Instructions, interfaces, and interoperable data: the RIMMF experience with RDA

Gordon Dunsire
Independent Consultant, Edinburgh, Scotland
E-mail address: gordon@gordondunsire.com

Deborah Fritz
TMQ, Inc., Melbourne Beach, Florida, USA.
E-mail address: deborah@marcofquality.com

Richard Fritz
TMQ, Inc., Melbourne Beach, Florida, USA.
 E-mail address: richard@marcofquality.com

Abstract:

This paper presents a case study of a software tool developed to bridge the gap between cataloguing rules based on the IFLA FRBR family of conceptual models of bibliographic entities and relations, now in the final stages of consolidation in the FRBR Library Reference Model, and cataloguers maintaining bibliographic data in systems based on inventory and text-processing applications. RDA: Resource Description and Access is the successor to the Anglo-American Cataloguing Rules and a major implementation of the FRBR model, but the data it is intended to produce is primarily maintained and used in MARC 21 encoding for use in applications that have not essentially changed in the past 40 years. In 2011 a small software company, TMQ Inc., began the development of a prototype cataloguing interface designed purely for RDA and the FRBR model. The RIMMF (RDA in Many Metadata Formats) software package has been tested in numerous field trials, the most notable being the "jane-athons" or hackathons for RDA data involving a wide range of practicing cataloguers. The paper describes how RIMMF interacts with the cataloguing guidance and instructions in the RDA Toolkit and the data elements and value vocabularies in the RDA Registry to present a set of cataloguer-friendly multilingual data input and editing interfaces, and discusses the additional IT infrastructure required to support future operational cataloguing systems.

Keywords: cataloguing, RDA, FRBR, multilingual linked data, terminology.

Introduction

Modern approaches to bibliographic and cultural heritage resource metadata began to emerge at the beginning of the 21st century, following the International Conference on the Principles

and Future Development of AACR held in 1997. AACR, the Anglo-American Cataloguing Rules, was used as a metadata content standard in harmony with the International Standard Bibliographic Description (ISBD) that specified the sources of data values and their formatting and display, and the MARC (machine-readable cataloguing) data exchange format used for storing AACR and ISBD data. The focus of these standards was the "resource", essentially a tangible information carrying object typified by the printed book. AACR was unable to cope with the proliferation of digital content resulting from evolution of information technology, the Internet and World-Wide Web, and the development of the third edition (AACR3) mutated in 2005 into RDA: resource description and access. RDA adopted the entity-relationship analysis provided in the Functional Requirements for Bibliographic Records (FRBR) conceptual model published in 1998 by IFLA (International Federation of Library Associations and Institutions) [1]. FRBR and ISBD were components of IFLA's development of Universal Bibliographic Control (UBC), in pursuit of standards that would improve access to information through metadata that could be shared at global level. FRBR was followed in 2009 by Functional Requirements for Authority Data (FRAD) and in 2010 by Functional Requirements for Subject Authority Data (FRSAD). By then UBC had been abandoned by IFLA, unable to impose standards for the global onto a network where everywhere was becoming local and everyone was becoming a supplier and consumer of online information. The technology was moving on, delaying the development of FRAD and exposing cracks in the façade [2]. The idea of the Semantic Web, "a common framework that allows data to be shared and reused across application, enterprise, and community boundaries", was proposed in 2001 and influenced the foundations of RDA before it was published in 2010. This Resource Description Framework (RDF) has also been used as the basis of the 2016 draft for world-wide review of FRBR-Library Reference Model (FRBR-LRM), a consolidation of FRBR, FRAD, and FRSAD into a single conceptual model. This joins initiatives beyond the library community, such as Google's schema.org, to exploit RDF and linked data for large-scale resource discovery.

There has been a profound shift in the conceptual basis of cataloguing and the maintenance of metadata for information resources. The idea of describing an item in a library collection using a monolithic record has been replaced with the more granular approach of FRBR that identifies four connected entities for accommodating data about different aspects of the resource, allowing greater flexibility in relating separate resources with common characteristics and reducing duplication. This was clearly reflected in the change of terminology in the titles of the Functional Requirements models. The IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR) developed FRAD. The change from "authority records" to "authority data" is intentional: "The model focuses on data, regardless of how it may be packaged (e.g., in authority records)" [3]. This approach is continued in FRBR-LRM [4]. There is also an increasing emphasis on relationships between bibliographic entities, going beyond the resources themselves to cover related entities. This is exemplified in FRBR-LRM, for example where the addition of the entity Place transmutes an attribute such as "place of publication" into a relationship between a manifestation and a place. Relationships support navigation between entities and allow description to focus on each entity in itself rather than in a specific context, allowing data from different communities and applications to be linked and re-used.

