CNI Spring 2018 Membership Meeting

Prototyping a Linked Data Platform for Production Cataloging Workflows

April 13, 2018

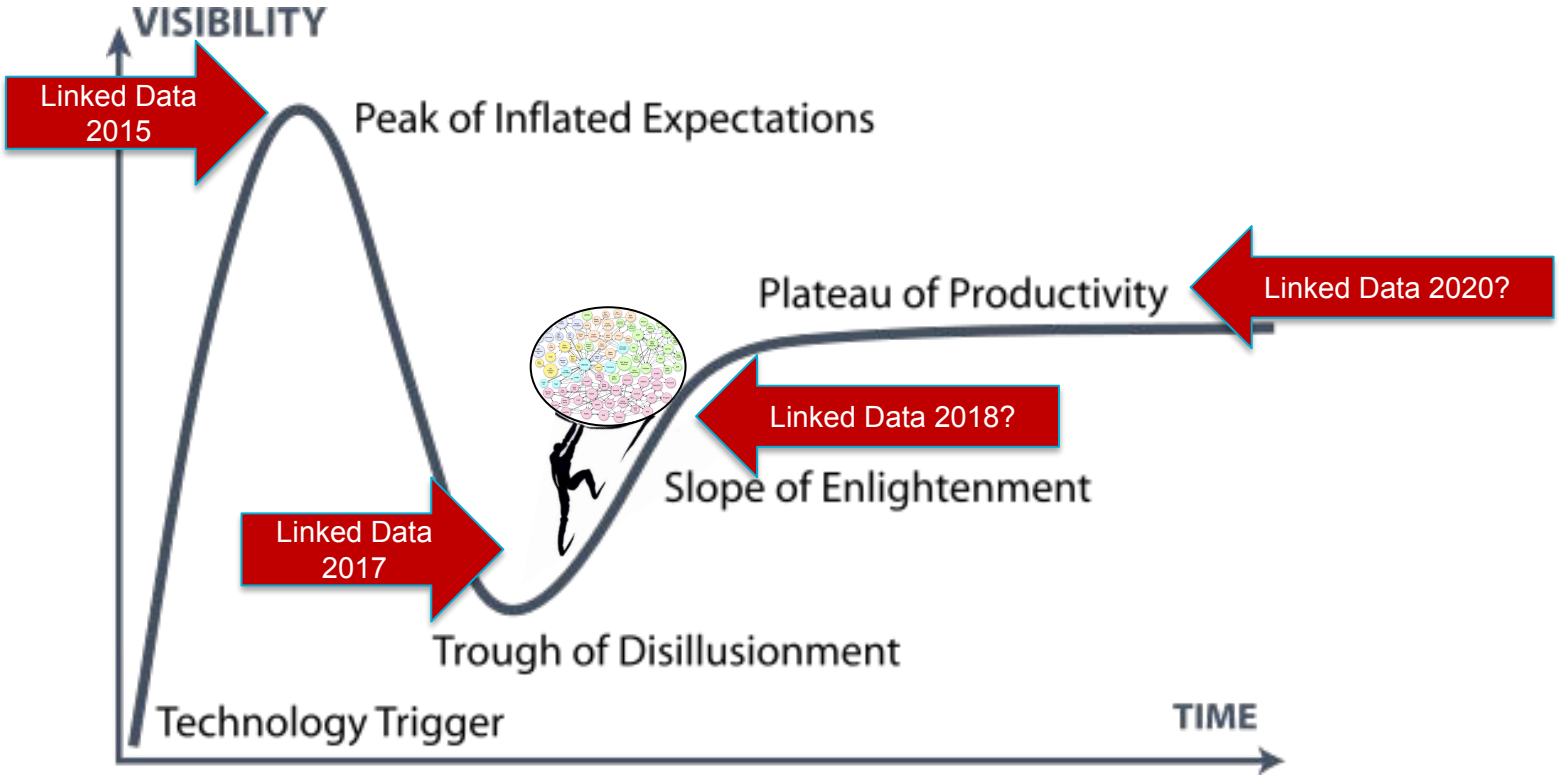
Andrew K. Pace, Executive Director, OCLC Research
Jason Kovari, Director of Cataloging & Metadata Services, Cornell University

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

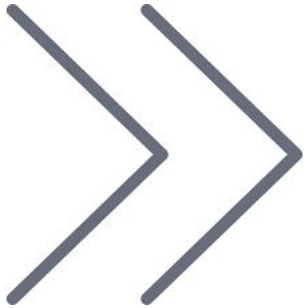
Gartner Hype Cycle of Emerging Technologies



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)

