Rare Books and Manuscripts Section
Controlled Vocabularies Editorial Committee
Linked Open Data Report
ALA Midwinter 2016

Amy Brown
Allison Jai O'Dell
Amber Billey

INTRODUCTION

Publication of the RBMS *Controlled Vocabularies for Use in Rare Book and Special Collections Cataloging* (RBMS-CV) as Linked Open Data will enable rare materials and special collections catalogs to integrate with and become discoverable on the Semantic Web. Publication as Linked Open Data will also streamline library workflows -- especially reducing the labor in authority control -- and support broader initiatives in the library profession, such as BIBFRAME.[1]

At the 2015 ALA Annual Conference and Exhibition, the RBMS Controlled Vocabularies Editorial Group charged a working group to investigate and recommend a solution for publication of the RBMS *Controlled Vocabularies* as Linked Open Data. This document summarizes the findings of the working group (Amber Billey, Allison Jai O'Dell, and Amy Brown, in consultation with Jason Kovari), and presents a solution for management and publication of library thesauri as Linked Open Data, within the context of a collaborative and dynamic editorial workflow.

Summary of recommendations:

- Host and publish the *Controlled Vocabularies* on a new domain or subdomain
- Migrate to a Linked-Data-friendly Content Management System
- Create meaningful Linked Open Data during the editorial process
- Provide multiple points of access to the Controlled Vocabularies as Linked Open Data, including a human-readable, searchable, and browsable front-end interface, data export options in Linked Data formats, a SPARQL endpoint, and periodic data dumps to the Library of Congress Linked Data Service.

BACKGROUND

The RBMS *Controlled Vocabularies*

"The *Controlled Vocabularies for Use in Rare Book and Special Collections Cataloging*, [is] developed and maintained by the Bibliographic Standards Committee of the Rare Books and

[1] For an explanation of BIBFRAME, see: http://www.loc.gov/bibframe/

Manuscripts Section (ACRL/ALA). [The] thesauri provide standardized vocabulary for retrieving special collections materials by form, genre, or by various physical characteristics that are typically of interest to researchers and special collections librarians, and for relating materials to individuals or corporate bodies."[2]

Linked Data

Linked Data refers to a set of principles that confer machine-actionable, semantic meaning to data using Web technologies. This is achieved with the Hypertext Transfer Protocol (HTTP) for data communication, Uniform Resource Identifiers (URIs) for a kind of authority control, and a simple, three-part data structure known as a triple (specified by the Resource Description Framework, or RDF[3]). A brief explanation of the principles and mechanics of Linked Data is given below. Benefits of Linked Data include connected information networks, ease of data merger, knowledge inferencing, search engine optimization, and more.[4]

The World Wide Web has historically relied on Hypertext Markup Language to encode and link documents. HTML encoding results in content that is semantically meaningful to humans, but not to machines. For instance, in an HTML document, the statement:

John Steinbeck is the author of The Grapes of Wrath.

can be interpreted by a human as semantically meaningful pieces of information (e.g., that John Steinbeck is a person and a writer, who authored a book called *The Grapes of Wrath*). Humans use context and semantics to derive information from this statement -- but machines must rely on mark-up formats to interpret data. HTML mark-up supplies formatting information alone. Linked Data formats supply the context and semantics that machines need to understand and use information. This is done by defining things with URIs and making assertions about them through relationships. Relationships are made explicit through the use of triples, which are three-part statements that contain a subject, predicate, and object, thereby relating the subject data to the object data. Using our example above, we could write the triple statement:

"John Steinbeck" "author" "*The Grapes of Wrath*".

And using URIs:

<http://viaf.org/viaf/96992551>
<http://schema.org/author>
<http://www.worldcat.org/oclc/289946>.

This tells a machine unequivocally the same information and meaning that a human interprets in reading the original sentence.

---

[2] http://rbms.info/vocabularies/
[3] http://www.w3.org/RDF/
[4] Library Linked Data Incubator Group, "Benefits." http://www.w3.org/2005/Incubator/lld/wiki/Benefits

In order for triples to make sense to a machine, the subject and predicate data must be defined by a URI. Furthermore, these URIs should be stable in perpetuity and dereferenceable, meaning that information will be returned when a human user or robot accesses that URI.

The Semantic Web is the desired outcome of Linked Data implementation -- a Web where the data in documents is discoverable, connectable, and re-purposable.[5] Using the simple triple structure and URIs, data can be understood and processed by machines, heterogenous data sources can be merged, and new inferences can be made from combined data sets.

