Extending the Digital Archives of Italian Psychology With Semantic Data

Claudio Cortese and Glauco Mantegari

Lombard Interuniversity Consortium for Automatic Computation (CILEA)
Segrate, Italy

Abstract. ASPI is a project that aims at creating a digital library of historical documents of Italian Psychology and extending it with semantic data. The extension makes it possible to retrieve archival documents not only on the basis of archival metadata, but also according to the connections the documents have with specific activities of researchers, groups and institutions, as well as with more general events in the history of Italian Psychology. The paper provides an overview of ASPI and discusses the approach and workflow we adopted in its development. In particular, ontology modeling according to CIDOC CRM, ontology population and the prototyping of a semantic search and browsing portal based on the ClioPatria platform are introduced.

1 Introduction and Background

Today, cultural heritage represents one of the most promising and challenging areas for the application of the Semantic Web and Linked Data principles and technologies [6] [8]. In particular, digital repositories of historical archives are increasingly paying attention to and taking advantage of the new technologies, especially for what concerns the creation of highly interoperable datasets and the improvement of search functionalities beyond traditional keyword-based approaches.[7].

Our working group has a consolidated experience in the field of digital technologies applied to cultural heritage, and notably in the areas of digital preservation and web-based systems[1]. In 2007, as a part of the "Open Library of Milan" (BAMI) project, we started investigating Semantic Web technologies through creating one of the first semantic digital libraries in Italy [1]. The main objective of BAMI was to offer online access to digitized documents of different libraries and archives held by prominent cultural institutions in Milan. In particular, we focused on a subset of the heritage, which is made up of musical documents of

[1] Since 2004 we have been involved in several projects, and we have been developing the CodeX[ml] system (http://codex2.cilea.it) for the management, preservation, fruition and dissemination of library and archival (meta)data. Today, the system is used by 17 prestigious Italian institutions, which include the Ambrosian Library (Milan), the Conservatorio "Giuseppe Verdi" (Milan), and the State Archives of Milan and Venice.

the 19th century. The semantic dataset we created is based mainly on FRBR[2], the Music Ontology[3], and FOAF[4]. Access to the semantic repository is possible by means of a web portal[5] that makes use of Longwell[6], a faceted browser for RDF datasets developed by MIT. Longwell has been extended in order to offer different search and browsing functionalities, according to different user needs and experiences. In particular, facet-based querying has been integrated with relation browsing, with visual exploration of the RDF graph, and with temporal navigation through an interactive timeline.

Despite the efforts we put in the deployment of a user-friendly system, users (who include archivists, music professionals and more general communities of people interested in the history of music) did not always give positive feedback, especially for what concerns browsing the dataset. For example, in FRBR the concept of "book" is split into four different classes (Work, Expression, Manifestation and Item) whose meaning was difficult to understand by non-specialized users when navigating in the repository. In addition, some users felt slightly uncomfortable with the faceted-browsing approach and the way search results are presented. Nevertheless, BAMI has been altogether a successful project, not only because it offered us the opportunity to test Semantic Web technologies in a real application case, but also because it helped diffusing knowledge of these technologies in the communities of Italian archivists and librarians. Hence, we decided to further investigate the application of the Semantic Web to digital libraries. This has been done with particular reference to the deployment of intelligent retrieval and browsing services built on top of semantic data.

The paper introduces a new project in this area and it is organized as follows: Section 2 introduces the general characteristics of the project and motivates the choice of using Semantic Web technologies. Section 3 describes the approach and the workflow we adopted concerning ontology modeling, ontology population, and the deployment of a semantic search and browsing prototype. Section 4 summarizes the results obtained so far and outlines possible directions for future work.