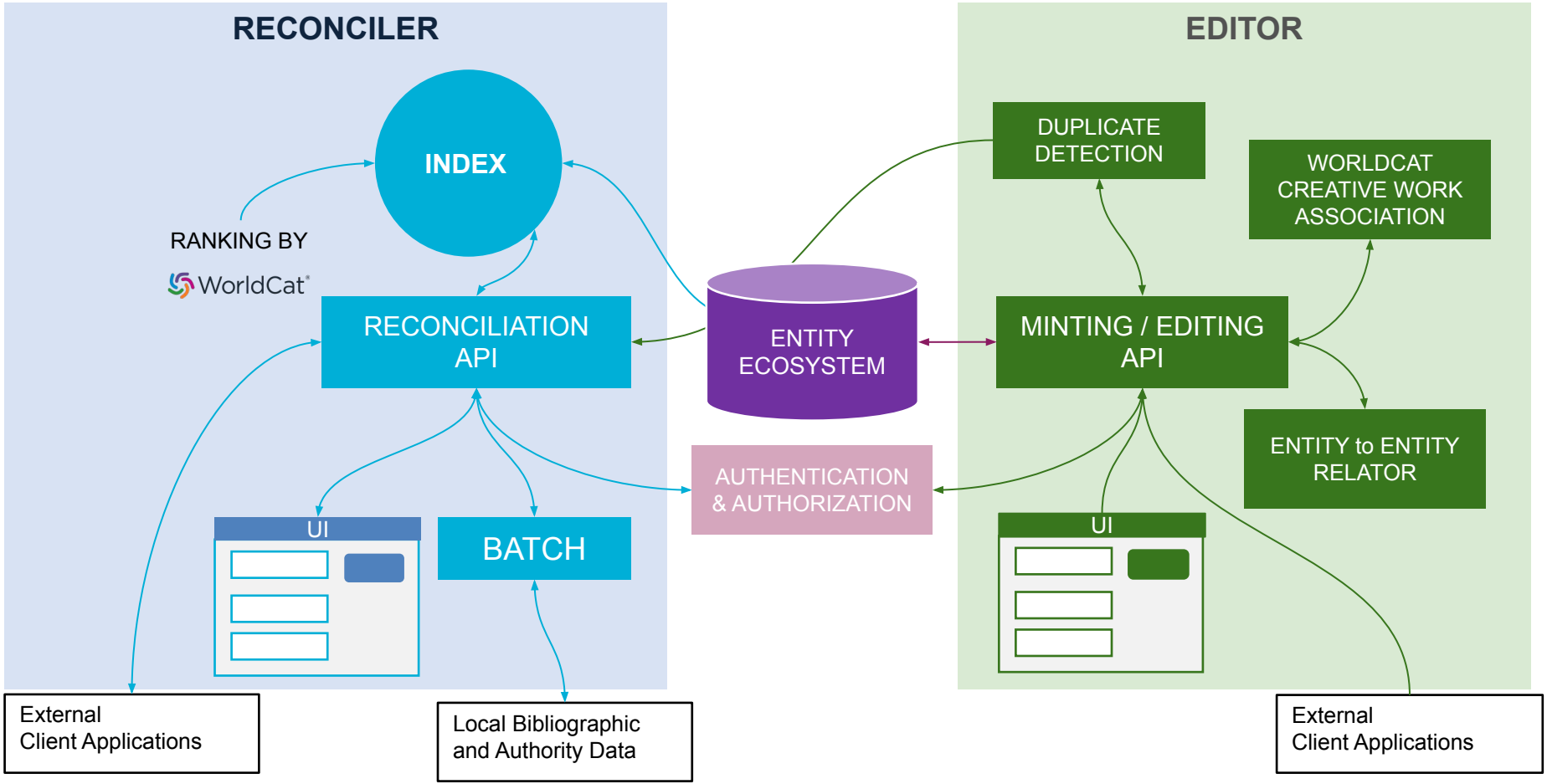
- American University
- Brigham Young University
- Cleveland Public Library
- Harvard University
- Michigan State University
- National Library of Medicine
- North Carolina State University
- Northwestern University
- Princeton University
- Smithsonian Library
- Temple University
- University of Minnesota
- 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
RANKING BY
WorldCat
RECONCILIATION API
UI
BATCH
External Client Applications
Local Bibliographic and Authority Data
ENTITY ECOSYSTEM
AUTHENTICATION & AUTHORIZATION
EDITOR
DUPLICATE DETECTION
WORLDCAT CREATIVE WORK ASSOCIATION
MINTING / EDITING API
ENTITY to ENTITY RELATOR
UI
External Client Applications

How: A few key technologies

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.

