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
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

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
<table>
<thead>
<tr>
<th>Linked Data Buzzword BINGO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silo</td>
</tr>
<tr>
<td>Resource</td>
</tr>
<tr>
<td>Data Hub</td>
</tr>
<tr>
<td>Directed Graph</td>
</tr>
<tr>
<td>OpenRefine</td>
</tr>
</tbody>
</table>
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
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

#### Labels

<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 Frauenrechtlerin</td>
<td>Amelia Mary Earhart</td>
</tr>
<tr>
<td>Spanish</td>
<td>Amelia Earhart</td>
<td>avacora estadounidense</td>
<td>La avacora</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>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>employer</td>
<td>Brigham Young University</td>
</tr>
<tr>
<td>place of death</td>
<td>Pacific Ocean</td>
</tr>
<tr>
<td>sex or gender</td>
<td>female</td>
</tr>
<tr>
<td>instance of</td>
<td>person</td>
</tr>
</tbody>
</table>
FUNCTIONAL USE CASES
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.
Revision history of "Jane Austen" (Q664501)

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

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

16:34, 13 March 2018 Admin (talk | contribs) (21,632 bytes) (+81) (Setting [en] alias: Austen, Jane)
18:24, 28 February 2018 Admin (talk | contribs) (21,551 bytes) (+428) (Created claim: notable work (P137): Persuasion (Q315999))
18:22, 28 February 2018 Admin (talk | contribs) (21,123 bytes) (-336) (Removed claim: ISNI ID (P40): 000000012283635X)
17:59, 15 February 2018 Admin (talk | contribs) (21,459 bytes) (+351) (Created claim: SHARE-VDE ID (P145): Agent/2568128)
16:06, 8 February 2018 Btwashburn (talk | contribs) (21,106 bytes) (+5) (Changed claim: death date (P10): 24 July 1877)
16:00, 7 February 2018 Btwashburn (talk | contribs) (21,103 bytes) (+760) (Changed claim: death date (P10): 24 July 1877)
23:50, 6 February 2018 Btwashburn (talk | contribs) (20,343 bytes) (+1,079) (Changed claim: death date (P10): 24 July 1877)
23:39, 6 February 2018 Btwashburn (talk | contribs) (18,952 bytes) (+448) (Changed claim: death date (P10): 18 July 1817)
18:59, 6 February 2018 Btwashburn (talk | contribs) (18,504 bytes) (0) (Changed claim: death date (P10): 18 July 1817)
18:58, 6 February 2018 Btwashburn (talk | contribs) (18,504 bytes) (0) (Changed claim: death date (P10): 19 July 1817)
05:45, 1 February 2018 ClaimAdder (talk | contribs) (18,504 bytes) (+1,515) (Changed an item: Updating timeclaims)
05:20, 13 January 2018 ClaimAdder (talk | contribs) (16,898 bytes) (+2,370) (Changed an item: Adding claims)
07:27, 10 December 2017 HelloWikiBot (talk | contribs) (14,619 bytes) (+14,619) (Created a new item: Creating entity)
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.
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.

Use case: Complex queries

In this example SPARQL query, items describing people born between 1800 and 1880, but without a specified death date, are listed.
Use case: Reconciliation

- 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.
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

PCC: Program for Cooperative Cataloging

&

ISNI: Minting person and organization identities
Motivation: Complementary Effort #3

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

URIs in MARC records

URI

\&

URL & URN

URLs in MARC records
Motivation: Complementary Effort #5

New ILS affords new opportunities
Hopes & Dreams

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