On the way to Library Linked Data: The view from OCLC Research

Jeff Mixter
Lead Software Engineer
OCLC Research

Bruce Washburn
Principal Software Engineer
OCLC Research
ABOUT OCLC RESEARCH
OCLC’s Research Division

- is a *community resource* for the research-driven advancement of libraries and archives,
- provides *advanced development and consultation* for the teams that create and maintain OCLC’s products, and
- enhances OCLC’s *member and partner engagement* to mobilize the community around shared concerns.
OCLC Research Areas

The Collective Collection
Providing evidence and models that help libraries optimize access to collections.

Research Collections & Support
Exploring new modes of researcher support and collection management.

Data Science
Analyzing bibliographic data to derive new meaning, insights, and services.

User Research
Studying trends in consumer technology and learning environments and sharing evidence.

Libraries as Community Catalysts
Equipping library staff to respond to ever-evolving community needs.

Library as Enterprise
Exploring how libraries organize internally, engage with their community, and act collectively.
For more information about the people and programs of OCLC Research:

www.oclc.org/research/
THE LINKED DATA VALUE PROPOSITION
“Linked Data” refers to …

a set of best practices for publishing structured data on the Web.

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

https://www.w3.org/wiki/LinkedData
Moving from records to entities
Moving from records to entities
Linked data models reflect common-sense and in the long-term can save money

- Describe real world relationships between things
- Evolve and improve the data model inexpensively
- Avoid complex and inefficient database processes
- Obtain new skill sets and software frameworks
Quality issues with legacy data can be found and fixed when viewed as linked data

- Reveal problematic data using identifiers and rules
- Set data constraints on the model’s classes and properties
- A relational database problem can become a linked data problem unless completeness and quality is addressed
- Significant resources are needed for data modeling
Attainable, incremental Linked Data projects produce solid results and lead to long-term goals.

- Build and test manageable subsets of the complete ontology in a graph-based data model
- While iterating, work towards a more comprehensive longer-term objective

Recommended reading: "Overcoming Resistance to Technology Change: A Linked Data Perspective", Tim Williams, UCB Biosciences Inc., PhUSE EU Connect 2018
LINKED DATA RESEARCH AT OCLC
VIAF, FAST, and WorldCat: Publish Linked Data on the web with a UI, API, and downloadable datasets
EntityJS: Explore how Linked Data improves discovery of related entities
Person Entity Lookup Pilot: Test use cases and interoperability for Linked Data as a web service
CONTENTdm Metadata Refinery: Evaluate shared tools to manage the Linked Data workflow
Project Passage: Think big... Build a complete system based on Linked Data

CONTENTdm Linked Data Pilot: Think "long tail". Attend to the issues around the special and unique.

Entity Management: Build a full entity management infrastructure at global scale.
ABOUT THE CONTENTDM LINKED DATA PILOT PROJECT
What is CONTENTdm?

• OCLC’s digital content management service.
• CONTENTdm builds and showcases digital collections.
• CONTENTdm securely manages digital materials in a cloud-based system.
Research Areas of Interest

1. Divergent practice and collection assessment
2. Shared data models for diverse collections and institutions
3. Automated data processing and human attention
4. Tools for subject matter experts
5. Discovery tools to reveal hidden connections
6. The Linked Data paradigm shift
Research Areas of Interest

1. Divergent practice and collection assessment
2. Shared data models for diverse collections and institutions
3. Automated data processing and human attention
4. Tools for subject matter experts
5. Discovery tools to reveal hidden connections
6. The Linked Data paradigm shift
BUILDING A LINKED DATA MODEL
Developing a Data Model

• CONTENTdm metadata is very heterogeneous, when comparing different institution’s practices.
• There can be variability in vocabularies and data dictionaries even for different collections at the same institution.
• A single data model was needed for the Linked Data pilot project, as a single system would manage metadata for all institutions and collections.
Developing a Data Model

• The pilot project tested a hypothesis:
  – through sampling current CONTENTdm metadata, and looking for patterns and convergence, a model could be driven by data and avoid speculation.

