Prototyping a Linked Data Platform for Production Cataloging Workflows

Andrew K. Pace, Executive Director, OCLC Research
Jason Kovari, Director of Cataloging & Metadata Services, Cornell University

April 13, 2018
Agenda

- OCLC: Why another linked data project?
- OCLC: What is it?
- OCLC: Who is building it?
- OCLC: How are we building it?
- Cornell: Why are we participating?
- Cornell: What use cases are we testing?
- Cornell: How could these services be potentially used?

http://oc.lc/linkeddatasummary
Gartner Hype Cycle of Emerging Technologies

- Peak of Inflated Expectations
- Trough of Disillusionment
- Plateau of Productivity
- Slope of Enlightenment

Linked Data 2015
Linked Data 2017
Linked Data 2018?
Linked Data 2020?
Why?--Efficient, impactful workflows

Today
- Searching
- Copy cataloging
- Original cataloging
- Authorities

In the future
- Amplified searching
- Adding relationships
- Entity management
- Library-sourced vocabularies
A project vision statement

Work with our members through a foundational shift in the collaborative work of libraries, communities of practice, and end-users—dramatically improving efficiency, embracing the inclusive, diverse, and earnest OCLC membership, and empowering a new and trusted knowledge work enabled by the web.
Who

Phase I Partners (Dec ’17 - Apr ‘18)
– Cornell University
– University of California, Davis

Phase II Partners (!!!!) (May ‘18 – Sep ‘18)
– American University
– Brigham Young University
– Cleveland Public Library
– Gale Cengage
– Harvard University
– Michigan State University
– National Library of Medicine
– North Carolina State University
– Northwestern University
– Princeton University
– Smithsonian Library
– Temple University
– University of Minnesota
– University of New Hampshire
– Yale University
WHAT & HOW
What

• Develop an Entity Ecosystem that facilitates:
  – Creation and editing of new entities
  – Connecting entities to the Web

• Build a community of users who can:
  – Create/Curate data in the ecosystem
  – Imagine/propose workflow uses

• Provide services to:
  – Reconcile data
  – Explore the data
What

RECONCILER

INDEX

RECONCILIATION API

LOCAL BIBLIOGRAPHIC AND AUTHORITY DATA

RANKING BY

EDITOR

DUPLICATE DETECTION

WORLDCAT CREATIVE WORK ASSOCIATION

MINTING / EDITING API

ENTITY to ENTITY RELATOR

External Client Applications

AUTHENTICATION & AUTHORIZATION

External Client Applications

Local Bibliographic and Authority Data

UI

BATCH

LOCAL BIBLIOGRAPHIC AND AUTHORITY DATA

UI

EXTERNAL CLIENT APPLICATIONS
How: A few key technologies

- Mediawiki
- Wikibase
- Refine
- Sparql
- Pywikibot
How: Disambiguating Wiki*

- **Wikipedia** – a multilingual web-based free-content encyclopedia
- **MediaWiki** - a free and open-source wiki software
- **Wikidata.org** - a collaboratively edited structured dataset used by Wikimedia sister projects and others
- **Wikibase** - a MediaWiki extension to store and manage structured data
How: MediaWiki Features

• Search/Autosuggest/APIs
• Multilingual UI
• Wikitext editor
• Change history
• Discussion pages
• Users and rights
• Watchlists
• Maintenance reports
• Etc.
How: MediaWiki+Wikibase Features

• Search/Autosuggest/APIs/Linked Data/SPARQL
• Multilingual UI
• Structured data editor
• Change history
• Discussion pages
• Users and rights
• Watchlists
• Maintenance reports
• Etc.
How: Wikibase advantages

• Open source
• An all-purpose data model that takes knowledge diversity, sources, and multilingual usage seriously
• Collaborative – can be read and edited by both humans and machines
• User-defined properties
• Version history
A few key terms

- **Entity** – the content of a page in the system that represents an item or a property.
- **Item** -- a real-world object, concept, or event that is given a unique system identifier together with information about it. E.g., the book titled “Sense and Sensibility” by Jane Austen is an item entity.
  - Items include an identifying "fingerprint" of labels, descriptions, and aliases. The main data part of an item is the list of statements about the item.
- **Property** -- each statement on an item page links to a property, and assigns the property one or more values. E.g., “author” is a property entity.
  - Property entity pages specify the property’s assigned datatype and other statements.
A few key terms

- **Statement** -- a piece of data about an item, recorded on the item's page.
  - A statement consists of a **claim**, and may be augmented with references (giving the source for the claim) and a rank (used to distinguish between several claims containing the same property).

