Ontology Overview

Linked Data for Libraries (LD4L) Workshop
February 23, 2015

Jon Corson-Rikert
The team
Goals for an LD4L ontology framework
(from our 2013 proposal)

• reuse appropriate parts of currently available ontologies while introducing extensions and additions where necessary
• be sufficiently expressive to encompass traditional catalog metadata from the 3 partners
• maintain compatibility with VIVO and research networking ontologies
• include usage and other contextual elements
Early discussions

• What library information is local vs. global?
• How do we add links to external identifiers, authorities, and real world objects (RWOs)?
• How do we link across our libraries?
• What existing ontologies should we use?
• What services and workflows do we need?
• Which services are (or aren’t) ready for prime time?
Other starting points

- LD4L is not about original cataloging
- LD4L is not about reconciling the different approaches of Schema.org and BIBFRAME
- Linking to data beyond library catalogs is a focus
  - Predisposed to re-use ontologies appropriate to the domain involved
- Ontology work has focused on our use cases
BIBFRAME and Schema.org in LD4L

• We considered the needs of our project and our libraries

• Our libraries will likely need the greater expressiveness that BIBFRAME offers
  – Technical services units are actively participating in BIBFRAME training and trials

• Libraries also want broader discoverability
  – Value in exposing bibliographic metadata on the web with Schema.org tags
Conversations continue on the BF list ...

- “Much of Bibframe’s problems seem to come from trying to keep everything from the past (MARC) while moving to something fit for the future, an ambition that has I think also afflicted RDA” (Thomas Meehan)

- “We can’t possibly formalize the rules if nobody can even describe his examples in natural language” (Jeff Young rephrasing Ronald Murray)

- “Why do the 'powers that be' think that we even want our local catalogs to be semantically connected to the web or have all of our data linked?!” (Michael Ayres)

- “The whole linking idea is great, but really, after 40 years using MARC21, some yahoo wants to unravel everything and bill me for it? I don’t think so.” (Jeffrey Trimble)
Open questions

• Limitations of a work-centric model for event-centric content
  – Abandoning the ‘record’ for independent entities and the architecture of the web
  – Not every library resource even has a work
  – “A resource may be of several types, and all may be specified.”

• Granularity distinctions
  – E.g., archiving and preservation vs. discovery
  – Where is the right point of crossover from BIBFRAME to existing (or modified) content standards?

• Tensions between consistency and flexibility
  – E.g., string/literal values vs. URIs vs. placeholders for future URIs
  – Availability of data

1. BIBFRAME AV Modeling Study: Defining a Flexible Model for Description of Audiovisual Resources, Kara Van Malssen, AVPreserve, May, 2014, p. 42
Local vs. global identifiers

• Establishing local identifiers allows libraries to make their own assertions about resources and authorities
  – Assigning stable linked data URIs are accessible anywhere

• Shared references to global identifiers enable both direct and indirect linkages

• OCLC, VIAF, ISNI, ORCID and others are addressing global identifiers at scale
Local linked data identifiers

• For library resources but also local or unique information on people, organizations, events, collections
  – There is value and credibility in the institutional namespace
• Supplement with locally-sourced and/or locally-targeted annotations
  – Not restricted to local generation or visibility
• Does not require exposing all operational data
From strings to things

- People
- Organizations
- Places
- Subjects
- Events
- Works
- Datasets

Image: designyoutrust.com
Reliance on strings alone is problematic

• Using reliable external URIs will improve data quality and connectivity, enabling interoperability
  – We will need to assign local URIs even to unknown things

• There will be challenges in resolving duplicate URIs
  – And owl:sameAs may not always be the appropriate relationship

• The development and dissemination of resolution services and tools to use them is paramount
  – To support new cataloging as well as legacy metadata
Converting MARC to RDF

• The LOC Converter continues to evolve and there will be other editors and converters
• We have focused on the data required for our use cases
• Three workflow phases
  – Pre-processing to adapt local anomalies or augment MARC with additional authorities
  – Conversion to BIBFRAME plus minimal extensions
  – Post-processing to add additional entity references and support interoperability with other linked data
Addressing complexity

• Ontologies such as PROV-O and VIVO support both simple and reified relationships
  – Reified relationships allow linking spatial or temporal extent, roles, outcomes, provenance, or multiple parties

• Levels of detail and quality vary widely in existing library metadata

• Search indexing and application display both need to accommodate variation in data structure and completeness
  – Balancing representational granularity and consistency
Working with non-MARC metadata

• Faculty research profiles (CAP, Faculty Finder, VIVO)
• Library guides and other library-sourced web content
• Digital collections that vary widely in size and complexity, and that encompass diverse subject domains
• Pilot projects
  – Cornell’s Hip Hop flyers
  – Harvard’s Visual Image Access metadata
Entity resolution

• Can potentially happen before, during, or after MARC to RDF conversion

• Can draw on existing authorities directly and indirectly
  – Local sources may involve custom workflows and services
  – Remote sources are likely shared and can more likely benefit from standardized services

• Assess potential to loop back
  – From non-MARC sources to other catalog resources
  – From external sources to local
Annotations and online collections

• Personal/individual annotations for integration into current local discovery systems
  – Use case 1.1
  – Persisted in triple store using ActiveTriples

• Online collection management
  – Use case 1.2
  – To improve the functionality of the Curated Lists of Library Resources (CuLLR) project at Cornell
  – For more general collection identification, management, and sharing by patrons and/or library staff

• Annotations can reference any URI on the web and link together physically disjoint collections
Usage data

• Inspiration
  – Harvard’s Stacklife

• Goals
  – To supplement library discovery interfaces
  – To inform collection review and additions

• Challenges
  – Data availability
  – Concerns for patron privacy

• Potential for a normalized stack score across institutions
Welcome to StackLife, a new way to browse the Harvard Library collection.

This is a prototype. We’re eager to hear from you about what works, what doesn’t, and what you’d like to see. Email us at lil@law.harvard.edu!
Continuing work

• Improving the quality and consistency of search by exposing more nuanced metadata in more consistent and controllable ways
  – Express important user-facing distinctions semantically rather than only through opaque, black box workflows
  – Facets such as genre, uniform title, online availability
• Further leveraging linked data
• Further cross-linking among our 3 institutions and other partners
Remaining challenges

• Scalability
• Moving beyond pilots to production services
• Entity resolution, both locally and at global scale
• ROI
• Sustainability