

# Discovering Connections

Subject Maps for Browsing Aggregated Collections

**John Mark Ockerbloom**

**Digital Library Federation Spring Forum**

**April 25, 2007**

# Discovering Connections

Subject Maps for Browsing Aggregated Collections

**John Mark Ockerbloom**

**(and some aggregated counter-voices)**

**Digital Library Federation Spring Forum**

**April 25, 2007**

# Conclusions (already?)

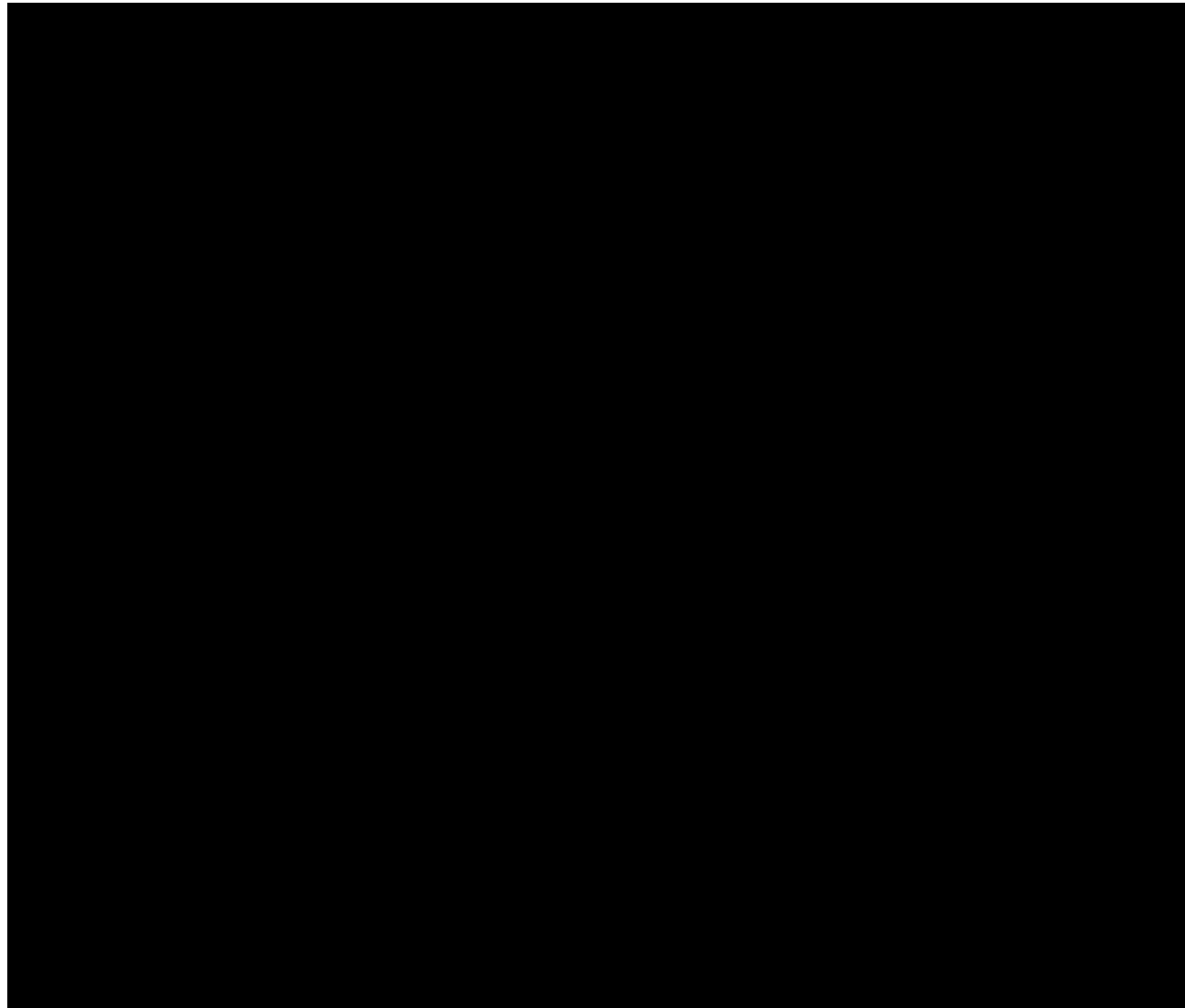
- **Subject browsing is an important mechanism for resource discovery across aggregated collections**
- **Detailed ontologies like Library of Congress Subject Headings can form the basis for useful browsing, if used with appropriate tools**
- **Subject maps are useful and practical tools for exploring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

