Truly shared cataloging ecosystem development

SWIB Conference Workshop

Facilitators: Steven Folsom, Jason Kovari, Simeon Warner
(Cornell University Library)
Introductions - Who are you?

- Name
- Institution
- Title/role
- What do you hope to get out of this workshop?

SWIB Code of Conduct: [https://swib.org/swib23/coc.html](https://swib.org/swib23/coc.html)
Agenda

Introduction
Norms and Expectations
Facilitator presentation
Break
Facilitator presentation continued
Discussion: Benefits & Impediments
Breakout: Defining shared environment
Break
Breakout: Boxes and Arrows
Discussion: Connections between shared environments
Conclusion
What are your backgrounds?

- Systems librarian
- R & D management
- Authority management
- Metadata librarian
- Ontology / Modelinng
- (2) IT management
- Software development

Languages:
- se (1)
- fi (3)
- nl (2)
- de (7)
- be (1)
- fr (1)
- ch (1)
- it (1)
- cz (1)
- us (1)
Where are we coming from?

- Jason, Steven, Simeon – We all work at Cornell University. We have experience of Cornell and other US university libraries
- Tom Cramer, Phil Schreur – Add experience from Stanford University and prior libraries
- Together, we have led a series of Andrew W. Mellon foundation grants exploring linked-data in libraries: the LD4L, LD4L Labs and LD4P grants spanning 2014-2023. I’ll briefly describe some of the influence from the work in these grants
- Recent discussions have involved U Penn & Library of Congress
- We have also engaged with many others, including the SWIB and LD4 communities

We have benefitted from lots of discussion, but our perspective is **US centric** and also **large-academic-library centric**
Experience from LD4L/LD4P Grants and related work
Ontology Development

- LD4L Ontology, bibiotek-o
  - Concerns about early versions of BIBFRAME, desire to explore extensions
  - Work ended in 2017, focus on BIBFRAME feedback instead

- BIBFRAME extensions
  - E.g. ongoing maintenance of Art & Rare Materials (ARM) extension

- BIBFRAME very closely aligns with MARC – this is a mix blessing:
  - Makes transition much easier and also reasonable fidelity ongoing conversion
  - Did not take full advantage of the opportunities of creating a new linked data format
  - “Weak” in terms of semantics/restrictions but inference etc. not a priority use case
MARC → BIBFRAME Conversion

- Early grant work on converters to new ontologies
- Collaboration with Share-VDE on bulk-conversions for LD4P cohort and for PCC
- Exploration of Library of Congress converters

Data pools created by conversion have been valuable: understanding, source for copy cataloging, discovery work…

… BUT, less useful in helping develop environments and workflows for ongoing BIBFRAME data management. Need clear understanding of primary vs derivative records and synchronization rules

(Also, BIBFRAME → MARC converters exist)
BIBFRAME RDF Editors

- Library of Congress
  - Developed “BIBFRAME Editor” (BFE), 2014–2021. Large scale tests
  - Rewritten as “Marva” starting in 2021
  - Being integrated with FOLIO by EBSCO for production use 2024 or 2025

- LD4L/LD4P
  - Early experiments with VitroLib (based in VIVO/Vitro editor) showed specialized facilities required for cataloging
  - Stanford team developed “Sinopia” editor, used by grant cohort, PCC and others
Without lookups it’s just data

- Editors need lookups to create linked data
- BFE and Marva focus on LC authorities
- LD4L lookup service developed based on Samvera “Questioning Authority” gem, used by Sinopia editor to support a range of authorities: direct and cached
- Needs: usability, speed, accuracy, reliability, sustainability

→ Steven will discuss linking experience more
→ Also, Steven’s talk **11am Tuesday “Supporting sustainable lookup services”**
Community

- SWIB & EuroBIBFRAME
- LD4 community:
  - Supported initially with help from LD4P grants
  - Affinity groups - Art & Design, Discovery, Ethics, Non-latin, Profiles, Rare Materials, Serials, Wikidata, Sinopia
  - LD4 Conference - Pandemic reinforced decision to have online not in-person LD4 conference - low cost, very broad participation
Much work to date copies these for every institution.

Real progress requires sharing.

Shared authorities, vocabularies & other linked entities (& a few strings)

Genuinely local data
Two key motivations

**Improved discovery**
- “Obvious” outcome of linked-data but needed practical demonstrations
- Experiments throughout the LD4L & LD4P grants, and by others
  - Specialized interfaces for particular data types or sub-domains
  - More general additions
  - User testing to verify utility / refine
- Improvements live in Cornell’s current discovery environment

→ Steven will share more details next

**Improved efficiency**
- Not obvious with new technologies and more complex models
- Experiments have verified feasibility of components and workflows but have not been performed to get at efficiency
- We argue that proper use of linked entities, avoiding duplicative work, and avoid maintaining unnecessary copies can be more efficient

→ Jason will lead us into discussion of an efficient ecosystem
Discovery: Lessons Learned
Discovery: Background

- “Decorating” our MARC-based discovery environment
- MARC to RDF conversion pre-indexing
- “Linked Data” mostly serialized in RDF (but not limited to RDF)
- Cornell’s MARC often doesn’t include $0’s or $1s (yet)
The wizard of Oz

"Illus. by Evelyn Copelman ; adapted from the famous pictures by W. W. Denslow."

