Linked Library Data and the Semantic Web

Leveraging Library Authority Control outside of MARC Applications

Presented 2008-09-17
At the National Library of Sweden
Corey A Harper

(MARC, Machine-Readable Cataloging, a communications format developed by the Library of Congress for producing and distributing machine-readable bibliographic records on magnetic tape.)
Topical Overview

• Linked Open Data and SemWeb
• Library Authorities and Controlled Vocabularies – Toward Library LOD
• Work in progress in these areas
• Metadata Normalization, Harmonization and Recombination
• Possibilities...
“The vast bulk of data to be on the Semantic Web is already sitting in databases ... all that is needed [is] to write an adapter to convert a particular format into RDF and all the content in that format is available.”

-Tim Berners-Lee in an interview with the Consortium Standards Bulletin
Linked Open Data

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information.
- Include links to other URIs. so that they can discover more things.

http://www.w3.org/DesignIssues/LinkedData.html
Growth of the LOD (Linked Open Data) Cloud: May 2007 (left) - May 2008 (right)

Created by Richard Cyganiak, blog: http://dowhatimean.net
Linked Library Data

- Resources get URI’s early in lifecycle
- Properties get URI’s
- Vocabularies get URI’s
- Everything is dereferenceable: Able to request meaning over http
Library Authority Data

“Include links to other URIs. so that they can discover more things.”

Short of providing and linking to URIs, this *is* authority data.

This is what our authority files are for.
Authority Information

• Controlled Vocabulary
• SKOS for LCSH, Dewey, LCC, Mesh, others
• Need a structure for Name Authorities
  – FOAF is only part of the answer
• Standard URI’s for concepts and agents
  – Possibly for FRBR Entities?
Library Controlled Vocabularies: Benefits

• Reputation - Trusted Tradition
• Mature - Time tested and carefully developed
• General & Comprehensive - Cover large knowledge spaces
Library Controlled Vocabularies: Drawbacks

- Overly Complicated - extraneous information
- Archaic Syntax - MARC Records
- Slow to evolve - authorities control the authority control
LCSH

World Wide Web [R S D]
[TK5105.888 (Telecommunication)] [B L S D]
[ZA4195-4235 (Information resources)] [B L S D]

UF W3 (World Wide Web)
Web (World Wide Web)
World Wide Web (Information retrieval system) [Former Heading]
WWW (World Wide Web)

BT Hypertext systems
Multimedia systems

RT Internet

NT Semantic Web [R]
WebDAV (Standard) [R]
WebTV (Trademark) [R]
LCSH in Dublin Core

- Encoding Scheme for DC Subject
- No easy way to draw on equivalent terms and cross-references
- Abstract Model, RDF and SKOS could enable applications to make use of the whole vocabulary
Vocabulary Encodings

- MARC - Great for Library Applications
- MARC-XML → Helping Get Library Apps online
- MADS
- SKOS - Designed for use with RDF
LCSH in SKOS

<skos:Concept rdf:about="http://example.com/lcsh#95000541">
  <skos:prefLabel>World Wide Web</skos:prefLabel>
  <skos:altLabel>W3 (World Wide Web)</skos:altLabel>
  <skos:altLabel>Web (World Wide Web)</skos:altLabel>
  <skos:altLabel>World Wide Web (Information Retrieval System)</skos:altLabel>
  <skos:broader rdf:about="http://example.com/lcsh#88002671" />
  <skos:broader rdf:about="http://example.com/lcsh#92002381" />
  <skos:related rdf:about="http://example.com/lcsh#92002816"/>
  <skos:narrower rdf:about="http://example.com/lcsh#2002000569"/>
  <skos:narrower rdf:about="http://example.com/lcsh#2003001415"/>
  <skos:narrower rdf:about="http://example.com/lcsh#97003254"/>
</skos:Concept>
FIG. 2. Semantic Relationships Between Concepts.