# The appeal of browsing

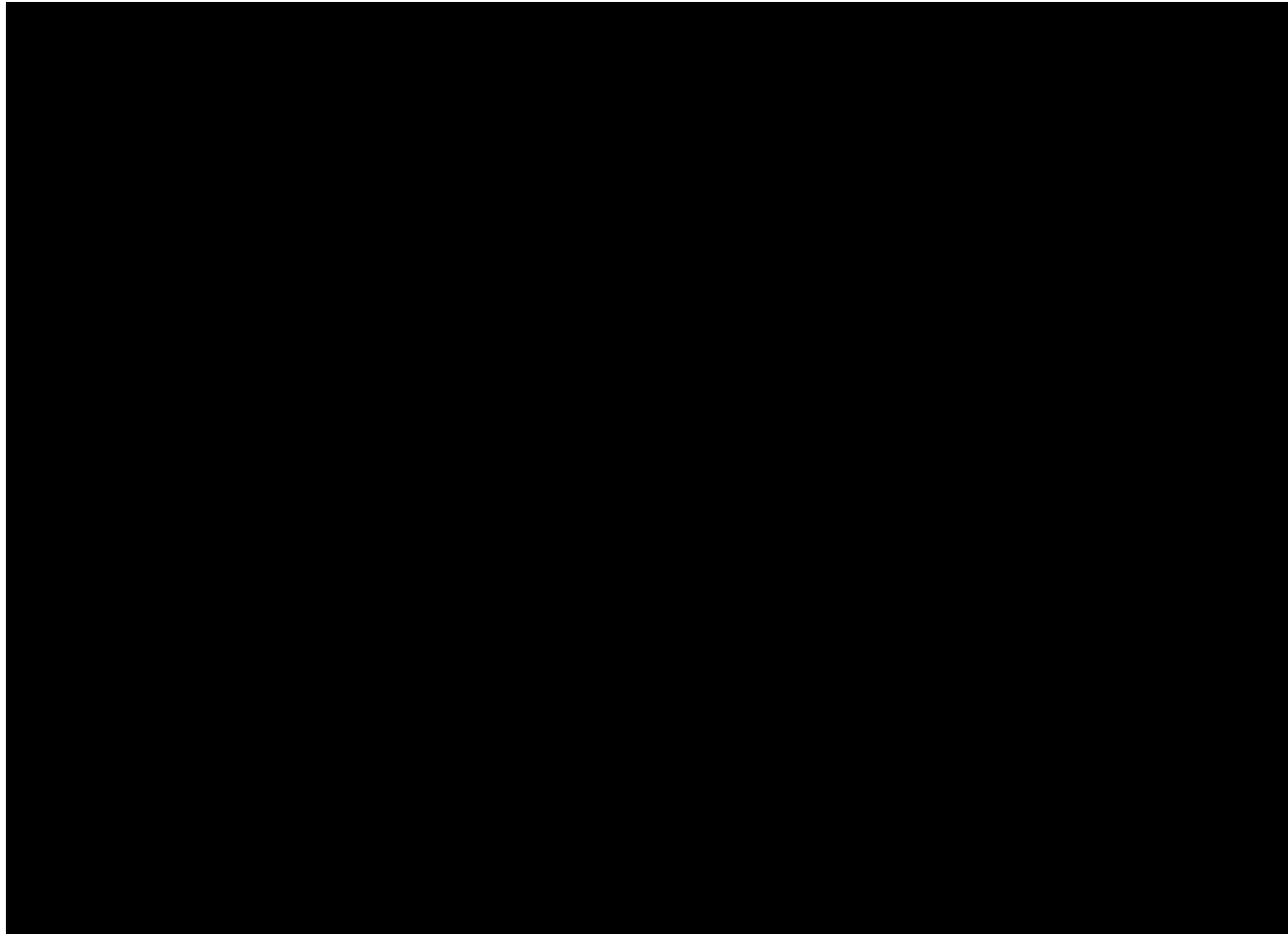


Green Apple Books (San Francisco), March 2007  
Photo by Dolan Halbrook (Copyright 2007. Creative Commons Licensed)