The harmony of 1990s has now been stretched to breaking point. Large-scale catalogue management systems have naturally lagged behind the development of new models and practices. A library cataloguer is typically expected to apply linked data concepts from RDA and FRBR in workflows based on AACR2 to create data in MARC21, a format designed in

the 1960s to automate the production of catalogue cards. The outcome is inefficient, ineffective, and confusing for cataloguer and end-user alike. The adaption of familiar labels for new contexts and meanings is fundamental in human discourse, so it is no accident that FRBR-LRM makes a clear distinction between a "nomen" that labels something, and the thing that is labelled; this distinction can be applied to the language of cataloguing itself.

RDA

RDA is defined as "a package of data elements, guidelines, and instructions for creating library and cultural heritage resource metadata that are well-formed according to international models for user-focused linked data applications" [5]. RDA Toolkit provides the user-focused elements, guidelines, and instructions [6]; the RDA Registry provides the infrastructure for well-formed, linked, RDA data applications [7].

Terminology issues were identified at the very beginning of the development of RDA [8]. Attempts to develop AACR3 failed in part because it proved impossible to reconcile the semantics of the terms "work" and "item" with FRBR. RDA started anew with the FRBR definitions and contexts, but other issues remain. The term "resource" carries two distinct semantics that need to be appropriately linked if RDA and ISBD data are to interoperate [9]. To confuse things further, the Semantic Web and linked data communities define the "resource" of RDF as synonymous with "entity" and include places, persons, and concepts [10]. Other terminology issues occur because of differences in technical vocabularies. The RDA element set documentation and RDA Toolkit use the entity-relationship language of FRBR: "entity", "element", "sub-element", and "element sub-type". The RDA Registry uses the language of RDF and linked data: "class" and "property" [11]. This can make it difficult for cataloguers and application developers to communicate.

There is a particular problem with recording data for related entities. RDA Toolkit (section 17.0) notes "Some encoding standards may not have a design that is suitable for recording the primary relationships. In these cases, primary relationships are not explicitly recorded though they may be inferred from other data elements in composite descriptions." Primary relationships are those that link the work, expression, manifestation, and item entities for a single resource. A cataloguer describing a "resource" in MARC21 may not fully appreciate the FRBR semantics underlying the RDA instructions. Relationships between different resources are recorded using RDA relationship designators that make an explicit distinction between the FRBR entities, so it is important that the correct designator is used to ensure coherency and consistency with the FRBR model. The labels used in RDA Toolkit reflect the RDF semantics of the RDA Registry, so a mistake can also affect linked data. The RSC Technical Working Group is developing RDA guidance and instructions for a range of data storage scenarios [12], using unstructured descriptions, structured descriptions that aggregate data from two or more elements, and text string identifiers [13]. This will introduce an additional layer of technical complexity and terminology that needs interpretation for the cataloguer using RDA in MARC21 and other "flat" encoding formats.

RIMMF

TMQ Inc. was started in 1992 as "The MARC of Quality" (TMQ) by two professional librarians, and has been successful in developing software products for improving the quality of MARC21 data, and training materials for cataloguers using AACR2 and MARC21 [14]. The company followed the development of RDA with interest because it was promoted as the

successor to AACR2. In 2011, following the publication of RDA, the company set out to see what impact the new standard might have on cataloguers and cataloguing processes by developing a prototype cataloguing interface designed purely for RDA and the underlying FRBR model.

The prototype is called RIMMF (RDA in Many Metadata Formats) [15]. From the beginning, one of the stated goals for it was to explore how much repetitive work could be shifted from the cataloguer to the application via the developer. A simple set of basic input forms for each of the RDA entities was quickly developed to support RDA training, with each form based on the elements assigned to the entity. Then context sensitive hyperlinks to appropriate instructions in the online RDA Toolkit were added, while the forms were connected to each other using navigation links based on the primary FRBR entity relationships. The selection and order of the elements in each form can be changed by the cataloguer without changing the semantics of the data that is generated. Each input form can be saved as a template that can guide the cataloguer through the process of identifying, describing and relating metadata for different types of resources or complex situations. An automatic process called "Build WEM" applies the appropriate RDA instructions to copy specific data provided for one entity and add that data to another entity form. For example, the value for Title proper of the manifestation is copied as the Preferred title for the work following RDA guidance in section 6.2.2. The interface thus integrates the RDA instructions, guidance, and examples with the application.

