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**
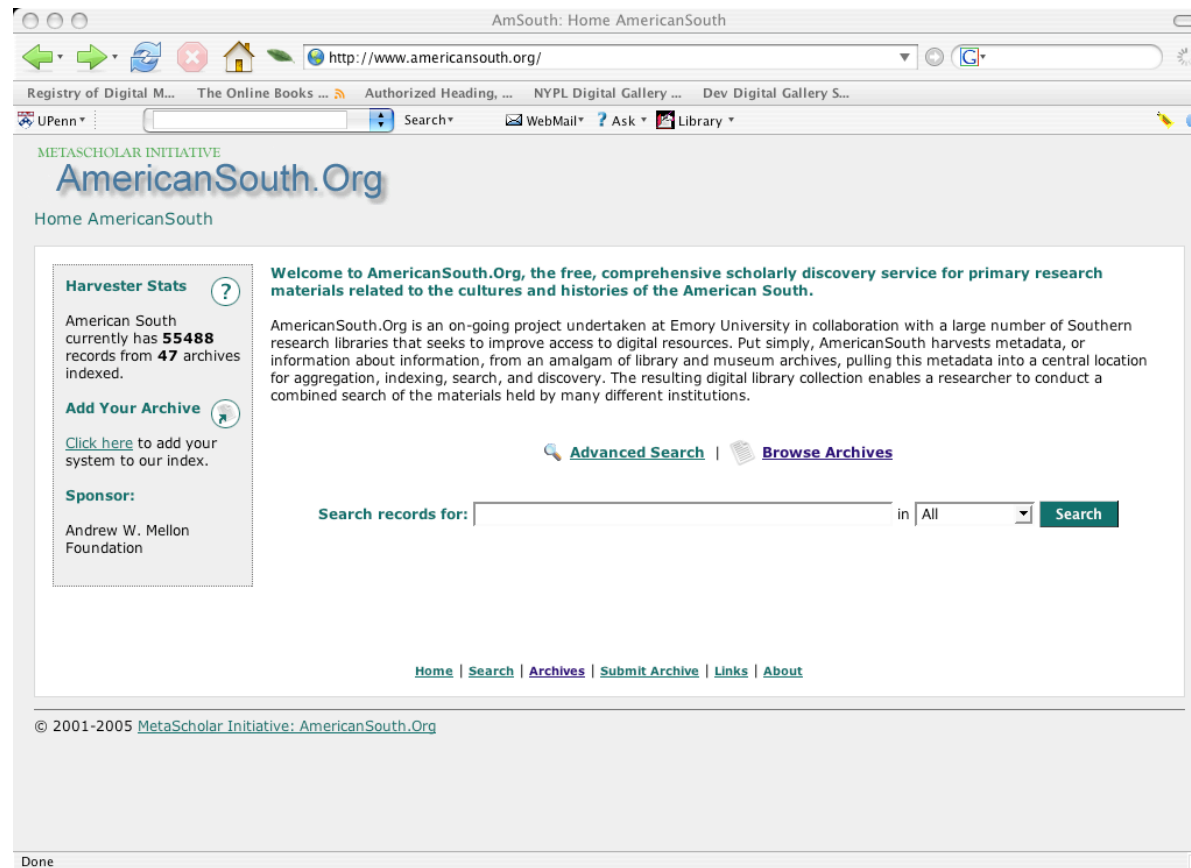
The appeal of browsing