2 ASPI: The Digital Archives of Italian Psychology

In 2009, a three-year project concerning the creation of a digital repository of archival documents produced by (or related to) the key figures in the history of Italian Psychology was launched. The project is coordinated by the University of Milano-Bicocca[7] and it includes several academic partners[8], each of which is

[2] http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records/
[3] http://musicontology.com
[4] http://www.foaf-project.org
[5] http://bami.cilea.it
[6] http://simile.mit.edu/wiki/Longwell/
[7] "Archivi Storici della Psicologia Italiana" resesarch group (ASPI).
[8] The University of Trieste, the University of Florence, the Catholic University of Milan, the University of Palermo and the University of Turin.

working on the study and cataloguing of important archives that are related to the history of Psychology. The technology partner of the project is the Lombard Interuniversity Consortium for Automatic Processing (CILEA), which is in charge of all the aspects concerning the development of the Digital Library.

The first phase of the project was mostly devoted to the creation of the Digital Library infrastructure, which integrates different applications offering the most important functionalities required by a modern system: accurate metadata creation and ingestion, search and browsing, interoperability and digital preservation according to international standards and protocols.

In particular, the CodeX[ml] system has been used to manage the digitized documents and ensure long-term digital preservation of both the scans and the associated metadata. CodeX[ml] is compliant with the recommendations of the OAIS model [3], and it constantly checks the validity and integrity of data and metadata during and after the ingestion phase[9] in order to prevent bit decay. CodeX[ml] is also able to provide metadata to harvesters according to the OAI-PMH standard[10], therefore enabling full interoperability with other existing repositories. Furthermore, thanks to the integration of the IIPImage server[11], high-resolution scans of the documents in the Tiled Pyramidal TIFF format can be viewed with extreme efficiency.

The AriannaWeb software[12] is dedicated to the browser-based visual navigation of a dynamically generated tree of XML-EAD[13] metadata describing the archival documents.

Finally, a web portal[14] developed with the Typo 3 Content Management System[15] allows the creation of both static and dynamic web pages. These pages provide information about the archival inventories and the historical researches carried out on them.

The Digital Library satisfies the most part of the expectations expressed by the project partners. However, it does not completely meet one of the requirements of the project, i.e. the possibility of retrieving documents on the basis of their relations to specific activities of researchers, groups and institutions, as well as to more general events that are related to the history of Italian Psychology. For example, a user may be interested in archival documents that have been produced by scholars whose activity was influenced by a specific research topic, such as "visual perception". EAD metadata do not make it possible to answer this kind of query. On the other hand, the unstructured information contained in the pages of the web portal (which may contain relevant data) is not suitable for automated processing. Therefore, we decided to extend the digital archives with structured data that could be linked to the documents, and processed by

[9] Controls on data are done through MD5 checking.
[10] http://www.openarchives.org/pmh/
[11] http://iipimage.sourceforge.net/
[12] http://www.ariannaonline.it/web/15390/11/
[13] http://www.loc.gov/ead/
[14] http://www.archiviapsychologica.org/
[15] http://typo3.org/

machines in an intelligent way, improving search and browsing functionalities. The choice of an approach based on Semantic Web principles and technologies appeared to be the most suitable solution for our needs.

3 Approach and Workflow

Our approach to extending the system with semantic data was based on an intense collaboration with the project partners. A preliminary activity concerned training archivists and researchers in the history of Psychology about the basics of the Semantic Web, and discussing the aspects involved in ontology modeling with them. The result of this activity highlighted the need of a model where the events that are associated with the authors of the documents (such as the affiliation of an author to a particular research institution, or the different interactions between two scholars who share some lines of research) play a central role.

Therefore, we focused our attention on event-centric models and, in particular, on CIDOC CRM[16], an upper-level domain ontology for cultural heritage that is strongly based on an event-centric perspective [5]. To our knowledge, no other domain-specific models having the same characteristics and scope of CIDOC CRM exist. CIDOC CRM was used both to link "contextual data" with the documents, and to provide a semantic description of the archives, as explained in Section 3.1.

In order to allow the project partners to populate the ontology, we built a relational database using PostgreSQL[17]. Data entry is possible through a web-based interface that supports collaborative work between the different research units. We excluded the possibility of using an ontology editor such as Protégé[18] (which also has an extension for collaborative ontology editing[19]) mostly because the archivists and researchers did not feel comfortable with the tool. However, using a relational database was not a big issue, since the database schema has been mapped on the ontology, and data extraction and transformation in CIDOC CRM-compliant RDF have been done through the D2RQ[20] mapping language. The schema of the database and its mapping to RDF are introduced in Section 3.2.