• Properties and Classes in the data model were associated with their counterparts in other ontologies and vocabularies, to support transformation, export, and linking.
Developing a Data Model

- Metadata analysis began with an inventory of CONTENTdm mapping of local data elements to Dublin Core.
- The most frequently encountered classes and relationships were identified.
- 13 million records were analyzed to find the most widely-used properties for creative works.
- The most frequently used terms were extracted to build an initial class taxonomy for creative works.
- Collaboration with project partners and colleagues revised this method.
CONTENTdm
Data Model
Class Hierarchy
WIKIBASE AS A LINKED DATA PLATFORM
Why Wikibase?

- OCLC Research is very familiar with the Wikibase ecosystem.
- Wikibase is well-suited for the Linked Data investigations for the CONTENTdm pilot project.
- Using Wikibase helped OCLC Research and the project participants dive right in.
Why not just use Wikidata?

• Testing new properties and classes, but the short time-frame of the pilot didn’t allow for the time needed to have these proposed, reviewed, and accepted in Wikidata.

• Participants would not have time to review and approve the metadata quality or be able to commit to maintaining it in Wikidata after the pilot project was over.
Managing the Data Model in Wikibase

• Applies a workflow used in Wikidata.
• We reused the Wikidata templates for new Class and Property proposals.
• Even with a very small team, it was beneficial to establish a review practice and to remind ourselves of prior decisions.
Talk Pages for Managing Collections

- Leveraged a built-in Wikibase feature.
- A useful tool for communication and collaboration with partners.
- Helped us track differences and similarities across collections.

### Item talk:Q148309

**John W. Mosley Photograph Collection**

#### OpenRefine Project

[http://164.49.101.140:8080]  

#### OpenRefine transformation notes

<table>
<thead>
<tr>
<th>CONTENT/source field</th>
<th>Wikibase data element</th>
<th>Transformation notes</th>
<th>Example value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>title (P124)</td>
<td>Decided on using the newly created title property as staff can't easily confirm if titles are cataloger-supplied or original without access to and evaluation of each of the physical materials.</td>
<td>1946 Cleveland Buckeyes</td>
</tr>
<tr>
<td>Date</td>
<td>date created (P21)</td>
<td>Use for single dates.</td>
<td><em>1954-06-11</em>, <em>1966, September, 09</em>, <em>August 1954</em></td>
</tr>
<tr>
<td>Date</td>
<td>approximate date (P122)</td>
<td>Use for single circa dates.</td>
<td><em>circa 1965</em></td>
</tr>
<tr>
<td>Date</td>
<td>earliest date (P857)</td>
<td>Use when there is a range of dates.</td>
<td><em>1940-1949</em></td>
</tr>
<tr>
<td>Date</td>
<td>latest date (P856)</td>
<td>Use when there is a range of dates.</td>
<td><em>1940-1949</em></td>
</tr>
<tr>
<td>Photographer</td>
<td>photographer (P88)</td>
<td>Same value in all records.</td>
<td>Mosley, John W.</td>
</tr>
<tr>
<td>Subject</td>
<td>about (P2)</td>
<td>Split concatenated values, reconcile, and verify that these are generally &quot;about&quot; relationships rather than depictions.</td>
<td>Schools; Segregation; Demonstrations; Civil rights; Girard College</td>
</tr>
<tr>
<td>Organization-Building</td>
<td>about (P2)</td>
<td>Split concatenated values, reconcile, and verify that these are generally &quot;about&quot; relationships rather than depictions.</td>
<td>Girard College; State Office Building (Philadelphia, Pa.)</td>
</tr>
<tr>
<td>Intersection</td>
<td>unmapped</td>
<td>Broad and Spring Garden Streets (Philadelphia, Pa.),</td>
<td></td>
</tr>
</tbody>
</table>
Implementing Authority Control

• CONTENTdm is a traditional, record-oriented system, with authority control based on the use of heading strings.
• Varying cataloging practices and sources can create discovery obstacles if searches do not use the preferred form of the heading.
• Wikibase takes multi-lingual data seriously and allows any number of variant forms to find matches, with an assigned unique identifier serving as the key.
Decreasing cataloging inefficiencies, increasing descriptive quality

