Summary of IMLS Shareable Local Authorities Forum, Library of Congress, April 10-11 2017

Report to the PCC on the Second IMLS Shareable Local Authorities Forum
Held April 10-11, 2017 and hosted by the Library of Congress
Submitted by Isabel del Carmen Quintana, PCC representative to the forum

The second IMLS Shareable Local Authorities Forum was a lively and thought-provoking exploration of how stakeholder communities can better share local authority information. It followed the first forum held last October in Ithaca, N.Y. and developed some of the key themes identified at that first meeting. The agenda featured a mixture of presentations and discussions, as well as follow-up actions that the group felt ready to proceed with. A full agenda is available here: https://confluence.cornell.edu/x/LKZ0F. More information about the project is available at this wiki: https://confluence.cornell.edu/display/sharedauth/IMLS+Shareable+Authorities+Forum+Home

Day 1 morning summary:

There was a demonstration of SNAC (Social Networks and Archival Context) by Worthy Martin and Daniel Pitti (University of Virginia)

- In this database they create “identity constellations” that connect individual to archival resources, in order to provide contextual information and holdings information for researchers.
- The data is ingested from archival records (EAC records), and the data stays connected to the archival sources. Entities can also link to other entities, such as Wikidata, VIAF, etc. For each assertion, the system tracks who made the assertion. They also have some geographic data linked to resources.
- Entities can be embargoed, in review, deleted, etc. In fact some of the problems have to do with maintaining the data, for example, merging records and creating permalinks.
- They would love to work with ArchiveSpace in such a way that if someone working in ArchiveSpace saw an entity in SNAC, that data could be sent directly to Archive Space.

The rest of the morning consisted of three lightning talks dealing with institutions and practices, and some facilitated discussion. The lightening talks were:

The Western Name Authority File: Building shared authorities for regional digital collections

(Anna Neatrour, University of Utah)

- 2-year IMLS pilot grant for cooperative authorities in Mountain West region
- This database consists of “local” entities from digital resources and special collections.
- The database is not available yet as they are still in a test phase.
- There were many names/entities that they collected, and had to dedeup from over 500,000 to 76,000. This effort was labor intensive and a lot of it was done manually. Therefore they are experimenting with ways to make this more automated.
- Need a way to onboard data, without as much manual work. Therefore they are currently investigating tools that could help with this process.
- The next phase will be to assess these tools, and hopefully be able to scale up their operations.

University of North Texas: Authorities: UNT Libraries Digital Collections Portal to Texas History (Mark Phillips, UNT)

- Aggregation of digital collections and some newspapers. There are millions of entities named in these collections.
- They use a UNT name app for local authors, and try to use ORCID or VIAF but there is very little overlap. So they are creating local authorities.
- They are internally storing links, not strings.
- They also had an open house for anyone to come and set up metadata records.
- They’ve controlled 7,000 names but they have over 350,000 in the system, and are not sure how to scale up. Therefore they want to collaborate with others on how to manage local authorities.

FOLIO project: FOLIO Kabalog overview (Peter Murray, IndexData)

- FOLIO is an open source library services platform; “Kabalog” is a term they have coined to describe a catalog mixed with a knowledge base
- It is metadata schema agnostic and supports multiple bibliographic utilities and knowledge bases.
    - It employs core metadata ”just enough for the other modules to understand” and native metadata “for apps that understand it”
        - Different apps will have different functional needs, therefore use different portions of metadata (e.g. circulation)
- Cataloging-by-reference--moving away from copy cataloging
    - Support multiple bib utilities and knowledge base
    - Keep a link to the source and list of local changes
    - Automated and semi-automated processes for updating local records with changes from source
- Not full embracing linked data but “taking on some of the promise linked data holds” and could move into it
- It could also support a local name authority service (although it does not yet). They service could create, update and delete names, and could be taken up by a special interest group. RESTful in nature; can leverage other protocols

There was then a discussion, facilitated by Jason Kovari (Cornell University), of some of the problems with “local” names, such as: key technical issues, whether we want to share them or not, and how we can leverage the work of others/work more collaboratively with others, especially if we could have automated means to do so.

Day 1: the afternoon began with three lightning talks on international services and providers:

ORCID (Simeon Warner, Cornell University)

- Why does ORCID have to be different from what we did before? About 8-20 million people who would be candidates for ORCID

- Most people have short active research lifespan: 3-6 years on average. (Many reasons for this.) Should be more people in this system.
- About 2 million journal articles published a year.
- Use researcher engagement to make sure identities are included in workflow rather than traditional curated process.
- Overlaps with VIAF/ISNI but tries not to duplicate
- Incorporate ORCID in submission workflow - “not disambiguation but ambiguity avoidance”.
    - Need to give researcher a degree of control we’re not used to.

