Evaluation of Mappings from MARC to Linked Data

Hyoungjoo Park (park32@uwm.edu) and Margaret E.I. Kipp (kipp@uwm.edu)

Northwest Quad Building B. 2025 E. Newport, Milwaukee, WI, 53211, USA
School of Information Studies
University of Wisconsin – Milwaukee

1. Introduction

The purpose of this study is to assess the quality and compatibility of library linked data (LLD) schemas in use or proposed for library resources. Linked Data (LD) has the potential to provide high quality metadata on the web with the ability to incorporate existing structured data from MARC via a mapping. Researchers selected representative libraries such as Harvard University Library, LC BIBFRAME (Library of Congress Bibliographic Framework), OCLC (Online Computer Library Canter) WorldCat, and National Library of Spain. For LD frameworks, four resources are matched into specific categories with MARC (MAchine-Readable Cataloging) tags so that it could be retrieved in both OCLC LD and BIBFRAME with the conversion tool at bibframe.org: (1) Classic, e-book, and fiction, (2) multiple authors and part of a series, and non-fiction, (3) varying title, translation, and fiction, and (4) sub title, non-fiction. This study revealed that the choices and elements of each library made in local decisions might bring interoperability issues for LD services due to the quality metadata creation issues.

2. Literature Review

Library data is a rich format which would enhance current Semantic Web projects, however, library data is stored in a format which is not compatible with current standards used on the web or the Semantic Web. Libraries have begun to examine possible schemas to convert MARC records into Linked Data formats, but such schemas are relatively untested and need to be studied for their long term data sharing potential. Such initiatives to share library data on the web could enhance the potential for discovery of library resources by linking bibliographic data into the web environment (Cole, 2013).

Several studies have been conducted which examine the mapping of MARC into LD (Styles et al. 2008; Vila-Suero et al. 2012; Kroeger 2013; Kumar et al. 2013). Styles et al. (2008) investigated the representation of MARC in LD format as part of a project to create an automatic mapping from MARC to Resource Description Framework (RDF) (Styles et al. 2008). Vila-Suero et al. (2012) studied mapping and transforming of MARC to RDF using the Functional Requirements for Bibliographic Records (FRBR) ontologies and a dataset from the Spanish national library (Vila-Suero et al. 2012). Kumar et al. (2013) used the Harvard Library Bibliographic Dataset to create a mapping from MARC to RDF (Kumar et al. 2013). OCLC provides linked data records using Schema.org and other metadata schemas (Fons et al. 2012) while the Library of Congress is currently developing a new linked data framework, BIBFRAME for Resource Description Access (RDA) (Godby 2013; Library of Congress 2011). Kroeger (2013) studied the need for the mapping of MARC into a LD metadata structure in a library environment.

3. Methodology

To analyse the quality of the metadata and its potential for use in libraries, the researchers transformed a sample set of MARC 21 records into four published metadata frameworks: Harvard Library (Kumar et al. 2013), National Library of Spain (Vila-Suero et al. 2012), OCLC WorldCat's Linked Data (Fons et al. 2012), and Library of Congress BIBFRAME. (Godby 2013).

The National Library of Spain provided open access to their mapping from MARC (and FRBR). The researchers used the MARC tags from the chosen records (Table 1) to create a preliminary mapping between the other three frameworks.

Characteristics	Resources
Classic, e-book, and fiction	Pride and Prejudice/ Jane Austen (ebook)
Multiple authors and part of a series, and non-fiction	Organization of Information 3rd edition written by Arlene G. Taylor and Daniel N. Joudrey
Varying title, translation, and fiction	Please look after mom written by Kyung-Sook Shin
Sub title, non-fiction	Managing cataloguing and the organization of information: philosophies, practices, and challenges at the onset of the 21st century written by Ruth C Carter

Table 1: Resources for comparison of LD frameworks

Each selected resource is held by an OCLC member library and the LC, so records could be retrieved in both OCLC linked data and BIBFRAME by using the conversion tool at bibframe.org. Each resource matches specific categories that will be of interest to librarians and users attempting to use linked data library resources as shown in Table 1.

The preliminary mappings from the four sample MARC records to each of the four schemas were then used to create a comparison chart of the use of elements between the four schemas.