- In CONTENTdm, information about a person related to a work needs to be repeated in every record related to that person.
- If the description changes, all the records need to be updated.
- In Wikibase, entities for works and related things are maintained separately.
- The description of the person can be entered and maintained in one entity.
- This efficiency gain could encourage richer descriptions.
Generating Data Visualizations

- The triplestore that maintains a linked data representation of Wikibase can be queried using SPARQL.
- The SPARQL Query UI provided by the Wikibase ecosystem makes it easy to query and visualize connections that would have been difficult or infeasible in a traditional record-oriented system.
Adding the Mirador Viewer

• The Wikibase ecosystem has a “Gadget” extension to develop and add custom features to the user interface.
• CONTENTdm items all have IIIF Presentation Manifests.
• Mirador is an image viewer that uses IIIF Manifests to present images and metadata.
• The Gadget can embed the Mirador viewer, making the entity view more dynamic and connected to digital content.
A key Linked Data value proposition: link to related things in other systems.

Pilot project entities sometimes have related identifiers for Wikidata entities.

A Gadget, combined with a server-based application, uses Wikidata identifiers to find content in Wikipedia and supplements the entity display with contextual descriptions and images.
TOOLS FOR SUBJECT MATTER EXPERTS
Tools for Subject Matter Experts

• Differentiating between what a work is “about” and what it “depicts”.
• Testing Wikibase as a platform for tool development.
• Help subject matter experts correct and supplement transformed metadata.
### Louis Armstrong and Jake Armstrong

Daniel Louis "Satchmo" Armstrong with his son, Jake Armstrong, at the Academy of Music.

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Louis Armstrong and Jake</td>
<td>Daniel Louis &quot;Satchmo&quot; Armstrong with his son, Jake Armstrong, at the Academy of Music.</td>
</tr>
</tbody>
</table>

#### Statements

**date created**
- 22 March 1966
  - 0 references
  - add reference
  - add value

**part of**
- John W. Mosley Photograph Collection
  - 0 references
  - add reference
  - add value

**photographer**
- John W. Mosley
  - 0 references
  - add reference
  - add value

**type**
- image
  - 0 references
  - add reference
  - add value

**process or format**
- Black and white prints
  - 0 references
  - add reference
  - add value
concept of trumpet and piano music—scores and parts
African American entertainers
Jazz musicians
African American men
Entertainers
Louis Armstrong

Selmer

Henri Selmer Paris K-Modified B-flat trumpet
Trumpet manufactured by Henri Selmer, Paris, in 1934

Henri Selmer Paris
musical instrument company, a predecessor of Conn-Selmer
"Think Aloud" user study results were mostly positive.

Subject matter expertise adds new access points and enriched relationships.

"About" vs. "Depicts" remains a puzzle.
TOOLS FOR DISCOVERY
Discovery Tools

• CONTENTdm sites are individuals
• Searched separately, not collectively
• Faceted searching has challenges
• Content overlaps separate sites
Discovery Tools

• Opportunity to test cross-site discovery
• Descriptions use a shared data model
• Browse a graph of linked entities
• Reveal related content from decentralized sources
The CONTENTdm Explorer

• Focus on the most frequently occurring connections.
• See relationships described by different institutions, for different items, in different collections.
• Look for thematically related content.
• Follow links to related entities.
Truck brings employees home during P.T.C. walkout

A man helps several women into the back of a truck at 25th and Hunting Park Avenue. The truck is picking up employees from Budd Manufacturing Company in order to get them home during the P.T.C. walkout. With clipping part of CONTENTdm Transportation Hub | George D. McDowell Philadelphia Evening Bulletin Photographs classification used Photographs about Philadelphia transit strike of 1944 | Philadelphia depicts Trucks | Persons