http:// oclc.url.org;/entity/Q585819

Amelia Earhart

American aviation pioneer and author
Amelia Mary Earhart

+ In more languages Configure

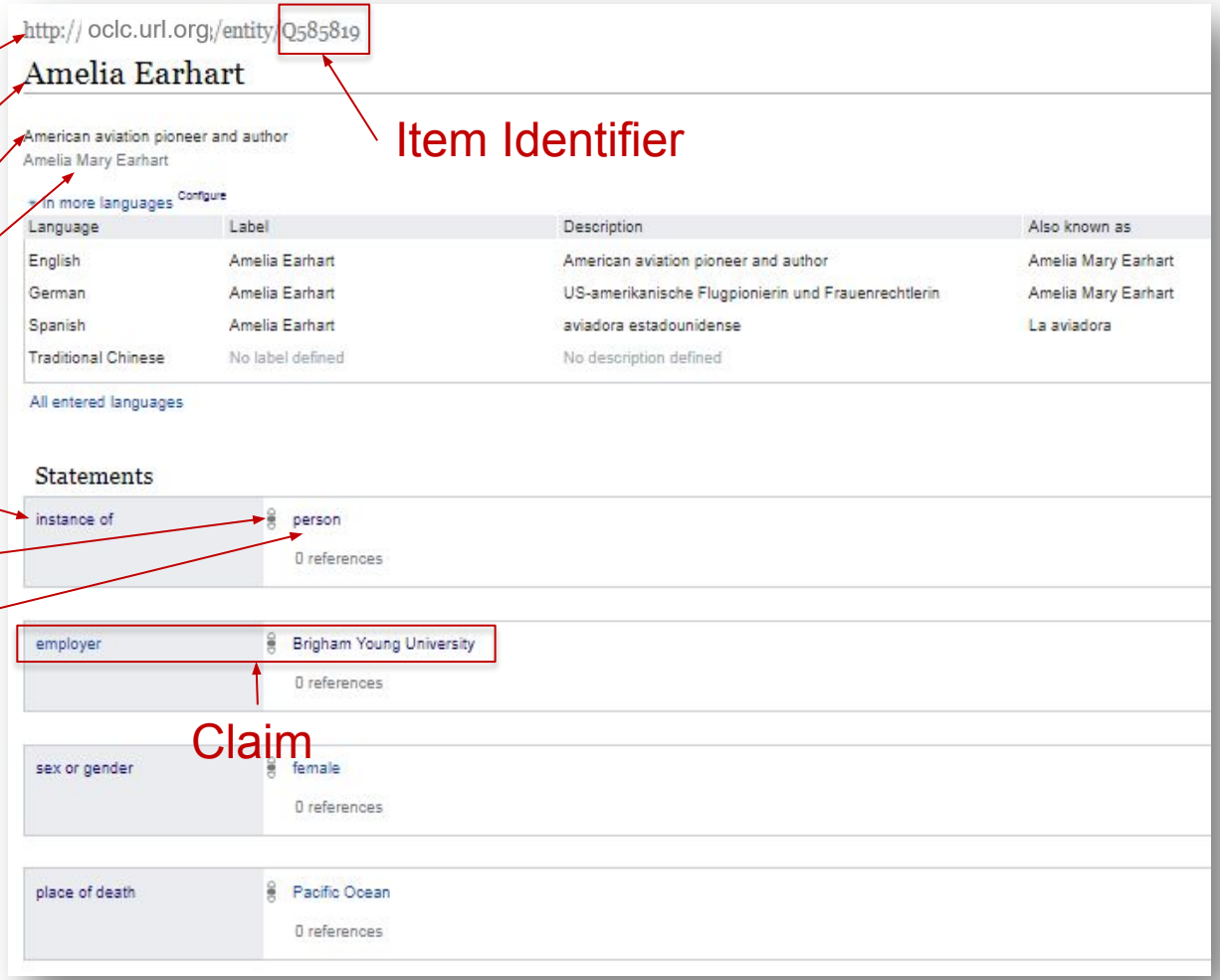
Language	Label	Description	Also known as
English	Amelia Earhart	American aviation pioneer and author	Amelia Mary Earhart
German	Amelia Earhart	US-amerikanische Flugpionierin und Frauenrechtlerin	Amelia Mary Earhart
Spanish	Amelia Earhart	aviadora estadounidense	La aviadora
Traditional Chinese	No label defined	No description defined	

All entered languages

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

OCLC



Item URL

Label

Description

Aliases

Item Identifier

Additional labels, descriptions, and aliases, in other languages.

