Improving the sharing of data: the Vocabulary Mapping Framework project

Metadata standards have proliferated over the last 20 years, as have the volume and diversity of data creation. We now need greater interoperability to allow the transfer of metadata to aid the retrieval of resources. **Helen Williams** reports on the Vocabulary Mapping Framework project, which aims to create a downloadable tool to support interoperability across communities.

The VMF is a first step towards understanding an appropriate balance between the use of standard schemas on the one hand, and interoperability mechanisms on the other, in order to allow greater sharing of data.

THE JISC-FUNDED VOCABULARY Mapping Framework (VMF) project aims to 'create an extensive and authoritative mapping of vocabularies from major content metadata standards, creating a downloadable tool to support interoperability across communities'.[1]

Stage one of the project was presented to representatives from both library and industry sectors at the British Library on 9 November 2009.[2]

Mark Bide, from EDItEUR, the trade standards body for the global book and serials supply chain, opened proceedings by outlining the prolific growth in metadata standards over the last 20 years. Such variety has meant that individual areas have been unable to make much use of metadata provided by other communities (e.g. libraries making use of publisher data) because the different systems have not been able to exchange meaningful information with one another.

Although each standard was designed to meet different requirements, there is a fair amount of overlap between them. The VMF is a first step towards understanding an appropriate balance between the use of standard schemas on the one hand, and interoperability mechanisms on the other, in order to allow greater sharing of data.

'Museum of the book' or 'network of knowledge'?

Before Gordon Dunsire, Head, Centre for Digital Library Research, and Godfrey Rust, Principal Data Architect for Ontologyx, Rightscom's ontology product initiative, gave details about what the VMF matrix is and how it is used, Alan Danskin from the British Library addressed the choice facing libraries – to be a 'museum of the book' or a 'network of knowledge'. A modern library is a hybrid facility consisting of both physical and digital collections.

The digital material (e.g. full text, data sets, web archives and moving images) may be born-digital or may have been digitised later in its life cycle, and different resources will provide access at different levels of granularity. Libraries have a long tradition of metadata creation to aid description, access and inventory control, but the digital age means that there is more material, with a greater level of complexity and detail than ever before. Increasingly, Anglo-American Cataloguing Rules, created for printed materials in linear sequences, are not meeting the requirements of this material. In contrast, Resource Description and Access (RDA), set to supersede AACR2 in the not too distant future, is intended to inform metadata creation for all types of resources, and using the underlying models of FRBR (functional requirements for bibliographic records) and FRAD (functional requirements for authority data) will support the use of linked data.[3]

On the bibliographic continuum, we might say that, historically, product (or identification) metadata has been produced by publishers, while controlled (or contextualised) metadata has been created by libraries. While a degree of feedback has been possible between these, the overwhelming result has been silos of metadata, partly because different viewpoints and end goals have prevented co-operation. Increasingly, however, the sheer volume and diversity of data creation requires greater interoperability to allow the transfer of metadata to aid the retrieval of resources.

The development of RDA and its outreach to other communities led to the creation of the RDA/Onix framework[4] in 2005/06 to address one important aspect of this issue. Alan described it as a semantic convergence creating a framework for categorising resources in all media to support the needs of both libraries and publishers. This was followed by the DCMI/RDA Task Group[5] which considered how RDA related to other metadata models.

The VMF is an extension of this principle, incorporating more standards and vocabularies to allow for richer, bilateral mappings and open the way for incremental convergence between metadata standards related to the Jisc community (with the ability to extend to other metadata sets). This work opens up opportunities for libraries to contribute expert metadata into the linked data pool.

