Date submitted: 03/08/2010

Initiatives to make standard library metadata models and structures available to the Semantic Web

Gordon Dunsire
University of Strathclyde
Glasgow, Scotland, United Kingdom

and

Mirna Willer
University of Zadar
Zadar, Croatia

Meeting: **149. Information Technology, Cataloguing, Classification and Indexing with Knowledge Management**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

This paper describes recent initiatives to make standard library metadata models and structures available to the Semantic Web, including IFLA standards such as Functional Requirements for Bibliographic Records (FRBR), Functional Requirements for Authority Data (FRAD), and International Standard Bibliographic Description (ISBD) along with the infrastructure that supports them.

The FRBR Review Group is currently developing representations of FRAD and the entity-relationship model of FRBR in resource description framework (RDF) applications, using a combination of RDF, RDF Schema (RDFS), Simple Knowledge Organisation System (SKOS) and Web Ontology Language (OWL), cross-relating both models where appropriate. The ISBD/XML Task Group is investigating the representation of ISBD in RDF. The IFLA Namespaces project is developing an administrative and technical infrastructure to support such initiatives and encourage uptake of standards by other agencies.

The paper describes similar initiatives with related external standards such as RDA – resource description and access, REICAT (the new Italian cataloguing rules) and CIDOC Conceptual Reference Model (CRM). The DCMI RDA Task Group is working with the Joint Steering Committee for RDA to develop Semantic Web representations of RDA structural elements, which are aligned with FRBR and FRAD, and controlled metadata content vocabularies. REICAT is also based on FRBR, and an object-oriented version of FRBR has been *integrated with CRM, which itself has an RDF representation. CRM was initially based on the metadata needs of the museum community, and is now seeking extension to the archives community with the eventual aim of developing a model common to the main cultural information domains of archives, libraries and museums. The Vocabulary Mapping Framework (VMF) project has developed a Semantic Web tool to automatically generate mappings between metadata models from the information communities, including publishers. The tool is based on several standards, including CRM, FRAD, FRBR, MARC21 and RDA.*

The paper discusses the importance of these initiatives in releasing as linked data the very large quantities of rich, professionally-generated metadata stored in formats based on these standards, such as UNIMARC and MARC21, addressing such issues as critical mass for semantic and statistical inferencing, integration with user- and machine-generated metadata, and authenticity, veracity and trust. The paper also discusses related initiatives to release controlled vocabularies, including the Dewey Decimal Classification (DDC), ISBD, Library of Congress Name Authority File (LCNAF), Library of Congress Subject Headings (LCSH), Rameau (French subject headings), Universal Decimal Classification (UDC), and the Virtual International Authority File (VIAF) as linked data. Finally, the paper discusses the potential collective impact of these initiatives on metadata workflows and management systems.

The authors

Gordon Dunsire is Head of the Centre for Digital Library Research at the University of Strathclyde, Glasgow, Scotland. He is chair of the IFLA Namespaces Task Group, a member of the FRBR Review Group, a member of the ISBD/XML Task Group, and a member of the IFLA Classification and Indexing Section. He is a co-chair of the DCMI RDA Task Group, a member of the CILIP-BL Committee on RDA, and a member of the RDA Outreach Group, and has made numerous presentations on RDA. He is also a member of the VMF project, chair of the European DDC Users Group Technical Issues Group, and a member of the CILIP Committee on DDC.

Mirna Willer has been Associate Professor at the University of Zadar, Zadar, Croatia since 2007. She worked from 1980 to 2007 as systems librarian, standards officer and senior researcher at the National and University Library in Zagreb, Croatia responsible for implementing the UNIMARC bibliographic and authority formats on Library's library management software, and for incorporating national cataloguing rules into the formats. Among other international body memberships, she was a standing member of the IFLA Permanent UNIMARC Committee from its establishment in 1991 until 2005 (chair of Committee from 1997 to 2005), since when she has been its consultant and honorary member. She was also a member of the IFLA Working Group on FRANAR (Functional Requirements and Numbering of Authority Records), the Working Group responsible for the conceptual model FRAD, as well as the ISBD Review Group, and ISBD Future Directions Working Group. Currently, she is a chair of the ISBD/XML Task Group.

Background and introduction

There have been a number of recent initiatives to make standard library metadata models, structures, and vocabularies developed by IFLA available to the Semantic Web. These have been accompanied by similar projects involving related standards maintained by other organisations.

Indeed, the impetus for IFLA's work in this area was initially stimulated by just such an external project, which evolved from a meeting between representatives of RDA: resource description and access[1], the Dublin Core Metadata Initiative (DCMI)[2], IEEE Learning Object Metadata (IEEE LOM), and Simple Knowledge Organization System (SKOS)[3]. The Data Model Meeting[4] held at the British Library in London in 2007 recommended several activities that would provide benefits including the library community getting "a metadata standard that is compatible with the Web Architecture and that is fully interoperable with other Semantic Web initiatives". These activities include developing RDA bibliographic entities, relationships, and controlled content terminologies to make them more compatible with Dublin Core and the Semantic Web. The meeting resulted in the creation of the DCMI RDA Task Group[5] to oversee and carry out the work, much of which involves the formulation of classes (equivalent to entities) and properties (equivalent to relationships and attributes of the entities) in Resource Description Framework[6] (RDF), the data model of the Semantic Web.

