Modelling Library Linked Data in Practice

Three Swiss Case Studies

Nicolas Prongué, René Schneider

Haute Ecole de Gestion, Geneva
7, rte de Drize, 1227 Carouge, Switzerland
{nicolas.prongue, rene.schneider}@hesge.ch

Abstract

In this paper we present reflections and recommendations concerning the conversion of library metadata into Linked Data. We will briefly describe the different data models that exist for this purpose and argue their strengths and weaknesses with a special focus on the entification issue, before illustrating this problem with the description of three different ongoing projects. As will be outlined afterwards, it is essential to distinguish at the start of a project between data-driven or user-driven design approaches. As an alternative, the realization of a Linked Data project for bibliographical data might also lead to a hybrid approach were the design process shifts reciprocally between the analysis of the data and the user needs.

Keywords: Linked Data, Open Data, Data modelling, Data conversion, User Centered Design, Data-driven design

1 Introduction

During decades, the library community has developed technologies and standards for the description of bibliographic information. At present, the fruit of that work comprises on the one hand complete sets of cataloguing rules,

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th–21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 118–128.

which are more or less similar from one country to another. On the other hand various record formats have emerged, also different but all of them based upon the MARC standard. These standards were harmonized more and more over the years.

Since the 1990s, the web has highlighted new opportunities for the publication of bibliographic information, and brought the library community to question the relevance of the actual standards on a more fundamental point of view. The underlying *model* of library data is rethought with the ambition of better integrating library resources with other resources into the web.

This article proposes a reflection around FRBR and BIBFRAME, two new frameworks for data modelling in libraries, and compares them with the traditional framework used until today. It identifies the problems and opportunities raised by this evolution in the context of the transformation, publication and consumption of library data in RDF, the semantic web standard promoted by the World Wide Web Consortium. It discusses the practical process of modelling illustrated by three concrete semantic web projects within the Swiss library landscape. As results, it proposes a generic decision schema for the modelling process in accordance with the factual context.

2 Data models in libraries

For the description of library metadata, new frameworks have been conceived over the past years by the library community, especially FRBR and BIBFRAME. Each of them has a different data structure, which can be represented as illustrated in figure 1.

2.1 The traditional model

The traditional data model is largely based upon the MARC format, developed at the end of the 1960s at the Library of Congress (USA) (McCallum, 2010). MARC is widely spread in the library world, and so is its underlying data model. It consists of one bibliographical entity for one physical resource. Over the years and according to the needs of libraries, this simple model has evolved into a slightly more complicate one, which differentiates item information and bibliographical description. An item record describes

one specific and localized document while a bibliographic record describes a conceptual document that can be materialized several times as items.

The result of this evolution is a two-level model for library data, which currently is the most used model in libraries.

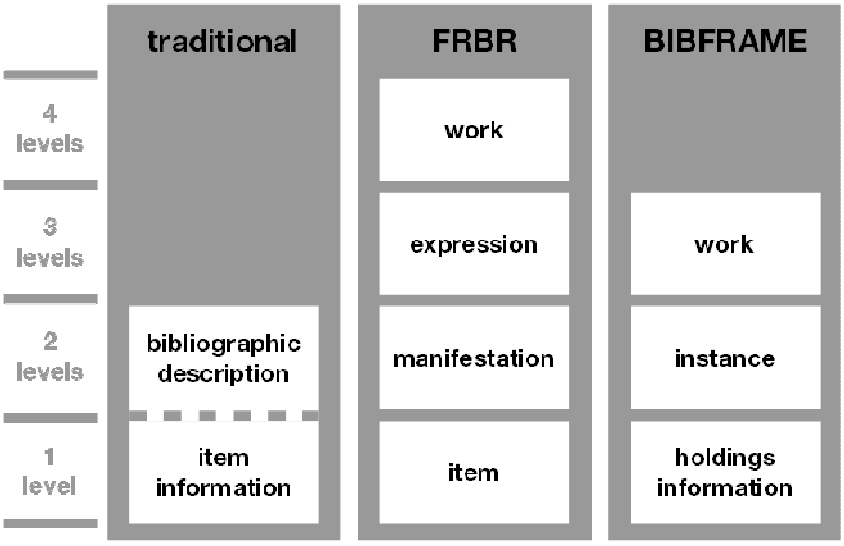

Figure 1. structure of library data models (inspired by Coyle, 2013)

