The OPAC and the Web: Harmonious Future or Perpetual Discord?

Heather Lea Moulaison and Rebecca Thompson
Missouri Library Association
2011 Annual Conference
October 6, 2011
(Awareness session: basic information on a topic)
Online library catalogs as information repositories

Discovery layer

- OPAC
- Other subscription resource
- Newspaper article index #1
- Journal article index #1
- Digitized local collection
Other information repositories

- Libraries are not alone in providing content to patrons

- Wikipedia
- VIAF
- Open Library
- LibAcat
Changes afoot: Library users

• Are not better researchers and/or searchers than in the past, despite the wealth of new sources
  – “At Illinois Wesleyan University, ‘The majority of students -- of all levels -- exhibited significant difficulties that ranged across nearly every aspect of the search process.’”
    --Steve Kolowich

• Are more hurried, have less money and time, and have more conflicts
  – “Convenience is the “hook” that draws users into VR services”
    --Lynn Silipigni Connaway, Marie L. Radford
    http://www.oclc.org/reports/synchronicity/full.pdf

Forthcoming book: College Libraries and Student Culture: What We Now Know. Edited by: Lynda M. Duke and Andrew D. Asher
Changes afoot: Library systems

• In our systems
  – Web-accessible content
  – System navigation
  – Community-generated content
  – Discovery model promoting “discovery” of resources

• Beyond our systems
  – Web-accessible curated collections of information
Is a new model in order?

Is it worth including curated content from the web into library results?
Agenda

• The intelligent web (i.e. the semantic web)
  – Definition
  – What it might mean to libraries and library users
  – Examples of interesting projects
• Transaction log analysis
  – Rationale
  – Method
  – Results
• Discussion
• Conclusions
The semantic web: An ideal

• Tim Berners-Lee’s vision: computers that could understand each other
  – Semantic web
    • The intelligent web
    • Web 3.0
    • The web of data
  – No need for human intervention at each step
    • Technology and protocols need to be in place

• As of right now, the semantic web isn’t fully functioning – but things are improving all the time
Semantic web and discovery

“The Semantic Web isn't just about putting data on the web. It is about making links so that a person or machine can explore the web of data. With linked data when you have some of it, you can find other, related, data.”

--Tim Berners-Lee

http://www.w3.org/DesignIssues/LinkedData.html
Linked data

“"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

http://www.w3.org/2005/Incubator/lld/wiki/DraftReportWithTransclusion
Some linked data projects of interest to the library world

An overview of selected linked data repositories:
• Open Library
• NNDB Mapper
• New York Times Linked Open Data
• Library of Congress (LOC) Names
• Virtual International Authority File (VIAF)
• DBpedia
• Freebase

For more information on many of these and on other vocabularies of interest to libraries, see the section on value vocabularies in W3C Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets: http://www.w3.org/2005/Incubator/lld/wiki/Vocabulary_and_Dataset

The OPAC and the web

H. L. Moulaison & R. Thompson

Presenter: Heather Lea Moulaison, @libacat

http://openlibrary.org/authors

- allows users to add content, merge records
- points users directly back to libraries at the book-level

• Connections between 34,000 individuals (and organizations) are made machine readable

• Users can create their own “maps”
- 10,467 entries for people, organizations, locations, & descriptors
- **http://id.loc.gov/authorities/names**
- Library of Congress name authority file as linked data
Virtual International Authority File (VIAF)

- [http://viaf.org/](http://viaf.org/)
- “joint project of several national libraries plus selected regional and trans-national library agencies”
About: John Van Seters
An Entity of Type: living people, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

John Van Seters (born Hamilton, Ontario, Canada, 2 May 1936) is a scholar of the Hebrew Bible (Old Testament) and the Ancient Near East. Currently University Distinguished Professor Emeritus at the University of North Carolina, he was formerly James A. Gray Professor of Biblical Literature at UNC. He took his PhD at Yale University in Near Eastern Studies (1965) and a ThD h.c. from the University of Lausanne (1999).

