ALA Annual 2018
New Orleans, LA
Report by Tracey Snyder, Chair, MLA Cataloging and Metadata Committee

Linked Library Data Interest Group

See abstracts and slides at:
https://connect.ala.org/communities/community-home/digestviewer/viewthread?MessageKey=047fa095-508d-4bb1-b65d-16720d0535c4&CommunityKey=e329ee56-0b9b-4189-ba75-66d4715562ba&tab=digestviewer#bm047fa095-508d-4bb1-b65d-16720d0535c4#bm0

Evidence-Building using Linked Administrative Data at the Census Bureau
Kate McNamara (Census Bureau)

Kate McNamara gave a presentation about the Census Bureau's Data Linkage Infrastructure, whereby the Bureau acquires, ingests, curates, links, analyzes, and archives data. The Bureau is expanding this infrastructure to improve access to this data for program evaluators, policy analysts, and researchers. In addition to issuing surveys (such as the population census every ten years and the economic census every five years), the Bureau collects administrative data related to programs such as SNAP, Medicare, and Medicaid to be used as evidence in evaluating program effectiveness and improving policy. The Bureau collects data and performs research as allowed by Title 13 in order to carry out its mission to measure the nation's people and economy, while protecting confidential individual data and maintaining a secure environment for analysis of linked datasets. Person records are matched across datasets based on name, date of birth, and Social Security number, using a unique identifier called a PIK (Protected Identification Key) assigned by the PVS (Person Identification Validation System). Kate named some specific pilot projects being conducted by Chapin Hall at the University of Chicago, including:

- Using linked data to examine the trajectories and service utilization of families and children experiencing homelessness
- Health at birth and later life outcomes: Evaluating the returns to policy-driven early health investments
- Program Utilization by Formerly Criminalized Youth: Linking Juvenile Recidivism Data to Data Held by US Census Bureau
- Adult labor market outcomes of Chicago Public School students: Pathways from K-12 to work

Kate asked attendees to disseminate information about the Data Linkage Infrastructure to faculty and colleagues and advised that researchers interested in using the infrastructure should contact the Census Bureau.

Who Will be Our bf: Comparing techniques for conversion from MARC to BIBFRAME

Sharon Farnel, Abigail Sparling, Ian Bigelow, Danoosh Davood (University of Alberta Libraries)

Speakers from the University of Alberta gave a presentation comparing two approaches to converting MARC to BIBFRAME: (1) tools developed by Casalini Libri and @CULT as part of SHARE-VDE (Virtual Discovery Environment) and (2) a process developed in-house based on LC's conversion tool and additional open-source resources. (See slides and speaker notes for technical details on both approaches.) Having a complex set of MARC data created according to various cataloging standards (RDA, AACR2, pre-AACR2, etc.) posed challenges in creating mappings for both conversion processes. For the most part, key data was mapped and converted accurately, although some encoding intricacies were lost in conversion. For example, square brackets around cataloger-supplied elements such as place of publication were lost in the LC-based conversion but retained in the Casalini-based conversion. Relationship designators were mapped appropriately and assigned a URI in both processes; however, in cases where no relationship designator existed in the MARC data, the LC-based conversion assigned "contributor" by default, whereas the Casalini-based conversion incorporated a process to more precisely detect an agent's role and assign a more specific relationship designator. Neither conversion made use of OCLC Work IDs (in MARC 758), but the Casalini process used MARC 130 and 240 and other work and instance data to generate new SHARE-specific work IDs. Similarly, the Casalini process generated SHARE-specific URIs for many elements rather than using URIs from existing vocabularies such as LC or VIAF, raising questions of accessibility and usability outside of SHARE-VDE. The presenters concluded that both conversion processes have strengths and weaknesses. While Casalini's SHARE-VDE offers vendor support, complex text analysis, and a prototype discovery platform, the LC-based process can provide a single openly available tool for conversion, reconciliation, and enrichment. Next steps at the University of Alberta include further review and assessment of converted data to identify areas for improvement in both processes, refinement and enhancement of in-house processes and development of new tools, ongoing staff development and training in RDF, BIBFRAME, and SPARQL, and ongoing and new partnerships and collaborations such as LD4.

