Date : 22/05/2007

Library of Congress controlled vocabularies, the Virtual International Authority File, and their application to the Semantic Web

Barbara B. Tillett and **Corey Harper**

Meeting:	**147 IFLA-CDNL Alliance for Bibliographic Standards (ICABS)**
Simultaneous Interpretation:	Yes

WORLD LIBRARY AND INFORMATION CONGRESS: 73RD IFLA GENERAL CONFERENCE AND COUNCIL
19-23 August 2007, Durban, South Africa
http://www.ifla.org/iv/ifla73/index.htm

Abstract:

This presentation is based on an article written by the author and Corey Harper for *Cataloging & Classification Quarterly* ("Library of Congress controlled vocabularies and their application to the Semantic Web," by Corey A. Harper and Barbara B. Tillett, v.43, no. 3/4, 2006). It reviews the Library of Congress controlled vocabularies (Library of Congress Subject Headings, Library of Congress Classification, and Library of Congress/NACO Authority Files) and describes the VIAF project (Virtual International Authority File. These controlled vocabularies hopefully will be useful as building blocks for the Semantic Web to internationally link the world's authority data from trusted sources to benefit users worldwide.

The 2001 *Scientific American* article by Tim Berners-Lee on the Semantic Web[1] set the stage for the idea of a linked universal Web of data. He mentioned that independent groups are working on similar concepts and need a common language to better interoperate. **Libraries and the developers of the Semantic Web share goals for naming concepts, naming entities, and bringing different forms of those names together**. Library tools have been developed over many decades and are very rich sources of connected data. We just need to now translate them into new tools to help the infrastructure of the Semantic Web. We seem to be reaching a critical mass of understanding and agreement that can help move us forward with developing tools for the future, and libraries have a great part to play.

One giant step will be to move our controlled vocabularies to Semantic Web standards, such as the Web Ontology Language (OWL), so they can be available for use in new ways. There is also the Semantic Web technology called Simple Knowledge Organization System (SKOS) Core for encoding the contents of thesauri. So for our authority data, both SKOS and OWL represent useful ways to translate the authority information. Also useful for the Semantic Web is to translate our bibliographic data into RDF (Resource Description Framework) and initial work to move us in that direction is taking place with the Dublin Core and RDA (Resource Description and Access) communities.[2]

The validity and trustworthiness of library controlled vocabularies is well-respected and acknowledged, which is beneficial in moving things forward.

Most of the work to date with digital libraries in particular has been to make library collections known on the Web – with digitization of the contents of selected collections. Libraries' authority data has an even greater potential of helping users find what they need on the Web by providing pathways that will connect them to relevant information.

Web 2.0 is often used to describe user interactive systems on the Internet – applications that enable users to customize things they find, adding their own comments, their own "folksonomies" or tags – that we'd call subject headings or access points – to make material they wish to use more easily findable and usable in their own space.

To give you examples of controlled vocabularies created by the library community, we have the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC). We also have the subject heading terms in controlled lists – the Library of Congress Subject Headings (LCSH), two Thesauri for Graphic Materials: Subject Terms (TGM I) and Genre & Physical Characteristic Terms (TGM II), Guidelines on Subject Access to Individual Works of Fiction, Drama, Etc. (GSAFD), and the Ethnographic Thesaurus. Currently there is no way for search tools to take advantage of the syndetic structures (cross-references) in these controlled vocabularies.

Controlled vocabularies could make the Semantic Web and Web 2.0 tools and services much more useful – providing structure and connections for topics and for names of persons, things, etc.

There is a benefit of Semantic Web technology to those controlled vocabularies as well. Representing the syndetic structures of such vocabularies using the common framework provided by OWL or SKOS makes cross-vocabulary interoperability more plausible. Representing and modeling vocabularies in this way makes it easier to identify and exploit relationships and equivalencies between concepts in disparate controlled vocabularies. This could, as an example, enhance some of the clustering capabilities showing up in emerging next-generation catalog systems.

OCLC has experimented with some uses of controlled vocabularies within existing software like Microsoft Word. In OCLC's Terminologies Project they can enable people

using Word to reach controlled vocabularies and find subject terms to use in their documents as they create them – without having to exit the Word software.[3] We could imagine catalogers having a similar tool to quickly access appropriate terms from controlled vocabularies in an efficient system that would save time and cataloging costs. The Terminologies Project is already laying the ground-work for these advancements by integrating the tools exposed through Microsoft Office into the OCLC Connexion cataloging software.

We also could make our existing name authorities available in a more Web-friendly structure so they could be used for Web applications. It is hoped the Virtual International Authority File could eventually evolve into such a resource.