Semantic search and browsing have been implemented with ClioPatria[21], a SWI-Prolog-based platform for Semantic Web applications that is also currently used as a research prototype by the Europeana project[22]. The choice of ClioPatria was motivated by the need to provide efficient means of browsing the semantic dataset, and by the lack of resources to develop our own solution. In

[16] http://www.CIDOC CRM.org/
[17] http://www.postgresql.org/
[18] http://protege.stanford.edu/
[19] http://protegewiki.stanford.edu/wiki/Collaborative_Protege/
[20] http://www4.wiwiss.fu-berlin.de/bizer/d2rq/
[21] http://e-culture.multimedian.nl/software/ClioPatria.shtml
[22] http://eculture.cs.vu.nl/europeana/session/search/

addition, using Prolog for Semantic Web applications offers several advantages, as it is discussed in [13] and [10]. ClioPatria provides different functionalities (such as semantic search, and faceted browsing) that can be easily configured and extended, thanks to the open-source license of the platform. Configuration and customization of ClioPatria according to the requirements of our project are outlined in Section 3.3.

3.1 Ontology Modeling

The ontology, which is based on version 5.0.2 of CIDOC CRM [4], was modeled through the continuous interaction with domain experts.

A fundamental part of the ontology concerns data that extend the digital archives with "contextual" information. These data take into account the following entities:

- Persons: birth, death, research activity, meeting with another person, writing of a book, writing of a paper, creation of a research instrument, participation in conferences, affiliation to a group, affiliation to an institution
- Groups: formation, dissolution, joining a group, disjoining a group, joining an institution, disjoining an institution
- Institutions: formation, dissolution, joining an institution, disjoining an institution, choice of a headquarter
- Gestalts: influence of a topic on one or more research activities

Thanks to the nature of CIDOC CRM, the identification of events and activities characterizing our domain was quite straightforward. Since we decided not to extend the model, we made an extensive use of the "E55 Type" class and the "P2 has type" property to identify different elements that are represented by the same class. For example, the "E7 Activity" class can represent both the participation in a conference and the research activity of a psychologist. Therefore, instances of E7 are associated to types that make it possible to distinguish the different activities and ease the retrieval of relevant data.

The second part of the ontology concerns mapping of some metadata of the archives to CIDOC CRM in order to link them to persons, groups, institutions, and gestalts, and the related events. Our initial intention was to map the entire EAD dataset to CIDOC CRM, following the proposals described in [12] and [11]. We soon realized that the effort required to complete the mapping was beyond the possibilities of the project, especially because of the consistent differences in the structure of the two models, as it is discussed in a very recent work [2].

The EAD elements we took into consideration concern basic metadata of archives, archival partitions, series, and single documents, such as their denomination and the date they were produced.

3.2 Populating the Ontology

In order to facilitate mapping and transformation of relational data in RDF, the database schema has been designed taking into consideration the structure of the ontology.