Demonstrating the Production Value of Linked Data Services

John Chapman (OCLC)

Xiaoli Li (University of California Davis)

John and Xiaoli gave a presentation about a joint research project (conducted by OCLC and institutional partners including UC Davis) prototyping a new suite of linked data services using Wikibase tools and multiple linked data sources. The project included the development of an "Entity Ecosystem," where users could create and edit entities and make connections between entities. It also provided services to reconcile and explore data. Project participants learned to use Wikidata-based terminology in entering data about persons and other entities and were able to enter conflicting statements (with indications of degree of certainty) and to track revision histories. An "Explorer" interface allowed users to query sources to find related entities (such as works by a person), and a SPARQL endpoint allowed users to perform complex searches for entities (such as persons born in a specified date range). John mentioned a few challenges encountered with the prototype (and corresponding solutions). Xiaoli summarized her team's experience as participants in the project. She showed examples of the inputting screens for a person entity, a resource entity (like a bibliographic record linking to a person entity--the author), and a place entity (comparable to an authority record for a place name but with more elements possible, such as the Chinese name for an American city). Three catalogers and one data scientist at UC Davis used and evaluated the editing tool and the user interface. The catalogers were sometimes unsure about how much data to add or what predicate to use, but they felt good about creating linked data without needing to understand in detail how the URIs and triples worked "under the hood."

Implementing Linked Open Data in the Real World

At the time of writing this report, slides for this session were not yet posted online. See abstracts at: https://www.eventscribe.com/2018/ALA-Annual/fsPopup.asp?Mode=presInfo&PresentationID=352572

Linked Data for Production: Pathway to Implementation (LD4P Phase 2)

Philip Schreur (Stanford University) gave a presentation about the new grant project funded by the Andrew W. Mellon Foundation, Linked Data for Production: Pathway to Implementation (LD4P Phase 2), beginning summer 2018 and following on the work of LD4P Phase 1 (2016-2018). The focus areas of LD4P 1 were:

- Explore working in a shared system (several institutions)
- Explore BIBFRAME and write extensions in specialized areas
- Develop linked data inputting tools
- Develop workflows to demonstrate that working in a linked data environment is possible

- Shift to using identifiers in authority work

LD4P 2 will be a collaboration among four institutions (Cornell, Harvard, Stanford, and the University of Iowa) and the Program for Cooperative Cataloging (PCC). Its focus areas will be (from the program's abstract):

"the creation of a continuously fed pool of linked data expressed in BIBFRAME; development of an expanded cohort of libraries capable of the creation and reuse of linked data through a cloud-based sandbox editing environment; the development of policies, techniques and workflows for the automated enhancement of MARC data with identifiers; the development of policies, techniques, and workflows for the creation and reuse of linked data as libraries' core metadata; better integration of library metadata and identifiers with the Web through collaboration with Wikidata; the enhancement of a widely-adopted library discovery environment (Blacklight) with linked-data based discovery techniques; and the orchestration of continued community collaboration through the development of an organizational framework called LD4."

Philip emphasized that in order to make the transition to a linked data environment, the implementation scenario needs to be practical. We have a complex ecosystem, and we will need to question some assumptions and dump some baggage. Philip said that in the past, he endeavored to motivate others to embrace this change through compelling reasoned arguments, and then later by appealing to their sense of self-interest, and that he now tries to inspire others with a sense of hope for the benefits that linked data will bring.