- **Property**: dbpedia-cw:abstract
  - **Value**: John Van Seters (born Hamilton, Ontario, Canada, 2 May 1936) is a scholar of the Hebrew Bible (Old Testament) and the Ancient Near East. Currently University Distinguished Professor Emeritus at the University of North Carolina, he was formerly James A. Gray Professor of Biblical Literature at UNC. He took his PhD at Yale University in Near Eastern Studies (1965) and a ThD h.c. from the University of Lausanne (1999). His honours and awards include a Guggenheim Fellowship, an NEH fellowship, an ACLS fellowship, and research fellowships at Oxford, Cambridge, Katholieke Universiteit Leuven, and National Research Fund of South Africa. His many publications include "The Hyksos, A New Investigation" (1965); "Abraham in History and Tradition" (1975); "In Search of History" (1983, for which he won the James H. Breasted Prize and the American Academy of Religion book award); "The Edited Bible" (2006); and "The Biblical Saga of King David" (2009). "Changing Perspectives: I: Studies in the History, Literature and Religion of Biblical Israel", a collection of articles from 1964 onwards with an introduction by Thomas L. Thompson, will appear in mid-2011.

- **Property**: dbpedia-cw:birthDate
  - **Value**: 1936-05-02 (xsd:date)

- **Property**: dbpedia-cw:wikiPageExternalLink
  - **Value**: http://www.bookreviews.org/bookdetail.asp?Title=2388&CodePage=3929,5237,2388

- **Property**: dbpprop:birthDate
  - **Value**: 1936-05-02 (xsd:integer)

- **Property**: dbpprop:name
  - **Value**: Van Seters, John

- **Property**: dct:subject
  - **Value**: category: Fellows_of_Gillespie_College_Oxford
  - **Value**: category: Fellows_of_Gillespie_College_Oxford

- **Property**: dbpedia-cw:wikiPageExternalLink
  - **Value**: http://dbpedia.org/

- **Property**: contents of Wikipedia in linked data format

- **Property**: (SPARQL builder: http://dbpedia.org/fct/)

Slide 17, The OPAC and the web
H. L. Moulaison & R. Thompson
Presenter: Heather Lea Moulaison, @libacat
- “An entity graph of people, places and things, built by a community that loves open data.”
- Entries may not be disambiguated,
- Links when possible to previously mentioned vocabularies plus many more
Linked data and libraries?

• The semantic web idea assumes certain kinds of data of interest to the library community can be encoded in a way that is machine-readable

• Currently library data is structured using MARC
  – MARC is not web-ready!
  – MARC encoding permits display in libraries
  – MARC isn’t used outside of the library community

• Online catalog queries are made using unstructured textual strings in particular indexes
  – This is not the most effective way to search or browse
Rationale for the current study

- Transaction log analyses have demonstrated that users do not query OPACs effectively
- Linked data initiatives are a way to link out to curated data on the web
- To limit the focus of the current study, it is advisable to concentrate on “authors” in general and personal names in particular
- The connection between library content, user searches, and linked data is worth exploring in the current library environment

- By investigating a certain class of queries and matching them against the contents of various linked data repositories, it will be possible to gauge the degree of effectiveness a semantic online library catalog might have.
Method

• Analysis of author index (left-anchored) queries to Millennium III Integrated Library System (ILS) at Missouri State University from 28 March – 6 April, 2011.

  ▪ Transaction logs
    ▪ showed the characters input minus punctuation, time of the search, and date of the search
    ▪ no information about patron ISP, number of results returned, number of searches in session, etc. were recorded

• Exactly 700 left-anchored author searches were carried out
  ▪ Because of the interest in “authors”, 1-term queries (n=238, 34%) were not considered unless they reasonably represented an artist or association acronym (i.e. Beck, ACM, etc.)
Method, continued

Two-part analysis:

1. To assess the quality of the retained set of OPAC queries, the search terms were analyzed for their context and construction

2. The 462 retained queries were re-run against the OPAC to assess hits
   - Queries yielding zero hits in the OPAC were then manually relaunched in seven linked data repositories
Results: Formulation of OPAC queries

- Corporate bodies (ALL acceptable left anchor) (n=16)
- Personal names: correct left anchor (n=337)
- Personal names: unable to tell if left anchor correct (n=2)
- Personal names: incorrect left anchor (n=99)
- Incorrect index; unable to tell (n=7)
Results: OPAC queries and linked data repositories

Success and failure (fail = 0 hits) in the OPAC and LD repositories

- "Successful" OPAC searches
- Failed OPAC searches, Failed in LD repositories
- Failed OPAC searches, "Successful" in LD repositories