In 2014 TMQ was asked to join the RDA Development Team that develops the RDA Registry. This has allowed the Registry to be tested against a "pure" RDA encoding format, and RIMMF to be developed to use the Registry linked data representations of RDA elements directly. The data input and maintenance forms conceal the technical terminology of the Registry and use the language of RDA Toolkit for element and relationship designator labels. Only elements and designators associated with the entity in focus are displayed on the form, removing potential confusion between similar labels assigned to different entities. Controlled data content terms specified in the Toolkit as vocabulary encoding schemes and the refined designators are presented as drop-down pick lists, making it easier for cataloguers to select the correct value or designator. RIMMF is now in version 3, and is available for free download. The software remains a prototype and training tool, and is not intended to be an operational data maintenance package. Large-scale systems will require a shared store for linked data triples with facilities for Uniform Resource Identifier (URI) management to evolve authority control into entity control. This will allow RIMMF and other data editors to search and download existing data for further augmentation without generating new identifiers for existing entities. Accommodation of provenance data for quality assurance is also needed to make appropriate choices from multiple sources in related communities. These might include the social media of end-users as well as metadata maintained by publishing and cultural heritage organizations. Collaborative management and open access to data is necessary for linked open data, and data maintenance interfaces need to be embedded as web services with access to the shared store.

Old and new data formats

RIMMF can import metadata in MARC21 bibliographic and authority records. The process carries out automatic parsing of the MARC encoding to separate out data belonging to each of the RDA entities. The parser uses a map from general and specific combinations of MARC tags and subfields to the RDA entities and elements, using the content where appropriate. The

map is effective, and the resulting RDA data usually requires little or no manual editing. Feedback from the import of rarer types of record is used to improve the map. The process also creates appropriate relationship links between the resulting RDA entities so that, for example, an imported person is linked as a contributor to an imported expression or an imported expression is linked to its existing work with a primary "work expressed" relationship. This makes it easier for a cataloguer to compare the legacy data with its representation in RDA and FRBR, note differences and similarities and improve their understanding of RDA. RIMMF can also export RDA data in MARC21 format, so that the effects of "FRBRization" from ISBD and AACR to FRBR and RDA can be demonstrated, as shown in Figure 1.

Figure 1: RIMMF display of entity relationships in a German translation of a Swedish text.

An additional aid to the cataloguer is the relationship tree, or "R-Tree", that presents the individual entities resulting from the FRBRization of a single MARC bibliographic record in a tree list display. The R-Tree may also be used to visualize relationships in the database as a whole, after the cataloguer has created their own local set of entity records. Figure 1 also shows the correct restriction of relationships to entities: the "Translator" relationship is intended only for an expression and the "Author" relationship for a work. The automatic generation of primary links between entities in both directions is demonstrated with the "Work expressed" and "Expression of work" relationships. The individual entity labels are access points automatically generated from RDA elements of finer granularity, selected and sequenced by the cataloguer.

RIMMF data can also be exported for linked data applications in RDF format, specifically an N-triples serialization. The facility is based on FRBR and RDA terminology, not RDF. The cataloguer does not require any technical skills to produce well-formed linked data, but can use the RDF export to improve their understanding of the complexities involved.

Multilingual cataloguing

The RDA Registry contains translations of the RDA elements and controlled vocabularies. Labels, definitions, and scope notes in multiple languages are linked to the same URI in RDF, allowing RIMMF to switch the language of its element labels and values for a non-English-speaking cataloguer with ease. A RIMMF user can set their current profile to display English, French, and Spanish. Figure 2 shows English labels, while Figure 3 shows the same data

displayed with French labels. The profile language is also applied to RDA vocabulary encoding schemes such as the Content type terms presented as a drop-down pick list in Figures 2 and 3. Other languages are under development according to the RDA Translations Policy, and include German, Finnish, and Italian [16].

Element Label	Text	RDA Rule	AAP
⊟ Expression			
RIMMF identifier °	viklun00000055	no rule	
Authorized access point °	**Lagerlöf, Selma, 1858-1940. Anna Svärd. Text. German**	6.27.3	
Language of expression	German	6.11	☑
Content type	Text ▾	6.9	☑
Variant access point °	Notated music / Other	7.4.5	
Variant access point °	Performed music / Sounds	7.4.5	
Source consulted	Spoken word / Still image		
Date of expression	Tactile image	)+	☐
Other distinguishing characteristic of the expression	Tactile notated movement / Tactile notated music / Tactile text	?	☐
Work expressed	Tactile three-dimensional form / Text	5+	☑
Manifestation of expression	Three-dimensional form / Three-dimensional moving image	)+	

Figure 2: An English view of a German expression by a Swedish author.

