Adopting Semantic Technologies in Public Health Documentation

Joffrey Decourselle, Frédéric Riondet

Adopting Semantic Technologies in Public Health Documentation

Joffrey Decourselle[1] and Frédéric Riondet[2]

[1] LIRIS, UMR5205, Université Claude Bernard Lyon 1, Lyon, France
joffrey.decourselle@liris.cnrs.fr
[2] Central Documentation - Hospices Civils de Lyon, France
frederic.riondet@chu-lyon.fr

Abstract. We present a success story on the adoption of semantic technologies for the library of the second biggest university hospital of France. This project was divided into three parts: *preprocessing*, *semantic enrichment* and *data integration*. This abstract introduces the research challenges faced in the project as well as the outcomes obtained so far.

Keywords: health documentation, semantics, migration, enrichment

The Central Documentation of the second biggest university hospital of France, the Hospices Civils of Lyon (HCL), holds about 500.000 online bibliographic records where each contains metadata about documents like article, journals, books or legislative texts in the health domain. The availability of such data is crucial for the work of at least 25.000 health professionals, researchers or students. Yet, while the latter become more and more demanding in terms of search and exploration features, improvements of the HCL digital library was impossible due to the limited capabilities of the old-fashioned library system and the lack of flexibility of the cataloging model. Thus, the decision was taken to adopt semantic web technologies for the management of the HCL's catalog.

The bibliographic domain inherits for two centuries of deep changes in cataloging formats from the old-fashioned cards to computer databases which, on the one hand, has mainly increased the complexity of the librarian's job while, one the other hand, did not reconcile the users with library services. Many catalogs from digital libraries are still isolated and inefficient due to the lacks of semantics from the digitized legacy cataloging models. Related studies agrees that necessary improvements in libraries should come from deep changes in cataloging models to adopt a more semantic way to represent bibliographic relationships and to improve knowledge discovery. Thus, new standards and visions from the bibliographic domain like FRBR[3] as well as from the Semantic Web community served as a basis to address HCL's issues. Our main objective was to automatically migrate and enrich the whole catalog towards semantic linked data before integrating the latter into a system based on semantic web technologies.

[3] https://www.loc.gov/cds/downloads/FRBR.PDF

Preprocessing. The first step of the project aimed at analyzing the catalog to tune a system for the automated semantic enrichment process. Such task mostly aimed at extracting all the valuable knowledge patterns from scattered data in the records. For HCL, this task was overwhelming due to the variety of documents in the catalog. Thus, we relied on early research studies[4] to automatically analyze the catalog to detect both inconsistencies and valuable knowledge patterns. The next step aimed to implement the rules extracted from the analysis into an enrichment tool. Yet, because most of existing solutions only provide basic mappings as rules, the latter was too limited to correctly handle all the high level knowledge patterns from the catalog without writing too complex and redundant mappings. Thus, we proposed a novel approach to encapsulate mappings and rules in a pattern-based migration model. The latter aimed to help domain experts to easily discuss and manage the migration rules thanks to the clear representation of high-level patterns and the graph model.
Semantic enrichment. The automated transformation of a catalog is a costly process where each record must be evaluated against the dedicated rules to both construct the new knowledge base and to extract additional knowledge from external sources. In the context of HCL, we chose to rely on our pattern-based model of rules which was implemented as an oriented graph of patterns where each pattern held conditions and mappings written from the first phase. Hence, we adapted a migration process to evaluate each record using this graph. The major improvement of this approach came from the inheritance of conditions between patterns of the oriented graph which prevented useless computation of mappings to finally enhance the global performance of the system.
Data Integration The company Progilone provided the documentation system Syrtis[5] to integrate the migrated and enriched data from the previous phase. Syrtis relies on a graph-based model to provide various features to manage semantic data like RDF graph visualization or semantic search. We just had to provide a pivot model of mappings between the data generated from the enrichment phase and the Syrtis's vocabulary. In the end, the whole HCL's catalog has been successfully migrated and integrated into Syrtis and is available online[6]. The new online catalog based on Syrtis records at least 800 visits in a month.

The use of Semantic technologies to manage the catalog offers new possibilities for users to navigate between bibliographic relationships to better discover the rich knowledge which was hitherto hidden in records. It also reduces the cataloging efforts for practitioners because intellectual or editorial information are now well separated into related entities instead of being aggregated in textual lists. Interoperability of the catalog is also improved thanks to the graph model and the standard vocabularies making it easier to integrate additional data from other repositories. Besides, our ongoing works focus on the automated detection of knowledge patterns from these sources to ease their integration.

[4] `https://hal.archives-ouvertes.fr/hal-01324529`
[5] `http://www.progilone.fr/en/syrtis`
[6] `https://documentationcentrale.docchu-lyon.fr/`