VRA Bulletin

Volume 41 | Issue 2 Article 6

May 2015

Using the Getty Vocabularies as Linked Open Data in a Cataloging Tool for an Academic Teaching Collection: Case Study at the University of Denver

Heather Seneff
University of Denver, hseneff@du.edu

Shea-Tinn Yeh
University of Denver, sheila.yeh@du.edu

Fernando Reyes
University of Denver, fernando.reyes@du.edu

Using the Getty Vocabularies as Linked Open Data in a Cataloging Tool for an Academic Teaching Collection: Case Study at the University of Denver

Abstract
This case study examines the collaboration of two units at the University of Denver to create a new cataloging tool for the university's teaching and learning object management system. The Visual Media Director for the School of Art and Art History, the University Library's Digital Infrastructure and Technology Coordinator, and the Library's Senior Systems Analyst successfully developed the Art History Metadata Management System (MMS) in 2013. The collaborators were able to harness the power of Linked Open Data (LOD) from vocabularies from the Getty Research Institute and the Library of Congress to facilitate the creation of metadata in MMS. This case study examines LOD in the context of cataloging cultural objects using integrated controlled vocabularies to ensure metadata integrity. This study also demonstrates principles of agile software development that encourage frequent communication contributing to the success of a multi-departmental project.

Keywords
metadata, cataloging, collaboration, vocabularies, Linked Open Data, Semantic Web, agile software development

Author Bio & Acknowledgements
Heather Seneff is the Director of the Visual Media Center in the School of Art and Art History at the University of Denver. Shea-Tinn Yeh is the Library Digital Infrastructure and Technology Coordinator in the University Libraries at the University of Denver. Fernando Reyes is a Senior Systems Analyst in the University Libraries at the University of Denver.

Introduction

Linked Open Data (LOD) is an exciting component of the Semantic Web that has special impact in the realm of art and cultural objects. Using Linked Open Data, the context of objects and concepts is expanded and refined; the use of structured, defined data in a systematic manner improves searches and services on the Internet and exposes cultural objects in a new way. The Getty Research Institute has made the *Art and Architecture Thesaurus* (*AAT*) available as LOD, with plans to do the same for its *Union List of Artist Names* (*ULAN*) in 2015. These vocabularies are traditional mainstays of cataloging cultural objects for museums, archives, and academic teaching collections. What are the practical applications of LOD in the cataloging environment?

In December 2013, the University of Denver activated a second generation Art and Art History image cataloging tool developed by the University Libraries' Information Technology Department in collaboration with the School of Art and Art History Visual Media Center (VMC). This Metadata Management System (MMS) is a JavaScript, PHP, and MySQL based application utilizing a Fedora digital repository and Solr technology for asset management and full-text searching. The metadata format of choice is qualified Dublin Core. One significant feature that distinguishes this application from the previous cataloging tool is the integration of the Getty's *AAT* and *ULAN* in their pre-LOD form, as well as the Library of Congress Subject Heading (LCSH) and Name (LCN) authorities, for rigorous authority control. When the Getty vocabularies became available as LOD, MMS was updated to utilize this exciting Semantic Web component into the new cataloging tool.

This paper will discuss key success factors in incorporating LOD authority vocabularies into a complex yet sustainable tool in a cataloging environment. The discussion will be from the cataloger's and the programmer's point of view. Mostly importantly, this paper will highlight the implications of having such improved authority control for cataloging and searching a teaching collection.

The Semantic Web

The Semantic Web is the newest evolution of the Internet. In the Semantic Web, natural language is structured in such a way that it is machine-readable and logical connections can be made between "things." Sir Tim Berners-Lee (who is considered the inventor of the Internet and is now Director of the World Wide Web Consortium) envisions the developing Semantic Web as a "web of linked data" rather than a "web of hypertext" as it is now.[1] Linked data can, as Berners-Lee put it in his 2009 TED Talk *The Next Web of Open, Linked Data*, unlock enormous potential on the Internet. The content is there on the web, but the data about the content needs to be explored, added to, and structured to improve discovery and accessibility.

