

Summary of IMLS Shareable Local Authorities Forum, Cornell University, October 24-25, 2016

Summary drafted by Isabel Quintana

Meeting themes

Workflow

- Human versus machine processing
 - Finding right mix for scalability and sustainability.
 - Role of data science and AI.
 - For both human and machine resolution we need an ensemble of information about an entity.
- Administrative/provenance metadata
 - Provenance metadata very important for certain use cases, e.g. where identification of material is part of the scholarly enterprise. Examples: rare materials, archival data.
 - Not all consumers have the same needs, and tracking metadata provenance brings scalability issues. Needs must be defined by reference to real use cases and functional requirements.
 - We can keep the information at a basic level; or we can keep complex data but only expose a subset; or we could expose it all and allow the user to take only what they need.
 - Role of aggregator may be just that: to aggregate, and not necessarily to vouch for the data collected. But this makes recording provenance important.
 - How can we track this information? We don't have a well established or widely shared model, infrastructure, or process for doing this.
 - Intellectual property/business model questions also play into this issue.
- Keeping data in sync
 - Whose responsibility is this? What are the obligations of providers?
 - Even if one or two databases exchange data, that data does not usually make it back out to all the contributors to that database.
 - But heavily centralized and regulated processes (e.g. NACO) also impose costs.
- Linked data issues
 - Some skepticism voiced about linked data.
 - Bridging or elimination of silos brings great potential for enriching knowledge. Importance of communities who do not participate in library authorities but have names that appear in library authorities. See e.g. RelFinder: <http://www.visualdataweb.org/refinder.php>
 - It allows for very rich representation of relationships. What level of alignment is needed for interoperability and reuse?
 - What is the best way to do versioning?
 - Need better understanding of what data and services aggregators need to provide.
 - Local vs shared URIs
 - Mint at point of need vs requirements for stability.
 - Issues of trust/veracity of statements if we don't maintain data locally and just point to another place.
 - Risks involved in sameAs assertions.

Community

- Sharing remains important but may need to work differently from the way it used to.
- Different models of collaboration or sharing between various participants throughout the identity management lifecycle
 - Different communities of practice: national, institutional, domain, regional.
 - For example, ISNI and publishers exchange information, as does the Getty and museums.
 - Collaboration with other language communities.
 - Need for examples of successful cooperative models that others can learn from.
- Roles/responsibilities of various participants in the process (i.e. service providers, library community, publishers, standards community, etc.)
 - Publishers may care about identities because metadata sells, and they care about rights management.
 - Stakeholders looking for guidance from metadata community about requirements; CrossRef is a model.
 - Role of hubs like Getty in facilitating collaboration.
- Centralized versus distributed work, for example, VIAF aggregates data that could be done in one workflow.
- Strong interest in developing minimum viable product specifications for shared data.

Responsibility and/or ethical issues

- Issues of veracity of statements about entities; tracking the provenance of statements.
- Traditional authority control is driven by literary warrant, but in a wider identity management context this is a limitation.
- Privacy of data, e.g.
 - Cultural institutions stripping out data before sharing it in Europeana
 - Authors who do not want (or have not approved) certain data elements shared across databases
 - Providers (e.g. Ringgold) retaining proprietary rights over some data.

Tooling, infrastructure

- Along the continuum of human versus machine processing, we need better computer processes so that humans need to work with the data less.
- Strong interest in more effective sharing of algorithms and metadata profiles for matching. Consider clearinghouse of processes, workflows, data elements, etc., but sustainability is an issue.
- What are needs for shared discovery of local authority data? How does this fit into the overall infrastructure? Could we do e.g. AWS template? We should look at examples.
- Reconciliation as a service. Identities layer does not need to be ILS- or vendor-specific. Potential for cross-platform (including cross-vendor) solutions.
- How can we leverage infrastructure that already exists, e.g. with ISNI? Are there other potential models, e.g. CrossRef?
- Role of ISNI as the clearinghouse for names (the "ultimate ID")
 - Role of libraries in duplicate resolution.

Best practices

- Guidelines needed on creation and maintenance of URIs.
- We need best practices for updates, versioning, etc.; and discuss who is responsible for disseminating changes, and at what point.
- Need for mechanisms/services to enable identity management further upstream, e.g.
 - publishers could include identifiers when they send out their data
 - publishers could include enough information (full names, subject keywords) to facilitate disambiguation.
 - Enable authors to enter ID at the point when they submit the manuscript, then the ID can travel with the author/work through the workflow. What are the impediments to this workflow?
- Let institutions gather data, curate it, etc., and leverage that out globally.
- What are models for contribution/aggregation?

Community

- Perhaps NISO can shepherd some development of best practices.
- Document and share existing or pilot workflows as model for other users. Examples exist among older projects as well.
- What are social (e.g. membership) barriers to acceptance of existing authority infrastructure? E.g. SNAC is arguably more democratic than NACO.
- How can we get smaller libraries and publishers to do some identity management "without even realizing they are doing it"? In other words, make it as seamless, automated and low-barrier as possible.
- Need for sustainable business models.
- Consider other community models, e.g. <http://archesproject.org>, <http://vocab.getty.edu>.

Potential areas for future work

The forum suggested a number of directions for future work which are summarized below. Some of them will be investigated during the course of the IMLS project, while others may be taken up in other venues as appropriate.

Project outputs

- Draft of white paper and reference model that will serve as a public discussion document, and outlining key issues and recommendations for joint action
- Survey relevant communities on local authority creation
 - survey may include: workflows, entity types, stakeholder roles, sustainability plans, integration/interoperability efforts, etc.
- Compile actual or model workflows.
 - Share information on how Harvard is resolving issues with their local authorities, and model an identity management workflow
 - ISNI workflows and use cases (see Gatenby and MacEwan slides)
 - Workflows from BIBFLOW project
 - Identify patterns for sharing data
 - Stakeholder role analysis, e.g. publishers, service providers, libraries, academic institutions, scholars

Areas that may benefit from specifications

- Minimum viable product specification for identities (low-barrier entity creation)
- Draft a document on obligations of data providers
- Draft a specification for publishers to supply identity data
 - Examine CrossRef as a potential model
- Application developers' requirements

Potential collaborations and community actions

- Explore opportunities to work with publishers to acquire IDs with their data
- Bring issues concerning identities to the FOLIO project's UX team; encourage collaboration with community partners
- Seek use cases and requirements for reconciliation as a service (already initiated by Peter Murray in FOLIO forum)
- Propose ways to share information on matching algorithms
 - Sustainability is a consideration
- Utah IMLS-funded regional authority control project
 - Consider how such a service may or may not interact with a larger hub such as ISNI.
- Initiate discussion on lowering the barriers to NACO participation
- Ways of sharing "how to" information

Other issues for further investigation

- Outline of how ISNI and VIAF relate to local authority needs
- Use cases and models for administrative and provenance metadata
- Change management
- Sustainability and business models
- Linked data infrastructure
- Privacy
- Others?

Next steps and second meeting

- Future project discussions will be topic-oriented, with nominated discussion leaders.
- Asynchronous communications preferred, discussion leads will schedule periods for focussed discussion

- Additional participants on Google Group welcome
- Consider spinning off discussions in other forums if appropriate
- Partners may consider possibilities for further grant proposals
- Project leads in discussions concerning the possibility of holding the second meeting in DC
- Projected dates for second meeting in March or April 2017