2.2 The FRBR framework

In 1997 the *Functional requirements for bibliographic records* (FRBR) were published by the International Federation of Library Associations (IFLA, 2009). This document is the result of a study whose aim was to redefine the framework of bibliographic records according to the user needs. The entity-relationship model was chosen to be the "language" in which this conceptual framework was described. In FRBR, the bibliographic information is comprised of four core entities: work, expression, manifestation and item. The previous two-level schema does not formally distinguish the work, expression and manifestation information; it embeds them all – and often incompletely – into one entity. The FRBR framework is more complex and its implementation requires highly structured data.

Since its publication, FRBR has gained recognition within the library world. Accordingly, the new set of cataloguing rules RDA (Resource De-

scription and Access), first released 2010, is based upon the four levels of FRBR (RDA JSC, 2014). Various implementation scenarios of RDA exist; it is therefore not indispensable to create an entire FRBR structure as an entity-relationship model (RDA JSC, 2007). Thus it appears, paradoxically, that RDA is in a more advanced implementation stage than its underlying FRBR framework.

2.3 The BIBFRAME model

BIBFRAME is the name of an initiative led by the Library of Congress. This still ongoing project is trying to redesign the bibliographic description in libraries at the era of the web of data. It also aims to replace the current record format, MARC, with a new global standard. The actual work has resulted in a conceptual model, composed of two levels for bibliographic information: work and instance.

This model differs from the others in that it doesn't consider the holdings data as bibliographic information. This information is recorded in so-called *annotations*, which are not characteristics of the instance itself, but contain additional data about it (Library of Congress, 2012).

2.4 Modelling issues

The issues of the library data models concern several areas: data quality and its usefulness, implementation, interoperability between the various coexistent models, capacity to interact with other web resources and acceptance by the library community.

A more complex model like FRBR has the great advantage of enabling more features at the user interface level. The proof of this concept is *data.bnf.fr*¸ the semantic application of the Bibliothèque nationale de France (BnF, 2014). It exploits the potential of FRBR and other interlinking types to create aggregating pages about works, topics or persons among others, in order to increase the user experience and attract traffic coming from web search engines (Hügi & Prongué, 2014).

Nevertheless, the successful implementation of FRBR from MARC records requires a good data quality, as we start from a two-level model to end in a four-level model. The two additional entities, work and expression, and their attributes have to be extracted from groups of bibliographic descrip-

tions, mainly by means of alignment algorithms based on string comparison when the data quality is low. The identification of the existing expressions and of the relationships between them, for example, also needs precise data that is not included in all datasets, like role codes for contributors or the indication of the original title.

In the perspective of a FRBRization, it is not conceivable to add such information manually to the existing MARC-records, given the large amount of data. It may be relevant to start entering them only for new records, while bearing in mind that these new records will be handled in the same interface than the old ones.

In the situation described above as in most cases in libraries, the FRBR structure is not stored in the basis MARC records. It has to be recreated automatically at each database update or at a predetermined frequency. The newly created entities can therefore vary with the data modifications and additions. This can lead, in a Linked Data point of view, to persistence problems for identifiers.

A three-level or four-level model presents a clear benefit for library records, which describe published and widely disseminated documents. This is not the case for other cultural documents, most of them unique, like museum objects or archival collections. The interlinking between those various entities can therefore be challenging, because of the complexity of the data structure, especially when the model is defined in RDF with many constraints (Baker, Coyle, & Petiya, 2014). But on the other hand, the additional conceptual entities (levels 3 and 4 of figure 1) could also be propitious for link generation with web resources issued emanating from other fields of interest.