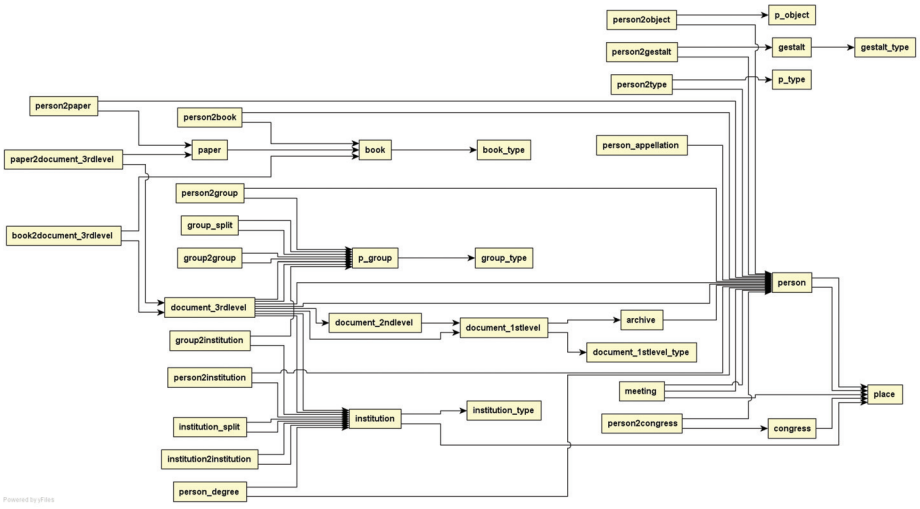

Fig. 1. A simplified representation of the database schema.

The schema (Fig. 1) includes six principal entities: persons, groups, institutions, archival documents, publications, research instruments.

Basic biographical data is represented by a series of entities and relationships that makes it possible to describe psychologists as well as other persons that fall outside the research community but can be considered relevant for the project. These include, for example, a psychologist's relatives or friends who, according to domain experts, may have played an important role in influencing research activities.

Persons are also connected to the books and papers they have written, and the scientific instruments they have invented.

The structure of the ontology greatly facilitated the development of the database, especially for what concerns the parts of the schema corresponding to events and activities such as conferences, meetings, or groups and institutions dynamics.

A part of the schema is dedicated to archival metadata and it has been populated automatically from the XML-EAD files. Thanks to the database, the documents can be annotated with the names of the persons, groups, and institutions they are related to, as well as with the papers or books for which they represent the draft version.

The web-based interface of the database (Fig. 2) allows an easy and collaborative data entry. Predefined values according to the E55 Type class instances are available in the drop-down lists.

Data extraction and transformation into CIDOC CRM-compliant RDF is very easy and efficient, thanks to the D2RQ platform. The mapping language provided by the platform has been privileged among other solutions [9] because it allows defining the mappings in a very modular and compact way using the RDF

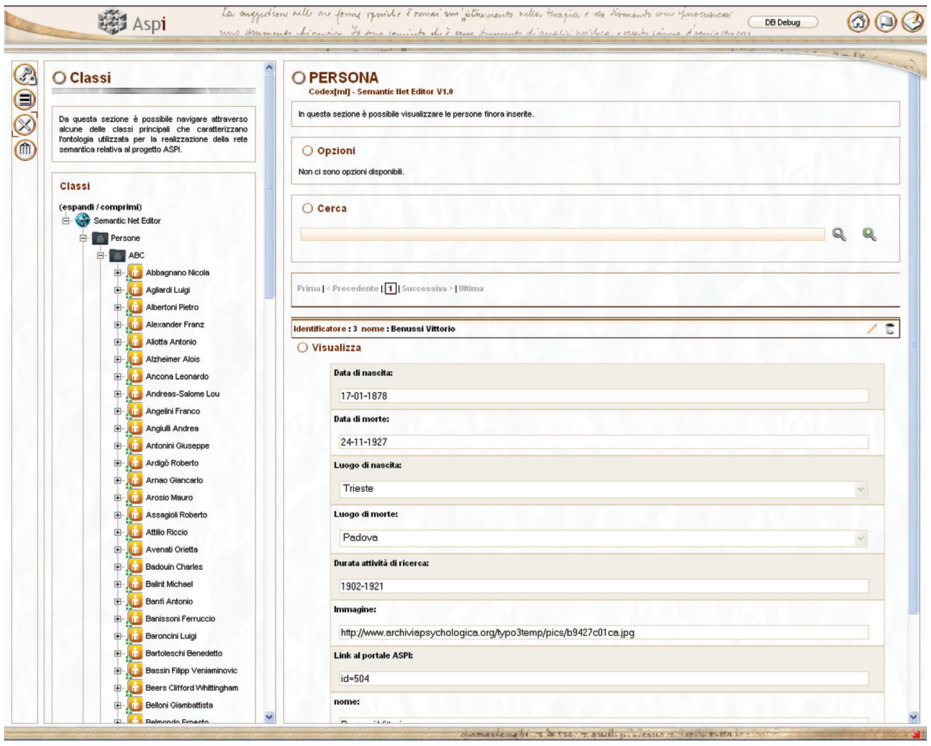

Fig. 2. A section of the web-based interface for data entry.

