Syrtis: New Perspectives for Semantic Web Adoption

Joffrey Decourselle, Fabien Duchateau, Ronald Ganier

Syrtis: New Perspectives for Semantic Web Adoption

Joffrey Decourselle, Fabien Duchateau and Ronald Ganier

LIRIS, UMR5205, Université Lyon 1 Lyon, France
jdecours@liris.cnrs.fr, fduchate@liris.cnrs.fr

Progilone SAS, Lyon, France
ronald.ganier@progilone.fr

Abstract
The last two decades have shown a huge interest in using Semantic Web technologies in Cultural Institutions. There are now a large number of Open Data projects such as Europeana, which aggregates millions of records, allowing us to find many relevant information. However, most of the effort to adopt Semantic Web principles has been spent in national libraries or at larger scope and most local libraries still store isolated data. Cultural information held in these local institutions is generally not available outside the library due to old-fashioned Integrated Library Systems (ILS). We advocate that improvements have to be done in the transformation and enrichment of the catalogs of any cultural institution to enhance the accessibility of data. The Syrtis project lies at the heart of the cultural heritage challenges and aims at providing relevant solutions for guiding librarians through the adoption of Semantic Web concepts. This paper presents an overview of the architecture built for the management of cultural records in the context of the Syrtis project. We also discuss briefly the current challenges faced by the project.

Keywords: Semantic Web, Cultural Heritage Catalogs, FRBR, MARC, Metadata Migration

Motivation

These last years, the emergence of new models and rules for cataloguing cultural records (e.g., FRBR[1], RDA[2]) raised new interests in using Semantic Web technologies in cultural institutions (Alemu, 2012). Europeana started publishing data in RDF using a model[3] based on FRBR (Doerr, 2010). OCLC clustered large amount of bibliographic records from Worldcat to group them according to the FRBR semantic levels (Hickey, 2002). In France, the national library also started a project to "FRBRize" its catalogs (Simon, 2014).

These projects have been mostly initiated at the national or higher level while local libraries still have to manage records in old formats (e.g., MARC) leaving the feel that the global movement towards Semantic Web of any cultural institution is still stuck. Indeed, the full adoption of these new principles requires to face important challenges such as transforming existing records in the new semantic models and developing new systems that handle them. Furthermore, such migration should be done mostly in an automated manner to process the thousands of existing records. Yet, the recent studies on the FRBRization process (i.e., a metadata migration of records to FRBR) have shown that the automated transformation of catalogs can be very complex due to the heterogeneousness of cataloguing practices in different institutions (Aalberg, 2013).

The Syrtis research project has been started in 2013 to bring new solutions for the adoption of Semantic Web principles in CH institutions. The project has released tools to migrate catalogs of records to the FRBR concepts and to propose all features of a complete Integrated Library System. The rest of this paper presents the current and future works done in the context of Syrtis.

[1] Functional Requirements for Bibliographic Records
[2] Resource Description and Access
[3] Europeana Data Model

Overview of the architecture

This part gives an overview of the global and theoretical architecture of the Syrtis research project. The system is based on three main steps for processing records: Metadata Migration, Validation & Enrichment and Exploitation.

Metadata Migration process
The process upstream the whole application takes in input a record-based catalog (e.g., MARC) and returns a FRBR-based catalog. Such process is done in two steps. It begins with a rule-based extraction where a set of rules is used to generate FRBR entities and relationships from data in input. Then, it deals with duplicate entities, due to repeated information in the records, by invoking a deduplication step where potential equivalent entities are matched and merged. The output objects are stored in a knowledge base that can be queried as a graph.

Validation and Enrichment process
To validate the extracted entities before their integration, an additional step of semantic enrichment is launched. First, it queries sources from Linked Open Data using attributes from extracted entities and it aggregates the results. Then, a process of deduplication which involves both ontology matching and entity matching algorithms group the equivalent entities and relationships from aggregates. Finally, a fusion process merges the new information to the initial entities according to an expert validation.

Exploitation of FRBR collections
Once the data is validated, the entities are integrated in our system based on the FRBR concepts. The solution provides various GUI[4] to manage the data both from the expert and from the end user points of view. A first interface deals with the representation of bibliographic families by clustering the data according to the FRBR semantic levels (Work, Expressions, Manifestation and Item). This allows the user to benefit from a Top-Down look at the data. Another possible interface is graph-based and allows the user to navigate between the main entities of the three groups of FRBR (Bibliographic, Agent and Subject information) to facilitate its discovery of knowledge.