The library community stands in a perpetual quest for international standards. The FRBR model, released 18 years ago, struggles to establish itself as a standard, and the BIBFRAME initiative, launched 2011, seems to lead the community in a slightly different but more flexible direction. In this context, the model adopted by a few leading libraries or by a majority of institutions will become a real standard. At the moment, there is no clear trend for one solution. The difficulty lies in predicting the emerging trend.

2.5 Beyond the bibliographic entities

The spectrum of bibliographic data is vast. Beyond the information concerning directly bibliographic objects, libraries also collect and create valuable data about various other entities within their authority files. OCLC in its en-

tification project identifies for example the additional entities *concept*, *organization*, *person*, *place, object* (Wallis, 2014). As they do not contain purely bibliographical information, these entities are referred to as "satellite entities" in this paper.

They are nonetheless fundamental for information retrieval, since they represent real-world things. As the user is usually not directly interested in the bibliographic object itself, but rather in its content and context, the search process always passes through the satellite entities. They are especially significant in a Linked Data environment, where the resources have to be interlinked. The benefit of using these entities lies in the existence of value vocabularies available as Linked Open Data, which enable the enrichment of the dataset. The encyclopaedic data provided by DBpedia, or GeoNames in the field of geographical data, are two typical examples of such value vocabularies used in library Linked Data projects. The use of common reference vocabularies on the web across different semantic platforms – in libraries as well as in other communities – leads to a globally interlinked network.

In the modelling perspective, the satellite entities have to interact with the bibliographical entities, by means of links based upon the authorship of a resource, the ownership, the publication place, the topics, the context of creation or exhibition for events, an adaptation or inspiration relationship, etc.

3 Three case studies

In order to further illustrate the problematic presented above, we will continue this reflection by describing shortly three case studies that are currently undertaken in Switzerland. The projects are described briefly because they are all at the beginning level and – in the sense of agile computing – it is still unclear where they will end. Nevertheless, the different initial settings of these three projects will show the diversity of conversion problems and the need for flexible modelling procedures. Despite the Swiss local context, we consider these different cases as sufficiently representative for studies placed in similar settings.

3.1 Case 1: RERO

RERO is the most important library network in the Western part of Switzerland, with a broad range of institutions, from public to academic libraries. In 2014, as a result of a strategic decision, RERO carried out a project whose aim was the publication of their library metadata as Linked Open Data (Prongué, 2014). The expected benefits are an increase in the visibility of library data and the promotion of innovating applications to access it.

The basic data of RERO is composed of two-level MARC records according to the traditional model. A reflection around the conversion of these records forces us inevitably to seek the new possibilities for structuring and modelling them. The question of adopting a more complex data model – like FRBR or BIBFRAME – has been raised; this would add value to the metadata. Nevertheless such an operation represents a substantial investment of time and resources, whereas the project was intended to be a first step towards the semantic web. Moreover, from a technical point a view, the conversion to a three-level or four-level model would be difficult because of the lack of precision in the data. Indeed, due to a trend towards data simplification in the past, some needed information like role codes for authors or uniform titles are no longer entered in the database. The RERO metadata pool shows good internal consistency, but presents a lower level of precision and homogeneity regarding, for example, a FRBRization.

Consequently, a simple model has been preferred at this first step, consisting of one core bibliographical entity, which is the equivalent of the manifestation level from the FRBR framework, interlinked with satellite entities. The objective was to create the best semantic model with the existent data quality, in order to make available the raw metadata to third parties.

3.2 Case 2: LOD-B

The Swiss city libraries possess a great variety of valuable metadata (library catalogue, archive collections, digitized manuscripts, pictures, etc.), stored in separated databases and in different formats. LOD-B (Linked Open Data en Bibliothèques)[1] is a small project which focusses on the diversity of these metadata. It attempts to transform them into a uniform RDF model, and to build innovative services upon it, for human beings as well as for machines.