RDA is based on the IFLA metadata models Functional Requirements for Bibliographic Records[7] (FRBR) and Functional Requirements for Authority Data[8] (FRAD), and makes frequent reference to the bibliographic entities, relationships and attributes which they describe. RDF formulations of RDA elements therefore must include terms and definitions from FRBR and FRAD which are controlled by the FRBR Review Group rather than the Joint Steering Committee for Development of RDA (JSC). The FRBR Review Group considered the implications of the Data Model Meeting at its meeting[9] during the 2007 World Library and Information Congress in Durban, South Africa, and decided to initiate a project "To define appropriate namespaces for FRBR (entity-relationship) in RDF and other appropriate syntaxes." A namespace is a method of publishing RDF formulations and making them available to other Semantic Web applications, so JSC would then have the option of using elements from the FRBR namespace instead of publishing and maintaining their own versions.

The IFLA Cataloguing Section’s ISBD Review Group has recently taken action on the recommendations from its Material Designation Study Group to develop an XML Schema for the ISBD. The ISBD Review Group considered it important to start researching into the possibilities of reviewing ISBD concepts and the standard itself by the application of web technologies, and eventually of evolving the standard into a tool open to the Semantic Web technologies and services. The ISBD/XML Study Group was formed during the 2008 World Library and Information Congress in Quebec. The project proposed by the Study Group was accepted in 2008 by the Professional Board, and is now in its second year. The primary objective of the ISBD/XML Study Group is to position the ISBD as a relevant factor in assessing structured bibliographic information in the global information environment. In order to do that the Study Group had been charged with defining XML Schema for ISBD; however, after considering the work on RDA and FRBR related to RDF, the Study Group decided to bypass a general XML mark-up, and go directly to the RDF/XML environment. The result of the project would be ISBD RDF/XML. At the Study Group’s meeting in Milan, during the 2009 World Library and Information Congress the following actions were identified and agreed upon: support the motion to form a task force/alliance working group across the Sections in Division III and beyond to position IFLA standards and models in the Semantic Web environment as authoritative documents for Semantic Web services and tools (see below for further information); approve the Study Group’s involvement in the Vocabulary Mapping Framework (VMF) project (see below); analyse technical and modelling issues of ISBD in the RDF/XML environment; define uses and functions of ISBD in the RDF/XML syntax; develop draft ISBD RDF/XML schema; analyse and define the functionalities of ISBD elements in relation to FRBR, (UNI)MARC, and DC/XML schemas, new cataloguing rules such as RDA, REICAT and the Finnish cataloguing rules; and analyse and support the concept of linked data, and promote its relevance to vendors in support of the development of new generation library information systems.[10]

The FRBR Namespace Project stimulated discussion within and between other IFLA bibliographic standards groups during the 2008 World Library and Information Congress in Québec City, Canada, and the 2009 World Library and Information Congress in Milan, Italy. As a result, the IFLA Classification and Indexing Section proposed a task group to prepare a requirements and options paper on the topic of IFLA support for the representation of IFLA standards in formats suitable for use in the Semantic Web. The proposal was supported by the IFLA Bibliography, Cataloguing, and Information Technology Sections, and the IFLA Namespaces Task Group was created in late 2009. The Task Group delivered its paper to IFLA in March 2010.

IFLA standards

Many of IFLA's standard bibliographic models and applications are inter-related on a formal or informal basis. The majority of formal relationships consist of direct references from one standard to another, or of mappings between elements from different standards which have been approved by the IFLA groups which maintain the standards. Most informal relations consist of mappings developed by non-IFLA groups and individuals. A significant issue is that of time: it can take years for an international and largely voluntary group, a typical situation for IFLA, to develop a standard, and more years to review and amend it and establish relationships with related standards which may be at a different stage of their development. An obvious example is what has become known as the "Functional Requirements" family or "FRBR family of models". FRBR was published in 1998, and describes itself as an "initial attempt"; it goes on to point out that the "model could be extended to cover the additional data that are normally recorded in authority records. In particular, further analysis is needed of the entities that are the centre of focus for subject authorities, thesauri, and classification schemes, and of the relationships between those entities." FRAD, which addresses the first of these suggestions concerning authority records, was published 11 years later in 2009. Functional Requirements for Subject Authority Data (FRSAD), which addresses the second suggestion about subject authorities, is expected to be published in 2010.