Green Apple Books (San Francisco), March 2007
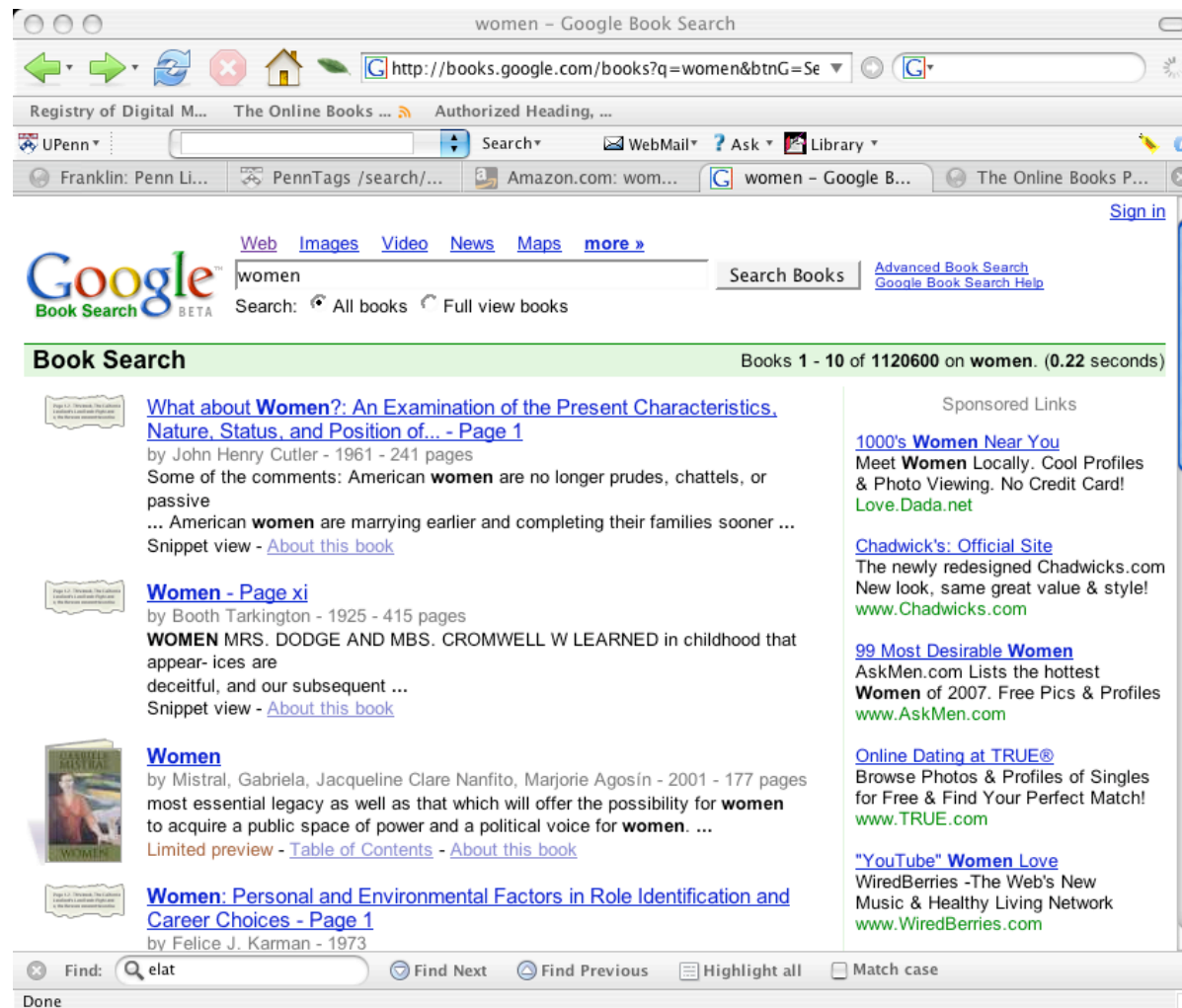**Photo by Dolan Halbrook (Copyright 2007. Creative Commons Licensed)**