Rare Materials Ontology Extension: from Modeling to Implementation

Jason Kovari (Cornell University) and Francis Lapka (Yale University) provided a presentation about one component of the recent (2016-2018) multi-institution grant project funded by the Andrew W. Mellon Foundation, Linked Data for Production (LD4P) Phase 1 — the creation of a BIBFRAME ontology extension for art and rare materials. Previously, Cornell and the RBMS Bibliographic Standards Committee (BSC) were collaborating on a rare materials ontology extension (RareMat) while Columbia and the ARLIS Cataloging Advisory Committee were collaborating on an art ontology extension (ArtFrame). These two projects were so similar that it was decided to merge them into a single ontology for art and rare materials (ARM). Areas modeled include awards, bindings, exhibitions, custodial history, and signature statements. Participants developed application profiles using SHACL and began testing the model and application profiles in VitroLib, an RDF-based cataloging tool developed at Cornell as part of a related grant project, Linked Data for Libraries Labs (LD4L-Labs). RBMS BSC has committed to maintain the ARM ontology and is looking to identify a long-term solution for hosting it. Other

next steps include further modeling and ontology development, further application profile development, and further testing of models.

Linked Data URI workflows at the Digital Virginias DPLA Hub

Jeremy Bartczak (University of Virginia) gave a presentation about adding URIs to metadata for collections represented in the Digital Public Library of America (DPLA). UVa. Library is a DPLA content hub and will become a regional administrator for the Digital Virginias hub (which will include partner institutions in Virginia and West Virginia). Members of this regional hub will increase the number of digital objects they contribute to DPLA and are eager to enhance their metadata for these objects with URIs. The DPLA Metadata Application Profile (MAP), which is based on the Europeana Data Model (EDM), and the DPLA Metadata Quality Guidelines encourage inclusion of URIs. UVa. enhanced their MODS metadata for their printing services collection and visual history collection with URIs. Jeremy discussed data modeling challenges for mapping URIs from MODS to Qualified Dublin Core XML.

PCC At Large

In introducing the session and its speakers, all from the Library of Congress, Judith Cannan announced that the new online PCC directory would be rolled out over the summer. Over 100 institutions already have access, and all member institutions will have access by September 30, 2018. The new PCC directory will be used for statistics and elections next year.

Manon Theroux gave a presentation explaining the difficulties involved in maintaining the naco@loc.gov email account and outlining a protocol for NACO members to use when sending messages to this address to request bibliographic file maintenance, deletion of duplicate name authority records, and deletion of undifferentiated name authority records with only one remaining identity. Although the staff members who monitor this account are grateful for the diligence and collegiality of NACO correspondents, they simply do not have time to send replies, as they are about three months behind in acting on requests due to heavy volume and limited staffing. Correspondents are asked to help speed up processing by observing the following:

- cite the LCCN (not the ARN)
- cite the *entire* LCCN
- cite the AAP
- cite the AAP *exactly*
- specify which authority record to delete
- update the preferred authority record as needed

- say whether or not a new authority record should be created for the last remaining identity on an undifferentiated name authority record being reported for deletion
- be concise
- do not send requests to additional addresses
- do not resend requests
- do not send messages if no action is being requested

Paul Frank gave an update on the question of recording gender in name authority records. In a recent survey, the PCC membership indicated support for the best practices recommendations of the PCC Ad Hoc Task Group on Gender in Name Authority Records. (See https://www.loc.gov/aba/pcc/documents/Gender_375%20field_RecommendationReport.pdf.) This summer, the task group will look more closely at the comments provided by respondents who did not agree with the recommendations. The membership can expect guidance to appear in DCM Z1 in the fall. Paul also advised that the draft guidelines for applying relationship designators in authority records were nearing final approval in revised form PCC.

In SACO news, Paul Frank announced that LC would discontinue a program whereby members would use an online form to propose a literary author number for an author not represented in LC's collections and LC would approve it and add it to the 053 of the author's authority record. The program is being discontinued due to relatively little use and unpleasant complications caused by thousands of suppressed "holding" bibliographic records in LC's catalog. Instead, members should code 053 as local. Multiple local 053s are acceptable.