Property

Rank

Value

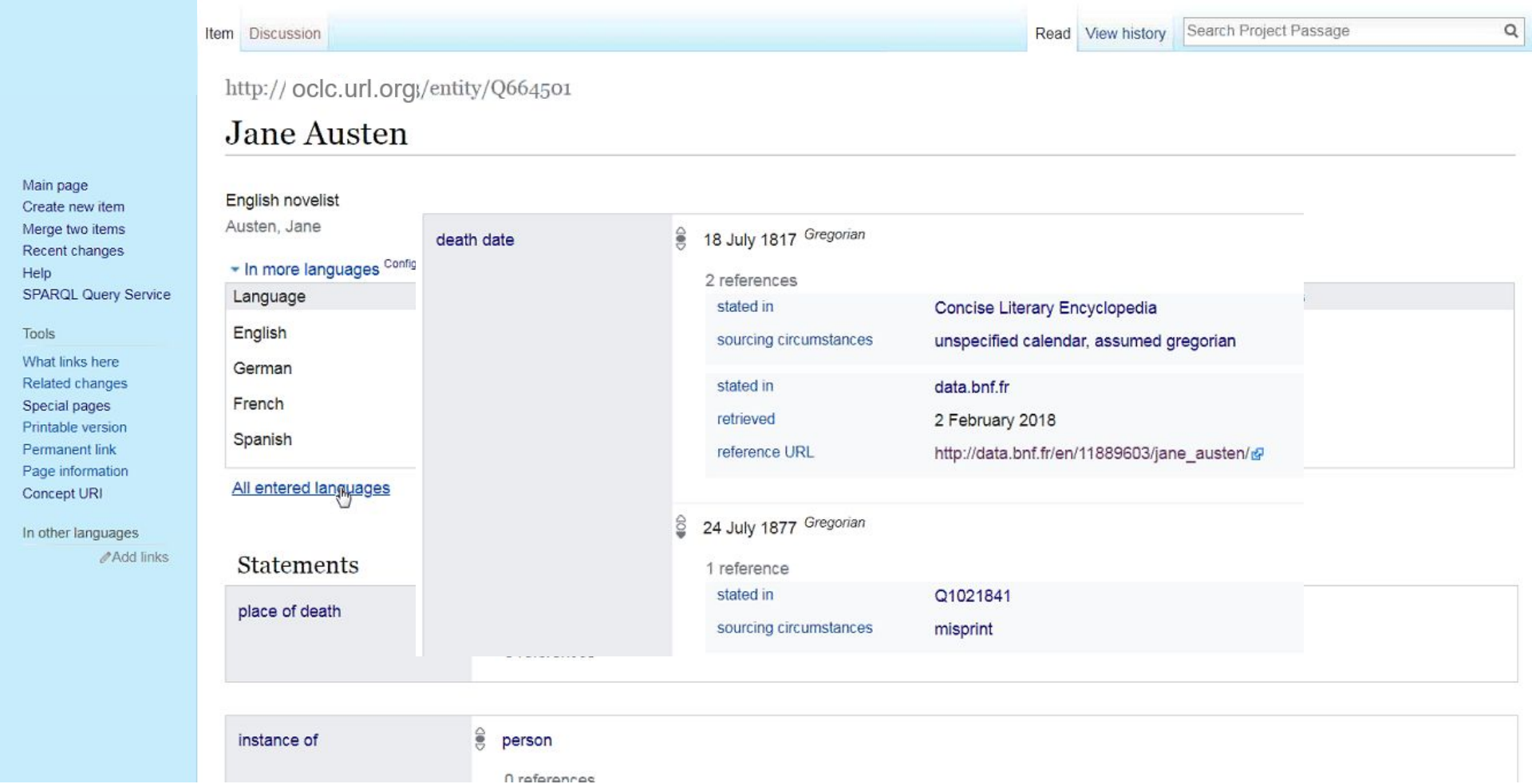
Statement

Claim

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.


Item
Discussion
Read
View history
Search Project Passage
http:// oclc.url.org/entity/Q664501
Jane Austen
English novelist
Austen, Jane
In more languages
Language
English
German
French
Spanish
All entered languages
Statements
place of death
death date
18 July 1817 Gregorian
2 references
stated in
Concise Literary Encyclopedia
sourcing circumstances
unspecified calendar, assumed gregorian
stated in
data.bnf.fr
retrieved
2 February 2018
reference URL
http://data.bnf.fr/en/11889603/jane_austen/
24 July 1877 Gregorian
1 reference
stated in
Q1021841
sourcing circumstances
misprint
instance of
person
Main page
Create new item
Merge two items
Recent changes
Help
SPARQL Query Service
Tools
What links here
Related changes
Special pages
Printable version
Permanent link
Page information
Concept URI
In other languages
Add links

Revision history of "Jane Austen" (Q664501)

Help

View logs for this page

Main page
Create new item
Merge two items
Recent changes
Help
SPARQL Query Service

Tools

What links here
Related changes
Atom
Special pages
Page information

Search for revisions

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

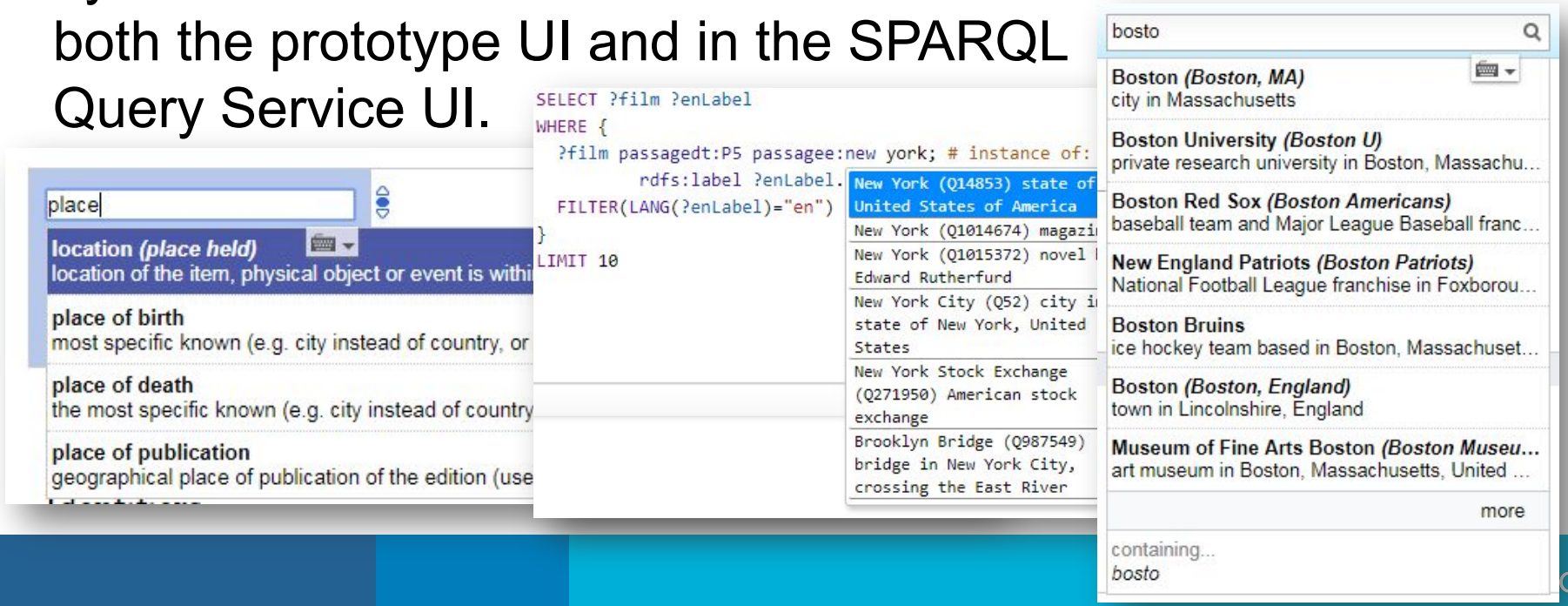
Diff selection: Mark the radio boxes of the revisions to compare and hit enter or the button at the bottom.
Legend: **(cur)** = difference with latest revision, **(prev)** = difference with preceding revision, **m** = minor edit.

