

Report to the ISBD/XML Task Group, World Library and Information Congress 2009, Milan, Italy

Vocabulary Mapping Framework and ISBD

The Vocabulary Mapping Framework (VMF) project is funded by the UK's Joint Information Systems Committee. It commenced in July 2009, and will report to a seminar at the British Library in London in early November 2009. VMF is an extension to the RDA/ONIX framework for resource categorisation. It aims to analyse bibliographic entities and relationships from a wide range of published standards to identify common semantic categories, develop an ontology for bibliographic relationships, and map appropriate elements of the source standards to the ontology. The result will be a spoke-and-hub set of mappings between source standards and the common ontology, which can be used to establish cross-walks and other mappings between any pair of source standards. The whole framework will be published openly in semantic web formats.

Although ISBD is not one of the source formats identified by the project, I have asked for it to be included in the analysis of entities and relationship. Developing a mapping from ISBD to the VMF relationship ontology and the existing RDA/ONIX ontology for content and carrier categories will require declaration of appropriate elements of ISBD in semantic web formats. These elements will include agent- and resource-type entities as well as relationships between them.

This work can be carried out as part of the ISBD/XML project, or as an extension to the VMF (I have already offered resources from my own organization to do this)

Further information is available from the VMF website at <http://cdlr.strath.ac.uk/VMF/>

XML vs RDF

Semantic web formats are based on Resource Description Framework (RDF). RDF breaks down metadata into simple three-part statements, for example thisReport:has Author:Gordon_Dunsire. This allows complex metadata structures and records to be created from multiple simple statements; moreover, the statements can be taken from a variety of sources such as publishers, libraries, booksellers, encyclopedias, users, etc.

"... XML is very suitable for data interchange between applications that both know about what the data is, but not in situations where the addition of new communication partners occurs frequently ... When it comes to *Semantic Interoperability*, RDF has significant advantages over XML: semantic units are given naturally through its object-attribute structure: all objects are independent entities." (The Semantic Web - on the respective Roles of XML and RDF / Stefan Decker ... [and others]. Available at: <http://www.ontoknowledge.org/oil/download/IEEE00.pdf>)

RDF statements can be expressed in a variety of syntaxes, including XML. The difference between XML and RDF/XML is that the latter must use RDF-specified markup. In general XML the ISBD Title attribute can define <Title></Title> as its markup without specifying the meaning of "Title". In RDF/XML the attribute must be given a Uniform Resource Identifier (URI) so that it can be linked, say, to a definition, scope note, super- and sub-classes, etc.; the links can be processed by machine to infer the semantics of "Title".

URIs are the basis for “namespaces” which define metadata structure and content terminology and semantics. The Appendix to this report outlines some of the utility to be gained from semantic web representations of metadata standards, and raises issues and questions to be addressed by IFLA as a whole. It has been included in a document tabled to the Cataloguing and Classification and Indexing Sections meetings at WLIC 2009. It is included in this report so that these issues can be discussed from the point-of-view of ISBD. It has also been included in reports to other IFLA groups, including the FRBR Review Group, for similar standard-specific discussion.

Recommendation: The ISBD/XML Task Group should consider declaring ISBD elements in RDF/XML (using RDF applications such as Simple Knowledge Organisation System (SKOS), RDF Schema, and Web Ontology Language (OWL)), rather than general XML.

Gordon Dunsire
6 Aug 2009

Appendix

Identification of potential requirements for managing IFLA namespaces

The FRBR Review Group has initiated a project to develop appropriate namespaces for allowing other communities to use the FRBR(er) model in a semantic web environment. The ISBD Review Group has initiated a project to develop an XML representation of ISBD; certain elements may be more usefully expressed in an XML syntax compatible with the semantic web. Informal interest in these initiatives, and interaction with the semantic web in general, has been expressed by a number of other IFLA standards groups, including FRANAR (for FRAD) and the Permanent UNIMARC Committee. The term “namespace” is used hereafter as a synecdoche for a namespace itself (which assigns a machine-processable identifier to metadata structural and value components expressed in the semantic web’s Resource Description Framework (RDF)) together with XML representations of its whole and parts and associated structural components.