PCC ISNI Pilot: an initiative of the PCC (Michelle Durocher, Harvard University)

- Proposal to have PCC have umbrella membership in ISNI endorsed at November PCC meeting
- Launch a pilot to see how this might work, to inform formal membership agreement beyond the pilot
- Why an ISNI Pilot for the PCC?
    - ISNI is a good match because: ISO standard; broad and growing participation; expert metadata practitioners
    - Was in the vision statement of PCC
    - Shift from authority control (away from strings) towards managing identities; AND broaden community participation in PCC
    - Complementary efforts: URIs in MARC Task Group; Identity Management in NACO Task Group
    - ISNI pilot participants: Brown, Columbia, Cornell, Harvard, Stanford, UCLC, Northwestern Medical, and open call got 3-5 more
    - Library use cases for identifiers: faculty, departments, institutional repository, archives, theses/dissertations, CIP workflow collaboration with publishers
    - Benefits to ISNI: creates additional pipeline, quality data from multiple institutions, more simplified channel for libraries’ batch projects (less infrastructure work)
    - Next steps: Get pilot libraries engaged with ISNI tools à identify needs, wishlist of enhancements; also exploit and understand the APIs use; leverage PCC collaborative network and training to be able to scale up involvement.

LC/NACO Authority File: Identity in Linked Data (Paul Frank, Library of Congress)

- NACO has been around for a long time, and has always been a clean file
- If we had NACO Lite: How much tolerance does a system have for duplication or conflation? How much tolerance does a cataloger have for duplication or conflation? Can’t ignore the standards. Also there is the training component.
- How resilient in the NAF?
    - Centralized authority file; MARC-based distribution
    - What about id.loc.gov?
        - BIBFRAME experimentation; can accept RDF or MARC XML; but what about the business agreements (who is responsible for it?); is this the decentralized authority file (that we need)?
        - In the future, can we work directly in id.loc.gov, or load into it so we can decentralize the data
        - Who owns the LC/NACO Authority File? (LC used to, but now more records get created by PCC)
        - It would be great if we could filter on searching (for example, could look at both NACO records and NACO lite records)
        - Role of the Authorities Librarian
            - Disambiguation vs. Crafting (= making this great record)
            - Flexibility
            - Tolerance (for dealing with ambiguity)
            - Identity Management SNAC-Style
                - Possible matching constellations; put them up together, and move data that they share into middle column so can either merge or disambiguate

These sessions were followed up with some discussion, facilitated by Jason Kovari (Cornell University), which focused on how we could leverage systems like ORCID, NACO and ISNI more effectively.

There were then several more sessions during the afternoon:

IPFS (Matt Zumwalt, Protocol Labs) : Shareable linked data with IPLD and IPFS: harnessing the power of the decentralized web

- IPFS Content-addressed protocol to replace HTTP. IPLD: Data-model for all hash-linked content.
- There is a Persistence layer, and an Index layer
- The focus of the talk is really the persistence layer, which might be described as 'datasets'. Persistence can be in a number of forms - files, RDBs, etc. Derivative datasets are created such as indexes, optimised for querying. The main dataset is for reading and writing (a source of authority); the derivative layer is really for efficient traversal.
    - Datasets can be version-controlled. It can also be exposed through numerous views. Github provides a useful model for the social tracking, control, and collaboration of data.Example: https://github.com/whosonfirst-data/whosonfirst-data
    - Hash-linked data structures are what underpin this kind of architecture.
    - IPLD gives you the tools you need to address all the hash-linked data

SHARE-VDE (Michele Casalini and Tiziana Possemato, Casalini Libri)

Reconciliation as Authority Service

- Phase 2 = March-Sept. 2017, will convert 100 million records for participating libraries
- Reconciliation of entities (persons, works, etc.); reconciliation of resources (same work from different libraries); reconciliation of subjects (different languages, and different concepts, so can be difficult) à doing the first part now
- Names and heading came from different sources, created a clusters knowledge base; to assign a unique URI to the heading; also need to create a new cluster
- Massive clusters process: starts with a local file and analyzes the data; then data enrichment with external sources; MARC bibliographic process; entity detection (authors and co-authors identification process); name heading to authority names association (with algorithm weights)

Getty Vocabularies (Joan Cobb, Getty Research Institute)

- New Head of Technology appointed
- All Getty data will be available as linked open data in the future
- The URIs won’t change but some of the ontology definitions might change; getting a cloud architect, image architect, etc.; there will also be a data architect so that all will reconcile together

- Truly committed to developing tools that are open source; or at least the methodologies they used to reconcile the data structure
- Put out a survey on linked open data

Then there was a discussion, facilitated by Chew Chiat Naun (Cornell University), on whether the group should send out a survey on local authority work. OCLC Metadata Managers Group recently sent out a similar survey. Discussion centered on what information we would want to gather, and what the audience would be.