Element Label	Text	RDA Rule	AAP
⊟ Expression			
Identifiant de RIMMF	viklun00000055	no rule	
Point d'accès autorisé	**Lagerlöf, Selma, 1858-1940. Anna Svärd. Texte. Allemand**	6.27.3	
Langue de l'expression	allemand	6.11	☑
Type de contenu	texte ▾	6.9	☑
Variante de point d'accès	musique notée / Other	5	
Variante de point d'accès	musique exécutée / sons	5	
Source consultée	parole énoncée / image fixe		
Date de l'expression	image tactile		☐
Autre caractéristique distinctive de l'expression	mouvement noté tactile / musique notée tactile / texte tactile		☐
Œuvre exprimée	forme tridimensionnelle tactile / texte		☑
Manifestation d'une expression	forme tridimensionnelle / image animée tridimensionnelle		

Figure 3: A French view of the same data.

Figures 2 and 3 also show the links to context-sensitive RDA guidance and instructions in the "RDA Rule" column. Selection boxes for constructing authorized access points are displayed

in the "AAP" column. Controlled terms used in an access point are automatically translated, such as the Language of expression and Content type.

Hackathons

Starting in January 2015, RIMMF has been extensively tested "in the field" in a series of hackathons for RDA data, known as "jane-athons" from an initial focus on the works of Jane Austen. These events have been held in several countries on a range of topics with participants at various stages of RDA training, and have proved very successful. The resulting RDA data, in the form of linked entity clusters or "r-balls", along with event descriptions and feedback from cataloguers are freely available [17]. The data can be downloaded into RIMMF and used for further training and RDA skills development. It can also be downloaded into linked data applications and used to familiarize developers with RDA concepts and values.

Conclusion

RIMMF is an effective bridge between professional cataloguers and the technologies underpinning legacy and modern metadata. The traditional concepts are clearly separated from the radically different approaches supported by modern conceptual models, and much of the technical jargon is concealed. Primary links between the new entities are created automatically, and coherent semantic relationships are ensured by filtering out elements and designators that are not relevant to the entity being described. RIMMF also acts as a bridge between linked data application and system developers using the RDA Registry and the cataloguers who will use the resulting operational services.

References

[1] Gordon Dunsire. RDA and the Semantic Web. Lectio Magistralis in Biblioteconomia. Firenze : Casalini Libri, 2014. Available at: http://digital.casalini.it/9788876560132

[2] Gordon Dunsire, Diane Hillmann, and Jon Phipps. Reconsidering Universal Bibliographic Control in Light of the Semantic Web. In: Journal of Library Metadata, v.12, issue 2-3 (2012), pp. 164-176. DOI: 10.1080/19386389.2012.699831

[3] IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). Functional requirements for authority data: a conceptual model: final report. As amended and corrected through July 2013. Available at:
http://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf

[4] Pat Riva, Patrick Le Bœuf, and Maja Žumer. FRBR-Library reference model: Draft for World-Wide Review. 2016-02-21. Available at:
http://www.ifla.org/files/assets/cataloguing/frbr-lrm/frbr-lrm_20160225.pdf

[5] Committee of Principals Affirms Commitment to the Internationalisation of RDA. 5 June 2015. Available at: http://www.rda-rsc.org/node/235

[6] RDA Toolkit. Available at: http://www.rdatoolkit.org/

[7] RDA Registry. Available at: http://www.rdaregistry.info/

[8] Tom Delsey, Beth Dulabahn, Michael Heaney, Jean Hirons. The Logical Structure of the Anglo-American Cataloguing Rules. Part I. August 1998. Available at: http://www.rda-jsc.org/docs/aacr.pdf

[9] Gordon Dunsire, "Resource and Work, Expression, Manifestation, Item," amended October 6, 2013/ Available at:
http://www.ifla.org/files/assets/cataloguing/isbd/OtherDocumentation/resource-wemi.pdf

[10] W3C. RDF 1.1 Concepts and Abstract Syntax: W3C Recommendation 25 February 2014. Available at: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[11] RDA elements. April 2016. Available at: http://www.rda-rsc.org/sites/all/files/RSC-RDA-element-analysis-table.pdf

[12] Joint Steering Committee for Development of RDA. RDA Database Implementation Scenarios. 1 July 2009. Available at: http://www.rda-jsc.org/archivedsite/docs/5editor2rev.pdf

[13] JSC Technical Working Group. RDA accommodation of relationship data. 3 August 2015. Available at: http://www.rda-jsc.org/sites/all/files/6JSC-TechnicalWG-6.pdf

[14] The MARC of Quality. Available at: http://www.marcofquality.com/

[15] RIMMF3 home. Available at: http://www.marcofquality.com/wiki/rimmf/doku.php

[16] Joint Steering Committee for Development of RDA. Translation Policy for RDA and RDA Toolkit. 22 September 2015. Available at: http://www.rda-jsc.org/sites/all/files/6JSC-Policy-6.pdf

[17] R-Balls. Available at: http://rballs.info/