MARC	MARC Description	Harvard University Library	LC BIBFRAME	OCLC WorldCat	National Library of Spain
020 a	Standard Numbering	marcont:hasISBN	bf:isbn13 or bf:isbn10	schema:workExample	hasStandardIdentifier
100 a	Main Entry - Personal Name	marcont:hasAuthor	bf:creator	schema:author	hasNameOfPerson
245 a	Title Proper	rdagroup1elements:keyTitle	bf:titleStatement	schema:name	hasTitleProper
245 h	GMD	-	-	rdf:type	hasContentFormAndMediaTypeStatement
245 c	Statement of Responsibility	marcont:hasAuthor	bf:responsibilityStatement	-	hasStatementOfResponsibilityRelatingToTitle
260 a	Place of publication	rdagroup1elements:placeOfPublication	bf:providerPlace	library:placeOfPublication	hasPlaceOfPublicationProductionDistribution
260 b	Name of Publisher	rdagroup1elements:publishersName	bf:providerName	schema:publisher	hasTitleOfIndividualWorkBySameAuthor
260 c	Date of publication	-	bf:providerDate	schema:datePublished	hasDateOfPublicationProductionDistribution
300 a	Extent	dcterms:extent	-	schema:numberOfPages	hasSpecificMaterialDesignationAndExtent
300 c	Dimensions	rdagroup1elements:dimensions	bf:dimensions	-	hasDimensions
500 a	General Note	marcont:hasNote	bf:note	-	hasNote
650 a	Subject Added Entry - Topical Term	-	bf:subject	schema:about	-
700 a	Added Entry - Personal Name	-	bf:contributor	schema:contributor	hasNameOfPerson

Table 2. Comparison of LD Schemas

4. Findings

4.1 Choice of Schemas

Each institutional schema suggested for converting library data to the Semantic Web involved choosing whether to create a new schema or use existing schemas. Existing schemas in the Semantic Web may not provide library specific elements and my thus require the creation of new elements or the loss of specificity in data. The choice of existing schemas also involves examining available schemas for long term viability.

Harvard University Library

The project from the Harvard University Library used three schemas or ontologies for mapping MARC to the Semantic Web: marcont, rdagroup1elements and dcterms. MarcOnt was an integration ontology for bibliographic description format developed in 2005 at DERI (Digital Enterprise Research Institute). This ontology was based on MARC, DC and BIBTEX (Kruk, Synak & Zimmermann 2005). Marcont was maintained until 2008. The original website for the ontology is no longer available as of 2010, but the ontology can be retrieved using the Wayback Machine (https://web.archive.org/web/20071117023047/http://www.marcont.org/ontology/marcont-2.0.owl) or via the sourceforge page (http://sourceforge.net/projects/marcont/). The rdagroup1elements are from the RDA (Resource Description and Access) Vocabularies registered at the open metadata registry (http://rdvocab.info/). The rdagroup1 elements are now deprecated. These elements were entered between 2009 and 2012 and have not been updated since 2012. Finally, the Harvard University Library mapping used dcterms for the extent element. DCTERMS is the extended version of the Dublin Core schema, which has been continuously maintained since 1995 (http://dublincore.org/).

Library of Congress BIBFRAME

The BIBFRAME elements developed by the LC are based on a simplified model of work - instance developed by the Library of Congress for use with RDA (Library of Congress 2012). BIBFRAME has been in continuous development since 2012 (http://bibframe.org/).

OCLC WorldCat

OCLC used Internet standards such as Schema.org and RDF to conform to existing standards in use on the web and created a library specific ontology for elements that were not well served by schema.org or RDF, for example library:placeOfPublication (Fons et al. 2012). Schema.org was created by Bing, Google and Yahoo! as a common schema for use in web pages. It is recognised by each of these search engines and is continuously maintained. Schema.org allows for the creation of new related schemas (http://www.schema.org/docs/gs.html). RDF (Resource Description Framework) was created by the RDF Working Group of the W3C (World Wide Web Consortium) in 2004. RDF was intended to provide a standard data interchange model for the web and as the language of a semantically marked up web, in other words the Semantic Web. It has been continuously maintained and was last updated in 2014 (http://www.w3.org/standards/techs/rdf#w3c_all).

National Library of Spain

The project from the National Library of Spain made use of FRBR (Functional Requirements for Bibliographic Records) principles to map MARC records to a Semantic Web format. They used elements from the open metadata registry from the list of ISBD (International Standard Bibliographic Description) elements and the FRBRER (FRBR entity-relationship model) elements. The ISBD elements are based on the ISBD standard and map well to existing MARC tags. They were published by De Gruyter Saur in 2011 (http://metadataregistry.org/schema/show/id/25.html). The FRBRER elements were entered into the open metadata repository in 2010 and were based on the 2009 FRBR model developed by IFLA (International Federation of Library Associations) (http://metadataregistry.org/schema/show/id/5.html).

4.2 Elements

While all four frameworks differ in use of elements, there are obvious overlaps between them in that all have an element for title, author and publishers. However, we found that not all of the schemas have subjects and that none of the schemas dealt with series.

