Truly shared cataloging ecosystem development: A report from the workshop at SWIB23

Authors:
Jason Kovari (Acting AUL for Collections & Technical Services, Cornell University), Steven Folsom (Head of Metadata Design and Operations, Cornell University), Simeon Warner (AUL for IT, Cornell University)

Abstract:
This report documents the "Truly shared cataloging ecosystem development" workshop held at the Semantic Web in Libraries (SWIB23) conference in Berlin, Germany on 2023-09-11. Participants engaged in structured brainstorming to explore the idea of moving MARC-based cataloging practice from its current state to one where work is performed in shared data stores, using BIBFRAME linked data rather than copying.

Background and Workshop Organization

Workshop Proposal Abstract

Current cataloging practice entails copying data from shared pools into local environments, which enables local editing, but at the cost of data divergence and substantial complexity when trying to aggregate data or perform large-scale enhancement operations. Meanwhile, many BIBFRAME proofs of concept are simply switching MARC for BIBFRAME and thus continuing the practice of copying data. To fulfill the promise of linked data, institutions must stop copying data and instead move to shared source data where groups of institutions consider their data of record to live in stores outside their sole control. This transition is as much a social challenge as it is a technical one.

In this workshop, participants will collaborate to develop the idea of what it means to move cataloging practice from its current state to one where work is performed in shared data stores, using linked-data approaches rather than copying. Participants will directly engage in structured brainstorming and designing infrastructure components, workflows and metadata issues relevant to shifting to this model. The facilitators will outline some initial thoughts resulting from ten years of BIBFRAME and linked data for cataloging work, offer these for discussion, and then proceed with several rounds of breakout and discussion. The facilitators will collect notes throughout the workshop and compile a summary report to be placed in an open-access repository soon afterwards.

Document Shared with Attendees Prior to Workshop

Truly shared cataloging ecosystem development workshop: Vision and questions
Tom Cramer, Steven Folsom, Jason Kovari, Philip Schreur, Simeon Warner

Since 2014, Cornell University, Stanford University and partners have worked on several linked data grants funded by the Andrew W. Mellon Foundation. At issue were libraries' dependence on MARC-based systems for the communication, storage and expression of the majority of their bibliographic data. With each day of routine processing, libraries add to the backlog of MARC data that they must eventually convert and enhance as linked data.

MARC is an inherently records-based ecosystem. Although libraries can greatly reduce their original cataloging costs by making use of work done by others (i.e.: copy cataloging), they do so by making a copy of the record of interest and most often altering it in some way to meet local specifications. The result is hundreds, if not thousands, of records for the same resource that closely resemble each other but that are stored and maintained separately from each other.

An important element in libraries transition to linked open data has been the Library of Congress's development of BIBFRAME as a replacement for the MARC formats. BIBFRAME and other Linked Data approaches have great promise to free libraries from outdated formats and inefficient workflows, alongside shifting focus from local copies to shared metadata stores.

There have been many years of experimentation with linked data cataloging; however, most explorations with linked data still rely on copying data between environments. Instead, libraries must relinquish library-specific, individualized data in favor of truly shared data.

We imagine the development of new environments that would be cooperatively owned, eliminating the need for every institution to establish complex and expensive linked data tooling locally. Most importantly, this environment avoids the need to copy and synchronize data between systems, other than for caching and indexing to support searching in local systems. These environments would support a clear separation of shared bibliographic data (e.g. BIBFRAME Works and Instances) from locally managed, institutional data (holdings and administrative data stored in library service platforms like FOLIO and Alma).

Such new environments would form the core of library bibliographic ecosystems, linking to external entity stores and local library management systems; further, these environments would support production workflows and dataflows that meet the complex needs of library metadata including work with legacy systems..

During our workshop, we will imagine these environments, their components and workflows. Beyond technology, working in these environments would confront long-standing practices and expectations institutions have around their data.

For the workshop, please consider the following:

- What are the benefits of a shared environment?
- What are the impediments to a shared environment?
- What are the key components of a shared cataloging system and how does data flow between them?
- What is the right scale for shared environments?
- How do you ensure trust in a shared environment?
- What rules do users and institutions need to follow?