Notation3 syntax[23]. In creating URIs, we tried to be as compliant as possible with guidelines and recommendations suggested by W3C[24]. The only remarkable limitation of D2RQ with reference to our project is the impossibility of creating hierarchical URIs, which would have instead enhanced human readability and understanding.

The resulting RDF dataset is based on the OWL-DL 1.0 implementation of CIDOC CRM that is known as "Erlangen CRM / OWL"[25]. As of June 2011, our semantic repository is still small (about 45.000 triples) since it is based only on initial data entered by a single project partner. Nevertheless, it is destined to increase progressively along with data entry activities that will be carried out by the other project partners in the next months.

[23] http://www.w3.org/DesignIssues/Notation3.html
[24] http://www.w3.org/TR/cooluris/
[25] http://erlangen-crm.org/

3.3 Enabling Semantic Search and Browsing

Version 1.0 beta 2.5 of the ClioPatria platform[26] was used for the creation of a portal enabling semantic search and browsing on the RDF dataset. Thanks to the administrator web frontend of Cliopatria, the basic aspects involved in RDF management (such as RDF uploading, clearing single statements or the entire repository, and querying) are greatly simplified, and triple storage is managed efficiently. Moreover, the platform is able to provide additional functionalities, such as the evaluation of RDF data quality or alignment checking.

The settings concerning the behavior of the search engine can be configured via the administrator frontend as well, making it possible to obtain in a very short time a fully functional portal for semantic search and browsing of RDF datasets.

Our customization of the ClioPatria semantic portal concerned mostly the layout elements. Beyond extending or overriding the standard CSS files, we made minor changes in the Prolog code in order to modify the parameters that were not directly configurable using the administrator frontend. These include, for example, the removal of links to display options that were not considered relevant for our portal, or the creation of a personalized layout for the home page (Fig. 3). Moreover, we made minimal interventions on JavaScript code in order to manage a few unexpected behaviors of the interface components.

Figure 4 shows the role the semantic portal plays in the overall architecture of the system. Users can search for information either by means of the Typo 3 web portal or by means of the ClioPatria engine. Once the desired document is found, its high-resolution scan as well as its metadata can be visualized respectively in the CodeX[ml] and the AriannaWeb systems.

The web portal offers multilingual support with respect to the labels associated to the classes and the properties of the ontology that are shown during search and browsing. English and French versions of the labels were already available, while for Italian we took care of the translation, following the official guidelines provided by the CIDOC CRM working group[27].

Thanks to the semantic portal prototype, search and browsing through the digital library has been considerably extended. For example, now users can search for the name of a research group in the semantic portal and, among the results, see a list of documents that are in some ways related to scholars who, in a certain period of their activity, were affiliated to that group.

If a user search for "visual perception" (see the example query introduced in Section 2), the system displays also a list of the scholars whose activity was influenced by that research topic. Selecting the name of a scholar, users can obtain several data, including a list of the scholar's documents that are present in the archives. Each item of the list is a hyperlink that leads the user to get more data about that item. Included in these data is a link to the web interface where the image of the document (as well as the images of other documents belonging to the same scholar) can be visualized in high resolution.

[26] The platform we used is based on SWI-Prolog 5.9.3.
[27] http://www.CIDOC CRM.org/translation_guidelines.html

Fig. 3. The semantic search and browsing prototype homepage.

4 Conclusions

Extending archival datasets with semantic data represents an important opportunity for the creation of a new generation of digital libraries with improved search and browsing capabilities. Our project shows that encouraging results can be obtained by taking advantage of ready-to-use solutions and applications, and combining them with existing digital library systems. The preliminary feedback we received from the project partners seems to confirm we met their general expectation, i.e. extending the digital library's search and browsing functionalities with the definition of semantic relationships between the archival materials and events in the history of Italian Psychology.