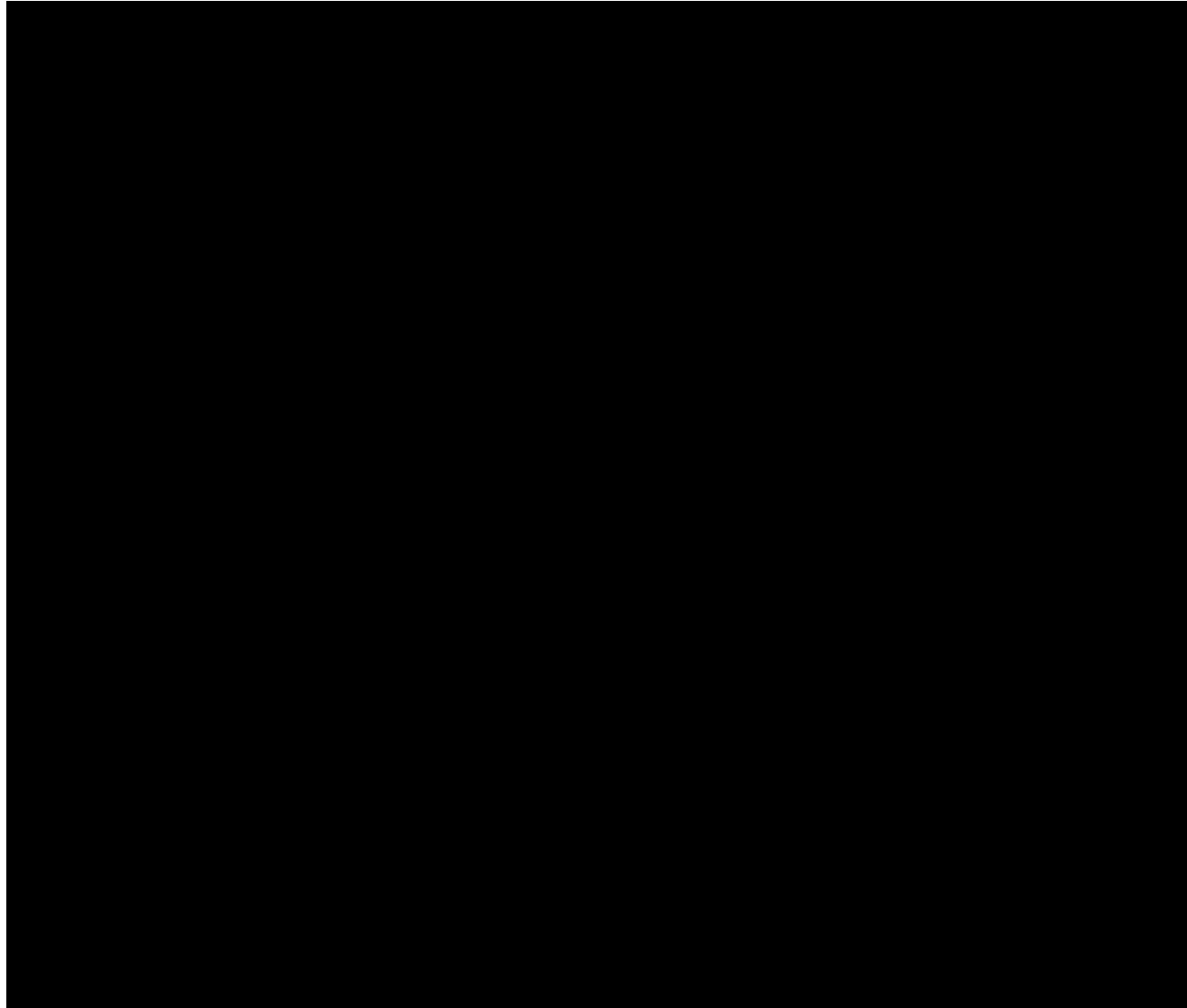
# More content not on shelves...



