Saturday, June 22, 2019

Speakers: Erin Dobias, Metadata Librarian, Google Books; Jackie Shieh, Descriptive Data Management Librarian, Smithsonian Libraries; Robert Chavez, Senior Content Applications Architect, New England Journal of Medicine, and instructor of RDF and Linked Data courses for Library Juice Academy.

The panelists (with some contributions from audience members) answered questions about their experiences interacting with and teaching about linked data.

Question: How do you interact with linked data? Erin works with incoming metadata that is mostly MARC (but also some ONYX), with URIs showing up in some MARC. Jackie works with MARC for the most part, and some non-MARC, looking for URIs and writing scripts for SPARQL queries to transform the data to RDF and get it ready for BIBFRAME. Robert works with METS, MODS, and Dublin Core, advising groups on the best way forward with RDF data from their metadata.

Question: What has been your experience teaching others about linked data? Answers emphasized the importance of knowing one’s audience, speaking their language, and understanding their needs. In Erin's experience, engineers are concerned about quality and want to know how they can trust incoming metadata that links out to Wikipedia and other sources, whereas administrators/managers and other librarians are an easier sell. Although library metadata is highly trusted, engineers want to know if URIs are generated by human or machine and make sure they are correct. In teaching for Library Juice Academy, Robert tries to understand everyone's use cases and scenarios for linked data, what services they are trying to build, and what systems they already have in place. People taking the course come from different angles with respect to linked data; they may need to know how to create it, store it, manage it, or code around it. Jackie finds that libraries, museums, and archives all have a different take on what linked data is. Also, she references MARC when speaking to catalogers and W3C when speaking to programmers (as examples of building trust in working with linked data, given in response to a later question). Erin noted that small incremental changes like interpreting URIs are useful; we don't have to flip a switch and make a total conversion from MARC to RDF. Robert noted that we can leverage the metadata that we have and apply it in a linked data context to make it useful; we don't have to throw it out and start from scratch.

Question: What can you say about the use of various standards and ontologies? Jackie said that BIBFRAME started out as a fairly simplistic substitute for MARC but has been transformed to accommodate more in-depth research data. Communities are testing BIBFRAME for different
types of materials and extending BIBFRAME to meet their needs. Erin said that MARC is very robust and gets simplified a bit in order to work well with Google's knowledge graph (which is a giant triplestore). Google Books uses a model that resembles FRBR's WEMI but adjusts rules as needed; for example, an abridgment is categorized as a new work. She looks forward to robust BIBFRAME data. Robert also spoke of adapting ontologies in order to suit different needs (of different content providers and aggregators in a publishing context). Paul Frank, Library of Congress, added that BIBFRAME will remain agnostic in terms of cataloging standards, but the Library of Congress is using it with RDA in its BIBFRAME pilot. He thinks that BIBFRAME will never be as robust as MARC.

Question: Who should be writing SPARQL? This was asked by a cataloger. Responses indicated that whether or not you need to learn SPARQL depends on where you are in the landscape — cataloging, coding, building user interfaces, etc. If you are building a linked data interface such as Linked Jazz, SPARQL may be something you need to learn (Robert), but catalogers do not necessarily need to become coders in order to use a linked data cataloging tool like LD4P2's Sinopia (Nancy Lorimer, Stanford).

Question: When administrators ask why linked data, what is an appropriate response? Xiaoli Li, University of California at Davis, tells administrators that users coming from outside the library's public catalog need to discover the library's resources, and linked data is one way to facilitate this efficiently. Jackie noted that how we communicate about linked data is important in ensuring administrative support of linked data efforts.

Question: As data becomes more machine-readable and less human-readable, how do we address the need for labels and how do we deal with outages? Jackie said that we can get URIs into our system and have developers set it up so that if one source is down, such as LCNAF, it goes to VIAF instead. Erin said that multiple URIs within a record helps the data be more robust. The more times that two particular URIs (such as from LCNAF and VIAF) co-occur, the more trustworthy this association becomes.

Question: What is the "killer app" for linked data? Robert answered that Google search results are more and more influenced by linked data and semantic data, but he thinks that rather than a single "killer app," we will see services pieced together in interesting ways.

Question: How do we balance trustworthiness with automation? Jackie answered that quality control is very important, but with limited staff we must do some things in bulk. A thousand Smithsonian records were enhanced with URIs for headings in all 1XX, 7XX, and 6XX fields, and 300 out of 4000 of the URIs added were incorrect and needed to be corrected.

Catalog Management Interest Group (abstracts and slides available at this link)

Saturday, June 22, 2019

The three presentations in this session were about institutions' experiences using tools such as MarcEdit, OpenRefine, Excel, and Python for batch processing and metadata reconciliation
projects (improving metadata for electronic resources, de-duping vendor records for e-books, etc.). In presentation #3, Brian Rennick spoke about querying Discogs (a database of crowdsourced metadata for audio recordings) for genre terms for audio recordings held at Brigham Young University. A recent exhibit at BYU promoted a large collection of LPs of popular music in various styles that had been cataloged using rather general subject and genre terms ("popular music" being the most common). More specific terms (such as soul music, glam rock, etc.) were desired for the interactive display in the exhibit. Brian designed a batch process using a combination of Python scripting to transform Discogs data and OpenRefine to identify problems and create clusters, enabling him to query the Discogs data using the catalog numbers from the MARC records (028 $a) to find matching records in Discogs and retrieve the genre terms found there. Brian designated additional match points to ensure the best match (title, artist, country, etc.). Although the genre terms retrieved from Discogs were only used for the exhibit's interactive display and not added to the MARC records in the catalog, the audience was made aware of Michigan State University's project to create mappings between Discogs genre terms and LCGFT.

