When Opportunity Knocks?

Challenges Associated with Implementing Linked Data in Libraries

Heather Lea Moulaison, PhD
@libacat
About me

Photo taken at the University of Missouri commencement last Saturday

I’m in my scarlet Rutgers robes in the middle

Dr. Jenny Bossaller (L), Dr. Heather Lea Moulaison, Dr. Denice Adkins
Poll #1

Your comfort level with linked data and your background
(this poll makes some big assumptions, so please just do your best)
linked data

...it doesn’t get better than this, right?
**Linked Data**

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, using the standards (RDF*, SPARQL)
4. Include links to other URIs so that they can discover more things.

--Tim Berners-Lee

http://www.w3.org/DesignIssues/LinkedData.html
""Linked Data" refers to data published in accordance with principles designed to facilitate linkages among datasets, element sets, and value vocabularies."

--Library Linked Data Incubator Group Final Report (2011)

http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/
Pundits, naysayers, interested librarians
Everybody talks....

http://www.youtube.com/watch?v=t0dEqgNKH0Y#t=1m14s
The future of libraries without linked data?
The waiting game

Yet, if linked data is such a great idea, why hasn’t it been implemented in libraries across the board?

And what do I need to think about when considering linked data in my library?
Agenda: When opportunity knocks

• Intro to the session
• What’s wrong with linked data?
  – Technical obstacles that libraries will have to overcome
• Where does linked data make sense?
  – Environments using linked data, and some spins on that notion
• Library systems and linked data
  – A library-land look at linked data projects
What’s wrong with linked data?
Problems with linked data

Vuk Miličić identifies some general issues with linked data (here are three):

– **Identity**: “The simple question of „What is exactly Linked Data?“ is not easy to answer.”

– **Publishing data**: “People find Linked Data hard to learn because of several steps along the way – certain things that are conceptually difficult to grasp.”

– **Consuming data**: finding content in LOD datastores is challenging
URIs

• Uniform Resource Identifiers (a kind of URL)
  – HTTP URIs are one of the foundational standards used in linked data
  – URIs are the nouns and verbs used in the RDF triples that form the basis for the semantics in linked data (Baker, 2012)
  – Identify real-world objects (web documents = non-information resources)
    • “On the Semantic Web, URIs identify ... real-world objects like people and cars, and even abstract ideas and non-existing things like a mythical unicorn.”
      (http://www.w3.org/TR/2007/WD-cooluris-20071217/)

Cool URIs ... not so cool?

• Ideally, each real world object and each concept would have one URI
  – This is the Cool URI for the Semantic Web concept
    (http://www.w3.org/TR/cooluris/)

• (Cool) URI issues in libraries:
  – Multiplicity of datasets (VIAF, DBpedia, NYTLOD, etc.)
    – Which one to use? There is no one organization that maintains URIs and selects THE ONE to use
      – In cases where URIs already exist, is it worth making a new one for library products?
      – What happens when URLs change and the URI isn’t as stable? (this happened to VIAF in early 2011)
  – Inability to automate the matching of URIs based on character string
    – VIAF record for Steven King the author might not be matched to the DBpedia entry for the same Steven King.
Datastore quality

• When thinking about quality in metadata, we tend to think about
  – Completeness (is there enough metadata to permit discovery and use?)
  – Correctness (i.e. accuracy of metadata supplied)
  – Consistency (i.e. comparability or coherence of that metadata across records, etc.)

• DBpedia is the linked data version of Wikipedia
  – Wikipedia InfoBox contents are crowdsourced
    • may well be lacking in all three indicators of quality...

• Information in DBpedia is presented as linked data, yet the data that is recorded may be incomplete or inconsistently applied.
  – Wikipedia
    • permits users to create a variety of attributes in entries (e.g. achievement, alma matter, astrological sign, buried place, and bust size for people: http://mappings.dbpedia.org/server/ontology/classes/Person)
    • Does not require this content to be entered
SPARQL and datastore accessibility

• SPARQL (recursive acronym for SPARQL Protocol and RDF Query Language)
  – Foundational way for LD (i.e. RDF) datastores to be searched
    • SPARQL can be searched by machines (or humans) as an “endpoint” to a database
    • “Both the formulation of the queries and the human-readable presentation of the results should typically be implemented by the calling software, and not be done manually by human users.” (emphasis mine)

http://semanticweb.org/wiki/SPARQL_endpoint
What's wrong with linked data?