Compare selected revisions

- (cur | prev) 16:34, 13 March 2018 Admin (talk | contribs) . . (21,632 bytes) (+81) . . *(Setting [en] alias: Austen, Jane)*
- (cur | prev) 18:24, 28 February 2018 Admin (talk | contribs) . . (21,551 bytes) (+428) . . *(Created claim: notable work (P137): Persuasion (Q315999))*
- (cur | prev) 18:22, 28 February 2018 Admin (talk | contribs) . . (21,123 bytes) (-336) . . *(Removed claim: ISNI ID (P40): 000000012283635X)*
- (cur | prev) 17:59, 15 February 2018 Admin (talk | contribs) . . (21,459 bytes) (+351) . . *(Created claim: SHARE-VDE ID (P145): Agent/2568128)*
- (cur | prev) 16:06, 8 February 2018 Btwashburn (talk | contribs) . . (21,108 bytes) (+5) . . *(Changed claim: death date (P10): 24 July 1877)*
- (cur | prev) 00:00, 7 February 2018 Btwashburn (talk | contribs) . . (21,103 bytes) **(+760)** . . *(Changed claim: death date (P10): 24 July 1877)*
- (cur | prev) 23:50, 6 February 2018 Btwashburn (talk | contribs) . . (20,343 bytes) **(+1,079)** . . *(Changed claim: death date (P10): 18 July 1817)*
- (cur | prev) 23:43, 6 February 2018 Btwashburn (talk | contribs) . . (19,264 bytes) (+312) . . *(Changed claim: death date (P10): 18 July 1817)*
- (cur | prev) 23:39, 6 February 2018 Btwashburn (talk | contribs) . . (18,952 bytes) (+448) . . *(Changed claim: death date (P10): 18 July 1817)*
- (cur | prev) 18:59, 6 February 2018 Btwashburn (talk | contribs) . . (18,504 bytes) (0) . . *(Changed claim: death date (P10): 18 July 1817)*
- (cur | prev) 18:58, 6 February 2018 Btwashburn (talk | contribs) . . (18,504 bytes) (0) . . *(Changed claim: death date (P10): 19 July 1817)*
- (cur | prev) 05:45, 1 February 2018 ClaimAdder (talk | contribs) . . (18,504 bytes) **(+1,515)** . . *(Changed an item: Updating timeclaims)*
- (cur | prev) 05:20, 13 January 2018 ClaimAdder (talk | contribs) . . (16,989 bytes) **(+2,370)** . . *(Changed an item: Adding claims)*
- (cur | prev) 07:27, 10 December 2017 HelloWikiBot (talk | contribs) . . (14,619 bytes) **(+14,619)** . . *(Created a new item: Creating entity)*

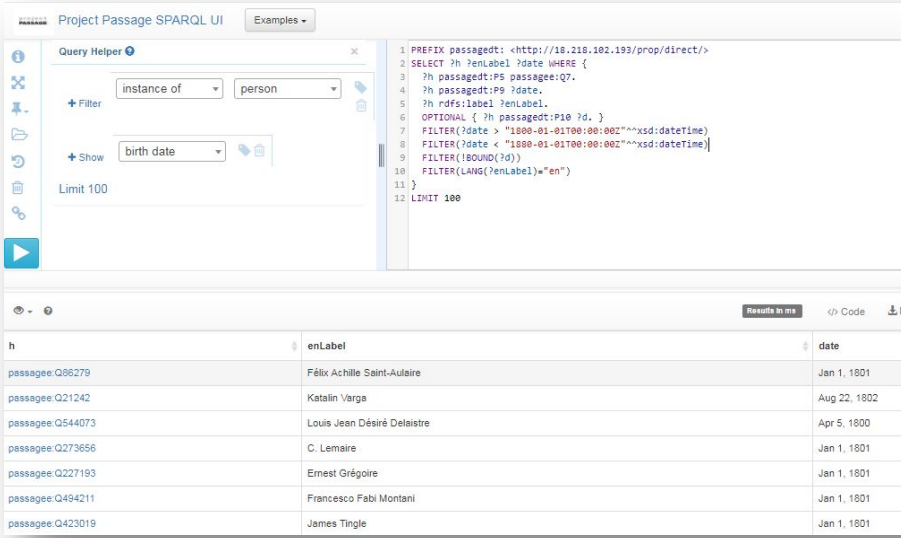
Use case: Autosuggest

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.



