Ya-Ning Chen

An FRBR-based approach for transforming MARC records into linked data

Abstract
In this study, a practical workflow is outlined for transforming MARC records into Linked Data according to classes and properties of FRBR and FRAD defined in the RDA Registry. Four bibliographic records of ‘Pride and Prejudice’ were selected as subject. One was in English and the others were Chinese records to illustrate the hidden relationships embedded in MARC records, including translation, version and reproduction, relationships between classes in FRBR group 1, and relationships between FRBR group 1 and FRAD for Linked Data. In total, the proposed workflow is composed of three stages and five tasks as follows: data preparation and selection (comprising two tasks, i.e., changing MARC tags into semantic labels with their data and selecting significant labels and their data for Linked Data), data modeling (comprising two tasks, i.e., aligning MARC labels and their data with semantic equivalent classes defined in the RDA Registry and building the relationships between selected classes through the properties of the RDA Registry), and reusing existing value vocabularies for external linking (comprising one task, such as linking terms from VIAF, TGN and so on). Issues related to de-duplication and collaboration of LD are also addressed.

1. Introduction
Traditionally libraries have provided access to scholarly information resources for scientific research and communication. Most of the accessible scholarly resources owned by libraries are represented by complex structures and attributes in a Machine-Readable Cataloguing (MARC) format, MARC having been used as a means to interchange bibliographic and authority records across various information systems within the library community. Although a huge volume of MARC records are valuable in the bibliographic universe, most are not included as part of the Google search engine which has limited their direct use by netizens on the Internet. Recently, Linked Data (LD) has become the preferred approach for the conversion of MARC-based legacy records into a part of the semantic web by libraries. Based on the principles of LD, MARC-based catalog records can be sliced into LD and then be aggregated with other external resources and their contexts. With the advancement of LD, official documents released by W3C (Hyland, Atemezing and Villazón-Terrazas 2017; Hyland and Villazón-Terrazas 2011) can be regarded as useful best practices for authoring and publishing LD.

With the application of the semantic web, the Functional Requirements for Bibliographic Records (FRBR) has been regarded as a conceptual reference model (CRM) for the bibliographic universe. Many projects and studies have argued that FRBR can feasibly be used as a CRM for the semantic web. On the other hand, the entities and relationships defined by FRBR and Functional Requirements for Authority
Data (FRAD) have been implemented as classes and properties into the RDA Registry (RDAR, http://www.rdaregistry.info/) with permanent namespace and identifiers (RDA Steering Committee, 2016); however, there is still a lack of common agreement on a best practice for the implementation and use of FRBR and FRAD in the library community. Therefore, a question worth investigating is how classes and relationships of FRBR and FRAD defined in the RDAR can be employed as a CRM to transform existing MARC records into LD in terms of workflow for LD generation.

2. Literature review

LD has become an accepted approach for combining related information with diverse viewpoints for the same resource. Libraries have taken advantage of LD to free MARC-based catalogue data so that they can enrich their contextual content with cultural heritage from the Internet. Although best practices have been released by W3C (Hyland and Villazón-Terrazas 2011; Hyland, Atemezing and Villazón-Terrazas 2017), it is still a challenge to know how to move from theoretical discussion into practical implementation of LD for libraries (Hanson 2014). Libraries need to know how LD are actually created and published in practice (Hanson 2014), and be able to understand many unexpected issues (Bowen, 2010), such as linking to the exact URI (Cole et al. 2013) and workflow (Bowen 2010; Cole et al. 2013; Di Noia et al. 2016; Hallo, Luján-Mora and Trujillo 2014; Hanson 2014; Lampert and Southwick 2013; Southwick 2015).