1 http://campus.hesge.ch/id_bilingue/projekte/lod-b/default.asp

The modelling process has to identify common entities in a selection of interesting datasets to generate bridges between them. Links can hardly be created on the basis of the pure bibliographical entities, because they are completely distinct across the databases. The satellite entities such as *People* or *Place* play a key role in this mechanism. These entities are described with various controlled vocabularies (thesauri, name heading lists, etc.) in the datasets. A core part of this project consists therefore in interlinking the data, e.g. with VIAF. The controlled vocabularies have to be aligned to achieve a consistent and common data model.

The added value lies here in the links between the databases rather than in the internal structure of the bibliographical entities. Consequently, a two-level model was preferred.

3.3 Case 3: linked.swissbib.ch

Swissbib is a platform that centralizes and merges most of Swiss library metadata. On this basis, it provides a search interface for humans and data services for other applications of the domain. To widen its service offer, primarily for human users, but also for machines, Swissbib has embarked at the end of 2014 on a Linked Data project called *linked.swissbib.ch*.

As in the first case, the existing data is of poor precision and the modelling aspect raises therefore the same issues. However, Swissbib does not own the metadata, but harvests it at various institutions participating in the project. As Swissbib cannot manage and adapt the basis data at source, all transformations or harmonization processes must be completely automated, so that they can be executed at least once a day. Consequently, in the case of a conversion into RDF, the generated Linked Data entities can vary, as well as their identifiers. In a reuse purpose, they should be persistent.

Nevertheless, the project perspective is different from the case of RERO. The main focus is not on data reuse, but on the creation of an added value directly for the end user, by an enrichment of the existing search service.

As a starting point for the modelling process, it was decided to adopt a simple one-level or two-level model, and to adapt it gradually according to the new features wanted for the interface. Because of the difficulty to create more structured bibliographical entities, the developed features have to be rather oriented on the satellite entities to generate added value for the end user.

4 Modelling procedures

As basis for the modelling process, very innovating theoretical elements like FRBR and BIBFRAME propose a new and original approach to bibliographic information which – at first sight – seem to suit perfectly to Linked Data projects. In practice, a good implementation of these models can be difficult to achieve and depends heavily on the existing data resources chosen for conversion.

From the three use cases presented, we can deduce two basic perspectives for a data model selection: a data-driven design and a user-driven design perspective (see fig. 2).

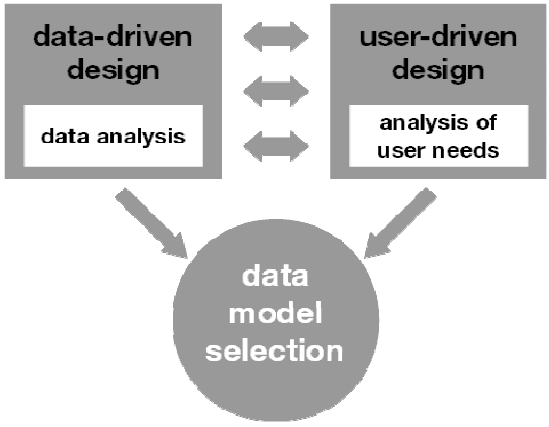

Figure 2. Two perspectives for the modelling process

The first two case studies described above are rather inspired by a data-driven design. They both aim to solve a data issue. The RERO case demonstrates a willingness to open the metadata and make it available for reuse by the widest possible range of third parties. The database merging and alignment goal of LOD-B case also starts from a data-related problematic: heterogeneous datasets that cannot communicate with each other. In the modelling process, such a design focusses on the analysis of the data.

While this approach does not focus directly on the end user, it presents the advantage of heightening the data quality and enabling new features. If the data is made available to third parties, even outside the spectrum of libraries,

innovative and original functionalities can be explored and developed, which the traditional library user may not have imagined.