159 total failed OPAC searches

- n=115 (24.9%)
- n=44 (9.5%)
- n=303 (65.6%)
Results: OPAC queries and linked data repositories, continued

Number of results in LD systems for failed OPAC queries (n=159)

- Open Library: 9
- Mapper: 8
- LOC: 44
- VIAF: 32
- NY Times: 12
- Dbpedia: 67
- Freebase: 106

No. entries in value vocabulary
Discussion

- Promote linked data in the OPAC when possible
  - Provide access to Freebase-like lists of current resources in response to patron queries through linked data
  - Help users disambiguate names based not only on library holdings, but on LD entries
  - Take advantage of linked information in LD repositories to help patrons explore authors further (access to videos, MySpace pages, etc.)
Summon is already displaying information from Freebase!

Observations:

- Difficult to manipulate from the regular search screen
- Fb blurb not hyperlinked in this iteration (cannot navigate to more info)
- Search results not sufficient for author searching in this instance
Discussion

Continue to work on providing a patron-centered OPAC

- To accommodate the over 20% of the retained author index queries that were incorrectly formulated, libraries might want to consider doing away with left-anchored searches in the author index
  - Avoid issues of searching the “right way” vs. the “natural way”
  - Make authority records and the whole MARC format more granular (i.e. implement RDA)

- Link out to LD resources (that link back to libraries!) when faced with providing zero hits or to supplement information in the OPAC

- Libraries have been doing authority work well for a long time
  - All locally created authority records should be exported and shared to enrich the data available in general
  - Some of the records won’t be perfect, but that’s ok
  - Consider crowdsourcing authority work (see the example at Open Library and Freebase)
  - Make authority records and the whole MARC format more granular (i.e. implement RDA)
Recommendations from the LLD XG

• “That library standards bodies increase library participation in Semantic Web standardization, develop library data standards that are compatible with Linked Data, and disseminate best-practice design patterns tailored to library Linked Data; That data and systems designers design enhanced user services based on Linked Data capabilities, create URIs for the items in library datasets, develop policies for managing RDF vocabularies and their URIs, and express library data by re-using or mapping to existing Linked Data vocabularies”
Recommendations based on this study

• Continue to strive to provide the resources patrons want at the point of need
  – Consider if linked data within library systems can assist in this mission

• Continue to innovate, while using standards that are accepted by the web at large
  – Do library-only standards like MARC still have a place?
A possible results screen for a "semantic OPAC"
Desirability of a semantic OPAC?

• Data quality for linked data portions would have to be vetted outside of the library
  – WWW domains would be key to decisions to trust, but would not be foolproof
  – Users who surf away from the library gateway *may* not come back
    • Yet we know they don’t begin their searches in the OPAC anyway, so is it wrong to “reward” with the info they want users who actually come to the library OPAC and find little?

• Machine matching inherently flawed
  – Human intervention still necessary for complete assurance of proper disambiguation
    • Human intervention can be crowdsourced! (c.f. Open Library)
Limitations to machine matching

- Freebase’s page for the horror novelist Stephen King (b. 1947) links to this LOC Names page... the page for the undifferentiated name heading “Stephen King” (i.e. the wrong heading!)
Limitations to the current study

• Transaction logs very limited in information provided
  – Unable to isolate librarian searches, in-library or on-campus searches, and off-campus searches
  – Results as seen by patrons not explicit
    • Over time, some of the books were acquired
  – Search sessions not identified

• Transaction log analysis cannot infer patron
  – Intent
  – Success and/or satisfaction
  – Knowledge
Future studies

• Future work should consider the need to
  – Build a prototype semantic OPAC and to test it on users
  – Continue analyzing patron OPAC queries
    • To ascertain the extent to which local collections meet user needs above and beyond the abilities of curated linked data repositories, it would be beneficial to examine all successful queries and to compare them with the results in LD repositories
Conclusion: The OPAC and the Web: Harmonious Future or Perpetual Discord?

• In the context of the rapidly changing information environment, it is recommended that libraries further embrace web resources as a way of providing patrons with the information they seek. This may mean integrating them into the OPAC, and removing the OPAC from the deep web to integrate it with the web at large.

• Libraries must continue to be a vital part of the information ecosystem, and exploiting existing resources like linked data is one way that can take place.