Use case: Complex queries

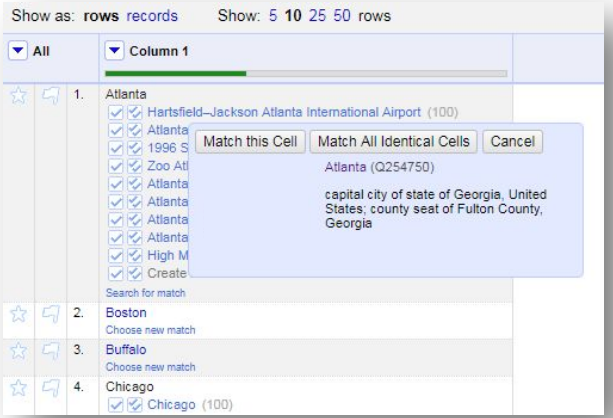
SPARQL (pronounced "sparkle") is an RDF query language … a semantic query language for databases. The prototype provides a **SPARQL endpoint**, including a user-friendly interface for constructing queries. With SPARQL you can extract any kind of data, with a query composed of logical combinations of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

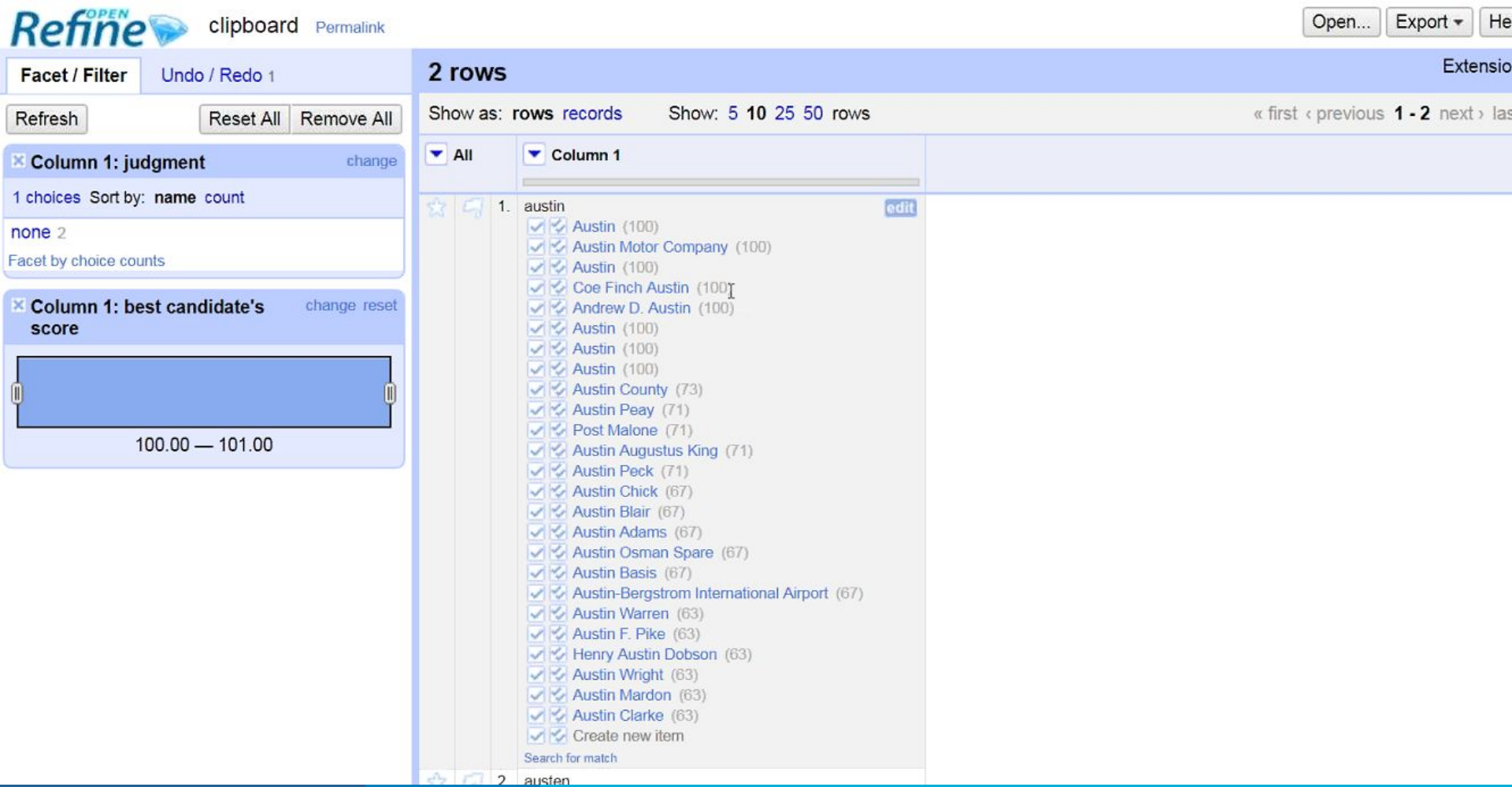


In this example SPARQL query, items describing people born between 1800 and 1880, but without a specified death date, are listed.