The Semantic Web relies on Uniform Resource Identifiers (URIs), Resource Description Framework (RDF), and ontologies and vocabularies.[2] URIs are unique strings of characters

[1] Ziyoung Park and Heejung Kim, "Organizing and Sharing Information Using Linked Data," *Library and Information Science, Volume 7 : New Directions in Information Organization* (2013): 62.
[2] Sharon Q. Yang and Yan Yi Lee, "Organizing Bibliographic Data with RDA: How Far Have We Stridden Toward the Semantic Web?" *Library and Information Science, Volume 7 : New Directions in Information Organization* (2013): 6.

identifying a resource (entity, "thing"). RDF triples (subject, predicate, object) are the syntax that describes the "thing." Ontologies are created by "subject domains" (specific disciplines) according to World Wide Web Consortium (W3C) standards in RDF schema language or Web Ontology Language (OWL) and are shared on the Semantic Web.[3] Ontologies define the properties and relationships of the "thing" and how it is described. Ontologies and collections that are freely accessible and shared without licensing on the Internet are called LOD; these are key to exposing content on the web. The W3C standard query language for the Semantic Web is known as SPARQL Protocol and RDF Query Language, which exposes RDF information in graph format. These "pillars" of the Semantic Web--URIs, RDF, and standardized ontologies--form the nexus of linked data.

The Semantic Web promises more meaningful searches and greater accessibility to resources for educators and researchers. In fact, the W3C specifically identifies "researchers, students, and patrons" and "librarians, archivists, and curators"[4] as two of four groups who will benefit most from the Semantic Web and linked data. The 2014 Library Edition of the *Horizon Report* identifies academic and research libraries as unique beneficiaries of Semantic Web technologies, leading to better bibliographic tools and more open access to "siloed" collections.[5] Examples of projects using LOD to the advantage of researchers, librarians, and curators include Europeana (http://www.europeana.eu/portal/) and the Digital Public Library of America (DPLA) (http://dp.la/).

The Europeana Foundation's goal is to connect and "publish data from Europe's cultural heritage and scientific communities."[6] 1500 institutions have contributed over two hundred million records to the Europeana Digital Library, and all metadata shared with Europeana is shared under a Creative Commons CCO 1.0 agreement.[7] The Europeana Foundation exposes its content as LOD through its professional portal (http://pro.europeana.eu/) and encourages reuse. The DPLA is a similar project and uses the Europeana Data Model (EDM) to aggregate the resources of a growing number of partners. Their "metadata model is based on RDF and uses Dublin Core as a central descriptive metadata standard."[8] In addition to an RDF-based data model, the DPLA also uses JavaScript Object Notation for Linked Data (JSON-LD) serialization.[9]

A majority of contributing partners to both Europeana and the DPLA are cultural institutions—museums, archives, and libraries. Such institutions usually have a wealth of metadata and description about their holdings since cataloging cultural objects has been a long-term and major activity. These cultural institutions benefit by having a number of controlled vocabularies structured specifically for the types of objects they describe. As an example of these vocabularies, "the Getty Research Institute has done remarkable work in developing

[3] Yang and Lee, "Organizing Bibliographic Data with RDA," 9.
[4] Park and Kim, "Organizing and Sharing Information," 71.
[5] L. Johnson, S. Adams Becker, V. Estrada, and A. Freeman, *NMC Horizon Report: 2014 Library Edition* (Austin, Texas: The New Media Consortium, 2014), 44.
[6] Erik T. Mitchell, "Building Blocks of Linked Open Data in Libraries," *Library Technology Reports: Library Linked Data: Research and adoption* 49 (2013): 38.
[7] Ibid., 39.
[8] Ibid., 35.
[9] Ibid., 35.

thesauri and controlled vocabularies"[10] and in making their vocabularies accessible. The Getty began working to expose their vocabularies as LOD, and in February 2014, the *AAT* was the first vocabulary to be released as LOD.

The potential to exploit the *AAT* as LOD was particularly enticing to the University of Denver's VMC in the School of Art and Art History. A new image cataloging tool had been designed in collaboration with the University Library in 2013. Two Getty vocabularies (the *AAT* and *ULAN*) had been integrated into the cataloging tool in their pre-LOD forms; the integration of the LOD version of the *AAT* was an exciting update.

Using images in the classroom at the University of Denver

In 2003, the University of Denver had developed an in-house, web-based, password-protected media management system to support teaching. Designed by the Office of Teaching and Learning, CourseMedia is a media learning management system designed for image, video, and audio files. Adobe ColdFusion, Flash, and MySQL database are the primary software technologies. An Adobe AIR component of CourseMedia allows for projection of media in the classroom, replicating the typical art historical pedagogy of dual projection.

