Lifting knowledge from the Medieval Age: CIDOC-CRM for the Nurcara Project

Matteo Lorenzini
matteo.lorenzini@oeaw.ac.at
Austrian Academy of Science, Austria

Luca Sanna
lucasanna@uniss.it
Università degli Studi di Sassari, Italy

Introduction

The structured repositories in cultural heritage have become the most used infrastructure for knowledge management towards different kinds of systems and platforms, ensuring a complete interoperability and reachability of data. Thanks to the semantic web paradigm, we are able to manage and enrich our data using formalisms and data standards: examples include digital libraries and digital archives, as well as SPARQL endpoints. However, the fragmentation of data produced by different kinds of mapping methodologies and different representations of the knowledge to be managed leads to some discrepancies between domains and results obtained during data retrieval.

A typical example is the study of an inscription, which will be addressed by linguists with regards to language; by philologists with regards to its text; by historians as a primary source; by archaeologists as material testimony of events and by conservationists as a piece of matter to be preserved and restored. In that scenario, semantic interoperability and standardization are two fundamental elements as they guarantee the circulation of knowledge inside a shared environment.

This paper aims to present the methodology followed in the Nurcara Project concerning resource integration and knowledge management. Nurcara consists of a dataset (almost 300 records) composed of textual Latin documents from the Medieval Age that provide historical context for the area of Monteleone Rocca Doria in Sardinia (Sassari, Italy) between the 11th and 15th centuries. Starting from an SQL database, our solution focuses on the development of a semantic framework solution able to both automatically map the relational dataset in CIDOC-CRM ontology on-the-fly and aggregate the knowledge from different repositories or endpoints. This method aims to semantically enrich Nurcara's dataset with external resources from the same domain, such as SPARQL-endpoint and Linked Data resources.

Nurcara project

Nurcara started from the following directive: to make accessible to the community the huge amount of published and unpublished medieval documents about Monteleone Rocca Doria that until now have been only accessible from the researchers.

Thanks to a collaboration between the Spanish Ministry of Cultural Heritage and the University of Sassari, it has been possible to start archival research of historical documents from the National Archive of Cagliari, the Historical Archive of Alghero and the General Archive of the Crown of Aragon in Spain. During the research almost 300 medieval documents in Latin, Catalan, Castilian and Sardinian were cataloged and stored in a MySQL database. The documents dated from the 11th to 15th centuries and related to Monteleone Rocca Doria (SS), which made them useful for defining the socio-economic context of the area during the medieval age.

Interoperability and semantic enrichment

The interoperability and semantic enrichment of the data represents one of the most important milestone of the project. It is crucial to publish the dataset in an accessible format with Semantic Web technology such as SPARQL or Linked (Open) Data.

In order to reach our goal, we started with the conceptualization of the entities from the relational database and the definition of the semantic schema. CIDOC-CRM has been chosen as the main reference model.

CIDOC-CRM is an ontology created in order to offer "definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation". The CIDOC-CRM model is a semantically rich model used to conceptualize the cultural heritage domain composed of 86 classes and 138 properties. The purpose of CIDOC-CRM is to provide a common definition for heterogeneous forms of information, and to enable their integration despite possible semantic and structural incompatibilities.

In order to translate data stored in the Nurcara relational database to the CIDOC-CRM ontology (expressed in RDF-Resource Description Framework) we have used the D2RQ tool that will provide an

automatic mapping between the relational database and the CIDOC mapping schema.

D2RQ is a popular mapping platform for publishing relational data as a virtual RDF graph. It enables legacy relational databases to be exposed on the Web, according to the principles of Linked Data, and to be included in the Semantic Web. D2RQ exposes relational databases as SPARQL endpoints. It translates SPARQL queries posed on a virtual RDF graph to SQL queries posed to the underlying relational database.