Day 2 morning sessions:

"Reconciliation as a Service" that was facilitated by Timothy Hill (Europeana) and Peter Murray (Index Data).

- This IMLS forum has put together a google doc with use cases for such a service.
- There are many issues that need to be further discussed such as:
    - What are we reconciling? URIs or local string data? And what are we reconciling against?
    - We could have different "confidence values" and have different "links" stating "this is the same" or "this is close, but can't be sure if is the same"
    - How to help people navigate the "almost same as" relationship
    - Should also tell us and store the information that "this is not the same as"
    - Broader issue of data exchange: if aggregator does enrichment of data, that needs to go back out to the hubs that put in data
    - There will also be an amount of human intervention, so how can we mitigate this?
    - Possible next steps:
        - People who are doing reconciliation should share algorithms; workshop of developers, etc.
        - Need success metrics to let management know what it takes to manually do clean up
        - Do another grant for a two day meeting to hammer out technical steps

Minimum Viable Product Specification (Isabel del Carmen Quintana, Harvard University) Discussion facilitated by Anna Neatrour (University of Utah) and Chew Chiat Naun (Cornell University)

- What is a minimal viable record?
    - An entity and some (any) assertion of that entity (not just that string)
    - The talk tackled records created in batch and manually, but focused on NACO records, although the concepts would be the same in other databases.
- NACO: Name & 670 (citation) is a viable record; easy, so why don't more people do it?
- What is time-consuming in NAR creation:
    - Determining usage (usually easy)
    - Determining string
        - Need to determine usage (can be tricky, but could be simplified)
    - Disambiguating name
        - Need to add something to string for differentiation
        - What if just add $c (Author of…)
    - String would be unique
    - But have to be sure there isn't another record
- Differentiation: we need to do it, so how can we do it more efficiently?
    - Better search programs to eliminate the most likely non-matches, so we only manually have to look at a few (could use data such as 046, 3XX, 670 information, etc.)
- Batch workflows
    - Can we do matching/non-matching algorithms?
- Even if we still need strings, we could set up the following workflow:
    - If find a match add $c "Author of …"
    - If batch load and therefore we are not sure if there is a match, also could add a $g Provisional or $g Batch loaded by Harvard (or whatever subfield)

This was followed by a discussion on what constitutes a minimal record, and how we could share these more readily. There was some discussion on whether we really need to do disambiguation, and how we could automate this process as much as possible. Some of the next steps to consider are: Exploring the name string assertion ; looking at how to search/merge data from local systems to other systems ; idea of differences based on workflows and systems ; and the problems of maintenance and duplication.

The afternoon featured two further sessions:

Data Provider Obligations (Janifer Gatenby, OCLC Leiden and Jean Godby, OCLC))

- Need as much data as we need to make it distinct now; aggregator needs to "seal" the data with an identifier
    - If done in batch mode, get reports to look at for possible matches, etc.
    - Disambiguation needs to be done and needs to be done at large file level; records aren't shareable if not disambiguated
- Crowd sourcing can work well
- Aggregator needs to send back out: new data, merges and splits, error correction, etc.
- Data creators: should know source, should have data in machine actionable fields, don't need another identifier for persons already in the system, should have standard compliance
- Requirements for publishing data on the web:
    - Most important for aggregators: access, enrichment, formats (open format, multiple formats, etc.),
    - Most important for the data creators: identifiers, quality (or say is a draft), resolvable on the web, provenance
    - Most important for both: vocabularies used to publish the data, licenses, sensitive data

The discussion focused on differentiation, crowd sourcing, and when to mint a new URI.

The next session was a discussion on outreach and community engagement, facilitated by Nettie Lagace (NISO) and Daniel Pitti (University of Virginia).

- We need to reach out to the cultural heritage folks, museum folks; be expansive and inclusive in bringing them into the conversation, Scholars, International players, Non-mainstream cataloging people in academic libraries (e.g. digital folks, etc.) which don't know or do any MARC cataloging, Publishers (for articles, if publishers gave us an ORCID, we would include it in our metadata), University presses, Scholarly SP, etc.; going to organizations of the types of associations we want to network with
- Propose projects for NISO
- Work on a paper on value of doing identity management. This IMLS forum will produce a paper.

The forum ended with a session on follow up projects and next steps. Various members volunteered to work on several key areas: survey, reconciliation as a service, data provider obligations, minimal viable product, and identifying patterns for sharing.