Records for art and cultural objects are added to CourseMedia by the VMC in the School of Art and Art History (video and audio files are added by the library) using cataloging software that provides the content for CourseMedia. Four graduate students and one half-time staff member catalog under the Director, who trains the students and edits each record before it is ingested into CourseMedia. The Office of Teaching and Learning developed the first generation, Flash-based cataloging tool in 2003; it relied on authority files compiled by the catalogers in lists within the tool. These lists were (for the most part) based on traditional vocabularies (Getty and the Library of Congress), but over the years became corrupted by erroneous entries.

When the opportunity to develop updated cataloging software arose in 2013, integrating controlled vocabulary authorities was an important consideration. There were other requirements as well. The software would have to accommodate fields from the first generation cataloging tool and would have to export metadata into CourseMedia. The VMC Director needed a system for reviewing records created by the catalogers, and for displaying metrics and statistics. The cataloging system should be modular and scalable to other disciplines, and should be user-friendly, since new graduate students in the VMC are trained every year. The VMC Director requested a qualified Dublin Core metadata schema and mapped the fields from the old cataloging tool and from CourseMedia. To ensure rigorous authority control, she recommended integrating the *ULAN* and LCN authorities for the "creator" and "repository" (DC: coverage: spatial) fields, and the *AAT* and LCSH for the "style/period" (DC: coverage: temporal) and "subjects" fields.

Collaboration

[10] Karen Koogan Breitman, Marco Antonio Casanova, and Walter Truszkowski, *Semantic Web Concepts, Technologies and Applications* (New York; London: Springer, 2007): 241.

At the time of the VMC's request for a second-generation metadata management system in December 2012, the focus of the Office of Teaching and Learning had shifted to support CourseMedia's interface rather than the backbone infrastructure. It was therefore reasonable that the University of Denver's Library, with its metadata understanding and expertise, should be responsible for developing the new metadata tool. The Director of the VMC in the School of Art and Art History, the Library's Digital Infrastructure and Technology Coordinator, and the Library's Senior Systems Analyst met to collect system requirements. The group focused especially on two objectives: (1) defining the boundaries of the system and acknowledging that not all desired features could be implemented due to either software or resource limitations (for example, the ability to update local subject vocabularies inherited from the previous cataloging system), and (2) prioritizing features to provide a basis for future development iterations (for example, adding a batch update feature). With this type of consensus, agile software development approach[11] was chosen for this project.

Using this approach, the project development would welcome reasonable changes in requirements and would pay continuous attention to modularized design for simplicity, sustainability, and reusability. The Library's development team also maintained steady communications with the VMC Director, including many face-to-face meetings, and utilized frequent deliverables as a measurement for feedback and progress, all important components of agile software development. As a result, communication and negotiation replaced the misunderstanding and frustration that are often experienced in a systems development project between departmental units.

Description of MMS

MMS was developed using JavaScript, PHP and MySQL. These technologies were selected because of their popularity among developers and the fact that they are part of the current curriculum in computer science education. The latter helps the sustainability of the project since MMS is not dependent on the specific skill set of one individual or a particular team but is based on technologies that are commonly understood by developers and programmers. In addition, the system leverages Solr search server and Flexible and Extensible Digital Object and Repository Architecture (Fedora). Solr is an open source search platform produced by the Apache Software Foundation. It is used in indexing metadata and providing a fast discovery layer. Fedora offers an accessible application programming interface (API) to manage XML metadata, in this case the qualified Dublin Core, in a flexible way. With Fedora digital repository, the modification of XML schemas can happen at the front end without affecting the stored data at the backend. Other popular systems such as Islandora and Hydra digital repositories are all Fedora based.

The VMC Director requested *ULAN*, *AAT*, and LCSH and LCN for authority control. The Library of Congress has long provided its subject headings and names to the community for authority control via API. However, the VMC Director had to secure funding of $3000 to purchase both *ULAN* and *AAT* vocabularies from Getty, so that they could be imported into a local data store for use. Soon after the data store implementation, Getty began to open up its

[11] "Manifesto for Agile Software Development," accessed January 23, 2015, http://www.agilemanifesto.org/principles.html.

vocabularies via API. A local JavaScript (Node.js) based Web service was therefore utilized to search both the Library of Congress's and Getty's Web services for real time controlled vocabularies. When using a data store, the entire updated data has to be reloaded, but with Web services, the vocabulary is always up to date and there is no need for reloading data. This is beneficial in terms of application maintenance and vocabulary currency. The only risk is the mandatory system downtime of Web services at either the Library of Congress or the Getty.