Diagram courtesy of Ed Summers
See upcoming DC2008 Paper
Expected Benefits

• Common RDF Semantics
• Many Possible Web Services
• Publish Vocabulary in Multiple Formats
  – Ease of re-use
• Entertainment
Name Authorities

• Many National Authority Files
• Separate records representing same author
  – Different Languages
  – Different Scripts
VIAF

- Virtual International Authority File
- First try - Merging
- Second try - Linking
  (then merging?)
- Why not just link....?
Same Entity/Variant Scripts

日本語

Japanese

Японский

ejapanisch
Linking Open Names

• Need an RDF Vocabulary for Names and Corporations
• FOAF is one piece of the puzzle
• DC Agents Application Profile
  – Quasi-Active DCMI Task Group
VIAF as LOD

• Use owl:sameAs to declare equality
• Every national authority file gets a SPARQL endpoint
• No need to merge authority files
• Applications can query, merging relevant sets locally
Renew, reuse, recycle

- Enable better sharing within Library community
- Share our data with other communities
- Reuse Authority Data in new and interesting ways...
This is only an example!!

- The Graph may not be entirely correct
- Tagging ontologies are very new
- May involve blank nodes &/or reification
Controlled Vocabularies Recontextualized

- LOD notion of “Information” vs. “Non-information” resources.
  - Info - documents on the web
  - Non-info - anything else: people, places, things, books
- Non-info resources have representations / descriptions
- These are info resources
Controlled Vocabularies Recontextualized

• Authority records are descriptions of non-information resources
• Bibliographic records are (usually) descriptions of non-information resources
• Other areas of Authority Control...
Image from the Getty Museum:
http://www.getty.edu/research/conducting_research/standards/cdwa/entity.html
FRBR

- Library community’s first formalization of our data model
- Untested
- Incredibly complicated
- Not reflected well in descriptive standards or practice
FRBR

“Simply by clustering your records into work sets, you have not moved your records into the FRBR model. FRBR is a complete data model that is a new way of looking at our data, not just taking existing records and identifying work relationships”

- J. Rochkind - bibwild.wordpress.com
...and Library data is extremely complicated
MARC Record Graph

- Does not include authority data
- Coins new URI’s any non-literal value
- Contains a few minor modeling errors

```xml
<modsrdf:Publisher modsrdf:value="Crowell"
    rdf:about="http://simile.mit.edu/2006/01/publisher/Crowell">
  <modsrdf:location>
    <modsrdf:Place modsrdf:name="New York"
        rdf:about="http://simile.mit.edu/2006/01/place/marcountry/nyu"/>
  </modsrdf:location>
</modsrdf:Publisher>
```
A Distinction

• **Metadata Harmonization:**
  - the “ability to use several different metadata standards in a single software system.”

• **Metadata Normalization:**
  - mapping several different metadata standards to a single schema or structure for use in a single software system.
Primo: A Case Study

- Normalization Rules
- Delivery templates
- Tight SFX and MetaLib Integration
- “Pipes” for different data sources
- Hourly Availability Checking
  - (Real Time in Version 2.0)
Harvesting

- Different Data Sources
- Different Normalization Rules
- All standardized on Primo
  - Normalized XML (PNX) Record
    - Very Flat, sections corresponding to
      Primo Functionality
Issues and Challenges

• Managing Deduplication
  – Dedup Data only out of box for MARC
  – Writing for OAI-PMH sources (EAD)

• Consortial Environment(s)

• Appropriate Delivery Options

• “Interpreting” Metadata
EAD Records

• Archivists Toolkit
  – Previously in Access, Notepad, Excel
  – Authority Control (sort of)

• OAI-PMH Overlay

• Multiple layers of Crosswalking

• Deduping
EAD / Aleph Dedup

• Aleph Title:
  – James E. Jackson and Esther Cooper Jackson papers

• EAD Title:
Value of Dedup

• Indexing the Best of Both Worlds
• EAD Records:
  – Inventory
  – Long Biographical / Historical Notes
• MARC Data:
  – Cross References for Access Points
It shouldn’t be this hard!

• Dedup Process shouldn’t be necessary
  – Authority files should be useable within non-MARC applications
  – Merging is easier with more granularity, more homogeneity, in data sets
Endless possibilities

• This barely scratches the surface
• Authority Data is only a small part
• With more soundly modeled bibliographic and authority data...
  – Mashups
  – Web Services
  – User Profiling
  – Collaboration tools
  – Terminology Services
  – Context sensitive searching
  – Customized interfaces
  – Customized exhibits
Thanks!

Questions?

corey.harper@nyu.edu
+1 212.998.2479