Figure 2. The expressivity–usability trade-off for querying over structured data. The blue dots indicate that an ideal query mechanism for linked data must provide both high expressivity and high usability. (This figure was adapted from previous work.)
Querying LD datasets

- Querying linked data in datasets isn’t straightforward
  - Programmers have to write the SPARQL code and run it in the correct datastore
  - Finding SPARQL endpoints isn’t always easy
  - And apparently, they don’t always work (see http://fileformats.wordpress.com/2012/09/06/sparql/)
Finding LD datasets

- It’s also necessary to identify LD datasets that can be queried
- At present, there is no registry of datasets that explains what each contains, how it is searchable, whether the SPARQL endpoint is freely available, etc.
  - [library aside: I’m reminded of the days of DIALOG, but then, at least, we had Bluesheets to consult!]

- In short, instead of breaking down silos as we’ve been told LD can do, we’re seeing the creation of silos with a very high barrier to use.
Closing thoughts on Part 1: The complexities of linked data

• “Linked Data is trying to follow the principles of the original Web, but instead of focusing on the most important one – simplicity, it insists on the implementation of various relatively complex and geeky technologies of the Web architecture.”

Poll #2

The biggest barrier to the implementation of linked data in libraries?
Questions (or debates) so far?

And the results from the poll.
Where does linked data make sense?
Linked data outside of libraries

• Projects featuring displays of web-based content can benefit from being part of the linked data web
  – Websites with specific content on a page and that want to offer dynamically updating targeted content can include information sourced from linked data datasets
  – Freebase (which was, for some time powering the Summon discovery product) aggregates data from a variety of LOD projects including DBpedia, NYTLOD, Musicbrainz, and others

Where does linked data incontestably make sense?
The Library of Congress is the research library of the United States Congress, the de facto national library of the United States of America, and the oldest federal cultural institution in the United States. Located in four buildings in Washington, D.C., as well as the Packard Campus in Culpeper, Virginia, it is one of the two largest libraries in the world by shelf space and number of books, the other being the British Library. The head of the Library is the Librarian of Congress, currently James H. Billington. The Library of Congress was established for Congress in 1800, and was housed in the United States Capitol for most of the 19th century. After much of the original collection had been destroyed during the War of 1812, Thomas Jefferson sold 6,487 books — his entire personal collection — to the library in 1815. After a period of decline during the mid-19th century the Library of Congress began to grow rapidly in both size and importance after the American Civil War, culminating in the construction of a separate library building and the transfer of all copyright deposit holdings to the Library. During the rapid expansion of the 20th century the Library of Congress assumed a preeminent public role, becoming a "library of last resort" and expanding its mission for the benefit of scholars and the American people. Wikipedia
Linked data makes sense... when money’s not an object

• To paraphrase Jain, Hitzler, Janowicz, and Venkatramani (n.d.): there’s been essentially no commercial uptake of linked data
  – LOD datastores aren’t necessarily open, and licensing is an issue
    • “You can do anything you want [with linked data from datastores] as long [as] it is a proof of concept or a research paper” (p. 2)


• No commercial adoption leads to slower uptake everywhere else...
Bypassing LD for search

- Schema.org is helping search engines understand the contents of web pages
  - Created by Bing, Google, Yahoo! and Yandex
  - Uses .html tags instead of RDF
    - Could be considered a competing standard to LD since it doesn’t follow the four principles
    - Again, I’m reminded of the rationale behind the creation of Dublin Core
  - Chosen by OCLC as the way of encoding WorldCat records
Projects that share *data*

- Remember the D-I-K-W model for thinking about the relationships between data, information, knowledge, and wisdom?
  - Generally, it’s considered that information is comprised of data, but that data, in and of themselves, are meaningless
  - Libraries have the mission, therefore, to connect users to information -- not to data
- Projects using linked data are, broadly speaking, using it to create links to *data*
Corrado (2013) asked:

- Why do I want my library’s data published as linked data?
- Isn’t it enough to have large datasets like OCLC WorldCat and vocabularies like LCSH/headings like VIAF as linked data?
  - Why make my data be linked?
  - Who even wants to link to my data?
  - How will my (local) patrons benefit?