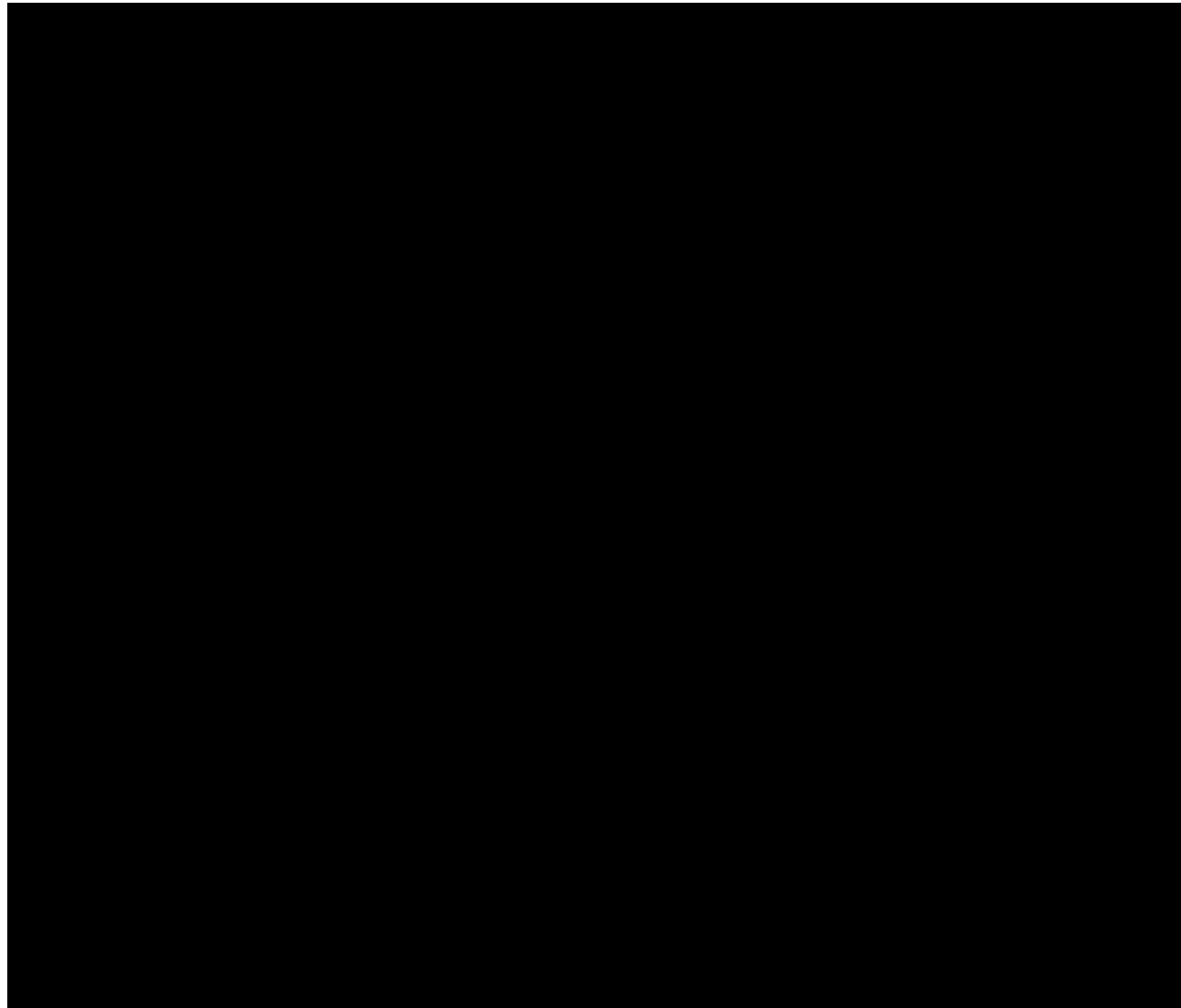
# ...and hard to browse



# Women