The *Controlled Vocabularies* can contribute to the Semantic Web by creating URIs for RBMS-CV concepts and by offering useful information in Linked Data formats when those URIs are dereferenced. Currently, the *Controlled Vocabularies* support human understanding of data and the relationships between data. Publication of the RBMS-CV as Linked Data will additionally enable machine understanding of the data contained, ultimately supporting such initiatives as BIBFRAME, automated resource description, global data integration, Web-based discovery, and the Semantic Web.

Impetus

A need arose for a new management solution to support production and publication of the RBMS-CV as Linked Data. The current content management system, MultiTes, is client-based and requires ongoing maintenance to publish new concepts. Static HTML pages are generated from a SQL database, and manually uploaded to the rbms.info site. We could upgrade to the MultiTes Online (cloud-based) version for a fee, however the hosting is maintained by MultiTes. This means that the vocabulary would reside at multites.org or on MultiTes servers. So, with MultiTes, there is no cloud-based option that will give us full control to manage the vocabularies. Most importantly, there is no way to openly query and access the metadata in MultiTes through the Web, as it does not have an API or SPARQL endpoint.

With the issues of MultiTes in mind, it was desirable to investigate other options -- especially considering the increase of open source Linked Data vocabulary services and software being developed by the library and information community.

Charge

A working group of the RBMS Controlled Vocabularies Editorial Team was charged to investigate and recommend a solution for publication of the RBMS *Controlled Vocabularies* as Linked Data. The group identified three primary options at the outset of the project. These options are outlined briefly here, in no particular order, and discussed in depth later.

- Allow the Library of Congress to host the vocabularies solely at their http://id.loc.gov authorities website. This option was attractive because it guarantees high visibility within the heavily-used and well-known Library of Congress network. However, it has significant workflow implications, and would necessitate RBMS losing some control of vocabulary maintenance.

[5] http://www.w3.org/standards/semanticweb/

- Host the vocabularies locally on a new subdomain, using either the software TemaTres or Vitro. This option allows for complete control to remain with the Controlled Vocabularies Group, but necessitates that we take sole responsibility for technical support.
- Partner with the Library of Congress Genre & Form Terms (LCGFT) and maintain the vocabularies only in partnership with the Library of Congress. We would serve in an advisory role as an expert community, but would not maintain oversight of the vocabularies.

Metrics

The solution must:

- Meet the research and cataloging needs of the rare materials and special collections community
- Allow publication of the RBMS-CV concepts as dereferenceable URIs in Linked Data format(s)
- Be simple enough for execution by a small volunteer organization

INVESTIGATION

The first route we considered was migrating the vocabulary to a central linked data vocabulary service. We contacted Nate Trail at the Library of Congress to explore moving the RBMS vocabularies to http://id.loc.gov -- the Library of Congress Linked Data Service (ID.LOC). ID.LOC would only be able to provide a front-end public access to the vocabularies. There is not web-based back-end management utility, and therefore it would not support the current RBMS workflows. As a result, it was recommended that we continue to maintain a separate instance of the vocabulary and send periodic data dumps to ID.LOC to publish the vocabularies through their service. While ID.LOC is not a one-stop solution for the RBMS vocabularies Linked Data problem, it will still provide a highly visible and accessible option for publishing and promoting the vocabularies. Duplication in two separate domains is not a concern as the terms can be linked with a sameAs relationship in their metadata.

Another option was to move the RBMS vocabularies to the Open Metadata Registry (http://metadataregistry.org/). After some consideration, this option was dismissed because it does not have the same management functionality as other tools so terms would have to entered manually/individually. We were also not confident in its lasting stability, since the RDA vocabularies no longer use it.

The last vocabularies service option considered was to merge the RBMS Vocabularies with the Library of Congress Genre and Form Terms vocabulary. We felt that this would be a tremendous loss of the unique RBMS vocabularies brand and product, and we cannot vote to dissolve ourselves.

With none of the Linked Data services meeting the desired requirements, we investigated software solutions to host and publish the Controlled Vocabularies ourselves.

SOLUTIONS

Host and publish the *Controlled Vocabularies* on the rbms.info domain
Currently, the *Controlled Vocabularies* are published at http://rbms.info/vocabularies/. Use of the rbms.info domain firmly establishes the identity of the RBMS-CV as an RBMS publication. To simplify the base URI pattern, we suggest using a subdomain, rather than a subdirectory -- that is, http://vocabularies.rbms.info. The RBMS Web Team is aware of, and has approved, this change.