Conversely, the Swissbib project design can be considered as user-driven, as it targets mainly an improvement of the human interface for better discovery and search services. Through the analysis of user needs, it puts the user in the forefront. Such an approach presents already well-known benefits in user-centered design. To strengthen this approach, several comparative usability tests are undertaken to evaluate the added value of Linked Open Data interfaces for the end user. During the development process, wireframing technologies will be used to assure the realization of a user-friendly interface.

Nevertheless, the Swissbib project also depends on the data available as the other two cases cannot ignore the user needs. The modelling process must be adapted to the context and to the different variables that can impact it.

For the sake of a better understanding, this circumstance may be illustrated metaphorically as the iterative perambulation of a two winged corridor whereas the person that walks through the corridor keeps an eye to both sides. This transition – indicated in figure 2 through the bidirectional pointers – can be seen as the essential procedure for the whole process of data conversion and system modelling.

5 Conclusions

Nowadays, more and more libraries face the challenge of converting and opening their data to the public and are urged to participate in conversion projects. Since out-of-box platforms for the realization of Linked Data projects are not yet at hand, libraries must in some cases develop solutions on their own and at all possible administrative levels (national, regional, municipal), without clearly knowing where they lead to. It is important to start these projects to gain expertise for further projects on a more complex level. Once, when more consolidation concerning models, cataloguing rules and formats will be achieved, this expertise will be valuable.

As explained in this paper, the concrete realization of such projects may vary considerably; it could depend heavily on the quality of the data given and on the data model chosen for their new representation. Besides this, it does also depend on the system objectives defined initially and on the user

needs inquired to build a system that allows a different use of the data for new forms of retrieval and mashing. Therefore, several strategic decisions have to be taken before the start of a conversion project which will lead either to a data-driven approach or a user-driven approach, or, in the ideal case, to a reciprocal shift between both approaches during the whole development process to ensure a maximal yield of both data and user requirements.

References

Baker, T., Coyle, K., & Petiya, S. (2014). Multi-entity models of resource description in the semantic web. *Library Hi Tech*, 32 (4), 562–582. doi:10.1108/LHT-08-2014-0081

BnF (2014). http://data.bnf.fr/ <16.8.2014>.

Coyle, K. (2013, June 29). FRBR and schema.org. http://kcoyle.blogspot.ch/2013/06/frbr-and-schemaorg.html.

Hügi, J., & Prongué, N. (2014). *Les bibliothèques face aux Linked Open Data*. Genève: Haute école de gestion de Genève. http://doc.rero.ch/record/209598/files/M7-2014_memoire_HUGI-PRONGUE.pdf.

IFLA (2009). *Functional requirements for bibliographic records: final report*. IFLA. http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

IFLA (2010). *Fonctionnalités requises des données d'autorité : un modèle conceptuel*. Paris: BnF. http://www.ifla.org/files/assets/cataloguing/frad/frad_2009-fr.pdf.

Library of Congress (2012, Nov. 21). Bibliographic Framework as a web of data: Linked Data model and supporting services. Library of Congress. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.

McCallum, S. (2010). Machine readable cataloging (MARC): 1975–2007. In: *Encyclopedia of library and information sciences* (3rd ed., Vols. I–VII; Vol. V, pp. 3530–3539). Boca Raton: CRC Press.

Prongué, N. (2014). *Modélisation et transformation des métadonnées de RERO en Linked Open Data*. Genève: Haute école de Gestion.

RDA JSC (2007, Jan. 14). RDA implementation scenarios. http://www.rda-jsc.org/docs/5editor2.pdf.

RDA JSC (2014, May 19). RDA: Resource Description and Access. http://www.rda-jsc.org/rda.html <9.12.2014>.

Wallis, R. (2014, Dec. 3). Entification: the route to useful library data. Presented at the SWIB14, Bonn. http://swib.org/swib14/slides/wallis_swib14_2.pdf.