Workshop Agenda

- Introduction
- Norms and Expectations
- Facilitator presentation (see accompanying slides)
- Discussion: Benefits and Impediments
- Breakout: Defining a Shared Environment
- Breakout: Boxes and Arrows
- Discussion: Connections between shared environments – [Not discussed, ran out of time]
- Conclusion

Workshop Participants

The following participants joined the workshop and contributed to discussion. Discussion notes are not attributed and views expressed in the workshop may not be shared by other participants,

- Adrian Pohl (Metadata Infrastructure Lead, hbz)
- Alexandra De Pretto (Systems librarian, Library Network Renouvaud)
- Andreas Mace (Systems Librarian, The National Library of Sweden)
- Anne-Kathrin Brandau (Metadata Librarian, Freie Universität Berlin)
- Antoine Isaac (R&D Manager, Europeana Foundation);
- Hadewijch Dekker (Metadata Librarian, University of Amsterdam)
- Jarmo Saarikko (Information Specialist, National Library of Finland)
- Katerina Sornova (Development Manager, The National Library of Finland)
- Lars G. Svensson (Editor-in-Chief of the Integrated Authority File, German National Library)
- Lynn Van Kerckhove (IT staff / project manager, Cultuurconnect vzw)
- Nils Berns (Software Developer, Kiel University Library)
- Philipp Weiß (Librarian, Bavarian State Library)

- Radek Světlík (IT and Digitization Manager, Education and Research Library of Pilsen Region)
- Sebastian Tilsch (SLUB - Saxonian State and University Library)
- Stuart Edelenbos (Product Analyst CBS, OCLC)
- Tiziana Possemato (@Cult and Casalini, for Share-VDE initiative)

Workshop Discussions

This section reports and summarizes facilitated discussions and breakouts during the workshop. Significant portions are text that was captured in a shared notes document and is *indicated with italics* in this report. Small edits to correct spelling, clarify grammar, or expand acronyms have been made without indication. American English spelling is used here for consistency, even if not in the original. Significant additions or changes in otherwise quoted content are indicated with *[square brackets]*.

Hopes for the Workshop

Participants expressed the following hopes in response to an email prompt prior to the workshop:

- *Better understand European issues/perspective*
- *Looking for others thinking about BIBFRAME and RDA*
- *Ideas for working with the GLAM sector*
- *Future ideas for linked data creation*
- *Better glimpse of library views of linked data. When catalog data published at linked data, what then?*
- *Shared project for catalog data in northern European libraries*
- *Sense of where we are moving as a library community*
- *Hope to get sense of direction for BIBFRAME cataloging and practices*
- *Consortium supporting small libraries in region*
- *Spying on the librarian community from the viewpoint of LAM intersection. Understanding decentralization*
- *Idea for future of metadata management if implement linked data*
- *Curious about thinking in shared linked data, how to move beyond MARC format*
- *Curious about new developments in linked data*
- *Want to help push the agenda forward*
- *Interested in how to work in a heterogeneous environment with traditional and linked data*

Discussion of Facilitator Presentation

The following questions and comments were made in response to the facilitator presentation:

- *What about public libraries?*

- *Why BIBFRAME and not also RDA?*
- *Is Marva releasing a new version? Yes, [the Library of Congress plans to release a new version] within the FOLIO environment.*
- *Sharing about the JCricket editor's place in the editor ecosystem.*
- *Vendor editors and where they might fit into the ecosystem.*
- *[European] National libraries have a responsibility to create and protect shared quality metadata, which puts requirements on who can edit.*
- *Focus on bibliographic data and not on holdings -- but a shared Instance needs to be kept even when a local holdings/Item is deleted, because others might refer to it.*
- *Recent use case where consortia of libraries looking for a linked open-data system for managing items and holdings - want to manage shared collection. Different from the focus being described, want to have better analytics etc.*
- *Where is the line between cataloging systems and domain-specific databases?*
- *What sort of scale are we walking about? A global collaboration would be quite different from a national one.*
- *In many countries the national libraries have a legal responsibility to register all publications within a country.*
- *It may be that every country has its own version of BIBFRAME and RDA.*
- *We need to work to overcome ontology differences. As we move from a bibliographic records system to entities we will have to deal with entities in different ontologies.*