Janis Young rounded out the SACO portion of the program with a presentation on matters related to LCSH and the other vocabularies maintained by LC. After a SACO member submits a proposal for a new or revised vocabulary term, the Policy and Standards Division's assistant editors schedule the proposal on a monthly tentative list and publish the list for comment. PSD would appreciate more public comments. The PSD policy specialists review the proposed terms, accompanying notes, cited research, and comments received from the community, and meet monthly to discuss proposals and assign each one a status (approved, not approved, not necessary, may be resubmitted, or withdrawn). These meetings, held the Friday before the third Monday of the month, are open to the public, so contact Janis if you would like to attend. After the meeting, the policy specialists write and distribute a summary of decisions and give revisions to the assistant editors. Approved lists are then published.

Janis also discussed LC's decision to cancel "multiple" subdivisions in LCSH. The LC update for ALA (see https://www.loc.gov/librarians/american-library-association/annual/lc-update/) describes the situation:

"Multiple" Subdivisions

The better to support linked-data initiatives, the Library's Policy and Standards Division in the ABA Directorate will cancel "multiple" subdivisions from *Library of Congress Subject Headings* (LCSH) beginning in fall 2018. "Multiple" subdivisions are a special type of subdivision that automatically gives free-floating status to analogous subdivisions used under the same heading. In the example **Computers—Religious aspects—Buddhism, [Christianity, etc.]**, the multiple subdivision is **—Buddhism, [Christianity, etc.]**.

Over 2,200 multiple subdivisions are established in LCSH, and they can be identified by the presence of square brackets. They generally appear in LCSH itself, as in the heading **Computers—Religious aspects—Buddhism, [Christianity, etc.]**, but some appear in lists of free-floating and pattern subdivisions. The multiples permit catalogers to "fill in the blank" and substitute any word, phrase, or other information that fits the instruction. For example, catalogers can create **Computers—Religious aspects—Hinduism** because Hinduism is a religion, just as Buddhism and Christianity are.

Staff in PSD will create authority records for each valid heading string that was created based on a multiple subdivision and delete the authority record for the multiple subdivision.

As of July 1, 2018, PSD will stop approving proposals for new multiple subdivisions. Instead, catalogers will propose the heading string that is needed. That is, instead of proposing **Paleography—Religious aspects—Buddhism, [Christianity, etc.]** for a resource about Muslim views on paleography, the cataloger would propose **Paleography—Religious aspects—Islam**. Proposals that were submitted before July 1, 2018, and that are already under editorial review will be revised to follow the new policy.

Catalogers should continue to use existing multiple subdivisions as usual until PSD creates individual subject authority records for each heading string that has been assigned. The multiple subdivision will then be cancelled and catalogers must propose each new use of a subdivision that was formerly authorized by a multiple.

Subject Headings Manual instruction sheet H 1090, Multiple Subdivisions, will be revised to reflect the new policy, as will other instruction sheets that refer to multiple subdivisions (e.g., H 1998, Religious Aspects of Topics). The lists of pattern and free-floating subdivisions (instruction sheets H 1095-H 1200) will be revised as the multiple subdivisions are removed from LCSH.

Additional details about the project will be announced later in summer 2018.

Finally, Janis reminded attendees that PSD placed a moratorium on proposals for new and revised terms for LCDGT in February. The moratorium is still in place while PSD re-examines LCDGT's structure, so SACO members should set aside ideas for proposals until a later date.

Program for Cooperative Cataloging (PCC) Participants

See select presentation slides at:
https://www.loc.gov/aba/pcc/documents/PCC-Participants-Meeting-2018-Presentations.pptx

PCC chair Lori Robare announced PCC's new Communication Board (see https://www.loc.gov/aba/pcc/taskgroup/PCC-Communication-Board.pdf), which will issue its first bulletin in August, and introduced the session and its speakers. The session's overall title was "Transitioning to Identity Management: Experiences with the PCC ISNI Pilot."