Mapping Solution

The mapping process started with the definition of the main entities useful for the definition of the conceptual model directly from Nurcara's database:

- Document (E84 Information Carrier)
- Author (E39 Actor)
- Issue place (E53 place)
- Issue date (E50 Date)
- Type (E55 Type)
- Title (E35 title)

We chose these entities as a starting point for the mapping activities because from a conceptual point of view, they cover the minimum knowledge representation useful for the description of the documents from Nurcara's database. Furthermore, with the aim of defining a more suitable (event-oriented) semantic model with respect to CIDOC-CRM ontology, we defined some "abstract" entities that are outside the structure of Nurcara's database, but are present as classes in CIDOC ontology:

- Dataset (E73 Information Object)
- Event (E5 Event)
- Activity (E7 Activity)

Here, the use of the abstract classes as an intermediate layer allowed us to guarantee a more detailed and coherent conceptualization of the proposed model with respect to CIDOC-CRM ontology. Then, thanks to definition of the "Dataset" as E73 Information Object, it is possible to extend the proposed model with CRMdig (developed as a compatible extension of ISO21127) application profile from CIDOC-CRM ontology, which allows us to encode metadata about the steps and methods of production of digitization products.

The defined conceptual schema has been directly implemented as a mapping file (using D2RQ syntax) in D2RQ in .ttl format, in order to ensure the live mapping process between Nurcara dataset and CIDOC-CRM ontology.

In order to enrich and extend the dataset, we have also considered in the mapping.ttl schema other Linked Data end-points such as DbPedia or Geonames, which are specified by prefix and reachable via a proper SPARQL query directly from D2RQ SPARQL end-point.

Data accessibility

As previously stated, after the mapping process, the data are expressed and exposed as RDF graph based on CIDOC-CRM structure.

Hyperlinking

The D2R server supports hyperlink navigation by providing links on the RDF and XHTML levels. Any RDF triple whose object is a dereferenceable URI can be seen as a hyperlink. This is how resources published by the D2R Server are interlinked with other databases and external RDF documents. To aid discovery of related resources, D2R Server includes an rdfs:seeAlso triple with every resource description that points to an RDF document containing links to other resources produced by the same ClassMap (In our case, DbPedia or Geonames). If resources are identified with external URIs, then an additional rdfs:seeAlso link points to a local RDF/XML document that contains everything the database knows about the resource. By dereferencing the external URI and by following the rdf:seeAlso link, RDF browsers can retrieve both authoritative and non-authoritative information about the resource.

Search

The D2R Server allows the users to query non-RDF databases using the SPARQL query language over the SPARQL protocol. Queries are executed against a virtual RDF graph representing the complete database. Query results can be retrieved in the SPARQL query result as XML or in SPARQL/JSON serialization.

Conclusion

In spite of progress in the area of RDF storage, a large quantity of data is still stored in non-semantic repositories or relational databases.

Nevertheless, especially in the domain of the humanities, we are seeing growth in the use of semantic platforms for the management of digital data as a common method of knowledgement management. The CIDOC-CRM ontology is one of the most used in

digital humanities. In this proposal we want to show how a specific dataset as Nurcara can be successfully integrated under the much more generic CIDOC-CRM structure, and how the knowledge can be easily enriched thanks to LOD integration. In order to reach the goal, the scalability and the full customization offered by the D2RQ server has played a crucial role.

The future development of this project concentrates on the integration of the Description Logic algorithm and reasoning system with the purpose of increasing resource discovery with respect to the domain utilized by the Nurcara project.

Bibliography

Bizer, C., and Seaborne, A. (2004) "D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs." Paper presented at the meeting of the ISWC2004 (posters), 2004.

Jannaschk, K., Rathje, C. A., Thalheim, B., and Förster, F. (2011) "A Generic Database Schema for CIDOC-CRM Data Management.." Paper presented at the meeting of the ADBIS (2).

Lourdi, I., Papatheodorou, C., and Doerr, M. (2009) "Semantic Integration of Collection Description: Combining CIDOC/CRM and Dublin Core Collections Application Profile." D-Lib Magazine 15 , no. 7/8.

Soddu, A. (2013) “Incastellamento in Sardegna:L’esempio di Monteleone.” Castra Sardiniae. Quaderni Vol 1, ed. Lulu.com.