Migrate to a Linked-Data-friendly Content Management System
In order to publish the RBMS *Controlled Vocabularies* as Linked Data, it is prudent to use a content management system (CMS) designed for Linked Data formats and links to external resources. This will allow us to create rich data during the research and editorial stages. The working group has identified two platforms for managing the *Controlled Vocabularies* as Linked Open Data:

TemaTres is a free, open-source content management system for knowledge organization systems (KOS) – such as library thesauri, taxonomies, ontologies, glossaries, and controlled vocabulary lists. TemaTres runs on a Web-server, and requires only PHP, MySQL, HTML, and CSS. The RBMS Web Team is prepared to run TemaTres in our existing hosting solution. Thus, using TemaTres requires no additional cost. Additionally, it is simple to install and straightforward to use.

A major benefit of TemaTres is that back-end users can have varying privileges to add, edit, or suggest concepts. This facilitates the RBMS-CV workflow wherein the Editorial Group drafts concept documentation, and opens up these drafts for public comment. Currently, this workflow is facilitated by three tools (the MultiTes CMS, a pbworks wiki,[6] and the RBMS *Controlled Vocabularies* Community Discussion[7]). TemaTres would allow us to centrally manage this information and save time.

Vitro was developed to support the VIVO project for connecting researcher information. Vitro is a generalizable RDF instance editor and can be configured for a variety of purposes, including thesaurus production. Vitro runs on a Java servlet, and would require an additional server environment at approximately $10/month.

Both TemaTres and Vitro provide a back-end administration and editing interface, as well as a front-end user interface for searching and browsing the *Controlled Vocabularies*. Both are cloud-based, and run on common Web technologies. Both will output to Linked Data formats (SKOS,

[6] http://rbmsthesauri.pbworks.com/
[7] http://rbms.info/cv-comments/

JSON-LD, etc.) and both offer a SPARQL endpoint for querying the data. Both TemaTres and Vitro will aid the Editorial Team's workflows and publication of the RBMS-CV as Linked Data.

Create meaningful Linked Open Data during the editorial process
Using either TemaTres or Vitro, output to Linked Data formats is seamless. And both CMSs will generate stable and dereferenceable URIs for concepts. However, the Editorial Team needs to choose a base URI pattern. We recommend either using the subdomain http://vocabularies.rbms.info/, or else choosing a new domain, such as http://www.rbmscv.org.

When the RBMS-CV data is migrated, existing relationships between resources (broader term, narrower term, related term, see, and see also) will be used to inform links within the *Controlled Vocabularies*. However, to create more beneficial, five-star Linked Open Data, which connects the *Controlled Vocabularies* with other Web resources,[8] the Editorial Team should begin capturing links to external vocabularies. As part of the research process for each new RBMS-CV concept, the Editorial Team identifies related concepts in other vocabularies. We recommend capturing this data within the RBMS-CV Linked Data set. We also recommend reviewing existing concepts and adding links to external datasets -- in particular, the Library of Congress Linked Data Service and the Art & Architecture Thesaurus.

Provide multiple points of access to the *Controlled Vocabularies* as Linked Open Data
Access to the RBMS *Controlled Vocabularies* as Linked Open Data may be granted through several means. Both TemaTres and Vitro offer a human-readable, searchable, and browsable front-end interface, data export options in Linked Data formats, and a SPARQL endpoint.

Additionally, we recommend periodic ingest of the *Controlled Vocabularies* into the Library of Congress Linked Data Service (ID.LOC). This will increase exposure for and use of the RBMS-CV. But doing so will require maintenance of relationships between the RBMS-CV and the Library of Congress authorities and vocabularies. Since external links are already recommended (to produce five-star Linked Open Data), this activity is not a burden, but a bonus. Nate Trail at the Library of Congress recommended this approach, and is aware of our intention to submit data dumps to ID.LOC.

NEXT STEPS

1. Choose domain or subdomain and base URI pattern
2. Install test instances of Vitro and TemaTres
3. Assess usability of both tools and implement preferred option
4. Review and enhance links in existing data, including links to external vocabularies
5. Simplify Editorial Group workflow around the new CMS and to create five-star Linked Data.

[8] For an explanation of the five-star system, see: http://www.w3.org/DesignIssues/LinkedData.html

CONCLUSION

Because both TemaTres and Vitro provide the necessary components for releasing the RBMS-CV as Linked Data, we recommend installing an instance of both on a newly established controlled vocabularies subdomain, which will necessitate an additional $10 per month hosting cost. The Controlled Vocabularies Editorial Group will test both software installations and implement the best option based on usability, ease of support, and overall functionality. Once an implementation decision has been made, we will migrate the latest data from MultiTes and dismantle the current rbms.info/vocabularies site. At this point, we will also begin work on developing the relationships necessary for five-star Linked Data.