Poll #3

Timeline to linked data.... is it in our future, and if so, when?
Questions or comments?

Some of that from the second section might have been a little provocative.

And, results of our poll.
Library systems and linked data
An example: *Out of the Trenches: A Linked Open Data Project*

- Partners of the [Pan-Canadian Documentary Heritage Network (PCDHN)](http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN%20Proof-of-concept_Final-Report-ENG_0.pdf) have developed a “proof-of-concept” to showcase a sampling of the network’s wealth of digital resources using “linked open data” and principles of the semantic web.
- The overall approach to the “proof-of-concept” was to make use of existing metadata about the resources and repurpose it without loss of context and meaning.
- Analysis: linked data CAN be created from metadata libraries currently have
- Results: the project was successful, but is small in scale and not available online

- Report:
An example: BIBFRAME Redis datastore (Colorado College)

• Tutt Library at Colorado College:
  – FRBR-Redis datastore that begun pre-BIBFRAME
  – When the BIBFRAME draft was published, work revised to align with it

• In partnership with the University of Denver’s Penrose Library, they are "prototyping a peer-to-peer BIBFRAME Datastore" in order "to test the usability of a shared BIBFRAME Redis-based union catalog with two institutions to prepare for scaling up to a consortium-level service."

An example:
Réseau francophone numérique

- Réseau francophone numérique
  http://www.rfnum.org
  - Content can be exported in OAI-PMH
    http://oai.rfnum.org
  - Has a SPARQL endpoint in RDF
    http://data.rfnum.org

Other library-land content has been encoded using linked data, including Europeana (http://www.europeana.eu/)
Curb appeal

- [http://id.loc.gov/](http://id.loc.gov/)
- The Linked Data Service provides access to commonly found standards and vocabularies promulgated by the Library of Congress.

- [Hungarian National Library (NSZL) catalog](http://id.loc.gov/)
- OPAC and Digital Library and the corresponding authority data as Linked Open Data.

- [medline](http://id.loc.gov/)
- RDF representation of the Medline catalog. Information about 19 million articles linked to http://dx.doi.org/ with article identifiers and http://crossref.org/ with journal identifiers.

- [Sudoc bibliographic data](http://id.loc.gov/)
- Sudoc is the French academic union catalog, maintained by ABES. It contains 10 million bibliographic records.

- [Linked Data Service der Universitätsbibliothek Mannheim](http://id.loc.gov/)
- Publishes RDF for a number of bibliographic resources: Bibliographic data of the Südwestdeutscher Bibliotheksverbund, Bibliographic data of the Hessisches Bibliotheksinformationssystem, and others.

- For these and more: [http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/](http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/)
Laying out the welcome mat?
Barbara Walters's participation in events, professional work, and personal relationships
(http://mapper.nndb.com/start/?id=23371)
Summary: LD is the right future solution for libraries

LD in libraries will make sense...

• When LD practices have been firmly established
  – and the kinks outlined in Part 1 and Part 2 of this presentation have been worked out

• When library metadata is LD-ready
  – When BIBFRAME has been solidified

• When library systems can handle complex patron searches in library LD datastores on the fly
  – This will truly be a long-awaited moment, finally permitting discovery in the library catalog
Conclusion

• Library linked data is:
  – Not a cure-all for library issues
  – Not going to be implemented in libraries tomorrow
  – Already working (to a limited extent) in selected leading libraries
  – An interesting consideration for the direction your library will be taking in the near future.
Questions and comments?