- Harvard University Library
    - 245 a (Title proper): Harvard University Library used rdagroup1elements:keyTitle instead of title. This schema appears to be for continuing resources such as serials, therefore, this might not be the best mapping for a universal system. keyTitle is a subproperty of Title, which would provide a better universal mapping and match the other schemas
- Library of Congress BIBFRAME
    - 260 a (Place of Publication): LC BIBFRAME has a substantially different element name for placeOfPublication than the other schemas
- OCLC WorldCat
    - 245 c (Statement of Responsibility) is missing from the OCLC mapping
    - 300 c (Dimensions) is also missing from the mapping
- National Library of Spain
    - 260 b (Publisher Name): HasTitleOfIndividualWorkBySameAuthor is used for publisher name in the Spanish schema, however, another element exists in the ISBD element set which would map better to publisher HasNameOfPublisherProducerDistributorStatement (http://iflastandards.info/ns/isbd/elements/P1169)

5. Discussion and Conclusions

This study revealed that libraries which are developing LD frameworks are working alone and using different ontologies and metadata schemas to convert the same library data into a linked data format. The use of very different schemas and frameworks means that records are not immediately interoperable. The use of different underlying conceptual models – some records made use of FRBR principles while some did not – also impacts on interoperability as separate work and item descriptions would need to be converted to a single record and vice versa.

The Harvard schema concentrated on serials, but could be adapted to other bibliographic materials with some modifications (e.g. using title rather than keytitle). The other schemas were all designed for more universal bibliographic control. The researchers noted that none of the schemas dealt with series titles. This is likely due to the LC's decision to cease series authority control in 2006 (Library of Congress 2006). However, Dunham et al. (2014) suggested that this could impact user ability to retrieve series items. This could be even more important in a public library environment where series may play an even greater role than an academic library. Additionally, subject was not dealt with in all of the schemas. The Harvard and Spanish National Library schemas did not provide suggested mappings for subject terms, while OCLC and LC did. Again, subject is an important access point for retrieval and subject authority control is limited on the web, which is thus a potentially important contribution which LD library records could make to the wider web community.

As noted in Table 2, some elements differ significantly between schemas. The researchers suggest that it is important to determine which schema elements provide the most specificity while retaining usability for users since all four suggested schemas are different. If competing schemas are eventually adopted, automated crosswalks will be necessary. Consistency in bibliographic records with quality metadata would enhance interoperability and improve access and retrieval in an online environment.

Based on our findings, it is clear that one of the most important aspects of creating a library schema for the Semantic Web is the maintenance of the component parts of the schema or schemas over the long term. Marcont has not been maintained since 2008 and the rdagroup1elements schema has been deprecated. In order to gain traction in the wider web community, library schemas must continue to be maintained. Schemas such as Dublin Core are already used in the Semantic Web along with Schema.org and RDF, which suggests that a well-maintained and extensible schema will be most likely to be used for the long term. Library catalogue records contain rich data which could enhance existing efforts to provide rich metadata on the web. Further research is necessary to determine the best method for providing access to this data.

6. References

Cole, T. (2013). Library Marc Records Into Linked Open Data: Challenges and Opportunities. *Journal of Library Metadata*, (2), 163-196

Dunham, B., McGurr, M., & el-Sherbini., M. (2014). Series Authority Control: Potential Effects of Library of Congress' Decision on Users at the Ohio State University. http://library.osu.edu/staff/cataloging/intrep_series_authority_control.php

Fons, T., Penka, J., & Wallis, R. (2012). OCLC's Linked Data Initiative: Using Schema.org to Make Library Data Relevant on the Web. *Information Standards Quarterly*, 24(2/3), 29-33.

Godby, C. J. (2013). The Relationship between BIBFRAME and OCLC's Linked - Data Model of Bibliographic Description: A Working Paper. http://oclc.org/content/dam/research/publications/library/2013/2013-05.pdf

Kroeger, A. (2013). The Road to BIBFRAME: The Evolution of the Idea of Bibliographic Transition into a Post-MARC Future. *Cataloging & Classification Quarterly*, (8), 873-890.

Kruk, S. R., Synak, M., & Zimmermann, K. (2005). MarcOnt--Integration ontology for bibliographic description formats. *In International Conference on Dublin Core and Metadata Applications, p. 231 - 234*

Kumar, S., Ujjal, M., & Utpal, B. (2013). Exposing MARC 21 Format for Bibliographic Data As Linked Data With Provenance. *Journal of Library Metadata*, 13(2), p.212-229

Library of Congress. (2006). Library of Congress to Cease Series Control," *Information Bulletin* 37, no. 3 (2006): 151-152.

Library of Congress. (2011). *Bibliographic Framework Initiative General Plan*. Retrieved from http://www.loc.gov/marc/transition/news/framework-103111.html

Library of Congress (2012). Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services. Retrieved from http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

Styles, R., Ayers, D., & Shabir, N. (2008). Semantic Marc, MARC21 and the Semantic Web. In *Linked Data On the Web*. Retrieved from http://events.linkeddata.org/ldow2008/papers/02-styles-ayers-semantic-marc.pdf

Vila-Suero, D. & Gomez-Perez, Asuncion. (2013). datos.bne.es and MARiMbA: aninsight into library linked data. *Library Hi Tech*, 31(4), p. 575-601.