FRBR family

Preliminary work by the FRBR Namespace Project used the sandbox or testing area of the National Science Digital Library Metadata Registry (NSDL)[11] to become familiar with RDF concepts and assess how the FRBR elements could be formulated. A report[12] on this work, including a recommendation that the FRBR Review Group should develop its own namespaces with some form of branding or indication of ownership, was considered by the Group at its meeting at the 2008 World Library and Information Congress. The IFLA Web Master agreed at the 2009 Congress to obtain a suitable Web domain to act as the basis of all IFLA namespaces, and "iflastandards.info" was subsequently registered for this purpose. At the same time a methodology for allocating the domain to specific IFLA standards was proposed and later used for creating the RDF formulations of FRBR in the production area of the NSDL Registry. The FRBRer model element set[13] is now mainly complete, although a few issues remain to be resolved by the Review Group before it is approved. For example, several of the attributes in the model are assigned to sub-types of the FRBR Group 1 elements, such as "reduction ratio" which only applies to manifestations which are microforms. Representing this in RDF involves creating the property "has-reduction-ratio" and the class "Microform-Manifestation" which is a sub-class of "Manifestation", but FRBR does not actually offer a definition of the entity "microform manifestation". Another issue is the variation in the labels used in FRBR. In RDF, these labels form the basis of the preferred label attached to the URI of the class or property, which is intended for human rather than machine consumption. Although the labels do not affect the computer-processing of the RDF representations, variations may reduce confidence in the authority of the model, and may cause problems for translations of English labels into other languages. These and other issues mostly arise when translating a human-readable document, with necessary variations in presentation to improve its readability, into a machine-readable "document" or set of RDF formulations which require complete and consistent definitions and labels. The namespace uses "http://iflastandards.info/ns/fr/frbr/frbrer/" as the basis of the uniform resource identifiers (URIs) of each RDF class and property in the model. The namespace follows the wording of FRBR as closely as possible, and relies solely on the FRBR final report as a source of information to ensure that the RDF formulation is not distorted by the benefit of hindsight or adaptation by other communities. The RDF file output by the Registry[14] does not represent the complete FRBR model, even after all elements are approved by the Review Group, because the Registry does not support the Web Ontology Language (OWL)[15] properties needed to declare complex relationships between the RDF classes and properties derived from FRBR's entities, relationships and attributes. These include constraints on classes, such as ensuring that something declared as a FRBR Work should not also be declared or inferred to be a FRBR Expression, Manifestation, or Item, and relationships between properties, such as which properties form reciprocal pairs so that one triple can be inferred from another. An additional set of formulations created outside of the Registry will be used to represent these more complex aspects of the FRBR entity-relationship model. The Review Group intends to combine these with the file from the Registry and publish the complete RDF version of the FRBR model on the IFLA website.

FRBRoo[16], the object-oriented version of FRBR, was developed between 2003 and 2009 to be compatible with and extend the CIDOC Conceptual Reference Model (CRM)[17] created for metadata in the museums community. One of the goals of this initiative was to enable FRBR to be used with RDF applications and other Semantic Web technologies. Although a partial representation of the CRM in RDFs (RDF Schema) was published in 2009[18], it does not include the FRBRoo extension. The mapping between elements of FRBRoo and FRBRer, published with FRBRoo, will be used in due course to compare the FRBRer and FRBRoo elements within the RDF environment and improve links between the FRBRer, FRBRoo, and CRM models.

The RDF representation of FRAD has been started, and is taking a similar approach to FRBR. A separate namespace based on "http://iflastandards.info/ns/fr/frad/" is being used to avoid confusion with the FRBRer model. Entities, relationships and attributes described in FRAD as FRBR elements are not being represented in RDF again; instead, the FRBR Review Group will consider re-use of classes and properties from the FRBRer namespace as appropriate. However, some FRBR elements are extended in definition or scope in FRAD, and where this results in a semantically significant difference, the FRBR element is being represented with a new class or property in the FRAD namespace. Areas of derivation, overlap and conjunction between the FRBRer and FRAD models is being noted to inform planned work by the FRBR Review Group to consolidate the FRBR family of models.

The last model in the family is FRSAD. FRSAD is in the final stages of its development, and when published it is likely to use the same approach as FRBR and FRAD to develop an RDF representation. Again, it is expected that this will eventually inform future consolidation work.

ISBD

A draft version of the ISBD consolidated edition was launched for worldwide review on 10 May 2010,[19] thus providing the ISBD/XML Study Group with a considerably stable text for the beginning of testing of RDF representation in the NSDL Registry's sandbox. During this process, it will be necessary to identify and investigate issues arising from the application of RDF, and in that process to draw on the RDA work where appropriate, while at the same time it will be of primary importance to position the ISBD in the environment of FRBR and (UNI)MARC and DC/XML schemas.

The positioning of ISBD in the environment of IFLA documents is a strategic (and political) question: what draws from what? On the application level, FRBR is a conceptual model built on the Entity-Relationship methodology which is intrinsically applicable to representation in RDF, while ISBD is a data standard. Consequently, the FRBR relationship to ISBD, or, more accurately, ISBD's relationship to FRBR, caused considerable discussions within ISBD, and also the wider community, which resulted in decisions relevant to this relationship as well as the eventual declaration of ISBD vocabulary in the RDF.

> "The ISBD Review Group considered that it was essential for IFLA to clarify the relationship between the ISBDs and the FRBR model. In trying to adapt ISBD terminology to the FRBR terms "work", "expression", "manifestation" and "item" and to replace terms such as "publication", the group encountered difficulties, owing in large part to the fact that the terms used in FRBR were defined in the context of an entity-relationship model conceived at a higher level of abstraction than the specifications for the ISBDs. As a report from the Frankfurt IFLA Meeting of Experts on a International Cataloguing Code (IME-ICC) had cautioned, "FRBR terminology should *not* be merely incorporated such as it stands into the ISBDs and cataloguing rules, but these should keep their own specific terminology, and provide accurate definitions showing how each term in this specific terminology is conceptually related to the FRBR terminology". The review group agreed with the advice from the IME-ICC and decided, in 2003, to avoid using FRBR terminology in the ISBD."[20]

Another, strategic issue is the mapping of the linear structure of the ISBD into RDF, specifically the characteristic composite or aggregate metadata statements used in the nine areas (Area 0 to Area 8) of ISBD. Thus the issues of considerable concern and relevance to the design of the RDF representation of ISBD are:

1. the treatment of aggregated statements in a defined number of elements within the areas to determine what is an RDF class and what a property;
2. the treatment of mandatory and optional elements and areas;
3. the order of areas and elements within an area;
4. the repeatability of areas and elements to some of which apply fixed pre-defined ISBD rules, while they can appear voluntarily (depending on the resource information);
5. the treatment of punctuation and its double function. The first one is the identification of elements within an area; however, as areas are not identified uniquely by specific punctuation, the "meaning" of the areas and subsequently elements cannot be inferred from the punctuation. The second function is the display of elements – bibliographic data; due to the voluntary appearance of data on the resource described, the display is tightly linked to the ISBD mechanism of defining mandatory and repeatable elements, as well as to the display order which itself controls the recording of data.

It was already recognized that some of these issues, specifically rulings of mandatory elements, punctuation and display definitions, need to be resolved outside the RDF declaration process by using XML transformations and application profiles to be applied during the process of creating or assembling an ISBD record.[21]

UNIMARC

UNIMARC is a carrier format for the exchange of bibliographic metadata between systems used by national libraries and other agencies. It does not specify any metadata structure or content to be used in specific systems, but it is closely aligned with ISBD. This relationship leads to secondary alignments and mappings with several of the other standards and models discussed in this paper, including FRBR and RDA via their relationships with ISBD, and in turn with FRAD and CIDOC CRM[22]. Many national library catalogues and bibliographies, constituting the "official" and authoritative (and therefore trusted) version of national bibliographic metadata, are based on UNIMARC and it is often used as a local record structure within their systems. As such, there is considerable value in developing an RDF representation of UNIMARC as a metadata structure schema for the purpose of extracting the content of records as linked data for the Semantic Web, as described below. An RDF representation would also help to update the alignments between UNIMARC and other standards which have been amended significantly in recent years. These issues have been brought to the attention of the Permanent UNIMARC Committee.

Related standards

The DCMI RDA Task Group has three goals: define RDA modelling entities as an RDF vocabulary of properties and classes; identify in-line value vocabularies as candidates for publication in RDFS or SKOS; and develop a Dublin Core Application Profile for RDA based on FRBR and FRAD. The Task group is using the NSDL Metadata Registry to develop RDF representations of the RDA vocabularies[23]. Work on the second of these is nearly complete and all of the controlled vocabularies for metadata content, from applied materials to video formats, have been represented in SKOS. The first goal has proved less straightforward because it needed a methodology to deal with composite or aggregated metadata statements which are a feature of RDA as well as ISBD. An example of an aggregated statement is "publication statement" which is composed of more granular attributes such as "place of publication" and "publisher". The Task Group developed a methodology which was published in a paper reviewing decisions and outcomes of the work of the Task Group up to the end of 2009[24]. This approach and other issues encountered during the representation of RDA entities as RDF classes and properties will be considered by the ISBD/XML Study Group in due course. The Task Group has not yet made significant progress with the third goal of developing an Application Profile. The initial idea was to produce something that might augment or replace the Dublin Core Library Application Profile (DC-Lib)[25], but it may be more useful to develop profiles for specific types of bibliographic resource; this is still under discussion.

Another issue encountered by the DCMI RDA Task Group is the relationship between RDA and FRBR and FRAD in RDF, and specifically whether the RDA namespace should contain its own representations of FRBR and FRAD classes and properties used by RDA or just make reference to the representations in the FRBR and FRAD namespaces. It was necessary to include FRBR and FRAD elements in the RDA namespace at the outset, simply because the FRBR namespace took longer to develop than anticipated, and the FRAD namespace has not yet reached a stable state. However, delays in the publication of RDA have allowed FRBR and FRAD to achieve a significant catch-up, but the Task Group is still proposing to include separate representations of relevant FRBR and FRAD elements within the RDA namespace to retain flexibility and reduce reliance on the IFLA initiatives. In due course, the RDA and IFLA representations can be either made equivalent using OWL properties, or the RDA versions deprecated in favour of the IFLA ones.

The National Library of Sweden has developed a methodology for representing MARC21 records in RDF and implemented it for LIBRIS, the Swedish Union Catalogue[26]. Similar proposals for making MARC21 metadata available to the Semantic Web have been made[27]. MARC21 is in widespread use, particularly in national and other libraries in Anglophone countries. There are large numbers of records encoded in MARC21; OCLC's WorldCat alone has several hundred million.

The Italian cataloguing rules REICAT[28] are the first cataloguing rules to be published that apply FRBR, FRAD and the Statement of international cataloguing principles[29]. The rules are based on the concept of work and uniform title, and the reference to FRBR concepts work and expression is being made however they do not follow the FRBR definitions and terminology closely [30] [31]. There is no information as to the Commissione RICA's intentions to declare REICAT in RDF yet.

The Vocabulary Mapping Framework (VMF) project[32] was funded by UK's Joint Information Systems Committee (JISC) to develop a major expansion of the RDA/ONIX framework for resource categorization[33] to create a tool to support the automated mapping of vocabularies from metadata standards of use to the JISC community, which includes research, teaching, and learning environments. The original RDA/ONIX framework was developed by representatives of RDA and the publishing community, and consists of an ontology of attributes associated with bibliographic resource content and carriers and a methodology for using the ontology to create high-level human-readable labels for content and carrier types[34]. The RDA carrier type, content type, and media type vocabularies are based on the framework.