Integration of the LOD *AAT*

The Library eagerly took advantage of the opportunity of accessing the LOD *AAT* when it was made available by Getty. The local JavaScript (Node.js) Web service was modified to search the *AAT*, LCSH, and LCN using SPARQL query language in a uniform manner. Getty facilitates the use of its LOD very professionally with a forum for users to ask technical questions of all kinds (http://answers.semanticweb.com/questions/ask/?tags=Getty,AAT). The programmers at the Library initially encountered ambiguous headers from the dataset returned from the Web services, for example, but were able to quickly gain insights from reading posts from the forum. The forum is also helpful when LOD changes the way their Web services return data, helping the programmers accommodate the changes to service. A future goal for MMS is to integrate the LOD *ULAN* service as soon as it is made available by Getty.

The integrated vocabularies appear in a pull-down list on the left side of the cataloging window in MMS. Catalogers can chose a vocabulary from this list and enter a search term. The results appear in a window below the search field, and the cataloger can click on one of the results to populate the appropriate field in the record. When the cataloger searches the LOD resources, a URI appears below each search result allowing the cataloger to click and access the actual *AAT* and LCSH and LCN web resources for each term. This ensures accuracy and strict adherence to controlled vocabulary terms.

Conclusions

The collaboration between the library and the VMC was very productive and rewarding. In addition to achieving a more modern cataloging tool for the VMC, the collaboration resulted in an exploration of LOD for the future of metadata. Integrating the Getty and Library of Congress vocabularies into the cataloging tool will encourage the catalogers in the VMC to be more attentive to authoritative terms, resulting in better searches in the University of Denver's learning media management tool, CourseMedia. While the importance of controlled vocabularies in cataloging had been emphasized in the legacy cataloging tool, the potential for human error was enormous because the vocabularies were created and contained in a local data store. Access to real time integrated controlled vocabularies has improved metadata accuracy and increased productivity in the VMC since the time it takes to ensure the accuracy of a name or subject term is greatly reduced. LOD can be exploited by collections to improve discovery in the Semantic Web. It can also help collections of all kinds improve the metadata they are compiling by provided access to standardized controlled vocabularies.

Bibliography

Breitman, Karen Koogan, Marco Antonio Casanova, and Walter Truszkowski. *Semantic Web Concepts, Technologies and Applications*. New York; London: Springer, 2007.

Hendler, James, Jeanne Holm, Chris Musialek, and George Thomas. "US Government Linked Open Data: Semantic. Data. Gov." *IEEE Intelligent Systems* 27, no. 3 (2012): 0025–0031.

Hooland, Seth van. *Linked Data for Libraries: How to Clean, Link and Publish Your Metadata*. Chicago, IL: Neal-Schuman, 2014.

Johnson, L., S. Adams Becker, V. Estrada, and A. Freeman. *NMC Horizon Report: 2014 Library Edition*. Austin, Texas: The New Media Consortium, 2014.

Manifesto for Agile Software Development. "Manifesto for Agile Software Development." Accessed January 23, 2015. http://www.agilemanifesto.org/principles.html.

Mitchell, Erik T. "Building Blocks of Linked Open Data in Libraries." *Library Technology Reports: Library Linked Data: Research and adoption* 49, no. 5 (July 2013): 11–25.

Mitchell, Erik T. "Three Case Studies in Linked Open Data." *Library Technology Reports* 49, no. 5 (July 2013): 26–43.

Park, Jung-ran, and Lynne C. Howarth, eds. *Library and Information Science, Volume 7 : New Directions in Information Organization*. Bradford, GBR: Emerald Insight, 2013.

Park, Ziyoung, and Heejung Kim. "Organizing and Sharing Information Using Linked Data." In *Library and Information Science, Volume 7 : New Directions in Information Organization*. Bradford, GBR: Emerald Insight, 2013.

Yang, Sharon Q., and Yan Yi Lee. "Organizing Bibliographic Data with RDA: How Far Have We Stridden Toward the Semantic Web?" In *Library and Information Science, Volume 7 : New Directions in Information Organization*. Bradford, GBR: Emerald Insight, 2013.