Question: Are all or most IFLA groups responsible for the development and maintenance of bibliographic standards likely to develop namespaces?

After a namespace has been created and metadata elements declared within it, there are a number of issues associated with the management and deployment of RDF and XML representations of the namespace.

The simplest approach is to make those representations available to other communities and individuals in an open environment. This can be achieved by storing the files on a web server with open access. The representation files can be copied for local processing in third-party services; such services might include the generation of display lists of vocabulary terms, for example the FRBR(er) Group 1 entities, construction of entity-relationship diagrams, translations between machine-readable identifiers and human-readable terms and between different language versions of terms, etc.

One of the more important services is dereferencing (see the Wikipedia article at http://en.wikipedia.org/wiki/Dereferenceable_Uniform_Resource_Identifier). The most effective use of namespaces in the Semantic Web is the substitution of machine-readable identifiers (URIs) for text values in metadata structures and record instances. This speeds machine-processing and removes the ambiguity in labels such as FRBR’s “Work”. However, at some point it may be necessary to replace the identifiers with human-readable labels through a dereferencing service; the URI is passed to the service, and the appropriate label returned.

Conversations with DCMI communities have raised the problem of rogue dereferencing requests, caused by improperly constructed link resolvers and crawlers. These may misinterpret a URI to be a URL and attempt to harvest the linked document (which is actually the namespace representation) for further automatic link checking. This can result in overloading the server containing the namespace representation. For this reason, it is highly recommended that such servers be isolated from other organisational web servers and use a different domain so that, for example, document services (such as the current IFLA website) can continue to operate satisfactorily.

Question: Do IFLA standards bodies wish to provide dereferencing and other terminology services based on their namespaces?

Conversations with colleagues working on the development of the RDA online product have raised a number of issues related to maintenance of namespaces, including:

- Version control. Even small, relatively static vocabularies require occasional amendment, and it is important that users of those vocabularies can be certain of the version they are employing, and of the existence of later versions.
- Change alerts. A method of alerting vocabulary users to new amendments and versions, for example via RSS, can help promote usage.
- Translations. The NSDL Metadata Registry service, currently used for the FRBR namespace project, has the capability of maintaining controlled vocabularies in translation. An example of this can be seen in the RDA content type vocabulary at <http://metadataregistry.org/vocabulary/show/id/45.html>. The service supports any number of languages, and in any script. It is highly likely that non-English communities will be interested in authoritative translations of vocabulary entries; note that the URI for the term remains the same, thus ensuring automatic interoperability between metadata records created in different countries.
- Feedback and constituency involvement. Social networking space for providing a means to manage comments, suggestions and queries from users, user groups, and constituencies is now becoming an expected feature of open standards administration.

Question: Do IFLA standards bodies wish to offer or utilise any of these functions, or others to be identified, as part of their management of existing and future namespaces?

Another set of issues is associated with promotion and utilisation of the namespace. These include:

- Provision of an Application Programming Interface. This would allow external developers to incorporate IFLA namespace entities using simple programming structures, rather than download the namespace files for local program integration. It would also improve control and co-ordination of use of the namespace, as well as potentially incorporating the version control and translations issues raised above.
- Exposure of the namespace as linked data (see the Wikipedia article at http://en.wikipedia.org/wiki/Linked_data). This overlaps with the dereferencing issues raised above. Adding IFLA data to the linked data universe will ensure maximum re-use, and potentially have a very significant impact on the development of the semantic web.
- Provision of (authoritative) mappings between namespaces. These namespaces may be IFLA ones, or external, such as the outputs of the Vocabulary Mapping Framework (VMF) (<http://cdlr.strath.ac.uk/VMF/>) project.
- Provision of links to (authoritative) RDF schema using IFLA namespace entities. This is similar to the idea behind the bibliographies maintained to track the interest and take-up of IFLA standards, by acting as a showcase for standards' utility, but could also be part of a clearing-house or directory of external services which have made use of a specific IFLA standard.