More content not on shelves…
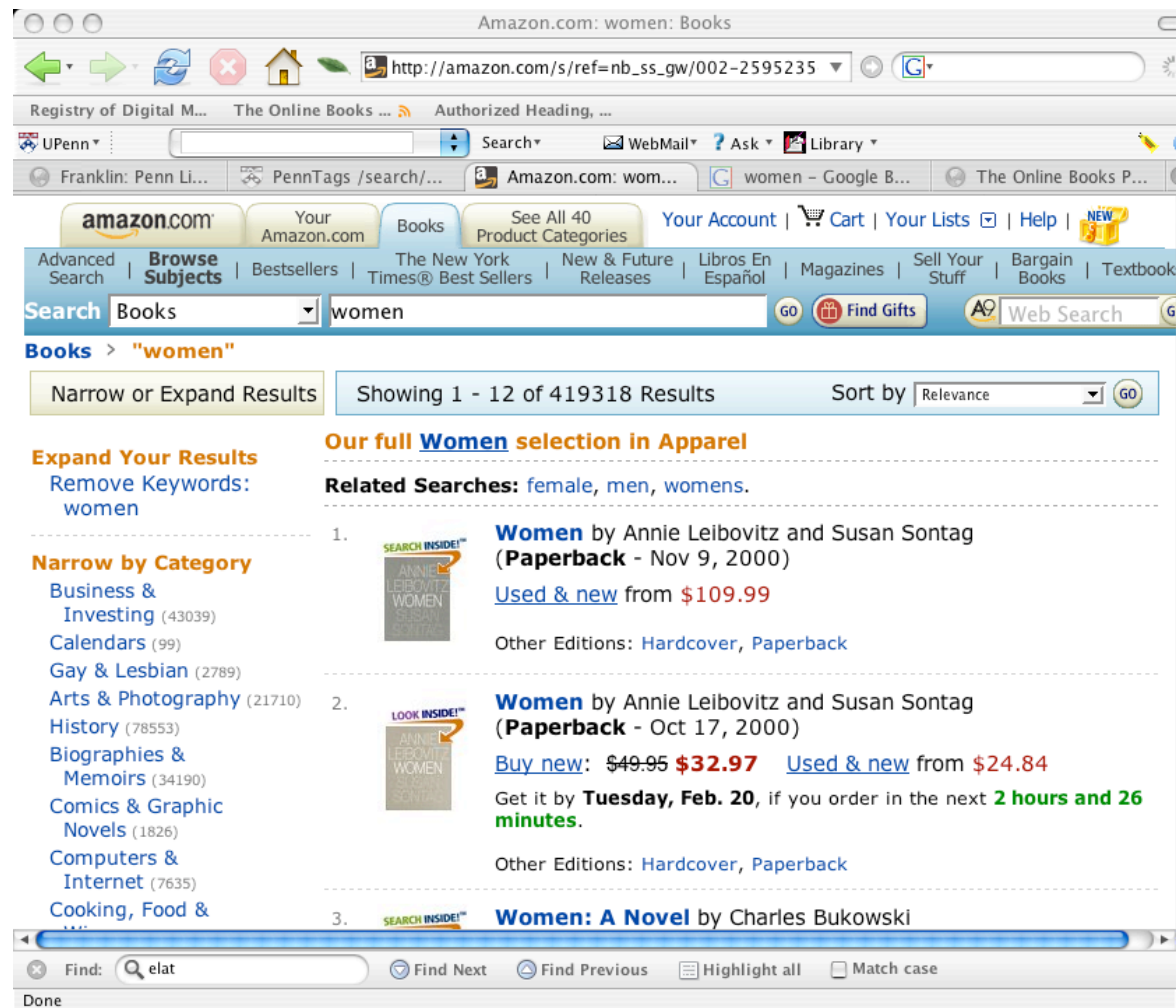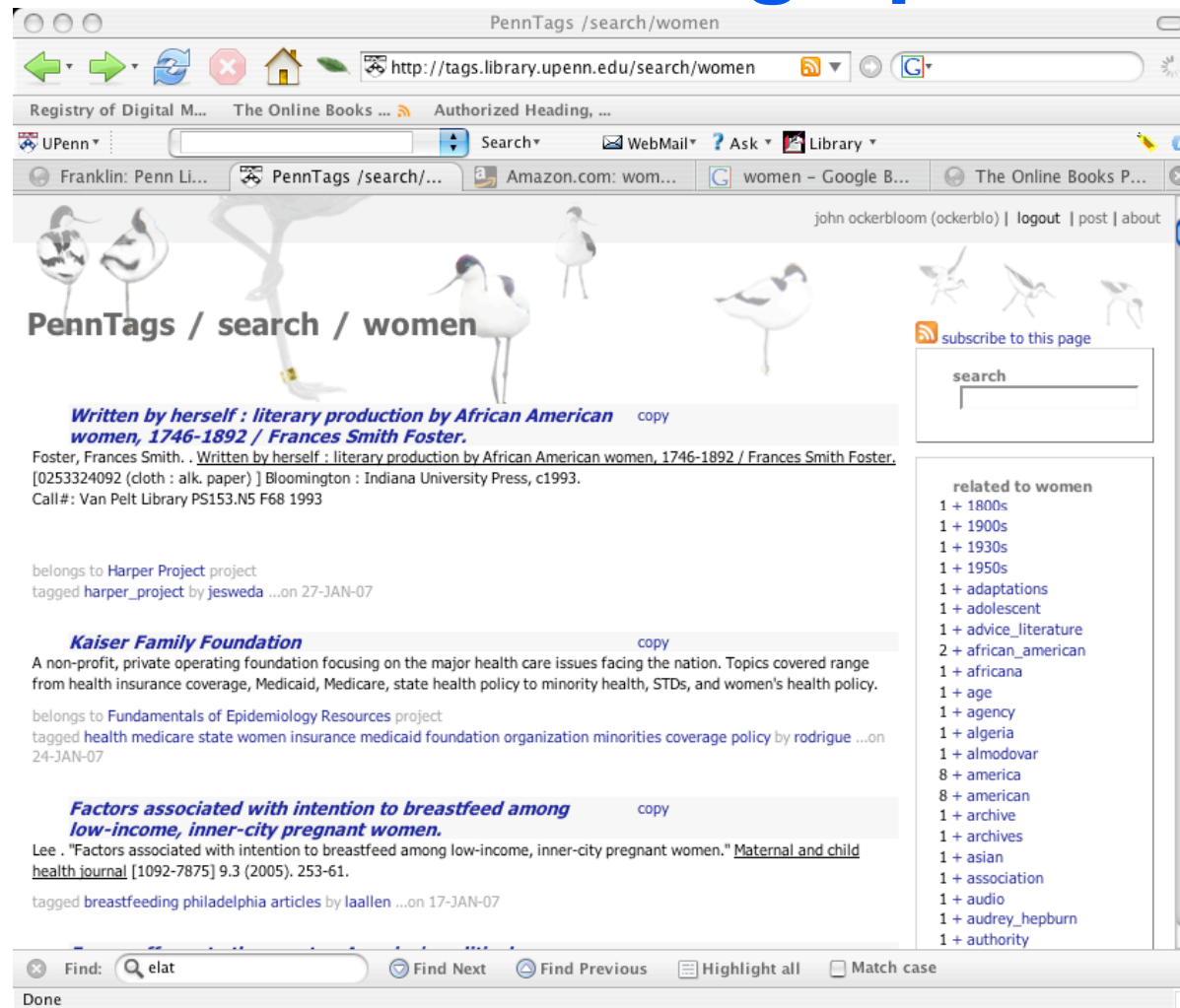
…and hard to browse