This architecture involves several research domains such as data integration, information retrieval and ontology modeling, where each implies to face specific challenges. In our project, the quality of the migration processes and the enrichment of data are at the central point of our reflections. The remains of this paper discuss our progresses and remaining challenges.

Progresses and Results

New approach for interpreting records
We have studied the projects and tools about automated FRBRization from the last decade and realized an original classification (Decourselle, 2015a). This work showed us that several improvements scattered in the different solutions might be merged to create an enhanced process. Furthermore, some steps such as the pre-analysis of input data and evaluation of the process have been less explored.
We assume that the weaknesses of most FRBRization tools concern the way they manage their rules. First, the latter are generally represented as mappings based on XML and XSLT transformations which requires specific skills to understand. Then, the models used for handling the rules are mostly entity-centered. This can be a limitation for interpreting all the complex bibliographic patterns that are represented in input catalogs (Cole, 2013). To improve these two

[4] Graphical User Interface

steps, we have started by building a case-based model for FRBRization rules where each case describes a specific part of a bibliographic pattern and the whole, modeled as a graph, represents the global structure of the expected FRBR output model. This approach eases the representation and documentation of all bibliographic patterns that must be extracted from a bibliographic catalog. We have also formalized metrics and built specific datasets for evaluating the quality of a FRBRization solution. This material allows us to refine gradually our own process and may help other practitioners in building and evaluating approaches for metadata migration of cultural records.

Fully FRBR-based Integrated Library System
A software that handles the FRBR concepts for managing FRBRized collections is operational in the Syrtis project. The large experiences of Syrtis stakeholders on former ILS helped to build a solution more in agreement with the user needs. The cataloguing specifications inside the tool can be fully edited to fit each requirement of an institution using the application. The extensibility of the FRBR model and the flexibility of the solution make the tool very customizable and intuitive.

Future Works

One of our works in progress relates to the domain of information extraction. Even as it becomes more and more convenient to extract data from Linked Open Data, most of fresh information (e.g., about recent cultural content) are still scattered on the web or on isolated databases. Thus, we plan to extend our external sources to those without a structured schema such as microblogging. A large variety of studies already exist about extracting information from such kind of sources, but we advocate that new enhancements can be done by using the potential of new semantic models. For instance, relationships already proposed by FRBR can refine the way we extract additional knowledge and facilitate the selection of relevant entities to keep.

Another part of our research focuses the models used in knowledge bases. On the one hand, the conceptual models from the FRBR family are open to interpretation. On the other hand, the interests of users on a specific aspect of a cultural entity may vary according to the resource explored. Thus, knowledge bases must adapt to the user requests which implies a flexibility in the ontologies used. We plan to provide a practical model for automatically build Thematic Knowledge Bases which implies to face the challenges from Ontology Modeling, Information extraction and Data Integration. We have already published a theoretical approach of such a system (Decourselle, 2015b).

Conclusions
The Syrtis research project brings together professionals of cultural institutions and researchers working on Cultural Heritage Data. There is still a long way to go for a full adoption of Semantic Web concepts, but we hope that our initiative will help both communities to move forward towards new experiments.

References

Alemu, G., Stevens, B., Ross, P., & Chandler, J. (2012). Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World*, 113(11/12), 549-570.

Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., & van de Sompel, H. (2010, August). The europeana data model (EDM). *In World Library and Information Congress: 76th IFLA general conference and assembly* (pp. 10-15).

Hickey, T. B., O'Neill, E. T., & Toves, J. (2002). Experiments with the IFLA functional requirements for bibliographic records (FRBR). *D-Lib magazine*, 8(9), 1-13.

Simon, A., Di Mascio, A., Michel, V., & Peyrard, S. (2014). We grew up together: data.bnf.fr from the BnF and Logilab perspectives. *In IFLA 2014*.

Aalberg, T., & Žumer, M. (2013). The value of MARC data, or, challenges of frbrisation. *Journal of Documentation*, 69(6), 851-872.

Decourselle, J., Duchateau, F., & Lumineau, N. (2015). A Survey of FRBRization Techniques. *In Research and Advanced Technology for Digital Libraries* (pp. 185-196).

Cole, T. W., Han, M. J., Weathers, W. F., & Joyner, E. (2013). Library marc records into linked open data: Challenges and opportunities. *Journal of Library Metadata*, 13(2-3), 163-196.

Decourselle, J., Vennesland, A., Aalberg, T., Duchateau, F., & Lumineau, N. (2015). A Novel Vision for Navigation and Enrichment in Cultural Heritage Collections. *In New Trends in Databases and Information Systems* (pp. 488-497).