January 2008

Mapping the library future: Subject navigation for today's and tomorrow's library catalogs

John Mark Ockerbloom
University of Pennsylvania, ockerblo@pobox.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/library_papers
Part of the Library and Information Science Commons

Recommended Citation

Presented at ALA Catalog Form and Function Group, January 2008.

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/library_papers/64
For more information, please contact repository@pobox.upenn.edu.
Mapping the library future: Subject navigation for today’s and tomorrow’s library catalogs

Abstract
My ALA Mindwinter 2008 presentation slides on subject maps. For more details on how subject maps are created, see the New Maps of the Library white paper from 2006.

Disciplines
Library and Information Science

Comments
Presented at ALA Catalog Form and Function Group, January 2008.

This presentation is available at ScholarlyCommons: http://repository.upenn.edu/library_papers/64
Mapping the library future

Subject navigation for today’s and tomorrow’s library catalogs

John Mark Ockerbloom

ALA Catalog Form and Function Group

January 12, 2008
Take-away points

• LC Subject Headings form a rich basis for discovery, crippled by poor discovery tools, slow evolution
  – A deep ontology with many detailed terms and relationships
  – Network of concepts not captured by keywords or facets alone

• Subject maps enable better subject browsing with ontologies like LCSH
  – Multi-dimensional browsing, highlighting relevant connections
  – May also make it easier to maintain, evolve big, messy subject ontologies

• They are practical for present and next-generation catalogs
  – Build on legacy data, but accommodate new collecting practices
  – Both large (research library) & small (special collection) scales
  – With ordinary web browsers, servers, and ILS, and without expensive proprietary software
  – Can work well with facet browsing, tagging

• We can build on library strengths and Web innovations, to better connect our users with the resources they need
The appeal of browsing

Green Apple Books (San Francisco), March 2007
Photo by Dolan Halbrook (Copyright 2007. Creative Commons Licensed)
More content not on shelves…

**Google Is Adding Major Libraries to Its Database**

By JOHN MARKOFF and EDWARD WYATT
Published: December 14, 2004

Google, the operator of the world's most popular Internet search service, plans to announce an agreement today with some of the nation's leading research libraries and Oxford University to begin converting their holdings into digital files that would be freely searchable over the Web.

It may be only a step on a long road toward the long-predicted global virtual library. But the collaboration of Google and research institutions that also include Harvard, the University of Michigan, Stanford and the New York Public Library is a major stride in an ambitious Internet effort by various parties. The goal is to expand the Web beyond its current valuable, if eclectic, body of material and create a digital card catalog and searchable library for the world's books, scholarly papers and special collections.

...
...and hard to browse
Women in Google Books

[Image of Google Books search results]

John Mark Ockerbloom

UNIVERSITY OF PENNSYLVANIA LIBRARIES

Jan. 12, 2008
In Amazon

Our full Women selection in Apparel

Related Searches: female, men, womens.

1. Women by Annie Leibovitz and Susan Sontag
   (Paperback - Nov 9, 2000)
   Used & new from $109.99
   Other Editions: Hardcover, Paperback

2. Women by Annie Leibovitz and Susan Sontag
   (Paperback - Oct 17, 2000)
   Buy new: $49.95 $32.97 Used & new from $24.84
   Get it by Tuesday, Feb. 20, if you order in the next 2 hours and 26 minutes.
   Other Editions: Hardcover, Paperback

3. Women: A Novel by Charles Bukowski
In academic tag-space
In our catalog

Franklin: Penn Library Catalog

Search Request: Subject = women
Search Results: Displaying 1 through 50 of 50 entries.

<table>
<thead>
<tr>
<th>#</th>
<th>Titles</th>
<th>Headings</th>
<th>Headings Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>121 Women.</td>
<td>Moving image genre-form guide</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>377 Women.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>161 Women.</td>
<td>MESH</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1 Women--20th century--Biography.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1 Women--20th century--Fiction.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1 Women--20th century--Pakistan.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1 Women abolitionists--Fiction.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2 Women abolitionists--Great Britain.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2 Women abolitionists--Great Britain--History.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1 Women abolitionists--Massachusetts.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1 Women abolitionists--Massachusetts--Boston--History--19th century.</td>
<td>Library of Congress</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1 Women abolitionists--Pennsylvania--Biography.</td>
<td>Library of Congress</td>
<td></td>
</tr>
</tbody>
</table>
The strength and the weakness of the subject catalog

• “One researcher... was interested in linguistic studies of the Cockney dialect. He simply typed ‘Cockney’ as a keyword into our catalog... [and] missed most of the linguistic studies..... The proper LC subject heading ‘English language -- Dialects -- England -- London’ rounds up in one categorical grouping all such works scattered by variant keywords-- and variant languages”

  – Thomas Mann, “Research at Risk”, in Library Journal online site, article dated June 15, 2005
How do we get there?
Cataloging is messy

Search Request: Subject = english language dialects england
Search Results: Displaying 1 through 50 of 50 entries.