Women in Google Books

John Mark Ockerbloom   UNIVERSITY OF PENNSYLVANIA LIBRARIES   Jan. 12, 2008

In Amazon

John Mark Ockerbloom   UNIVERSITY OF PENNSYLVANIA LIBRARIES   **Jan. 12, 2008**

In academic tag-space

In our catalog

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

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**
  - **(Now only available for limited times like this talk)**
- **Or see screenshot on the next slide**

Subject maps on Franklin

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

John Mark Ockerbloom   UNIVERSITY OF PENNSYLVANIA LIBRARIES   **Jan. 12, 2008**

A larger repertoire of discovery tools

Subject maps   **Facets**

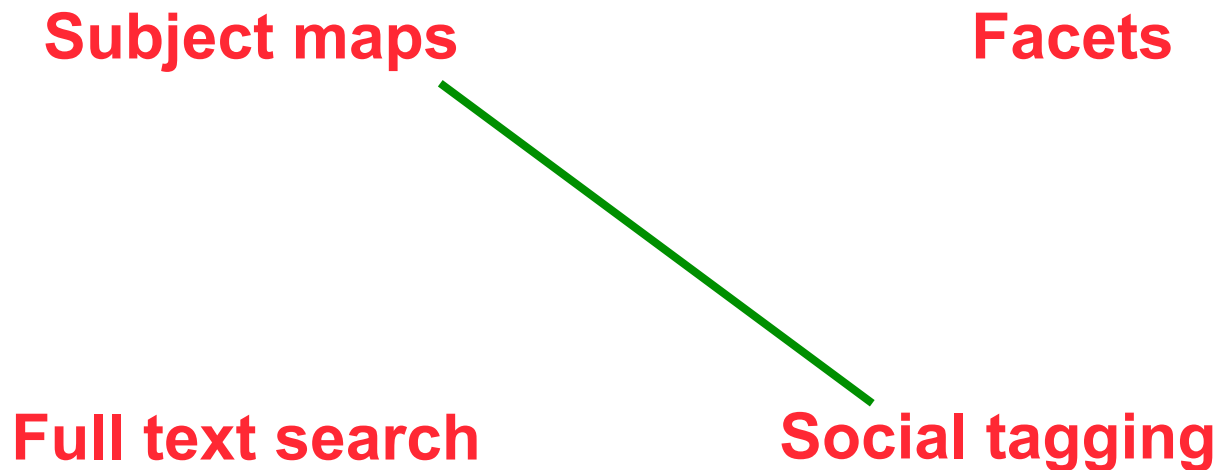**Full text search**   **Social tagging**

• 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!**