Deep semantic representation from metadata descriptions: a linked data perspective

Andrew K. Pace
Executive Director, Technical Research
OCLC
Agenda

• Research and Findings: a decade of linked data research
• A shared Entity Management Infrastructure
• Applying a linked data infrastructure to distinctive collections
• Shared Infrastructure & Local Entity Management: a juxtaposition

Acknowledgements: OCLC Colleagues Jeff Mixter, Bruce Washburn, Shane Huddleston, Jeff Young, John Chapman, et al; and the dozens of libraries who have participated in OCLC Research prototypes, experiments, and research efforts.
A decade with Linked Data

Publish linked data - FAST, VIAF, WorldCat (2009 - )

EntityJS Research Project (2013)

Person Entity Lookup Pilot (2014)

CONTENTdm Metadata Refinery (2015-16)

Project Passage (2017-18)

CONTENTdm Linked Data Pilot (2019-20)

Shared Entity Management Infrastructure (2020-21)

oc.lc/linkeddataresearch
VIAF and FAST: Publish Linked Data on the web with a UI, API, and downloadable datasets
EntityJS: Explore how Linked Data maximizes the discovery potential for sets of related entities (related by an event, a literature domain, etc.)
**Person Entity Lookup Pilot:** Test use cases and client interoperability for Linked Data as a web service
**Metadata Refinery**: Evaluate shared tools that help institutions take control of the Linked Data creation workflow
Publish linked data - FAST, VIAF, WorldCat (2009 - )

EntityJS Research Project (2013)

Person Entity Lookup Pilot (2014)

CONTENTdm Metadata Refinery (2015-16)

Project Passage (2017-18)

CONTENTdm Linked Data Pilot (2019-20)

Shared Entity Management Infrastructure (2020-21)

**Project Passage:** Think big... Build a complete system based on Linked Data, and see how workflows change
Publish linked data - FAST, VIAF, WorldCat (2009 - )

EntityJS Research Project (2013)

Person Entity Lookup Pilot (2014)

CONTENTdm Metadata Refinery (2015-16)

Project Passage (2017-18)

CONTENTdm Linked Data Pilot (2019-20)

Shared Entity Management Infrastructure (2020-21)
Current Events with Linked Data
SHARED ENTITY MANAGEMENT INFRASTRUCTURE
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 to The..."
For linked data to move into common use, libraries need reliable and persistent identifiers and metadata for the critical entities they rely on. This project begins to build that infrastructure and advances the whole field.

**Lorcan Dempsey**
OCLC Vice President, Membership and Research, and Chief Strategist
Shared Entity Management Infrastructure (2020-21)

Project goals

• Address infrastructure needs identified by libraries
  – 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.
Entity Management

Methods

• 24-month project, six-month increments
• Leverage Wikibase for 12+ months
• Multiple communication channels for input and iteration
• Division-spanning project including staff from engineering, UX research, architecture, systems, and technical research
• Multiple “workstreams” represent coherent teams
Entity Management

Communication channels

- Ad-hoc with libraries, groups (ex: PCC)
- Presentations and reports
- Ongoing with LD4P
- Entity Management Advisory Group
  - Monthly meetings
  - “Breakouts” / focus groups
  - Testing
Advisory Group Members

Shared Entity Management Infrastructure (2020-21)
Entity Management

- First increment was recently completed
  - Basic functionality
  - API and UI
  - Process, procedures, cadence

- “Findings” so far
  - Need focus: creative works and persons
  - Internal communication (especially now) takes effort
  - Scaling is a challenge
LINKED DATA FOR DISTINCTIVE COLLECTIONS

- Project Passage (2017-18)
- CONTENTdm Linked Data Pilot (2019-20)
Project Passage: Linked Data Wikibase Prototype

Project goals

• Evaluate a framework for reconciling, creating, and managing bibliographic and authority data as linked data entities and relationships.

• Build a community of users who could create and curate data in the ecosystem and imagine or propose future workflows.

• Not originally planned: Evaluate Wikibase and Wikidata as a technical platform
Findings (Part 1)

- Wikibase can be used to create structured data with a precision that exceeds current library standards.
- The platform enables user-driven ontology design but raises concerns about how to manage and maintain ontologies.
- The platform, supplemented with OCLC's enhancements and stand-alone utilities, enables librarians to see the results of their effort in a discovery interface without leaving the metadata-creation workflow.

[oc.lc/linkeddatawikibase]
Project Passage: Linked Data Wikibase Prototype

Findings (Part 2)

• Robust tools are required for local data management.
• To populate knowledge graphs with library metadata, tools that facilitate the import and enhancement of data created elsewhere are recommended.
• The pilot underscored the need for interoperability between data sources, both for ingest and export.
• The traditional distinction between authority and bibliographic data can disappear in a linked data description.
Project goals

• Developing the scalable methods and approaches needed to produce richer, state-of-the-art machine representations of entities and relationships to make visible connections that were formerly invisible.