The goal of the VMF is automatically to compute the 'best fit' mappings between any two vocabularies whether they are from the library/museums/archives

world, the music industry or the education sector. The VMF has so far dealt with nine[6] schemas and mapped 53 vocabularies, resulting in more than 500 concept families, and more than 30,000 RDF (Resource Description Framework) triples. [RDF triples are subject-predicate-object expressions which link resources. Take the statement 'Helen Williams is the author of the article "Improving the sharing of data: the Vocabulary Mapping Framework project"', for example; in RDF terms the person 'Helen Williams' and the article 'Improving the sharing of data: the Vocabulary Mapping Framework project' are resources linked by the relationship 'is the author of'. The challenge is for the VMF to accommodate the data models of not only all the vocabularies already mapped to it, but also as yet unknown vocabularies in the future. In older metadata schema, such as AACR2, ISBD and Marc, the significance of relationships has not been fully exploited. Relationships may be implicitly stated in attributes rather than represented explicitly. RDA defines a rich vocabulary of terms to express relationships defined in FRBR and FRAD. This rich vocabulary is expected to provide a good fit with RDF.

The matrix has been developed to allow machine processing of connecting terms in different vocabularies, and as such is not for human use (though in the longer term the matrix could be used to publish recommended mappings for specific schemas). Using SPARQL queries, the terms, which have been mapped into the matrix through a hierarchical event-based concept ontology, can be mapped to terms from other vocabularies in terms of parent, child, sibling or homogenous relationships. The matrix data is prepared in Excel and extracted automatically to RDF triples with RDFS and OWL axioms providing the core logic.

The advantage of the VMF is that, as a framework specifically designed for mapping, it has a 'hub and spoke' approach: any newly added schemas will need only to be mapped to the VMF matrix (rather than to each of the schemas contained in it) in order to get the 'best fit' mapping to any other schema already mapped.[7]

The matrix is now ready for use, though it does need to be refined by extending the mappings and by validating them with the authorities for each participating schema, so that each agrees with the VMF terms to which its vocabulary has been mapped. This means each mapping will be 'authority-controlled' and can be considered reliable by the communities which will be using them. What the VMF cannot do, of course, is ensure that those who have created the original data, say with Marc 21, have used the format correctly to input the metadata. In theory though, once each schema has been validated by its governing body, the VMF will be an authoritative source for public vocabulary mappings, allowing metadata crosswalks between different vocabularies such as publishers and libraries.

The next step is for some projects to test the matrix, so that errors can be fixed and further refinements made.

Keeping VMF current

Alongside that, a governance model, with support from participating standards, needs to be established, together with a long-term maintenance proposal or review mechanism to ensure the VMF will be kept current as individual schemas are amended or developed. This raises various issues currently under discussion, including the need for technical roles, funding, assurance of continuity, authorisation, intellectual property considerations and marketing. There was much discussion about, although no conclusions on, possible models.

Use cases for the VMF can be seen in detail on the project webpage[8] and include Onix publishing data being transferred to Marc for library use; mapping a local or bespoke schema to VMF so that the data can more easily be exposed through sources using other metadata schema; linking related works across different domains; improved cross-search services; and the transfer of preservation metadata.

It is important that the library community recognises the pressure from the outside world for greater sharing of interactive metadata. The VMF is helping to achieve the solution. 🅄

References

1 www.jisc.ac.uk/whatwedo/projects/vocab-framework.aspx

2 http://cdlr.strath.ac.uk/VMF/seminar.htm

3 www.ifla.org/en/publications/functional-requirements-for-bibliographic-records

4 www.dlib.org/dlib/january07/dunsire/01dunsire.html

5 http://dublincore.org/dcmirdataskgroup/

6 http://cdlr.strath.ac.uk/VMF/documents/VMFProjectAnnouncement.pdf

7 http://cdlr.strath.ac.uk/VMF/background.htm

8 http://cdlr.strath.ac.uk/VMF/documents.htm

Helen Williams is Assistant Librarian, Bibliographic Services, London School of Economic & Political Science Library (h.k.williams@lse.ac.uk).

Acknowledgements

This article is based on the presentations given at the VMF seminar at the British Library, and credit should be attributed to Mark Bide, Alan Danskin, Godfrey Rust, Gordon Dunsire and Norman Paskin.