Studying Conceptual Models for Publishing Library Data to the Semantic Web

Sofia Zapounidou(✉)

Department of Archives, Library Science and Museology,
Ionian University, Corfu, Greece
ll2zapo@ionio.gr

Abstract. This thesis studies the library data and the way that linked data technologies may affect libraries. The thesis aims to contribute to the research regarding the development and implementation of a framework for the integration of bibliographic data in the semantic web. It seeks to make sound propositions for the interoperability of conceptual bibliographic models, as well as for future library systems and search environments integrating bibliographic information.

Keywords: BIBFRAME · Conceptual models · EDM · FRBR · FRBRoo · Interoperability · Linked data

1 Research Area

Libraries generate and preserve bibliographic data by describing the resources they provide access and the items they keep in their collections. These bibliographic descriptions have been specified by standards, such as the ISBD family of standards and the AACR2 rules. Codification and sharing of bibliographic data is currently realized using the MARC21 and UNIMARC standards. The MARC structure and format originates in the 1960s and was developed in pace with the technology of the time. Technology has evolved facilitating the delivery of services using structured data from one or more domains. Integration of library data into the semantic web (SW) demands a shift in conceptual data models and data format according to the SW principles and standards. Thus, the library community has started either (re)using existing data models, such as the Functional Requirements for Bibliographic Records (FRBR) and FRBR Object-Oriented (FRBRoo) or developing new models, such as the British Library Data model and the Bibliographic Framework (BIBFRAME) for representing bibliographic information according to the new needs and formats. There are currently projects applying these models and publishing bibliographic data in linked data format (e.g. British National Bibliography, National Library of Spain Catalog). There are also other efforts considering the description of bibliographic entities, broader in scope or related to other domains, such as the Bib.schema.org for the web community, the CIDOC-CRM, the Europeana Data Model and the DPLA Metadata Application Profile for the cultural heritage domain. Other projects related to different aspects of the transition of library data into the SW are the Bibflow project examining future cataloguing workflows, the Linked Data for Production (LD4P) project

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 652–655, 2017.
DOI: 10.1007/978-3-319-67008-9_63

investigating both changes in workflows and extension of BIBFRAME to serve scholarly communication and description of special materials needs. All these initiatives call 'library data' the universe of bibliographic data and their relationships and agree that the shift from legacy to library linked data needs a new framework for bibliographic data definition, representation and interoperability. The existing relationships between the bibliographic data and how they should be preserved in the SW environment is an issue that this research effort focuses on. In general the research is initiated into the context of the following research questions:

- How could library data and their relationships be modeled respecting library domain principles and linked data prerequisites?
- What changes are expected regarding existing tools, practices and services? How will internal workflows and end-user services (existing or potential) be affected?
- How shall libraries collaborate and interoperate with each other, and with third parties, such as archives, museums, publishers, information resource providers?

2 Aim and Objectives

The aim of the thesis is to contribute to the evolution and integration of library data into the SW. The thesis' objectives lie on a framework defined by three axes:

- **New data models.** The thesis seeks to participate in the identification of library data representations that may support enhanced information services within the SW. Specific library data use cases and the way they are represented by library data models shall be examined. Specific element sets and value vocabularies shall be also identified for the expression of certain types of bibliographic information.
- **Workflows.** Adoption of linked data shall affect the overall operation and management of a library. What new services shall or may libraries offer? What kind of data (internal of from third parties) may facilitate the delivery of these services? The thesis seeks to study the requirements for these services in a SW environment by focusing on tools, new workflows and interoperability of data shared between libraries, or between libraries and other organizations.
- **Testbed.** To propose the design and development of a testbed within the scope of assessing the thesis' findings. On the basis of good practices [1, 2] this testbed shall include: (a) test scenarios, (b) test data, (c) workflows and use of data in the scenarios, (d) evaluation criteria regarding the data expressiveness, ease of implementation and use for both libraries and end-users, (e) proposed software tools to be used in implementing the thesis' findings and in managing the test data.

3 Approach and Preliminary Results

The methodology's first step has been a literature survey regarding new library data models and their semantics. The well-known bibliographic conceptual models FRBR, FRBRoo, BIBFRAME, as well as the Europeana Data Model (EDM) have been

studied so far. The expressiveness of the models with regard to specific use cases has been the thesis' first object of investigation. The cases are of varying complexity, from single-volume monographs to aggregates, involving also the representation of content relationships (derivation, adaptation, equivalence, whole-part) and large bibliographic families. The term bibliographic family, as defined in [3], refers to a group of related bibliographic works that somehow derive from a common progenitor. To cover all these cases, selected records from library catalogs have been used.