• Prototype an application for library staff to:
  – convert existing record-based metadata into linked data by replacing strings of characters with identifiers from known authority files and local library-defined vocabularies
  – manage and publish the resulting entities and relationships
MASS AGGREGATION
Data Harvest & Explorer

• Developed a prototype of the IIIF Change Discovery API for 13 Million CONTENTdm items

• Harvested the metadata

• Reconciled string headings to linked data URIs
  – Only looked at strings that were associated with Dublin Core fields
  – Limited those strings to ones that occurred more than 2000 times

• Built a prototype application to search and explore the 13 Million harvested CONTENTdm records
  – https://researchworks.oclc.org/iiif-explorer/
Findings

• It was exciting to find unexpected things in unexpected places

• Reconciliation was limited based on
  – Scale of the data
  – Heterogenous nature of metadata
  – Algorithmic approach to matching

• Exercise in harvesting and mapping record-based data for discovery
LINKED DATA FROM SCRATCH
Data Harvest & Explorer (part 2)

- Worked with 5 CONTENTdm users and selected 3 collections from each
- Manually reviewed, mapped, and reconciled metadata
- Imported the data into a Wikibase instance for management
- Built a new prototype application to search and explore the data ingested into the Wikibase instance
1 to 10 of 10 results for louis armstrong.
Satch," was widely recognized as a founding father of jazz. His influence, as an artist and cultural icon, is universal, unmatched, and very much alive today.

date created
1944

height
8 inch

width
10 inch

part of
John W. Mosley Photograph Collection

classification used
Photographs

process or format
Black and white prints

about
Philadelphia | Entertainers | Musicians | Singers | Jazz | African Americans |
Composers | African American entertainers | Jazz musicians | African American men |
African American pioneers | African American icons | Composers—United States |
Innovators

depicts
Louis Armstrong
Louis Armstrong (Q161624)
American jazz trumpeter, composer and singer
Satchmo | Pops | Armstrong, Louis

Context and Background
Louis Armstrong (August 4, 1901 – July 6, 1971), nicknamed Satchmo or Pops, was an American trumpeter, composer, singer and occasional actor who was one of the most influential figures in jazz. His career spanned five decades, from the 1920s to the 1960s, and different eras in jazz. Coming to prominence in the 1920s as an "inventive" trumpet and cornet player, Armstrong was a foundational influence in jazz, shifting the focus of the music from collective improvisation to solo performances. With his instantly recognizable gravelly voice, Armstrong was also an influential singer, demonstrating great dexterity as an improviser, bending the lyrics and melody of a song for expressive purposes. He was also skilled at scat singing. Renowned for his charismatic stage presence and voice almost as much as for his trumpet-playing, Armstrong's influence extends well beyond jazz music, and by the end of his career in the 1960s, he was widely regarded as a profound influence on popular music in general. Armstrong was one of the first truly popular African-American entertainers to "cross over", whose skin color was secondary to his music in an America that was extremely racially divided. He rarely publicly politicized his race, often to the dismay of fellow African-Americans, but took a well-publicized stand for desegregation in the Little Rock Crisis. His artistry and personality allowed him socially acceptable access to the upper echelons of American society which were highly restricted for black men of his era.

Sources: DBpedia | Wikipedia | Wikimedia Commons
<table>
<thead>
<tr>
<th>About</th>
<th>Philadelphia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depicts</td>
<td>Entertainers</td>
</tr>
<tr>
<td></td>
<td>Musicians</td>
</tr>
<tr>
<td></td>
<td>Singers</td>
</tr>
<tr>
<td></td>
<td>Jazz</td>
</tr>
<tr>
<td></td>
<td>African Americans</td>
</tr>
<tr>
<td></td>
<td>Composers</td>
</tr>
<tr>
<td></td>
<td>African American entertainers</td>
</tr>
<tr>
<td></td>
<td>Jazz musicians</td>
</tr>
<tr>
<td></td>
<td>African American men</td>
</tr>
<tr>
<td>Category</td>
<td>References</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>African American entertainers</td>
<td>0 references</td>
</tr>
<tr>
<td>Jazz musicians</td>
<td>0 references</td>
</tr>
<tr>
<td>African American men</td>
<td>0 references</td>
</tr>
<tr>
<td>Louis Armstrong</td>
<td>0 references</td>
</tr>
<tr>
<td>Trumpet</td>
<td>0 references</td>
</tr>
</tbody>
</table>
Findings

• It takes a lot of human effort to create the structured data
• Wikibase is a powerful and flexible infrastructure for creating, managing, and curating structured data
• There is a lot of potential for enhancing existing metadata about cultural heritage items
A semantic continuum

Shared, homogeneous, and centralized entities...

...accounting for the reality of localized, heterogeneous, de-centralized collections

Machine-matching, highly automated reconciliation...

...with tools for hand-matching, semi-automated reconciliation

Well-accepted context:
Persons & Works

Granular context:
About, Depicts, Annotations, Notes

Blurs the line between bib and authority work

Blurs the line between object and context description

Custom applications and interfaces needed

This is the new Knowledge Work
Next Steps

• **Shared entity management infrastructure**
  – Targeting millions of entities in the infrastructure by December 2020
  – Complete the project! — December 2021

• **CONTENTdm Linked Data Pilot**
  – Evaluate how to better balance algorithmic record conversion with domain knowledge expertise
  – Determine how to pull apart contextual metadata and descriptive metadata
  – Explore how to leverage the new contextual metadata in end-user applications
  – OCLC Research Report forthcoming later this year

• **More Research**
  – More contextual linked data being added to the infrastructure (e.g. Places)
  – Linked data for Concepts and Subjects
  – Federated Linked Data and Local Subject description: Linking and Localization
  – Entity Alignment: knowledge graph integration from multiple sources
  – Archives & Special Collections Linked Data: moving forward from CONTENTdm prototype and the findings of OCLC’s recent report, “Archives and Special Collections Linked Data: Navigating between Notes and Nodes”
Thank you!

Andrew K. Pace
Executive Director, Technical Research
pacea@oclc.org
@andrewkpace
https://www.oclc.org/research/people/pace-andrew.html