<table>
<thead>
<tr>
<th>#</th>
<th>Titles</th>
<th>Headings</th>
<th>Headings Type</th>
</tr>
</thead>
</table>
Better browsing with facets
Better browsing with facets
But what are you missing?
Problems with LCSH

• Subject headings are expensive and difficult to assign, maintain
  – And are inconsistently assigned across time and collections
• They are difficult to use
  – a >100 year old legacy system, with all the backward compatibility issues, mismatches with expectations that implies
  – They tend not to keep up with ongoing terminology well
• Our patrons aren’t using subject headings-based discovery much
  – Hard to find all the terms that are relevant, or even the first one
  – With the tools we give them, can you blame them? [see critiques by Karen Schneider and others]

• Are the costs worth the benefits?
Some benefits of LCSH

• It’s the metadata and ontology we have at hand
  – That’s the flip side of it being a >100 year old legacy system
  – Many resources of interest are only described by it

• Enables great precision in identifying subject areas
  – Millions of terms (including 200,000+ authorized headings), careful definitions, lots of explicit and implicit relationships
  – Can pinpoint items with subject as main topic more precisely than one can with keyword search
  – Much larger array of effectively controlled areas than tags or keywords

• Ontology can support much richer subject browsing, with good tools
  – Users like to browse (e.g. in stacks) when they can easily do so
  – Allows level of specificity not easily found with facets (e.g. general works on a topic)
  – May also allow useful browsing, adaptation when assigning headings

• Can we revive focused topical browsing with dispersed collections?
Subject maps: It’s the network

- Organized networks of well-defined subject terms and relationships between them
  - Applied and customized to particular collections of items
- Clustered, side-by-side displays of subjects, items, and relationships in ordinary text-based Web browsers
  - Other display options also possible
- Designed to function like geographic maps
  - Get users to a [subject] area they’re interested in
  - Let them see what’s there, and in nearby areas, at a glance
  - Show them routes to get to nearby areas, so they can home in on what’s most useful to them
  - Detail and layout helps compensate for
    » Imprecise cataloging (“It’s near the underground station”)
    » Differences in concepts and names (“Subway? Trolley? Regional rail?”)
- Work well with detailed ontologies like LCSH
Subject maps on a small collection

• The Online Books Page
  – 30,000+ items
  – 25,000+ subjects
  – Subject maps in production there since 2006

• Live online demo:
  – http://onlinebooks.library.upenn.edu/subjects.html

• Live laptop demo:
  – http://127.0.0.1/books/subjects.html
  – (it’s just as fast, on this MacBook)
Building subject maps

Build them automatically:

• First, create a collection-independent topic map based on authorities
• Second, tweak it based on local needs and data
  – “Tweaks” can overlay, inform centrally maintained authorities
  – Sources include usage data, domain-specific knowledge bases
• Third, adapt it to a particular [set of] collection[s]
  – Start by mining bibliographic data
  – Add terms not already in collection-independent version
  – Prune “dead ends” from collection-independent version
  – Apply rules to create additional subject relationships
Where do terms and relationships come from?

- **From bibliographic data**
  - Terms used in a catalog generally assumed to be valid
  - If terms not present, normalization may be possible
  - For relationships, consider assignment patterns: e.g. name as first subject followed by "<topic> -- Biography"

- **From authority files and other “canonical” data**
  - "BT", "NT", "UF"; lists of state names and abbreviations

- **From user data**
  - Look at search logs, tag usage and correlations, etc.

- **From facet analysis**
  - Which facets can be added or dropped? Which can have broader or narrower terms substituted? Permutations?

- **From lexical and domain analysis**
  - If “X and Y” relates to Y, it probably also relates to X
  - Recognizing geographic terms and abbreviations

- **Inclusion rules can be added or removed as required**
Subject maps on a large collection

- Franklin, the Penn Library catalog
  - Over 3 million items
  - Over 1 million subject terms
  - Subject maps built, demonstrated, now being tested
    » Can be newly generated from harvested data in ~40 minutes on up-to-date hardware with 8GB memory

- Live online demo:
  - [http://devplw.library.upenn.edu/cgi-bin/sdi/vbook/browse?type=lcsubc&key=Women](http://devplw.library.upenn.edu/cgi-bin/sdi/vbook/browse?type=lcsubc&key=Women)
  - (Now only available for limited times like this talk)

- Or see screenshot on the next slide
Subject maps on Franklin

Browsing subject area: Women (About this browser)
You can also browse an alphabetical list from this subject or from:

Women

Here are entered works on the human female. Works on female organisms in general are entered under Females.