However, the inherent characteristics of the ontology we used and the lack of resources to develop a completely custom presentation layer may limit the usability of the current system.

The event-centric nature of CIDOC CRM, combined with the way the standard ClioPatria interface shows search results, makes it sometimes difficult to easily obtain the desired information. For example, the title of a document created by a particular scholar can be retrieved only passing through a class that represents the activity of writing of that document. Expert users (who represent the main target of ASPI) may get easily familiar with the data structure, while more general and non-expert users may feel disoriented. A more detailed user

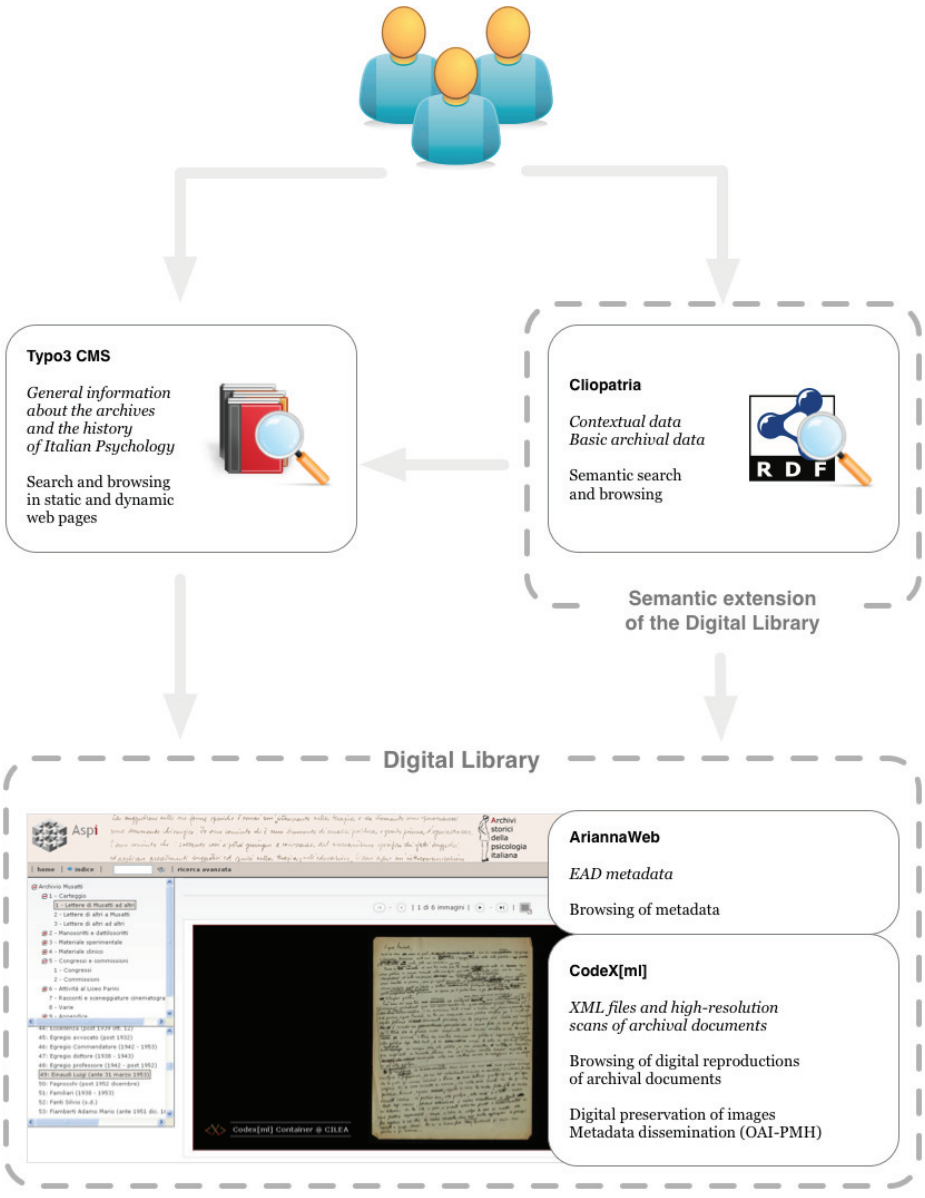

Fig. 4. The components of the system and the principal modalities of search and browsing in the digital repository.

study would help us identify the most critical aspects of the current system and define new strategies for improving the usability of the portal.