**MarcEdit: Past, Present, and Future**

Sunday, June 23, 2019

Terry Reese, The Ohio State University, discussed the 20-year history of MarcEdit, the widely used software he designed for creating and editing library metadata. Terry invented MarcEdit in 1999 to create MARC records for maps to load into OCLC Passport, and he has developed new capabilities (and hidden games such as Asteroids!) over the years, informed by input from the international community of users (which number about twenty thousand, judging by the number of downloads of the most recent version!). Terry retains ownership of MarcEdit, and there is a succession plan in place.

Mike Monaco, University of Akron, talked about some batch projects that made use of MarcEdit, including updating headings in bibliographic records to conform with the LCNAF, making routine edits to bibliographic records for ETDs and other e-resources such as Early English Books Online, and adding subject headings to bibliographic records for works of fiction.

Bryan Baldus, OCLC, talked about using MarcEdit at OCLC to examine and evaluate a batch of records (for example, checking for fields that may be absent, such as 007), transform vendor records (from MODS or MARCXML to MARC21), and perform cleanup in WorldCat (for example, to help with diacritics issues not addressed by the Connexion Client QC macro). In the future, a Connexion macro converter may be developed to allow Connexion macros to be used directly in MarcEdit.

Terry mentioned ideas for future development of MarcEdit, including built-in XML editor, individual record editor, expanded XML/JSON Wizard, and updated tutorials.

**PCC at Large**
Judith Cannan, Library of Congress, announced that although the PCC Directory requires frequent password changes, it is possible that they will become less frequent in the future. She also announced that there had been a merger of the Library of Congress divisions formerly known as Policy and Standards Division (PSD) and Cooperative and Instructional Programs (COIN). The new merged division is called Policy, Training, and Cooperative Programs Division (PTCP), and Judith is the chief.

Janis Young, Library of Congress, gave some SACO-related updates. She presented the same content at Authority Control Interest Group; see the slides and report from that session, as well as the report from Subject Analysis Committee. Janis also gave a demonstration of the new interface for ClassWeb that will be rolled out in the summer or fall. It will be easier to search and browse various vocabularies and LCC schedules and tables in the new interface.

Paul Frank, Library of Congress, gave some NACO-related updates. A Wikidata workshop was held in May at the PCC Operations Committee meeting (agenda and recording available at this link). The PCC Task Group on Identity Management is taking feedback and making plans for a Wikidata pilot. A pilot using various identifiers in $0 and $1 is also being planned. The day is coming where it will be common for headings in bibliographic records to link out to vocabularies other than LCNAF.

Program for Cooperative Cataloging (PCC) Participants Meeting

Sunday, June 23, 2019

PCC celebrated its 25th anniversary (1994-2019) with brief talks, plentiful cupcakes, and a competitive, team-based PCC trivia game. After opening remarks from PCC Chair Xiaoli Li, the membership was addressed by past chairs Lori Robare and Chris Cronin, past chair of PCC Standing Committee on Standards Becky Culbertson, and PCC Policy Committee members Cynthia Whitacre (OCLC) and Beacher Wiggins (Library of Congress). Lori acknowledged the 20th anniversary of Terry Reese’s MarcEdit and its automation of routine aspects of cataloging. Terry has been involved with the PCC Task Group on URIs in MARC, and MarcEdit now enables the population of MARC records with URIs. Terry is developing tools to help with changes to punctuation practice. Cynthia spoke about the relationship between PCC and OCLC. Before the PCC was established, CONSER and NACO were established in the 1970s. (BIBCO and the PCC were established in the 1990s, and SACO became the fourth PCC program in the early 2000s.) The CONSER database resided in OCLC’s Online Union Catalog (now WorldCat) from the start, and OCLC has been a longtime NACO participant. OCLC is a NACO node, allowing members to use a copy of the LCNAF via OCLC, and OCLC staff are independent NACO contributors for names and series. Cynthia spoke of PCC’s value to OCLC (by virtue of its emphasis on quality and its standards and guidelines on hybrid records etc.) and OCLC’s investment in PCC (through financial contributions and time and effort of OCLC staff on PCC task groups and committees). Beacher spoke about the relationship between PCC and Library of Congress (specifically in its supportive role as PCC Secretariat), expressing gratitude for the leadership of PCC chairs in
efforts such as the testing and adoption of RDA. Becky reviewed PCC’s past, highlighting some major contributions such as the BIBCO Standard Record and the CONSER Open Access Journal Project, and the work of PCC standing committees. She also reminded the membership that PCC, OCLC, and RLIN made it possible for libraries to share their cataloging, thereby reducing their backlogs, without waiting for or relying on Library of Congress. Chris spoke about the future of PCC after first reminiscing about the information landscape of 1994 when PCC was established, noting the democratization of information and PCC’s place in this process. In 2014, at age 20, PCC engaged in strategic planning, wanting to leverage our expertise and assert our influence. Looking to the future, Chris would like to see advocacy for technology and systems that serve our aspirations for our data, increased cooperation such as through sharing language expertise, and simpler rules for staff to navigate and apply.