Amber Billey (Bard College) gave a presentation about shifting from authority control to identity management. She spoke about her background as a research assistant at the Field Museum and a metadata specialist for CollectiveAccess (open-source software for managing Museum and archival collections), contrasting those environments, in which a person's works could be pulled together easily and updated automatically, with the LC NAF environment and its limitations (e.g., authority control is complex and time-consuming, requires extensive training and expensive software, and forces a cataloger to choose a single preferred label that must be unique). Amber mentioned the PCC Task Group on Identity Management in NACO (see https://www.loc.gov/aba/pcc/documents/PoCo-2017/Report-TG-IdentityMgmtNACO.pdf), which discussed use cases for identity management and options for moving in that direction. She named three options: (1) developing a "NACO Lite" program, which would lower barriers to participation, but could water down the quality of our data; (2) participating in ISNI, which has its own benefits (for instance, YouTube has adopted ISNIs for contributors, which speaks to its prominence and acceptance beyond the library world) and drawbacks (such as a participation fee); and, (3) working with Wikidata, the data source for Wikipedia. Amber raised the possibility of a PCC Wikidata pilot project.

John Riemer (UCLA) provided an update on the work of the PCC Task Group on Identity Management in NACO (see https://www.loc.gov/aba/pcc/documents/PoCo-2017/Report-TG-IdentityMgmtNACO.pdf), outlining five questions that would arise in a move away from traditional authority control toward identity management, with its emphasis on compiling contextual information for disambiguation:

- What if text strings become secondary? (We will still have text strings, but perhaps not a requirement of uniqueness.)
- How full fledged do authority records need to be (initially)? (There will need to be at least enough data elements present to confirm which identity it is.)
- How to utilize identifiers from other registries? (Decisions will need to be made regarding the coding and display of identifiers from multiple registries.)
- What files do we have the option to work in? (Options include LC NAF, ISNI, and Wikidata.)
- Single search environment, for multiple files? (Perhaps LC NAF data will be visible from within ISNI, or ISNI data will be visible from within LC NAF.)

In addition to exploring these issues, the task group is working on defining what "NACO Lite" could be, as mentioned in the PCC Strategic Directions document. (See https://www.loc.gov/aba/pcc/about/PCC-Strategic-Directions-2018-2021.pdf.) John also wrote a *Technicalities* column this year exploring the "NACO Lite" concept (see https://escholarship.org/uc/item/3ss1t4xx), tentatively defining it as the focusing of authority work on identity management, differentiation of entities, and relationships between entities. The creation of identifiers would be emphasized over the construction of text strings.

Isabel Quintana (Harvard University) provided an overview of the 2017-2018 PCC ISNI pilot (see https://wiki.duraspace.org/display/PCCISNI/PCC+ISNI+Pilot+Home). Libraries need identifiers that are stable, persistent, discoverable, shareable/reusable on the Web, and rich enough for matching/disambiguation. ISNI is a good match in this context because it is an ISO standard, has a team of experts responsible for data integrity and error resolution, has OCLC as a technical partner, and has broad participation from domains in and outside of libraries. Many libraries volunteered for the pilot, and all were included. Harvard coordinated the pilot, and training materials included videos and a wiki. Pilot participants used ISNI tools, documented what worked well and what could be improved, identified policy-related questions, and more. Subgroups of participants explored areas such as documentation and training, best practices for ISNI maintenance, best practices for workflows and tools, workflows and templates for batch processing, and API review and testing. As the pilot wraps up, PCC will work with the ISNI Board to make use of pilot participants' feedback, paving the way for PCC membership in ISNI beyond the pilot.

Three pilot participants spoke briefly about their experiences with ISNI. John Hostage (Harvard University) showed an example in ISNI, using Michelle Obama, of duplicate records that had to be merged, and he displayed a screenshot of the template for creating a new record. Chris Long (University of Colorado Boulder) talked about differences between NACO work and ISNI work (for example, in ISNI work, not all information needs to be accompanied by a justification) and highlighted one important similarity — the importance of searching carefully to prevent creating duplicate records. Chris's institution plans to create ISNIs for faculty via a batch loading process without creating a corresponding NACO authority record for each person. Jeanette Norris (Brown University) oversaw a small-scale project to create and maintain ISNIs for several Brown departments. She reported that her participants' most significant achievement was developing skills, expertise, and comfort in a new area.