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.

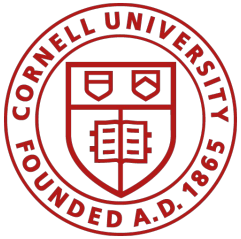



OpenRefine
clipboard
Permalink
Open...
Export
Facet / Filter
Undo / Redo 1
Refresh
Reset All
Remove All
Column 1: judgment
change
1 choices Sort by: name count
none 2
Facet by choice counts
Column 1: best candidate's score
change reset
100.00 — 101.00
2 rows
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 2 next ›
All
Column 1
1. austin
edit
Austin (100)
Austin Motor Company (100)
Austin (100)
Coe Finch Austin (100)
Andrew D. Austin (100)
Austin (100)
Austin (100)
Austin (100)
Austin County (73)
Austin Peay (71)
Post Malone (71)
Austin Augustus King (71)
Austin Peck (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 Mardon (63)
Austin Clarke (63)
Create new item
Search for match
2. austen

Use case: Batch loading

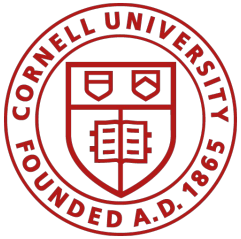
- For batch loading new items and properties, and subsequent batch updates and deletions, OCLC staff use **Pywikibot**.
- It is a Python library and collection of scripts that automate work on MediaWiki sites. Originally designed for Wikipedia, it is now used throughout the Wikimedia Foundation's projects and on many other wikis.

Lessons Learned and concerns so far	Next Steps
The Mediawiki-based API is not sufficient for reconciliation	Provide an OpenRefine API for matching by class and properties
The prototype data model for dates is capable but not user friendly	Document techniques for entering dates, mapping to LC's EDTF patterns
The prototype UI doesn't highlight connections to more information on the web	Prototype a UI that uses system data to connect to Dbpedia, Geonames, etc.
Autosuggested links aren't working well for personal names in indirect order	Add more aliases to the Wikibase to improve autosuggest matching, based on headings in VIAF
It's not yet clear how to handle creative works and editions in the prototype	Provide guidance and examples, beginning with works and translations
Will Wikibase / Wikidata scale to billions of entities?	Fruitful discussions with Wikimedia Deutschland started



The Why:

Cornell's Motivations and Potential Uses



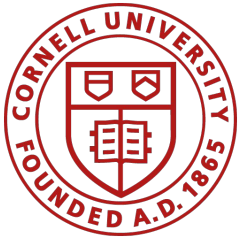
Motivation : Complementary Effort #1

- Local authority management system

vitro

- National Strategy for Shareable Local Name Authorities National Forum

Local entities



Motivation : Complementary Effort #2

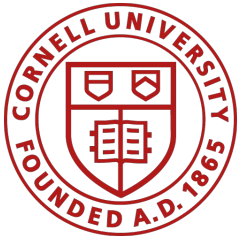
PCC

Program for Cooperative Cataloging

&

isni

Minting person and organization identities



Motivation : Complementary Effort #3

LD4L Labs

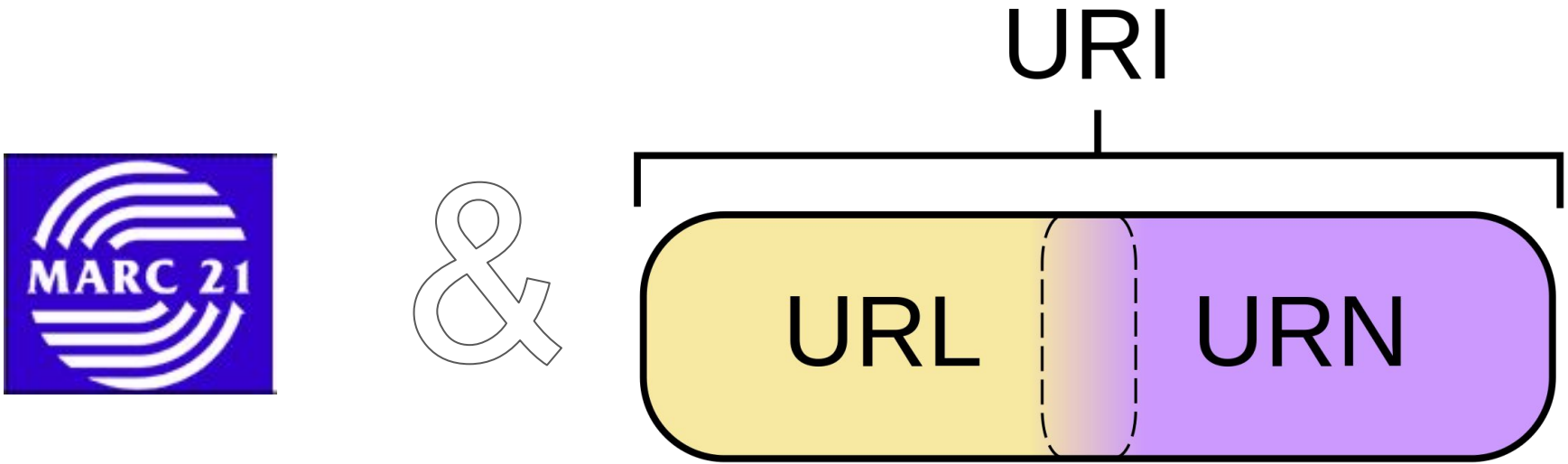
LD4L Linked Data for Libraries

THE ANDREW W. MELLON FOUNDATION

LD4P Linked Data for Production

Look-up services within cataloging environments

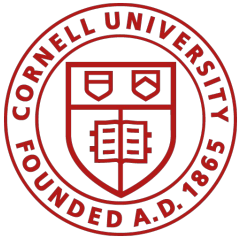
Motivation : Complementary Effort #4



URIs in MARC records

Motivation : Complementary Effort #5

New ILS affords new opportunities



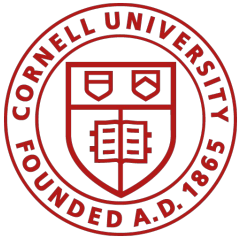
Hopes & Dreams

Low-threshold entity creation

Streamlining workflows across processes

Reconciliation services in MARC-2-RDF conversion

Data exchange questions in LD environment



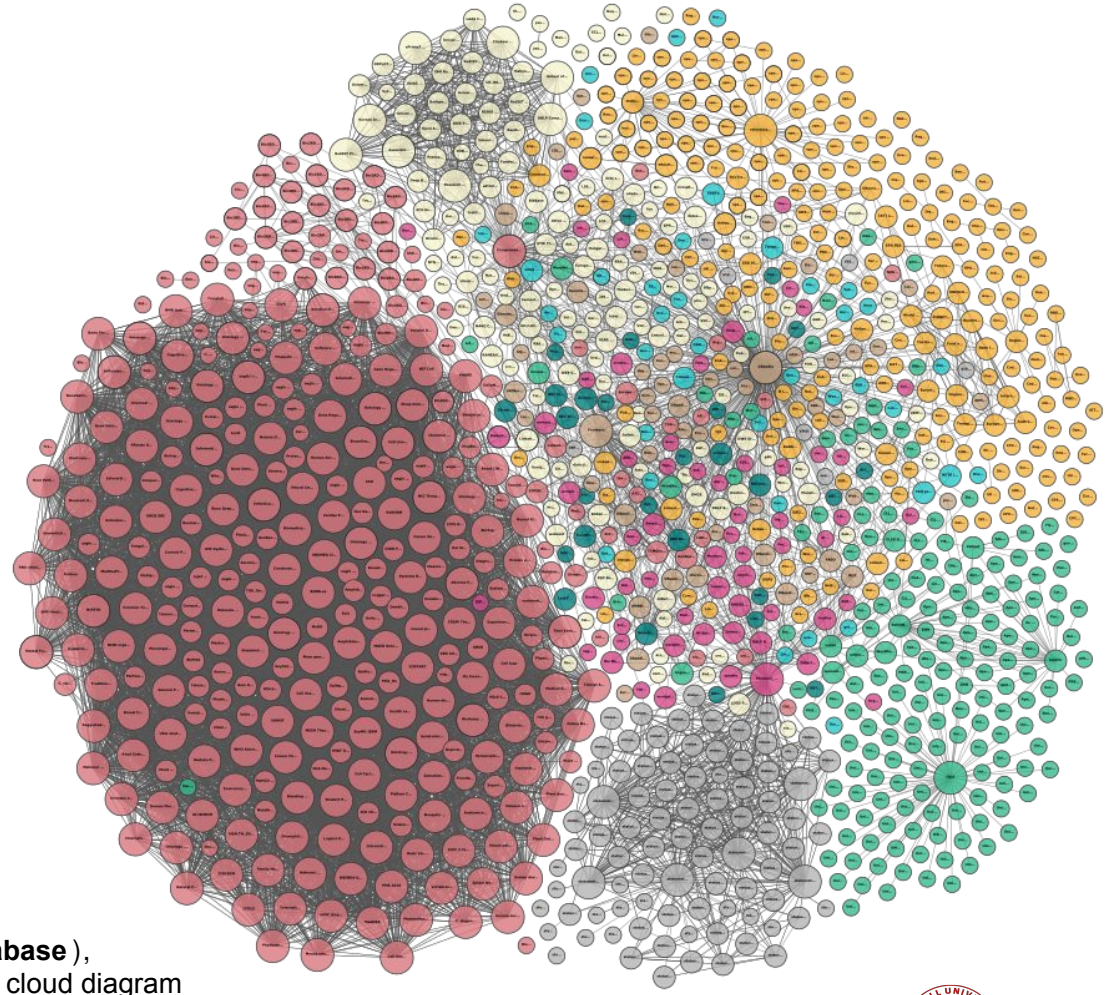
Finally...

What's in it for us (condensed)?

Questions?

Jason Kovari
jak473@cornell.edu

Andrew K. Pace
pacea@oclc.org



Massive Linked Open Data Cloud (**Reference Database**), **under-exploited by Publishers**. (Linking Open Data cloud diagram 2017–08–22, CC-BY-SA by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch, and Richard Cyganiak. http://lod-cloud.net/)