- **Claim** -- a piece of data about the entity on whose page the claim appears.
  - A claim consists of a property (such as “author”) and either a value (e.g., “Jane Austen”) or one of the special cases "no value" and "unknown value". A claim can have qualifiers, such as temporal qualifiers saying that the claim is valid within a specific time frame.
### Amelia Earhart

**American aviation pioneer and author**

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
<th>Also known as</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Amelia Earhart</td>
<td>American aviation pioneer and author</td>
<td>Amelia Mary Earhart</td>
</tr>
<tr>
<td>German</td>
<td>Amelia Earhart</td>
<td>US-amerikanische Flugpioniere und Frauenrechtler</td>
<td>Amelia Mary Earhart</td>
</tr>
<tr>
<td>Spanish</td>
<td>Amelia Earhart</td>
<td>avacora estadounidense</td>
<td>La avadora</td>
</tr>
<tr>
<td>Traditional Chinese</td>
<td>No label defined</td>
<td>No description defined</td>
<td></td>
</tr>
</tbody>
</table>

### Statements

- **instance of**
  - person
    - 0 references

- **employer**
  - Brigham Young University
    - 0 references

- **sex or gender**
  - female
    - 0 references

- **place of death**
  - Pacific Ocean
    - 0 references
<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
<th>Also known as</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Amelia Earhart</td>
<td>American aviation pioneer and author</td>
<td>Amelia Mary Earhart</td>
</tr>
<tr>
<td>German</td>
<td>Amelia Earhart</td>
<td>US-amerikanische Flugpionierin und Frauenrechtler</td>
<td>Amelia Mary Earhart</td>
</tr>
<tr>
<td>Spanish</td>
<td>Amelia Earhart</td>
<td>avvadora estadounidense</td>
<td>Amelia Mary Earhart</td>
</tr>
<tr>
<td>Traditional Chinese</td>
<td>No label defined</td>
<td>No description defined</td>
<td></td>
</tr>
</tbody>
</table>

**Statements**

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>instance of</td>
<td>person</td>
</tr>
<tr>
<td>employer</td>
<td>Brigham Young University</td>
</tr>
<tr>
<td>sex or gender</td>
<td>female</td>
</tr>
<tr>
<td>place of death</td>
<td>Pacific Ocean</td>
</tr>
</tbody>
</table>
Use case: Manual data entry

- For manual creation and editing of entities, **Wikibase** is the default technology.
- It has a powerful and well-tested set of features that speed the data entry process and assist with quality control and data integrity.
Jane Austen

English novelist
Austen, Jane

Death date
18 July 1817

Place of death

References
1 reference

Statements

Place of death

References

Instance of
person

References

All entered languages
Language
English
German
French
Spanish

In more languages

All entered languages

In other languages

Add links

References

2 references

Stated in
Concise Literary Encyclopedia

Sourcing circumstances
unspecified calendar, assumed gregorian

Stated in
data.bnf.fr

Retrieved
2 February 2018

Reference URL
# Revision history of "Jane Austen" (Q664501)

**Search logs for this page**

- **From year (and earlier):** 2018  
  - **From month (and earlier):** all  
  - **Tag filter:**

---

**Diff selection:** Mark the radio boxes of the revisions to compare and hit enter or the button at the bottom.

**Legend:** (cur) = difference with latest revision, (prev) = difference with preceding revision, m = minor edit.

**Compare selected revisions**

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>User</th>
<th>Action</th>
<th>Difference</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.34, 13 March</td>
<td>2018</td>
<td>Admin</td>
<td>(talk</td>
<td>contribs)</td>
<td>(21,632 bytes) (+81)</td>
</tr>
<tr>
<td>18.24, 28 February</td>
<td>2018</td>
<td>Admin</td>
<td>(talk</td>
<td>contribs)</td>
<td>(21,551 bytes) (+428)</td>
</tr>
<tr>
<td>18.22, 28 February</td>
<td>2018</td>
<td>Admin</td>
<td>(talk</td>
<td>contribs)</td>
<td>(21,123 bytes) (-336)</td>
</tr>
<tr>
<td>17.59, 15 February</td>
<td>2018</td>
<td>Admin</td>
<td>(talk</td>
<td>contribs)</td>
<td>(21,459 bytes) (+351)</td>
</tr>
<tr>
<td>16.08, 8 February</td>
<td>2018</td>
<td>Btwashburn</td>
<td>(talk</td>
<td>contribs)</td>
<td>(21,106 bytes) (+5)</td>
</tr>
<tr>
<td>00.00, 7 February</td>
<td>2018</td>
<td>Btwashburn</td>
<td>(talk</td>
<td>contribs)</td>
<td>(21,103 bytes) (+760)</td>
</tr>
<tr>
<td>23.50, 6 February</td>
<td>2018</td>
<td>Btwashburn</td>
<td>(talk</td>
<td>contribs)</td>
<td>(20,343 bytes) (+1,079)</td>
</tr>
<tr>
<td>23.43, 6 February</td>
<td>2018</td>
<td>Btwashburn</td>
<td>(talk</td>
<td>contribs)</td>
<td>(19,264 bytes) (+312)</td>
</tr>
<tr>
<td>23.39, 6 February</td>
<td>2018</td>
<td>Btwashburn</td>
<td>(talk</td>
<td>contribs)</td>
<td>(18,952 bytes) (+448)</td>
</tr>
<tr>
<td>18.59, 6 February</td>
<td>2018</td>
<td>Btwashburn</td>
<td>(talk</td>
<td>contribs)</td>
<td>(18,504 bytes) (0)</td>
</tr>
<tr>
<td>18.58, 6 February</td>
<td>2018</td>
<td>Btwashburn</td>
<td>(talk</td>
<td>contribs)</td>
<td>(18,504 bytes) (0)</td>
</tr>
<tr>
<td>05.45, 1 February</td>
<td>2018</td>
<td>ClaimAdder</td>
<td>(talk</td>
<td>contribs)</td>
<td>(18,504 bytes) (+1,515)</td>
</tr>
<tr>
<td>05.20, 13 January</td>
<td>2018</td>
<td>ClaimAdder</td>
<td>(talk</td>
<td>contribs)</td>
<td>(16,989 bytes) (+2,370)</td>
</tr>
<tr>
<td>07.27, 10 December</td>
<td>2017</td>
<td>HelloWikiBot</td>
<td>(talk</td>
<td>contribs)</td>
<td>(14,619 bytes) (+14,619)</td>
</tr>
</tbody>
</table>
Use case: Autosuggest