Broader terms:
- Anthropology
- Females
- Man
- Sociology

Related term:
- Misogyny

Narrower terms:
- Women -- 19th century
- Women -- 20th century
- Women -- Abstracts
- Women -- Abuses of
- Women -- Abuse of

Filed under: Women

<table>
<thead>
<tr>
<th>Full Title</th>
<th>Author</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solas [videorecording] = Alone / una producción / una Maestra Filma (Sevilla) con la participación / Alten de Villar Digital Canal Sur TV ; producida por Antonio P, Pollet ; escrita y dirigida por Benito Zambrano.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devaki [videorecording] / presented by K.D. Singh ; produced by Alchemist Ltd ; directed by Bapadiya Ramapadhyaya</td>
<td></td>
<td>2006</td>
</tr>
<tr>
<td>Novia que te vea [videorecording] = Like a bride / producida por Instituto Mexicano de Cinematografía / Pondo de Fomento a la Cultura Cinematográfica / Atenea Producciones Arte Nuevo ; screenplay by Hugo Hiriart ; Gatta Schyfter film.</td>
<td></td>
<td>2006</td>
</tr>
<tr>
<td>Female brain / Louise Brizendine, 1952-</td>
<td>Brizendine, Louise, 1952-</td>
<td>2006</td>
</tr>
<tr>
<td>What Lacan said about women : a psychoanalytic study / Colette Solé ; translated by John Holland, Androgyny and reproduction series / edited by</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

John Mark Ockerbloom  UNIVERSITY OF PENNSYLVANIA LIBRARIES  Jan. 12, 2008
Integrating subject maps with a library catalog

- Our ILS (Voyager) gives us a few important tools
  - SQL: Data-mine and query authority and bibliographic metadata
  - JavaScript: Rewrite catalog web pages to link out to map tools
  - URLs with fielded queries: Link back to the catalog from maps

- We’ve already used these techniques to integrate PennTags with our existing Franklin catalog
- Better: Subject map-aware catalogs
- Better still: ILS’s with well-defined APIs providing data and services to subject map and other displays
  - That’s the subject of an ongoing Digital Library Federation initiative
A larger repertoire of discovery tools

Subject maps

Facets

Full text search

Social tagging
A larger repertoire of discovery tools

Subject maps  Facets

Full text search  Social tagging

* Subject maps, non-subject facets (e.g. media type, language, availability), can be used simultaneously
* Subject maps can be annotated when built so only terms that apply under arbitrary facet limitations get shown
A larger repertoire of discovery tools

Subject maps

Facets

Full text search

Social tagging

• Correlated with formal ontology terms (perhaps automatically?)
• Useful both as leadin to catalog, leadout to user-tagged items
A larger repertoire of discovery tools

Subject maps

Facets

Full text search

Social tagging

- Search of full text (or metadata) can be used to identify frequently occurring subjects to browse
- Or, full text search limit results to items falling within a particular subject cluster
Some early promising signs and speculations

• Since maps introduced to The Online Books Page, users making substantially heavier use of subject browsing
  – And going deeper into the collection than before
  – Adding basic geographic smarts was easy

• Demos with Franklin, multiple sources show feasibility of building, using maps with larger collections
  – Noise filtering / normalization an issue with any large collection
  – Scaling issues seem to be increasingly tractable
    » As memory gets bigger and cheaper, map building for larger collections can be done on cheaper hardware
    » Map navigation is computationally cheap already (since it’s only looks at local portions of map)
    » We’re looking at more advanced techniques for presenting heavily populated map portions more effectively to users

• Could we scale this up to the global level?
  – Why not include what your users can get via ILL, other means?
What would we like to do next?

• Put full Penn catalog maps in front of users
  – Will require some refinement of maps, tweaking of Voyager, possibly normalization of obsolete/improper terms
  – User studies could be useful as well, not sure we’ll do this

• Try combining it with ongoing facet work
  – Test idea of subject maps that adapt with facet restrictions
  – See if we can combine it with Lucene/Solr views

• Develop smarter map building and display techniques

• Create a subject map builders toolkit
  – Configurable for different authorities, collections, rulesets
    » Authorities web service could be useful here
  – Make it easier to fix headings, add tweaks, correlate with tags…

• Think about how our users will discover the best resources (ours or others’) in a networked, highly digital world
Conclusions

• Subject browsing is an important mechanism for resource discovery
• Detailed ontologies like Library of Congress Subject Headings can form the basis for useful browsing, if used with appropriate tools
• Subject maps are useful and practical tools for exploring collections with complex ontologies, maintaining those ontologies

Subject maps build on library strengths to better connect our users with the resources they need in a distributed, digitized information environment

• For more information (demos, whitepapers...):
  – Web: http://labs.library.upenn.edu/subjectmaps/
  – Email: ockerblo@pobox.upenn.edu

• Thanks!