<table>
<thead>
<tr>
<th>Author, etc.</th>
<th>Baum, L. Frank (Lyman Frank), 1856-1919.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format:</td>
<td>📖 Book</td>
</tr>
<tr>
<td>Language:</td>
<td>English.</td>
</tr>
<tr>
<td>Published:</td>
<td>New York, Grosset &amp; Dunlap, [1956]</td>
</tr>
<tr>
<td>Subject:</td>
<td>Fairy tales.</td>
</tr>
<tr>
<td>Description:</td>
<td>204 p., [10] leaves of plates. ill. (some col.) 22 cm.</td>
</tr>
<tr>
<td>Other contributor:</td>
<td>Copelman, Evelyn</td>
</tr>
<tr>
<td>Series:</td>
<td>Illustrated junior library.</td>
</tr>
<tr>
<td>Notes:</td>
<td>first published in 1900 under title: The wonderful wizard of Oz.</td>
</tr>
<tr>
<td>Dedicated to:</td>
<td>Maude Gage Baum</td>
</tr>
</tbody>
</table>

Some of this information comes from Wikidata

Highlight the Wikidata data.
Discovery: Caching/indexing to support local searching

- Gathering use cases for indexing until we can support updated caches
  - Advanced searching based on specific properties
  - Browsing based on related entities
Discovery: Attribution, Trust, and Data Quality

- Honoring licensing and supporting information literacy needs
- Making informed decisions about what to integrate
- Data quality
  - Ability to disallow/block data for certain entities and/or properties
  - Data improvements for data sources we now rely on

Notes: first published in 1900 under title: The wonderful wizard of Oz.
Dedicated to: Maude Gage Baum

Some of this information comes from Wikidata
Remove the Wikidata highlighting.
Linking to and Reusing RDF: Lessons Learned
(Data Creation and Management)
Libraries produce a lot of data, but rarely from scratch

- Deriving new descriptions from existing descriptions
  - Using similar works/publications/etc. to create similar descriptions for new works/publications/etc.

- Reuse and enhance existing descriptions
  - Finding an existing description and sometimes improving it, “copy cataloging”
Linking moments: Lookups when creating new data
Linking moments: Reconciliation

- Reconciliation processes for strings
  - Adding identifiers to existing date, e.g. MARC $0s
  - Converting to RDF
    - May include blank nodes or new URIs where there are no matches
Linking moments:
Supervised conversion (hybrid)

- Converting a single record into RDF
  - Some automated matching, some lookups
  - The cataloger is in control, overseeing the outcome
- E.g. Oslo Public Library cataloging tool
Linking moments: At load time in applications

- As an application renders a page, a query is made to an API using a string match (implicit link) and brings back data.
WOLFcon takeaways from the BIBFRAME in FOLIO panel

- With preliminary tools in place we need to define the ecosystem for BIBFRAME for cooperative cataloging
  - What the ecosystem looks like will likely differ depending on the region and libraries involved
  - Where is data coming from?
  - Where is does it need to go so other libraries can find it?
Break
10 minutes
Shared Cataloging Ecosystem
Scope

- Bibliographic data
- Not context entities, aka "authorities"
Position of libraries in the ecosystem

- We are coming at this from a very US-centric view
- Cultural differences in ways that libraries organize
MARC's legacy

- Copying records as historical practice
- Duplication allows for institutional-specific variation
Linked Data development - if done wrong

- Reinforcement of old models of copying data
- Institutions unable to "give up" their local practices
What do we feel is needed?

- Shared datastore
- Support for operations
- Cooperative decision making & ownership
How do we shift the paradigm?

- Define & develop the components of a new ecosystem
- Drop idea that local bib data needs supercede shared requirements
- Build trust
Benefits

- Shared enhancements (3)
- Avoid duplication (2)
- Efficiency
- Big player help smaller ones (2)
- Improved quality in shared data (3)
- Focus in essentials, not exceptions
- Better collaboration
- Services on shared data
- Reduced complexity
- Better quality assurance, data cleaning
- No more MARC!

Impediments

- Trust (3)
- Who can edit?
- Letting go of control
- Resourcing central services
- Loss of prestige
- Need shared cataloging rules & practices (3)
- Data ownership
- Complex problem resolution
- Multilingualism
- History
- Including vendor metadata
Breakout

Further define a shared environment based on the discussions including the right scale of this environment and what is the work that should happen in this shared environment?

At what scale?

- Institutional (TOO SMALL)
- Group of institutions
- Consortium
- Regional (cf. some German networks, Flanders)
- National (cf. Finland)
- Multi-national
- Global (cf. wikidata)

Do institutions share one platform entirely (for item data) or just for the bibliographic data?
Breakout - Defining a shared environment

Prompt questions:

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

Logistics:

● 30 minutes for discussion, 15 minutes to report out
● Select a notetaker who will report out
● Etherpad link: https://etherpad.wikimedia.org/p/SWIB2023-TSC
Break (Cake)
30 minutes
Breakout - Architecture: Boxes and Arrows

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?

- If there are divergent views within the breakout group, feel free to draw multiple pictures
- Feel free to use a drawing program of your choice, or draw on paper
- Please be ready to talk through the picture(s) afterwards
Discussion

- What are key similarities and differences between our pictures?
- How do different shared environments connect/collaborate?
- We assume most data will be open. Should and how could such systems support closed data too?