Searching for entities as you type is supported by the Mediawiki API. This feature is found in both the prototype UI and in the SPARQL Query Service UI.
Use case: Complex queries

SPARQL (pronounced "sparkle") is an RDF query language ... a semantic query language for databases. The prototype provides a **SPARQL endpoint**, including a user-friendly interface for constructing queries. With SPARQL you can extract any kind of data, with a query composed of logical combinations of triples.

In this example SPARQL query, items describing people born between 1800 and 1880, but without a specified death date, are listed.
Reconciling strings to a ranked list of potential entities is a key use case to be supported.

We are testing an OpenRefine-optimized Reconciliation API endpoint for this use case.

The Reconciliation API uses the prototype’s Mediawiki API and SPARQL endpoint in a hybrid tandem to find and rank matches.
<table>
<thead>
<tr>
<th>Column 1: judgment</th>
<th>Column 1: best candidate's score</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Name, Count</td>
<td>100.00 — 101.00</td>
</tr>
</tbody>
</table>

2 rows
Show as: rows records
Show: 5 10 25 50 rows

1. Austin (100)
   Austin Motor Company (100)
   Austin (100)
   Coe Finch Austin (100)
   Andrew D. Austin (100)
   Austin (100)
   Austin (100)
   Austin County (73)
   Austin Peay (71)
   Post Malone (71)
   Austin Augustus King (71)
   Austin Puck (71)
   Austin Chick (67)
   Austin Blair (67)
   Austin Adams (67)
   Austin Osman Spare (67)
   Austin Basis (67)
   Austin-Bergstrom International Airport (67)
   Austin Warren (63)
   Austin F. Pike (63)
   Henry Austin Dobson (63)
   Austin Wright (63)
   Austin Martin (63)
   Austin Clarke (63)

Create new item
Use case: Batch loading

- For batch loading new items and properties, and subsequent batch updates and deletions, OCLC staff use **Pywikibot**.
- It is a Python library and collection of scripts that automate work on MediaWiki sites. Originally designed for Wikipedia, it is now used throughout the Wikimedia Foundation's projects and on many other wikis.
<table>
<thead>
<tr>
<th>Lessons Learned and concerns so far</th>
<th>Next Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Mediawiki-based API is not sufficient for reconciliation</td>
<td>Provide an OpenRefine API for matching by class and properties</td>
</tr>
<tr>
<td>The prototype data model for dates is capable but not user friendly</td>
<td>Document techniques for entering dates, mapping to LC's EDTF patterns</td>
</tr>
<tr>
<td>The prototype UI doesn't highlight connections to more information on the web</td>
<td>Prototype a UI that uses system data to connect to Dbpedia, Geonames, etc.</td>
</tr>
<tr>
<td>Autosuggested links aren't working well for personal names in indirect order</td>
<td>Add more aliases to the Wikibase to improve autosuggest matching, based on headings in VIAF</td>
</tr>
<tr>
<td>It's not yet clear how to handle creative works and editions in the prototype</td>
<td>Provide guidance and examples, beginning with works and translations</td>
</tr>
<tr>
<td>Will Wikibase / Wikidata scale to billions of entities?</td>
<td>Fruitful discussions with Wikimedia Deutschland started</td>
</tr>
</tbody>
</table>
The Why:
Cornell's Motivations and Potential Uses
Motivation: Complementary Effort #1

- Local authority management system
- National Strategy for Shareable Local Name Authorities National Forum

Local entities
Motivation : Complementary Effort #2

Program for Cooperative Cataloging

Minting person and organization identities
Motivation: Complementary Effort #3

Look-up services within cataloging environments
Motivation: Complementary Effort #4

URLs in MARC records

URI

URL

URN
Motivation: Complementary Effort #5

folio
future of libraries is open

New ILS affords new opportunities
Low-threshold entity creation

Streamlining workflows across processes

Reconciliation services in MARC-2-RDF conversion

Data exchange questions in LD environment
Finally...

What's in it for us (condensed)?
Questions?

Jason Kovari
jak473@cornell.edu

Andrew K. Pace
pacea@oclc.org