The research begun with an initial study [4] that revealed similarities and divergences between the models. More specifically, it exhibited that there is "more common ground between FRBR and FRBRoo, and between EDM and BIBFRAME". The similarity between FRBR and FRBRoo was expected, since the latter extends the semantics of the former. The common ground between EDM and BIBFRAME seemed interesting for interoperability reasons. Earlier in the same year the EDM-FRBRoo application profile was presented [5]. Therefore, the research proceeded [6, 7] focusing on a possible BIBFRAME-EDM application profile. The study presented in [7] used four different approaches for the representation of monographs in EDM: according to the library alignment report, using the *ore:Proxy* class, using the *edm:InformationResource* class, and using both *ore:Proxy* and *edm:InformationResource* classes. All three studies [4, 6, 7] revealed that for the case of monographs semantic interoperability between the models is desired and possible.

This finding was further studied in the four models (FRBR, FRBRoo, BIBFRAME, EDM) using more bibliographic records as test data. The bibliographic records were selected to study content relationships also. The content relationships studied are derivation, equivalence and the whole-part relationship. This study [8] identified that for the description of each case, there can be more than one alternative representations enabled by each model's semantics. The research has also revealed similarities and divergences between the models that may facilitate or hinder interoperability and sharing of library data. This last finding is currently tested in mappings from FRBR to BIBFRAME using bibliographic records as test data. All bibliographic records describe works from the same bibliographic family. The mappings are also studied for the preservation of the explicit and the implicit relationships between members of a bibliographic family. Control of bibliographic families and explicit linkages between their members would enhance navigation in a linked data environment. During these tests, it has been discovered that there are conditions enabling mappings, e.g. the existence of a specific attribute of a class or a specific value to an attribute, and that different mappings may be required for the preservation of content relationships. The findings may influence future cataloguing policies, workflows, software interfaces, as well as prospective mappings for sharing or integration purposes.

4 Planned Work and Future Directions

The research plans to develop more mappings to transform data from FRBR and BIBFRAME and vice versa, and to evaluate them using records from library catalogs. The consolidated FRBR model (FRBR-LRM) shall be approved in 2017 and mappings will take under consideration any changes with regard to Group 1 entities, on which

this thesis focuses. Since preservation of bibliographic families is an important issue, probably separate mapping algorithms are going to be developed for each bibliographic case. At this stage, it will also be evaluated whether existing MARC21 records include explicit statements of content relationships that may enable preservation of bibliographic families after transformation of data. The mapping algorithms may be expressed in a mapping language, such as the x3ml [9] or the RML [10]. Another issue to be investigated is whether available software tools for automatic conversion of MARC21 data in linked data, produce conversions that ensure preservation of content relationships and bibliographic families.

References

1. Ogle, V., Wilensky, R.: Testbed development for the Berkeley Digital Library Project. D-Lib Mag. **2**, 1–6 (1996)
2. Strodl, S., Rauber, A., Rauch, C., Hofman, H., Debole, F., Amato, G.: The DELOS testbed for choosing a digital preservation strategy. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 323–332. Springer, Heidelberg (2006). doi:10.1007/11931584_35
3. Smiraglia, R., Leazer, G.: Derivative bibliographic relationships: The work relationship in a global bibliographic database. J. Am. Soc. Inf. Sci. **50**, 493–504 (1999)
4. Zapounidou, S., Sfakakis, M., Papatheodorou, C.: Highlights of library data models in the era of linked open data. In: Garoufallou, E., Greenberg, J. (eds.) MTSR 2013. CCIS, vol. 390, pp. 396–407. Springer, Cham (2013). doi:10.1007/978-3-319-03437-9_38
5. Doerr, M., Gradmann, S., Le Boeuf, P., Aalberg, T., Bailly, R., Olensky, M.: Final Report on EDM–FRBRoo Application Profile Task Force. (2013)
6. Zapounidou, S., Sfakakis, M., Papatheodorou, C.: Integrating library and cultural heritage data models: the BIBFRAME - EDM case. In: Panhellenic Conference on Informatics Proceedings. ACM, Athens (2014)
7. Zapounidou, S., Sfakakis, M., Papatheodorou, C.: Library data integration: towards bibframe mapping to EDM. In: Closs, S., Studer, R., Garoufallou, E., Sicilia, M.-A. (eds.) MTSR 2014. CCIS, vol. 478, pp. 262–273. Springer, Cham (2014). doi:10.1007/978-3-319-13674-5_25
8. Zapounidou, S., Sfakakis, M., Papatheodorou, C.: Representing and integrating bibliographic information into the Semantic Web: a comparison of four conceptual models. J. Inf. Sci. (2016). doi:10.1177/0165551516650410
9. Kondylakis, H., Doerr, M., Plexousakis, D.: Mapping Language for Information Integration. Technical report 385. Institute of Computer Science, FORTH-ICS, Heraklion, Crete, Greece (2006)
10. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., de Walle, R.: RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In: Bizer, C., Heath, T., Auer, S., Berners-Lee, T. (eds.) Proceedings of the 7th Workshop on Linked Data on the Web, co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, 8 April 2014