Question: Do IFLA standards bodies wish to employ some or all of these methods of promoting their standards to semantic web developers, with the secondary result of improving bibliographic control at a global level?

Need for an IFLA standards namespaces framework

There are number of reasons why IFLA might consider developing an overarching framework for promoting and disseminating its standards.

There is a current framework implicit in the structure of IFLA divisions, sections and special interest groups. Standards produced by IFLA are constructed for a print-based environment, and most are made available in digital format as electronic surrogates of the print-based originals. There is no single place on the IFLA website where all standards are listed. Instead, retrieval requires knowledge of which area of the IFLA organisational structure is responsible, or enough of a citation to carry out a keyword search.

Question: Does IFLA wish to improve the identification and retrieval of its published standards?

However, the world in which libraries increasingly operate is digital. Non-IFLA communities and individuals expect a global information management community to present its outputs in formats that can be exploited more effectively by digital tools and techniques. One of the most significant developments in the digital environment is the creation of the semantic web, which offers huge potential for the utilisation of the standards and expertise of professional librarianship.

Question: Does IFLA recognise the potential role of its activities in developing the semantic web, and the potential role of the semantic web in furthering IFLA's aims? Does IFLA wish to engage more fully with the semantic web?

The semantic web architecture has nothing to say about the veracity of statements expressed in RDF. That is, false statements can be made, and there is no intrinsic way of determining whether a statement is true or false. Instead, extrinsic methods must be used: What is the source of the statement? Is the source trustworthy? Does the source ensure that statements remain true? There are no established protocols for answering these questions, but some assessment of veracity can be made from knowing which body created and maintains the namespaces used in RDF statements. If IFLA namespaces have "IFLA" as part of their URIs, preferably in the base domain used for all URIs in a namespace, then the brand itself can indicate a degree of reliability. Other techniques might involve human-readable background or explanatory documentation associated with a namespace.

Question: Does IFLA wish to capitalise on its brand and standing in the library world by extending it to namespaces for its standards?

Two major initiatives are underway in non-IFLA communities. These are the development of RDA: resource description and access as a digital publication exploiting the semantic web to express its metadata structure and value vocabularies, and the VMF as a hub-and-spoke approach to improving interoperability between many bibliographic standards in the semantic web. VMF is essentially an extension to the RDA/ONIX framework for resource description. Both initiatives involve IFLA standards: RDA is linked to elements of FRBR(er), FRBR(oo), and FRAD; VMF will include FRBR(er), FRBR(oo) via CIDOC CRM, and elements of FRAD, ISBD and UNIMARC. Both projects need to liaise with the bodies responsible for standards governance and maintenance, to ensure that their work is authoritative and "official".

Question: Would liaison between IFLA and such projects be more efficient and effective if there is a single IFLA body to contact, rather than the individual components of the IFLA organisational structure?

Many of the IFLA standards have interdependencies with each other, expressed as references and mappings. In addition, many IFLA standards have interdependencies with external standards, and in some cases the same external standard. Such dependencies can be chained together, showing indirect dependencies between pairs of standards (see my WLIC 2009 paper on UNIMARC, RDA and the Semantic Web available at <http://www.ifla.org/files/hq/papers/ifla75/135-dunsire-en.pdf>)

Question: Are the dependencies between IFLA standards, and between specific IFLA standards and external standards, properly understood by IFLA standards bodies, and sufficiently well understood to ensure semantic cohesion and avoid duplication of effort?

The potential requirements for managing an IFLA namespace identified above are common to all namespaces.

Question: Is it more effective or efficient for IFLA namespaces to be developed and administered as a whole, or separately by the relevant standards bodies?

Gordon Dunsire
4 Aug 2009