Benefits and Impediments

We discussed the benefits of a truly shared linked-data ecosystem might be, and what impediments there would be in moving toward it. Participants noted that a centralized system might suffer from increased risk because of a single point of failure and noted that with reductions in funding we expect to have fewer catalogers in the future and thus processes need to be more efficient. It was noted that analysis of current MARC data shows many errors and even some patterns of erroneous cataloging. How could a new approach improve this situation? Multilingualism was not seen as an impediment because linked data supports it better than the current MARC ecosystem.

Defining A Shared Environment Breakout

Participants worked in four breakout groups to consider these three questions:

- Where do you fall on the distributed versus hub model?
- What operations need to be supported?
- Within a shared environment, what governance/rules are necessary?

Group 1

Discussion of scale and distributed versus hub model:

- *The OCLC CBS system is used for various regional and national catalog systems. Germany has variation between PICA and MARC21 cataloging. Spain has a national Spanish library but there are also catalogs for other language regions with different viewpoints. Several systems and ways of cataloging in the Netherlands.*
- *Germany has federal vs regional, with different hierarchies between institutions and collectives (who does what). [There must be] coordination of a shared ecosystem:*
    - *There are now open data pools, which are at a more federated level, [along with initiatives] to collect/link open data repositories and local/regional/federal systems*
    - *There are collaborations between libraries and heritage institutions. National libraries, local, and municipal. Regional scopes of operations.*
- *Agree on collaboration between libraries and other GLAM institutions, many possibilities to collaborate.*

Discussion of functionality and operations:

- *In terms of scale, what should an ecosystem do? Should it do everything for a library?*
    - *A shared system would be beneficial if it can be managed from a central point, and then distribute this data to the local regions.*
    - *Different German library systems fulfill different needs.*
    - *Move shared cataloging to a shared system, keep i.e. ILS systems locally and differentiated.*
    - *Different types of collections (public, academic, etc.) have different needs*
- *For a shared dataset that covers both academic and public libraries we might have simple dataset because public libraries often include much less data (e.g. DOIs, arks not used much in public libraries)*
- *Currently, most [public libraries] do not do cataloging themselves.*
- *Might have models where some subject specialist libraries (e.g. medical) have full rights to edit descriptions of relevant material, whereas others would not have full edit access*
- *In Germany there are systems where there is a hierarchy of cataloging standards followed by different institutions as MARC records percolated up and down the chain*
- *How do we build systems and collaborations to support assessment of data quality?*
    - *Much is possible in MARC and how will this be done in RDF, how do we support expression of different viewpoints*
    - *RDF has benefits of generality and extensibility*

Summary:

- *On the topic of scale: two perspectives on scale --> 1) 'level of institution' 2) 'level of detail'*
- *In our regions (Germany, Czech Republic and the Netherlands) currently, library catalogs work similarly: local --> 'wider region' --> national, where there is also a difference in types of institutions: public libraries, academic libraries, national libraries, GLAM institutions, Medical, etc.*
- *At different levels, cataloging can occur, and is accumulated (and transformed) at the next level into a larger catalog.*
- *Different points of view and cataloging standards are at play*

- *Who do you accommodate with a 'shared dataset' and what data is shared? To accommodate everyone (equally), only a very basic representation will/can suffice (e.g. no DOIs for public libraries)*
- *Brief discussion on data quality and the use of RDF. Can implementation of RDF provide better ('factual') data than MARC?*

Group 2

Discussion of distributed vs hub models:

- *Finland: current software doesn't support it, we can only imagine most libraries can edit the central system and it gets downloaded so that local libraries can select which fields they want. What we want to do is to make library data authoritative.*
- *Germany: for the authoritative file, master copy at the DNB and other mirror copies (through OAI-PMH). All mirrors are up to date. Common data model for everyone Ideally we would add links and the system would pull the data for these links. For example for Wikidata, trust is necessary (case where someone changes the data)*
- *We all agree on a centralized system that keeps the authority files. What would happen with the items?*
- *In FOLIO there's an "instance" abstraction. We're talking about moving the source data out of FOLIO - that could be the centralized data store that feeds the cache and the admin data lives locally.*
- *Needs to support the case of finding items.*
- *In the US we don't manage that - it's in other systems.*
- *National Libraries don't do circulation as much.*
- *Finland: national collection is split from the local collection.*
- *The libraries using the Koha system have a central place, that feeds into the national database (don't know how the data flows the other way)*
- *Use cases:*
    - *Cataloging (circulations, inter-library loans are out)*
    - *Search and discovery on top of the central database?*
    - *Finna service (also for GLAMs) has all data for search. But they don't handle the central repository, they are 'client'*
- *Issues of scale?*
    - *Finland: 800 local public libraries, 180 databases.*
    - *Germany: no national library holding access system, 6 regional ones.*
    - *Several dimensions for the holdings.*

Discussion of options to move forward with shared cataloging:

- *Small consortium for international ones.*
- *We could use Wikidata as a shared catalog but [that raises[issues of trust. Also technical issues.*

- *Could Cornell start a consortium with Iceland? Sometimes geography is a weaker link than the closeness of collections.*
- *National libraries are a key case because of legal deposit [requirements].*
- *National libraries get metadata from publishers. (Onix)*
- *Others can come and correct errors, add subject headings and classifications. But which governance/rule?*
- *In Finland, the first [library] that receives a book can create the metadata. Metadata shows who created the statement. Emails are sent (because again some institutions don't automatically get all fields). All contributors are peers, the national library has an informal last word (people can change the data they create but it's not recommended). Everyone is identified: names = authority.*

Group 3

Discussion of the possibility of a hybrid environment where both models (hub and distributed) can work together:

- *The local level could be managed through application profiles, so that each node (a library, or a network and so on) can find benefit in working together but can also have its own local information. In this case some issues need to be solved, such as how to synchronize operations to avoid copying, but using linking to manage local data.*
- *We need to be able to access other resources and link to them easily.*
- *Within a shared environment, what governance/rules are necessary? Some suggestions:*
    - *define different levels to cooperate in a shared environment, to obtain a larger cooperation but in a controlled environment (agreeing on general rules, defined within a community).*
    - *In the Wikidata model many people can cooperate, but in a largely shared environment, so in a certain way controlling each other's [data]. This is an approach to reduce the gap between very active institutions and less active institutions: everyone can contribute with their own forces and skills.*
    - *We need to use as much as possible authoritative sources, vocabularies and ontologies, so that any assertion done for an entity can be trusted and reusable within the larger community. This is one of the advantages of a linked data environment, where you can link to trust your assertion (role of provenance to assure quality).*
- *How to involve everyone?: If there is a community, everyone should be involved and active, collaborating in relation to their own skills.*

Group 4

Discussion of scale:

- *[Should be] regional or national, global is not realistic but communication between the systems should be a focus; national there are a lot of difficulties because of different types of libraries, medical libraries want very specific subject headings, maybe only public libraries could work fine*

- *Shared rules are important, e.g. what is local, what is shared? Defining what's common and what's local isn't always clear or agreed.*
- *Even rules can be differently interpreted*

Discussion of distributed vs hub models:

- *[The choice] depends on how good your cataloging is, a system should also enable local subject headings.*
- *You always need the data in your local system, so you need to copy it. In the Swedish system they have Library of Congress subject headings, but they don't want to copy them to their system, they need to be cached. Caching data on a large scale is difficult. If you cache too much, you are copying.*
- *Libris:*
    - *In Sweden almost all libraries already take part in a national union catalog based on Bibframe (Libris), but they still have copy cataloging, they import data into their central system.*
    - *Each library can edit a record, diffs, including provenance (on an institutional level) are stored and can be reverted; edit wars can happen, e.g. when there are different opinions over correct subject indexing.*
    - *Libraries copy data to local system and often it is unclear what happens to it there*
    - *Big scale creates lots of problems.*
    - *[There is] not agreement nationally on which restrictions should exist.*

Discussion of rules that might be needed for a shared system:

- *It needs to be clear what is automatically created and what is the provenance (of everything).*
- *A rollback should always be possible for each record..*
- *Different levels of edit rights should exist, but not too complex a system: e.g. restrictions for bulk edits but not so strict restrictions for individual edits.*
- *Communication is central for shared cataloging.*
- *Even within a system based on a Bibframe data model, there are similar challenges to a MARC based network environment.*
- *A system could enable libraries with similar cataloging practices to benefit from each other, e.g. by subscribing to each other's enhancements. One could add an inbox to each record that is published by a central data provider. Reusers could submit patches and additions to it and others could decide whether to (automatically) apply them or not.*

Boxes and Arrows

In this breakout exercise, we divided the room into three groups: two groups represented individuals interested in the viewpoint of National Libraries and one group represented other cultural heritage institutions. All three groups used the following prompt:

What are the key components of a shared cataloging system and how does data flow between them?

- What scale of collaboration do you imagine?
- What components are shared vs distributed?
- What other systems does it connect to?

Each group drew a sketch of sketches of the system they imagine, and described the sketches to the whole group.

National Libraries Group A

This group thought in terms of services in the ecosystem and the relationships between participants and operations. The national library would occupy the position of greatest authority, operate a central metadata store, and be central to workflows. Metadata would be open and would be query services and APIs on it. There would be the ability to edit and describe data, including tools for curation and bulk edits. Rules would be needed to support a hierarchical system of edit rights to ensure quality and peer review may be needed. There would need to be reliable look-up services and that might require local caching, open data, query service, editing/describing data (incl. tool for curation). There should be a subscription mechanism to allow downstream users to track changes.

National Libraries Group B

This group thought in terms of linked-data objects. National libraries are in the center as a creator and publisher of linked data. Regional libraries consume data but might also push back to the national library for things like special collections. There would also be connections to other GLAM, publisher and commercial datastores. They envisioned a distributed system where links are created whenever someone “copies” a record (provenance data) and all diffs are recorded. The model thinks beyond local databases to instead consider pools of data that institutions can link to. Instead of records, the unit of information is an entity though this will be a significant semantic shift. For the time being at least, there would need to be conversion to and from MARC, and there might be other conversions such as from Onix for publisher feeds.

A cataloger works in three editors: the central one is for the bibliographic data, and also editors for shared authorities and for local databases. Bibliographic records contain links out to authorities. Although not part of the cataloging systems or library software, there will also need to be ontology editors to manage the shared ontologies.

This system will support discovery by end-users through search APIs that operate on the shared bibliographic data store.

Cultural Heritage Organizations Group C

This group started from the assumption that the cultural heritage libraries share the same goal. They saw the ideal as one central cataloging tool that includes user management and is

connected to a shared datastore for metadata. There would also be stores for holdings and item information. The environment would require conversion tools to support the ingest of publisher and vendor records into the cataloging tool since they come in different formats. There would be a discovery layer tool to aggregate data (including a cover image server and other meta-content). The role of cultural heritage aggregators was highlighted and there might be the ability to add annotations in these systems. There was incomplete discussion of whether annotations might be able to flow back to the central datastore, perhaps using notifications. It was imagined that anyone creating data outside the central cataloging tool would be building on the shared data but not editing the data itself.

Summary

The workshop engendered lively discussion of how we might move to a shared linked-data based environment for library metadata. The participants presumably self-selected based on an interest in this direction and it is thus unsurprising that focus was on “how to” rather than “why not” or “why we can’t”.

Discussions offered different perspectives on possible approaches and organizational structures to support shared cataloging. There wasn’t a single suggested approach although there was broad support for the desirability of linked open data. Repeated themes were the need to have effective data flows between distributed components, clear understanding of who can edit what, support for ensuring appropriate data quality (especially when certain organizations have specific legal requirements placed upon them), data import and export workflows involving format conversions, and shared rules to support shared editing.

From the perspective of US-based facilitators it was notable how the central role of national libraries in European countries influenced discussions. Two of the three groups involved in sketching ecosystems were composed of participants thinking from the national library perspective. The outputs from these groups place the national library in a central or authoritative position and raise questions, that we didn’t have time to explore, of how to enable collaboration beyond national boundaries.