In addition to BibFrame, FRBR is another popular CRM for LD adopted by libraries. FRBR, FRAD and Functional Requirements for Subject Authority Data (FRSAD) have been used for various applications, including as a design basis for the RDA and a CRM for data modeling during conversion of MARC into LD. In the cases of the National Polytechnic School in Ecuador (Hallo, Luján-Mora and Trujillo 2014) and the LIBRIS in Sweden (Malmsten 2008), FRBR group 1-2 has been employed, whereas in the case of the Bibliothèque Nationale de France (BNF) (Simon et al. 2013; Wenz 2013) and the National Library and Archives of Iran (NLAI) (Eslami and Vaghefzadeh 2013) FRBR 1-3 group has been used. Furthermore, the Biblioteca Nacional de España (BNE) has adopted FRBR, FRAD and FRSAD to delineate the relationship between LD embedded in bibliographic and authority data (Vila-Suero, Villazón-Terrazas and Gómez-Pérez 2012). During transformation, either data mapping (e.g., BNE) or data alignments (e.g., BNF, NLAI) were utilized to select or link to exact URI or terms from CRM classes or metadata elements. In 2014 the RDA Steering Committee (RSC) built up the RDAR composed by FRBR and FRAD with unique URIs for each entity and the relationship defined by the two aforementioned CRMs (RSC 2014). Existing cases have justified FRBR and FRAD as a feasible CRM for transforming MARC records into LD. Research is still lacking on how libraries can split MARC records into LD with a practical workflow based on RDAR.
3. Methodology

In this study, four bibliographic MARC records of ‘Pride and Prejudice’, authored by Jane Austen, were selected from the WebPAC of the National Taiwan University Library (http://tulips.ntu.edu.tw) as subject: one record was in English and the others were Chinese. Hidden relationships embedded between bibliographic MARC records including translation, version and reproduction, relationships between entities in FRBR group 1, and relationships between FRBR group 1 and FRAD were also included in this study. Three components provided by Hyland and Villazón-Terrazas (2011) were chosen as a basis to develop the workflow for transforming MARC into LD and examine the related issues (i.e., LD de-duplication and collaboration) for library-oriented LD as follows: modeling, naming with URIs, and reusing existing vocabularies.

4. Results and discussion

The workflow outlined in this study is composed of the following stages and tasks: data preparation and selection (including 1st and 2nd task), data modeling (including 3rd and 4th task), and reusing existing value vocabularies for external linking (including 5th task). The first task is to change MARC tags into semantic labels with their data. KO professionals need to change MARC tag numbers and subfield codes into semantic labels based on the MARC definition. The second task is then to select significant labels and their data for LD. Not all MARC tag numbers and subfield codes are, in fact, suitable for LD. Libraries can follow the principle “of interest” (Hyland and Villazón-Terrazas 2011) to select appropriate data from the significant MARC labels for LD. At the second stage, the most important task is to embody conceptual entities/classes and their relationships/properties in harmony with the classes and properties defined by the RDAR. The third task is, therefore, to use a crosswalk, or alignment table, to determine the semantic equivalence of selected labels regarding their MARC data and RDAR classes. Based on the results of the third task, the fourth task is to build relationships between selected classes through their RDAR properties (shown as Fig. 1). Furthermore, real instances of MARC labels and their data have been added to the aforementioned classes and their RDAR properties to validate the appropriateness of classes and their relationships. The final task is to reuse existing value vocabularies for external linking (as shown in Fig. 2). The purpose of this task is to enable library LD to become part of global LD on the Internet, and to expand the reuse or consumption by others as part of the knowledge graph generated, which is based on LD and related links.
5. Discussion

It is not an easy task to select an exact URI to distinguish individual LD from one another, simply based on URI, lexical meanings or appearances with insufficient information. Indeed, knowledge organization (KO) professionals need to perform a
series of knowledge checks with more related background information. Although LD is a useful basis for KO collaboration to seamlessly aggregate various external resources together, fundamental collaborative LD still relies on de-duplication and related issues, such as mapping and selection principles for reusing existing metadata element sets, value vocabularies and library datasets.

5. Conclusion

LD is a good approach for the practice of “thinking globally and acting locally” for KO. Although RDAR has paved the way for LD to be used by libraries with a combination of RDA and FRBR family members (e.g., FRBR and FRAD), a new approach for KO such as LD is also a paradigm shift. If the future trend of LD-based KO is to be distributed and aggregated over the Internet seamlessly, the formulation of new policies and guidelines is essential to obtain a common agreement on de-duplication, collaboration, mapping and selection for library-oriented LD.

References