The VMF project focussed on bibliographic resource and bibliographic agent (party) categories and the relationships between them, recognising that many apparent bibliographic agent roles, such as editor or manufacturer, are better treated as relationships, such as is-edited-by or is-manufacturer-of. The project analysed relevant vocabularies from a variety of communities and produced a tool, the VMF matrix, which can be used to automatically compute best-fit mappings between terms from different metadata scheme vocabularies. The matrix itself is a set of RDF representations of activities and roles which include all possible resource and agent relationships. Terms from external vocabularies can be plugged-in to the matrix using OWL properties, and because all possible pathways between any two terms are already represented in the matrix, it is possible to programme a method for computing the shortest "distance", and therefore best-fit, between those terms.

CIDOC CRM, FRAD, FRBR, MARC21 and RDA vocabularies were included in the core set used to create the matrix, while ISBD and UNIMARC interests were represented on the project. The matrix is available for download from the project website. While it is likely that IFLA vocabularies will be directly and explicitly linked by their maintenance groups during the process of RDF representation, consolidation, and re-alignment, the VMF matrix could be of significant benefit in aligning and relating IFLA vocabularies with those of other communities.

Linked data and the Semantic Web

Linked data are instance triples assigning values to specific properties of specific instances of entities. Triples come from multiple sources, and can be used to assemble aggregations of data about the same entity (that is, metadata) by using the entity's identifiers or URIs. Furthermore, RDF is designed to allow logical inferences to be made from sets of triples, which in turn generate new triples.

A traditional catalogue record is composed of the values of multiple properties associated with a bibliographic entity, for example its title and physical description. It is possible to decompose such a record into a set of instance triples, all using the same URI for the subject. The predicate part of the subject-predicate-object expression of each triple must be a URI identifying the property, as required by RDF. The object part contains the value of the property. This can be a literal value such as a character string or number, and will accommodate most instances found in catalogue records. The object value can also be another URI, such as for a term from a controlled vocabulary represented in SKOS, so authority-controlled headings can also be accommodated if the authority file is expressed in RDF.

A specific record might yield the triples:

- ?:thisResource ?:has-title "UNIMARC format for authority records".
- ?:thisResource ?:has-author <viaf:29776655>.
- ?:thisResource ?:has-publication-date "2004".

Note that these example triples are presented in a standard format where a URI is split into a generic namespace part followed by a specific identifier part, separated by a colon. A question-mark (?) is used to indicate an unspecified generic namespace. A URI with a pure numeric specific identifier is given in angle brackets. Literal values are given in quotation marks.

Another record for the same resource in a different catalogue might yield the triples:

- ?:thatResource ?:has-title "Unimarc format for authority records".
- ?:thatResource ?:has-publisher "Howarth Press".

But RDF allows the fact that both these resources are the same to be expressed also as a triple:

- ?:thisResource owl:sameAs ?:thatResource.

Taking all these triples together, the following triple can be inferred:

- ?:thisResource ?:has-publisher "Howarth Press".

If the first catalogue record is reassembled by aggregating all the triples with :thisResource as a subject, it will contain a value for the publisher which was not part of the original record.

So disaggregating catalogue records into RDF triples has clear benefit to the completeness of bibliographic metadata. Adding these triples to the "soup" of the Semantic Web and establishing equivalences between different URIs for the same individual bibliographic entities will increase this benefit further by including triples from other sources, such as publishers, booksellers, online encyclopaedias, and social networking sites in any re-aggregation of the metadata. The process of re-aggregation can be programmed to exclude unwanted properties and their values, as well as triples from un-trusted or irrelevant sources. Some aspects of re-aggregation might also be controlled by the end-user, such as which properties to display or a choice of standard record formats.

A pre-requisite for converting records to triples is a URI for each property to be extracted. RDF representations of metadata models and formats provide URIs for the properties used to define attributes and relationships of entities within the format. In order to process records based on ISBD, for example, each attribute defined by ISBD needs to be represented as an RDF property with its own URI; this is intended to be an outcome of the work of the ISBD/XML Study Group, with URIs taken from a namespace specific to ISBD.

RDF representations of formats have other utilities. They provide a structural map which can be used to disaggregate existing records in the format and re-aggregate instance triples to create records in the format. They can also be used in error-checking for missing or wrongly encoded fields

It is important that each format is represented in its own namespace, to avoid confusion between properties and classes representing similar attributes, relationships, and entities. For example, a triple using the FRBR attribute"date of publication or distribution" might be :

- ?:thisResource frbrer?:has-date-of-publication-or-distribution "1973".

The constraints on this property declared in RDF allow the computer to infer that:

- ?:thisResource rdf:type frbrer:Manifestation

This triple states that the resource must be a FRBR manifestation (and not a work, expression or item). If a similar property from a different namespace was used, the inference would be different or non-existent.