More Like This
Protest against Philadelphia Transportation Company

A Philadelphia Transit Company (it would eventually become SEPTA) protest in support of African American trolley drivers on November 8, 1943. After the PTC chose eight Black men to drive trolleys in 1944, White transit workers went on strike, stopping the production of weapons and war materiel and leaving millions of IDers stranded. The crisis ended after President Franklin D. Roosevelt sent in 5,000 armed troops to force the strikers back to work, the first time since Reconstruction that the federal government intervened to protect the rights of Black workers.

date created
1943-11-08
height
8 inch
width
10 inch
part of
John W. Mosley Photograph Collection | CONTENTdm Transportation Hub
classification used
Photographs
process or format
Black and white prints
about
Philadelphia | African Americans | African Americans--Civil rights | Civil rights | Transportation | Trolley cars | Civil rights demonstrations | Segregation in transportation | Philadelphia Transportation Company | African Americans--Pennsylvania--Philadelphia | Protests and Social movements

More Like This
Public Square 1899 CP04179

Showing the crowd surrounding the first cars to reach the downtown section during a street car strike. These cars were on the Zucali Line. The clock at the extreme right housed the clocks of Wm. Taylor & Son, and Co.

date created
1899
height
6 1/25 inch
width
6 1/25 inch
part of
Cleveland Picture Collection | CONTENTdm Transportation Hub
classification used
Photographs

directs
Cleveland

Streetcars parked on street during a strike

Written on slide frame: "Two tram cars, 6th Street, havigna
part of
CONTENTdm Transportation Hub | Wilbur "Willy" Ginzberg Collection

classification used
Photographs
process or format
slide (photographic)

about
street car | Cubic | Halves | General strikes | Pedestrians

Halves
CONTENTdm Explorer Findings

Aggregation adds value.

The 80/20 rule was about right for automated reconciliation.

Completing the 20% will be challenging.
Testing the Linked Data value proposition
A Linked Data strategy can succeed when descriptions use a shared data model, headings are replaced with identifiers for entities, and entities and their relationships are presented in a single discovery system.

Evaluating a shared data model
A solid set of initial classes and properties can be generated from existing standards, and then responsively extended as more data needs arise.

Selecting and transforming metadata
Tools should be shared, and workflows decentralized. Domain expertise is needed for accurate data transformations.

Continuing the journey to Linked Data
Substantial resource commitments are required. Benefits accrue long before the journey is completed. It’s a paradigm shift; takes time and long-term strategies.
Working partnerships represent strength in numbers
Resources and Contacts

- The Pilot Project report
  - oc.lc/transform-linked-data

- Jeff Mixter, OCLC Research
  - jeff_mixter@oclc.org

- Bruce Washburn, OCLC Research
  - bruce_washburn@oclc.org
Shared entity management infrastructure

• Address infrastructure needs identified by libraries
  – Start with Persons and Works…
  – Expand on “native” metadata management
  – Link library data to non-library data… and shared data to local data
  – Provide ID creation services to help “at the point of need”
  – Persistent and maintained entity URIs
• Operate at a large scale – and be sustainable
• Complement other efforts—LD4, PCC, DCMI, etc.

oc.lc/sharedentitymgmt

OCLC awarded Mellon Foundation grant to develop infrastructure to support linked data management initiatives

Entity Management Infrastructure will advance use of linked data and ultimately improve discoverability of scholarly materials on the web

DUBLIN, Ohio, 9 January 2020—OCLC has been awarded a grant from The Andrew W. Mellon Foundation to develop a shared “Entity Management Infrastructure” that will support linked data management initiatives underway in the library and scholarly communications community. When complete, this infrastructure will be jointly curated by the community and OCLC, and will ultimately make scholarly materials more connected and discoverable on the web.

The two-year grant, for $2.436 million, will support work on the project that will run from January 2020 to December 2021. The Mellon grant funding represents approximately half of the total cost of the Entity Management infrastructure project. OCLC is contributing the remaining half of the required investment.

“OCLC has been a leader in library linked data research for years, and we have developed prototypes, innovative pilot programs and partnerships that continue to inform our work,” said Skip Prichard, OCLC President and CEO. “OCLC enables libraries to work together to achieve economies, efficiencies, and consistency in metadata creation. We’re grateful for The
Convene, Understand, Share