Pia, 13 – 16145 Genova – Italy
giovanni.adorni@unige.it, marco.maratea@unige.it,
laura.pandolfo@edu.unige.it
[2] POLCOMING, Università di Sassari, Viale Mancini n. 5 – 07100 Sassari – Italy
lpulina@uniss.it

1 Context and Motivation

The digitalization of cultural materials is doubtless a key-enabler for increasing accessibility of cultural heritage documents, e.g., historical texts. In the last decade Semantic Digital Libraries (see, e.g., [1]) have attracted the attention of research communities coming from different research areas, such as Cultural Heritage, History, and Knowledge Engineering. In order to find more innovative methods to improve search and retrieval operations, recently portals and digital libraries concerning Cultural Heritage have been enhanced with Semantic Web [2] technologies. Such technologies can offer effective solutions about design and implementation of user-friendly ways to access and query content and metadata [1]. In this context, a prominent example is the Europeana project [3][3].

Particularly, cultural heritage documents are characterized by being syntactically and semantically heterogeneous, multilingual, semantically rich, and highly interlinked. They are produced in a distributed, open fashion by organizations like museums, libraries, and archives, using their own established standards and best practices [4]. Historical documents represent an important component of the cultural heritage field, and they have been digitized and published on the web by means of several applications.

In this extended abstract we present STOLE[4], an ontology-based digital archive collecting some of the most relevant journal articles published between 1848 and 1946 concerning the legislative history of public administration in Italy. In STOLE we leverage ontologies for describing domain knowledge and providing semantic information integration among data in order to support historians' research. The documents stored in STOLE are regarded as a valuable source of information for historical research since through the study of these texts it is possible to trace the course of Italian history and often to find out some unexplored but useful aspects about a particular event or person. Currently, these historical sources

[3] http://europeana.eu

[4] STOLE is the acronym of the Italian "*STOria LEgislativa della pubblica amministrazione italiana*", that means "legislative history of the Italian public administration".

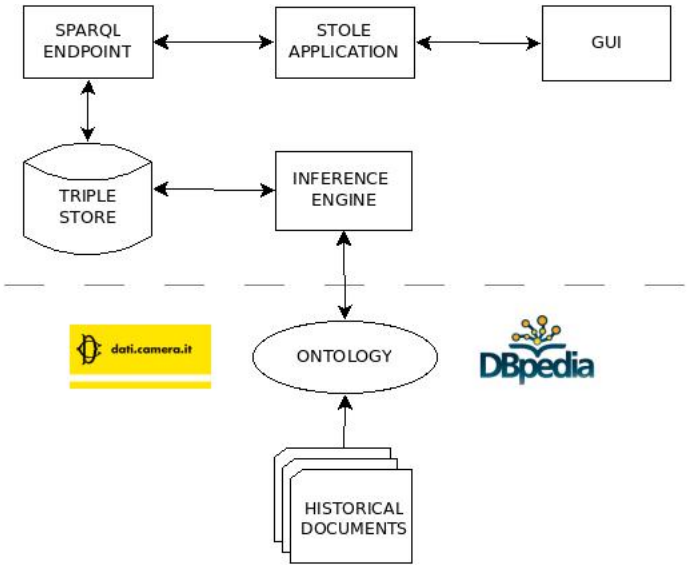

Fig. 1. The architecture of STOLE.

can be accessed going to the State Archives of Rome or to libraries that have a copy, but in both cases historians are not supported by information system during their research. Our ontology-based archive intends to handle this historical information providing search features and capabilities such that the user can search over this document collection. In the next section, we give details of the components of the system, while the abstract is concluded in Section 3 by listing the future work planned.

2 System Architecture

In Figure 1 we describe the architecture of STOLE; looking at the figure, we can see that it is composed of the modules listed in the following:

`Ontology` It represents the conceptual layer of our application. It is used to represent knowledge concerning journals, authors, cited people and events, as well as the relations between them. The DL expressivity if the STOLE[5] ontology is $\mathcal{ALCOIQ}(\mathcal{D})$, and it is composed of 2909 axioms. Actually, it is comprised of 342 individuals, its population is growing continuously. It has been developed building on existing standards and meta-data vocabularies, such as Dublin Core (`http://dublincore.org`), FOAF (`http://www.foaf-project.org`), Bio Vocabulary (`http://vocab.org/bio/0.1`) and the Bibliographic Ontology (`http://bibliontology.com`). Concerning the modeling ontology language, our choice falls to OWL2 DL [5]. We also

[5] The full documentation is available at `http://visionlab.uniss.it/STOLE_DOC`

considered some lightweight profiles, but we cannot avoid to introduce both cardinality restrictions and other role constraints, e.g., inverse object properties, in order to have proper expressivity for our application. The STOLE ontology has been populated leveraging on a set of annotated historical documents. Such semantic annotations were provided by a team of domain experts. Finally, we also developed a data integration layer in order to exploit information coming from external sources, such as DBpedia [6] and the Ontology of the Chamber of Deputies[6].

`Inference Engine` This module interacts with `Ontology` to check its consistency and to infer new knowledge to present to the user. Actually, we are using the HermiT reasoner [7].

`Triple Store` **and** `SPARQL Endpoint` They are the modules devoted to store and query the knowledge base, respectively. For these purposes, we are currently using Open Virtuoso[7].

`STOLE Application` In this module we implemented all functionalities related to query the data to `SPARQL Endpoint` and to process the answer in order to be presented to the user by means of the `GUI`.

`GUI` It is devoted to the user-system interaction and it aims at providing data representations widely used in the historical research field, such as timelines of historical events.

3 Conclusion and Future Work

We have presented an ontology-based archive and its application in the historical research domain, which includes numerous key aspects of Semantic Web. The current implementation of this framework can be extended in several ways. First of all, we are designing a Graphical User Interface to support the ontology population stage, in order to make this process of knowledge acquisition more interactive and dynamic. Finally, we are planning to introduce in STOLE a semantic indexing mechanism using Apache Solr[8], in order to extend keyword-based search functionalities.

[6] `http://dati.camera.it/data/en`
[7] `http://www.openlinksw.com/`
[8] `http://lucene.apache.org/solr`

References

1. Kruk, S.R., Westerki, A., Kruk, E.: Architecture of semantic digital libraries. In: Semantic Digital Libraries. Springer (2009) 77–85
2. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american **284**(5) (2001) 28–37
3. Haslhofer, B., Isaac, A.: data. europeana. eu: The europeana linked open data pilot. In: International Conference on Dublin Core and Metadata Applications. (2011) 94–104
4. Ahonen, E., Hyvonen, E.: Publishing historical texts on the semantic web-a case study. In: Semantic Computing, 2009. ICSC'09. IEEE International Conference on, IEEE (2009) 167–173
5. 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
6. 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
7. Shearer, R., Motik, B., Horrocks, I.: Hermit: A highly-efficient owl reasoner. In: OWLED. Volume 432. (2008)
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, ACM (2007) 737–746