Another form of inferencing can be applied to large numbers of triples with the same, or matched, subject URIs. Statistical methods can be used to determine a weight or trend of opinion about a value for a specific property of a specific bibliographic entity, where there are variations in the value coming from different catalogue records for that entity. For example, if nine out of ten triples from different sources say the value is "X" while one gives the value as "Y", it can be inferred that the true value is "X", and that "Y" is the result of misinterpretation, an entry error, or legacy from the application of different cataloguing rules. This technique is the basis of the experimental Classify[35] service offered by OCLC to suggest Dewey Decimal Classification and Library of Congress Classification notations for a specified work; the statistical analysis uses FRBR to group records for the same work and takes the values of the notations from the whole records rather than disaggregated triples, but the method is similar. The many millions of existing catalogue records can potentially generate billions of triples, so it seems likely that there will be many instances where statistical inferencing can be usefully applied.

Controlled vocabularies

Instance values from terminologies such as subject headings, classification captions and indexes, and thesauri can be represented in RDF using SKOS. SKOS caters for simple relationships between terms, for example broader, narrower, and exact match; these are usually sufficient to represent the internal structure of such terminologies. Representation in SKOS also results in the assignment of a URI to each concept represented by a term. SKOS allows language translations of a term to use the same URI, making the original term and its translations automatically interoperable in multilingual environments. Alternatively, the URIs for the same concept represented in different namespaces can be related in RDF to provide mappings from one terminology to another. Terms represented in SKOS can also be released as linked data, thus improving the potential for inferred triples.

Exposure of controlled vocabularies as linked data also encourages uptake by other communities. Infrastructure such as interfaces for maintenance or end-user services developed for one vocabulary can be used with any other.

As described above, the prescribed content vocabularies of RDA have been represented in SKOS and are likely to be published in 2010. These vocabularies are flat; there is no hierarchy of terms. The terms are in English, but some have been translated into German, keeping the same URI, by the Deutsche NationalBibliothek. The terms will be freely available for use by anyone. The base of each URI will show that the term is from RDA and not some other vocabulary. The common URI for each term and its German translation will improve federated searching in English-German bi-lingual environments.

Several other bibliographic vocabularies have been published in RDF and as linked data in the past year or so. Most are taken from subject heading and classification schemes.

Library of Congress Subject Headings (LCSH) as linked data can be downloaded from the Library of Congress authorities and vocabularies service[36] They can also be used interactively in OCLC's pilot terminology service[37]. The pilot also includes Faceted Application of Subject Terminology (FAST), Medical Subject Headings (MESH), Form and genre headings for fiction and drama, and Thesaurus for Graphic Materials (TGM).

The French RAMEAU subject headings have also been published as linked data[38]. Mappings between RAMEAU and LCSH headings, created by the MACS project[39], are also freely available as linked data.

OCLC's experimental service for the Dewey Decimal Classification (DDC), Dewey.info[40], returns linked data for the caption associated with a submitted DDC notation. The URI of each topic embeds the notation as part of its unique component, and the URIs follow a pattern that can be used to infer the URI of the topic if the DDC notation is known. The scope of the service is currently confined to the DDC summaries, which are the top thousand notations (that is, 3-digit notations with no decimal point). The service enables user-friendly captions to be displayed in place of notations which are so opaque to end-users that they are usually not exposed for subject retrieval. Linked data has the potential to obtain the value of rich subject metadata from the historical investment in processes to order physical library materials on the shelf.

The UDC Consortium has published a selection of around 2,000 UDC classes in 16 languages online as the UDC summary[41]. The Consortium intends to make the UDC Summary available in RDF as well as other formats, along with mappings to other knowledge organisation systems[42]. The content is released under a Creative Commons license.

The Virtual International Authority File (VIAF) is a set of controlled vocabularies (authority records) of personal and corporate names maintained by national bibliographic agencies. Names in one vocabulary are related to corresponding entries in the other vocabularies, and the whole set of names and relationships is available as linked data[43].

ISBD prescribes vocabulary control for the data in the first area 0 for content form and media type. Terms for the three elements, content form, content qualification and media type, are taken from closed lists.[44]

Metadata management

The availability of bibliographic metadata conforming to professional standards in the environment of the Semantic Web is likely to have a significant impact on cataloguing workflows and metadata management.

There will be shift of focus from the whole bibliographic record, the set of metadata for a resource specified by a particular format and cataloguing rules, to the single metadata statement. Current practice is to carry out cataloguing activity at the level of the bibliographic record. The whole record is manipulated if a single amendment is required, for example to add metadata mandated by cataloguing rules, add local metadata such as a subject classification, or correct a mistake. The record is typically loaded into an editing interface, the amendment is carried out to just one part, and the changed record used to replace the original.

Shared cataloguing services and consortia are based on the aggregation and exchange of entire records rather than component parts. This can result in large quantities of duplicate metadata within the aggregation, as it is difficult to automate the identification of records for the same bibliographic resource with sufficient accuracy to meet the needs of participating organisations. As a result it is usually only those records containing a consortium-wide or global identifier, such as an International Standard Book Number, that are de-duplicated. There are also difficulties in synchronising amendments so that changes made to aggregated records are not transmitted to local copies. The local and consortium versions of the record become different, although they describe the same resource and nearly all of the metadata is duplicated. The focus on the whole record also forces the duplication of metadata for different resources sharing a common characteristic. A new edition of a book will usually have the same author, title, subject and publisher as the previous edition. A digitised image has the same content as the original, and so on. The duplication is so great that it is common practice to create metadata for the new resource by copying the record for the prior resource and amending it rather than create a record from scratch.