The Virtual International Authority File (VIAF) is a project that has been a dream for a long time. One model is being tested with the Deutsche Nationalbibliothek, the Library of Congress, the Bibliothèque nationale de France, and OCLC. The initial stages have been tested, but we still need to develop the end-user applications and ways to make the data more Web-accessible. I would refer you to several IFLA publications and other reports on the VIAF project rather than repeat them here. The basic idea is a system that would be internationally shared, linking the major authority files of the world in a freely accessible system that links the authority records for the same entity and makes al the authority records available for re-use for many applications. This shared file would help reduce cataloging costs in individual libraries, offer a tool to display a user-preferred language or script for the names of people (and eventually corporate bodies and geographic names). A single system like this would be far more useful than having to access all the world's authority files separately and could be used by Web systems to add precision to searches that is now sadly missing from search engines like Google.

This view of authority records means that any of the variant forms of names clustered together in the authority record and by extension the linked authority records, could be searchable and used for displays for end-users. In the past, libraries have chosen only one form of name to be the authorized display form, and such a default would still be needed when a user didn't select a preference for a language or script.

We are hoping eventually to extend the VIAF project to non-roman scripts.

In addition if we extend VIAF to Semantic Web communities, the power of our authority system can be used in many different types of applications. More pathways open up to the resources by or about the person being searched or to connect to other interesting pieces of information – their blog, a wikipedia entry, holdings by or about that person in your local library, biographical information available through dictionaries and encyclopedias, journal articles written by them or about them, things they have available through e-commerce, and much more. These interconnections could help in evaluating the trustworthiness or authenticity of what you find on the Web or to give the end-user the context of the persons' own personal biases that might be influencing the documents they write.

So how could we get there from our current authority records? First our authority records need to be in machine-readable format, especially in formats designed for the Semantic Web, such as OWL or SKOS. Another step is to identify each term with a URI or instead to map the XML version of authority records to SKOS. Another option is to convert to the format on the fly from the original format. For example with the VIAF stored as MARCXML, the data can be converted as needed into RDF or SKOS to be used in each particular application. It also could augment systems that develop folksonomies.

The information included in authority files could then be repurposed in a variety of other applications in a very ad-hoc manner. Identifying authority information with URIs allows those URIs to be re-used to tie other descriptions of people to authority records, which in turn link to their works. Emerging standards for encoding relationships between people, such as the Friend of a Friend project (FOAF), could leverage this information to great benefit.[4] FOAF provides mechanisms for documenting relationships between different people as well as between individuals and the various things they create. The ability to include URI's for VIAF records in FOAF descriptions helps extend the library community's role of documenting trustworthiness and credibility. It would become easier to identify the contributor to a blog or Wikipedia entry as a reputable authority on a given topic.

As we noted in the 2007 article: "Part of the Semantic Web vision is about aiding resource discovery by creating tools to help searchers refine and develop their searches, and to aid in the navigation of search results."[5] The controlled vocabularies that libraries create could be a powerful and wonderful resource to help improve the end-user's experience on the Web.

An important meeting held April 30-May 1, 2007 in London brought together representatives from the SKOS, Semantic Web, Dublin Core, IEEE/LOM and RDA (Resource Description and Access) at the invitation of the co-publishers of RDA. Initially the meeting was to review various data models (including FRBR and FRAD) and to explore the usefulness of the RDA content standard for some of the metadata communities. The results of this meeting were posted immediately and recommendations were made to seek funding to develop an RDA Application Profile that could:

Develop an RDA Element Vocabulary

Develop an RDA DC (Dublin Core) Application Profile based on FRBR and FRAD and

Disclose RDA Value Vocabularies using RDF/RDFS/SKOS.

As the announcement said: "The benefits of this activity will be that:

- the library community gets a metadata standard that is compatible with the Web Architecture and that is fully interoperable with other Semantic Web initiatives
- the DCMI community gets a libraries application profile firmly based on the DCAM and FRBR (which will be a high profile exemplar for others to follow)
- the Semantic Web community gets a significant pool of well thought-out metadata terms to re-use
- there is wider uptake of RDA."

Bringing the library experience, standards, and controlled vocabularies into the Semantic Web will be beneficial to the library and metadata communities and the end-users will be the primary beneficiary.

---

[1] Berners-Lee. T., Hendler, J., & Lasilla, O. "The Semantic Web" [Electronic version]. *Scientific American*, 284, no.5, 2001, p. 34-43. http://www,sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[2] See meeting results announcement from the April 30-May 1, 2007 London "Data Model Meeting" at http://www.bl.uk/services/bibliographic/meeting.html

[3] Vizine-Goetz, Diane. "Terminology services: making knowledge organization schemes more accessible to people and computers." *OCLC Newsletter*, no. 266, 2004. Available from : http://www.oclc.org/news/publications/newsletters/oclc/2004/266/

[4] See homepage of the Friend of a Friend project. http://www.foaf-project.org/

[5] Harper Corey A. & Tillett, Barbara B., "Library of Congress controlled vocabularies and their application to the Semantic Web," *Cataloging & Classification Quarterly*, v.43, no. 3/4, 2006.