From a conceptual point of view we still think that CIDOC CRM represents a suitable model for our domain. Therefore, we are evaluating the possibility of creating a new version of the system based on a completely custom presentation layer hindering the complexity of the ontology. Version 2 of Cliopatria[28] might be a possible solution, since it provides great modularization and offers several JavaScript libraries that can be used for the design of flexible web-based interfaces.

In general, we think that ASPI is a step forward for us if compared to BAMI, especially because it offers improved searching and browsing capabilities that allow exposing the dataset in all its richness while providing a simpler user interface. However, a more detailed evaluation of the project outcomes and an extensive comparison with BAMI will be possible only with a bigger dataset integrating the cataloguing activities of the different research units.

To our knowledge our semantic dataset is the only one available today for the history of Psychology. For this reason, we are willing to define better modalities for sharing our data. To this respect, the creation of a SPARQL endpoint and the alignment of the dataset for Linked Data will be two major improvements we plan for the future, if the project will obtain additional financial support.

References

1. Barbera, M., Cortese, C., Zitarosa, R., Groppo, E.: Building a Semantic Web Digital Library for the Municipality of Milan. In: Mornati, S., Hedlund, T. (eds.) Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies - Proc. 13th International Conference on Electronic Publishing. pp. 133–154 (2009)
2. Bountouri, L., Gergatsoulis, M.: Mapping Encoded Archival Description to CIDOC CRM. In: First Workshop on Digital Information Management. pp. 8 – 25 (2011)
3. CCSDS: Reference Model for an Open Archival Information System (OAIS). Blue book, Consultative Committee for Space Data Systems (2002)
4. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: Definition of the CIDOC Conceptual Reference Model. version 5.0.2. ICOM/CIDOC CRM Special Interest Group (January 2010)
5. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
6. Hyvönen, E.: Semantic Portals for Cultural Heritage. In: Staab, S., Rudi Studer, D. (eds.) Handbook on Ontologies, pp. 757–778. International Handbooks on Information Systems, Springer Berlin Heidelberg (2009)
7. Kruk, S., McDaniel, B. (eds.): Semantic Digital Libraries. Springer (2009)
8. Nixon, L., Dasiopoulou, S., Evain, J., Hyvönen, E., Kompatsiaris, I., Troncy, R.: Handbook of Semantic Web Technologies, chap. Multimedia, Broadcasting and eCulture, pp. 901–965. Springer (2011)

[28] http://cliopatria.swi-prolog.org/home/

9. Sahoo, S., Halb, W., Hellman, S., Idehen, K., Thibodeau Jr, T., Auer, S., Sequeda, J., Ahmed, E.: A Survey of Current Approaches for Mapping of Relational Databases to RDF. Tech. rep., W3C RDB2RDF Incubator Group (2009)
10. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B.: Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. Web Semant. 6, 243–249 (November 2008)
11. Stasinopoulou, T., Doerr, M., Papatheodorou, C., Kakali, K.: EAD mapping to CIDOC/CRM. Tech. rep., Department of Archives and Library Science, Ionian University (2007)
12. Theodoridou, M., Doerr, M.: Mapping of the Encoded Archival Descripition DTD Element Set to the CIDOC CRM. Tech. rep., Institute of Computer Science, Foundation for Research and Technology - Hellas (2001)
13. Wielemaker, J., Hildebrand, M., van Ossenbruggen, J.: Using Prolog as the fundament for applications on the semantic web. In: S.Heymans, Polleres, A., Ruckhaus, E., Pearse, D., Gupta, G. (eds.) Proceedings of the 2nd Workshop on Applicatiions of Logic Programming and to the web, Semantic Web and Semantic Web Services. pp. 84–98