Duplicate data prevents efficiency in storage, access, dissemination and manipulation. It can also impede the effectiveness of services based on that data. FRBR's user-centred analysis shows that grouping metadata to correspond to the four bibliographic entities of Work, Expression, Manifestation and Item is more effective in meeting user needs. It allows the de-duplication of metadata displayed to the user, as demonstrated in the OCLC FictionFinder prototype[45]. There is therefore no need to maintain duplicate metadata in an information retrieval service built on the FRBR model, and the focus shifts from records describing the whole resource to records describing separate works, expressions, manifestations, and item. A new edition does not need a new work record, a digitised image does not need a new work or expression record, and so on. The level of granularity of the catalogue record increases from one to four in relation to the bibliographic record.

The Semantic Web increases the granularity of metadata to a much higher level. Every attribute and relationship described in a catalogue record yields a triple, so the number of triples constituting a bibliographic record will range from tens to thousands. Professional maintenance of bibliographic metadata would therefore expect to gain improved efficiency by treating the triple as the "catalogue record" and so reduce duplication to a logical minimum. Online catalogues in a linked data environment can assemble bibliographic descriptions from instance triples taken from all available sources, so there is no need for the library community to create its own triple if there is already one available from the publishing community. The archive, library, and museum communities can focus on maintaining metadata unique to the needs of their members, such as provenance, availability, suitability, and context. Metadata of more general interest, such as label, format, associated places and events, will be accessible as linked data from other communities.

Linked data comes from many sources and namespaces[46]. It is inevitable that triples with distinct URIs will be semantically identical, as a result of duplication between and within those sources. Identity management will become a key issue for presenting complete, consistent, and coherent bibliographic information to users. The provenance of linked data is also a key issue. There is nothing to indicate the accuracy of a triple, except another triple. Linked data may contain deliberate or accidental falsehoods, or result in contradictory inferences, not least because of inaccurate and incomplete legacy metadata from professional sources. It will be important to know the source of a triple and the date and context of its creation.

The new environment is evolving its own metadata management systems. Established vendor systems in the archive, library, and museum communities are monitoring developments, but have to overcome the inertia of their customer base which must balance cost against perceived benefit. There is increasing uptake of alternatives to the traditional cataloguing infrastructure in the form of open-source, social networking services for bibliographic information which encourage users to add and amend metadata, for example LibraryThing[47]. These systems are more likely to take early advantage of linked data and Semantic Web developments.

Conclusion

IFLA and its related communities will have an important, if not vital, role to play in the development of the Semantic Web through the promotion and application of standards and professional practice and the release of metadata to the linked data environment. IFLA and other communities will benefit from participating in the Semantic Web by improving services to users and increasing the efficiency and effectiveness of metadata.

References

All references checked 31 May 2010.

[1] Joint Steering Group for Development of RDA. RDA: resource description and access. Available at: http://www.rda-jsc.org/rda.html

[2] Dublin Core Metadata Initiative. Available at: http://dublincore.org/

[3] W3C. SKOS Simple knowledge organization system. Available at: http://www.w3.org/2004/02/skos/

[4] British Library. Bibliographic Standards. Data Model Meeting. Available at: http://www.bl.uk/bibliographic/meeting.html

[5] DCMI/RDA Task Group wiki. Available at: http://dublincore.org/dcmirdataskgroup/

[6] W3C. Resource Description Framework (RDF). Available at: http://www.w3.org/RDF/

[7] IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records. 1998, amended 2009. Available at: http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records

[8] IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). Functional requirements for authority data. 2009. Available at: http://www.ifla.org/publications/functional-requirements-for-authority-data

[9] IFLA. FRBR Review Group. Meeting Report Durban, August 21, 2007. Available at: http://www.ifla.org/files/cataloguing/frbrrg/meeting_2007.pdf

[10] IFLA. ISBD/XML Study Group. Available at: http://www.ifla.org/en/node/1795

[11] NSDL Registry. Available at: http://metadataregistry.org/

[12] Dunsire, G. Declaring FRBR entities and relationships in RDF. July 2008. Available at: http://www.ifla.org/files/cataloguing/frbrrg/namespace-report.pdf

[13] NSDL Registry. Element sets: show detail for FRBRer model. Available at: http://metadataregistry.org/schema/show/id/5.html

[14] Element set: FRBRer model . Available at: http://metadataregistry.org/schema/show/id/5.rdf

[15] W3C. OWL Web Ontology Language: overview. Available at: http://www.w3.org/TR/owl-features/

[16] International Working Group on FRBR and CIDOC CRM Harmonization. FRBR: object-oriented definition and mapping to FRBRer. Version 1.0. 2009. Available at: http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBR00_V2.0_2009_june_.pdf

[17] International Council of Museums. The CIDOC conceptual reference model. Available at: http://cidoc.ics.forth.gr/

[18] ICS-FORTH (ISL-ICS). CIDOC CRM v5.0.1 encoded in RDFS. 2009. Available at: http://cidoc.ics.forth.gr/rdfs/cidoc_crm_v5.0.1.rdfs

[19] IFLA. ISBD Review Group. Worldwide review of ISBD Available at: http://www.ifla.org/en/news/worldwide-review-of-isbd

[20] International standard bibliographic description (ISBD). Consolodated edition. Draft as of 2010-05-10. IFLA. ISBD Review Group. International standard bibliographic description (ISBD). Consolidated edition. Draft as of 2010-05-10. Available at: http://www.ifla.org/files/cataloguing/isbd/isbd_wwr_20100510_clean.pdf, pp. VI-VII.

[21] Escolano Rodríguez, Elena; Lynne Howarth; Mirna Willer; Boris Bosančić. News of ISBD. Project development of ISBD/XML schema: goals and objectives . Presented at World Library and Information Congress: 75th IFLA General Conference and Assembly, 23-27 August 2009, Milan, Italy. Available at: http://www.ifla.org/files/hq/papers/ifla75/107-escolano-en.pdf

[22] Dunsire, Gordon. UNIMARC, RDA and the Semantic Web. Presented at World Library and Information Congress: 75th IFLA General Conference and Assembly, 23-27 August 2009, Milan, Italy. Available at: http://www.ifla.org/files/hq/papers/ifla75/135-dunsire-en.pdf

[23] RDA (resource description and access) vocabularies. Available at: http://metadataregistry.org/rdabrowse.htm

[24] Hillmann, Diane; Karen Coyle; Jon Phipps, Gordon Dunsire. RDA vocabularies: process, outcome, use. In: D-Lib magazine, vol.16, no. 1/2 (January/February 2010). Available at: http://www.dlib.org/dlib/january10/hillmann/01hillmann.html

[25] Dublin Core Metadata Initiative. Libraries Working Group. Library application profile. 2004. Available at: http://dublincore.org/documents/library-application-profile/

[26] Malmsten, Martin. Exposing library data as linked data. Presented at the IFLA satellite preconference sponsored by the Information Technology Section "Emerging trends in technology: libraries between Web 2.0, semantic web and search technology", Florence, 19-20 August 2009. Available at: http://www.ifla2009satelliteflorence.it/meeting3/program/assets/MartinMalmsten.pdf

[27] Styles, Rob; Danny Ayers; Nadeem Shabir. Semantic MARC, MARC21 and the Semantic Web. Presented at Linked Data on the Web (LDOW2008). 2008. Available at: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper02.pdf

[28] Regole italiane di catalogazione REICAT / a cura della Commissione permanente per la revisione delle regole italiane di catalogazione ; [la redazione del testo è stata curata da Alberto Petrucciani]. Roma : Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni

bibliografiche, 2009. Bozza complessiva, Gennaio 2009. Available at: http://www.iccu.sbn.it/upload/documenti/REICA_bozza_complessiva_genn2009.pdf

[29] Tillett, Barbara B. and Ana Lupe Cristán. IFLA cataloguing principles: the statement of international cataloguing principles (ICP) and its glossary . München: K.G. Saur, 2009. Also available at: http://www.ifla.org/en/publications/statement-of-international-cataloguing-principles

[30] Petrucciani, Alberto. Every reader his work, every work its title (& author) : the new Italian cataloguing code REICAT. Presented at World Library and Information Congress: 75th IFLA General Conference and Assembly, 23-27 August 2009, Milan, Italy. 107 Cataloguing. Available at: http://www.ifla.org/files/hq/papers/ifla75/107-petrucciani-en.pdf

[31] Commissione RICA. L'applicazione del modello FRBR ai cataloghi: problemi generali e di impiego normative. Available at: http://www.iccu.sbn.it/upload/documenti/rica-frbr.pdf

[32] Vocabulary Mapping Framework. Available at: http://cdlr.strath.ac.uk/VMF/index.htm

[33] RDA/ONIX framework for resource categorization. Version 1. 2006. Available at: http://www.rda-jsc.org/docs/5chair10.pdf

[34] Dunsire, Gordon. Distinguishing content from carrier: the RDA/ONIX framework for resource categorization. In: D-Lib magazine, vol.13, no.1/2 (January/February 2007). Available at: http://www.dlib.org/dlib/january07/dunsire/01dunsire.html

[35] OCLC. Classify: an experimental classification web service. Available at: http://classify.oclc.org/classify2/

[36] Library of Congress. Authorities & vocabularies. Available at: http://id.loc.gov/authorities/search/

[37] OCLC. Terminology services: experimental services for controlled vocabularies. Available at: http://tspilot.oclc.org/resources/index.html

[38] RAMEAU subject headings as SKOS linked data. Available at: http://www.cs.vu.nl/STITCH/rameau/

[39] MACS Project. Available at: https://macs.hoppie.nl/pub/

[40] Dewey Decimal Classification: summaries. Available at: http://dewey.info/

[41] UDC summary. Available at: http://www.udcc.org/udcsummary/php/index.php

[42] IFLA Classification and Indexing Section. Newsletter, no.40 (December 2009). Available at: http://www.ifla.org/files/classification-and-indexing/newsletters/ifla-newsletter-classification-40_rev.pdf

[43] Hickey, Thom. VIAF as linked data. 2009. Available at: http://outgoing.typepad.com/outgoing/2009/09/viaf-as-linked-data.html

[44] IFLA. ISBD Review Group. International standard bibliographic description (ISBD). Consolidated edition. Draft as of 2010-05-10. Available at: http://www.ifla.org/files/cataloguing/isbd/isbd_wwr_20100510_clean.pdf, pp. 0.1-1-0.3-2

[45] OCLC FictionFinder. Available at: http://fictionfinder.oclc.org/

[46] W3C. SWEO Community Project: Linking Open Data on the Semantic Web. Statistics on data sets. Available at: http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

[47] LibraryThing. Available at: http://www.librarything.com/