in Google Books

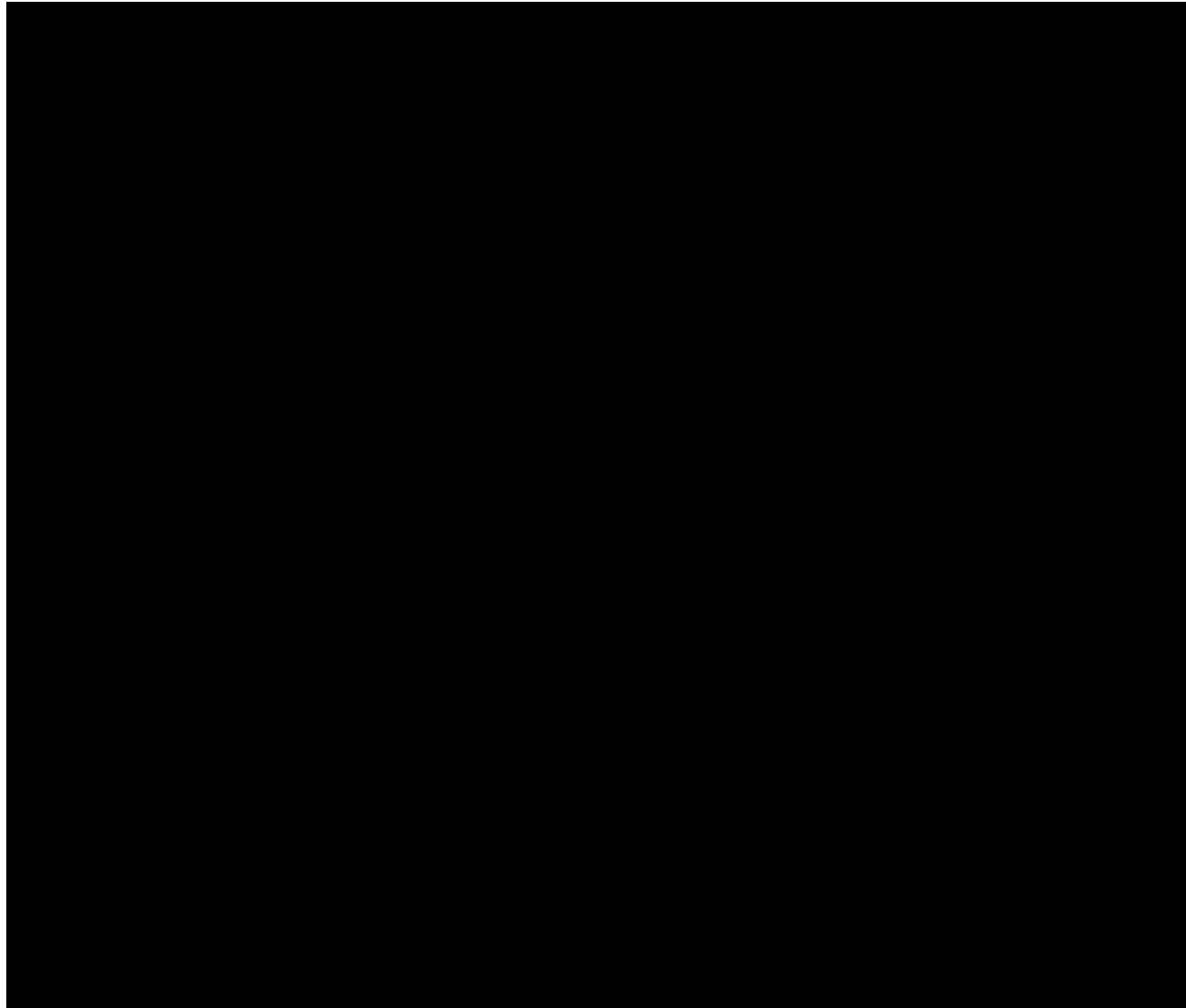


# In Amazon

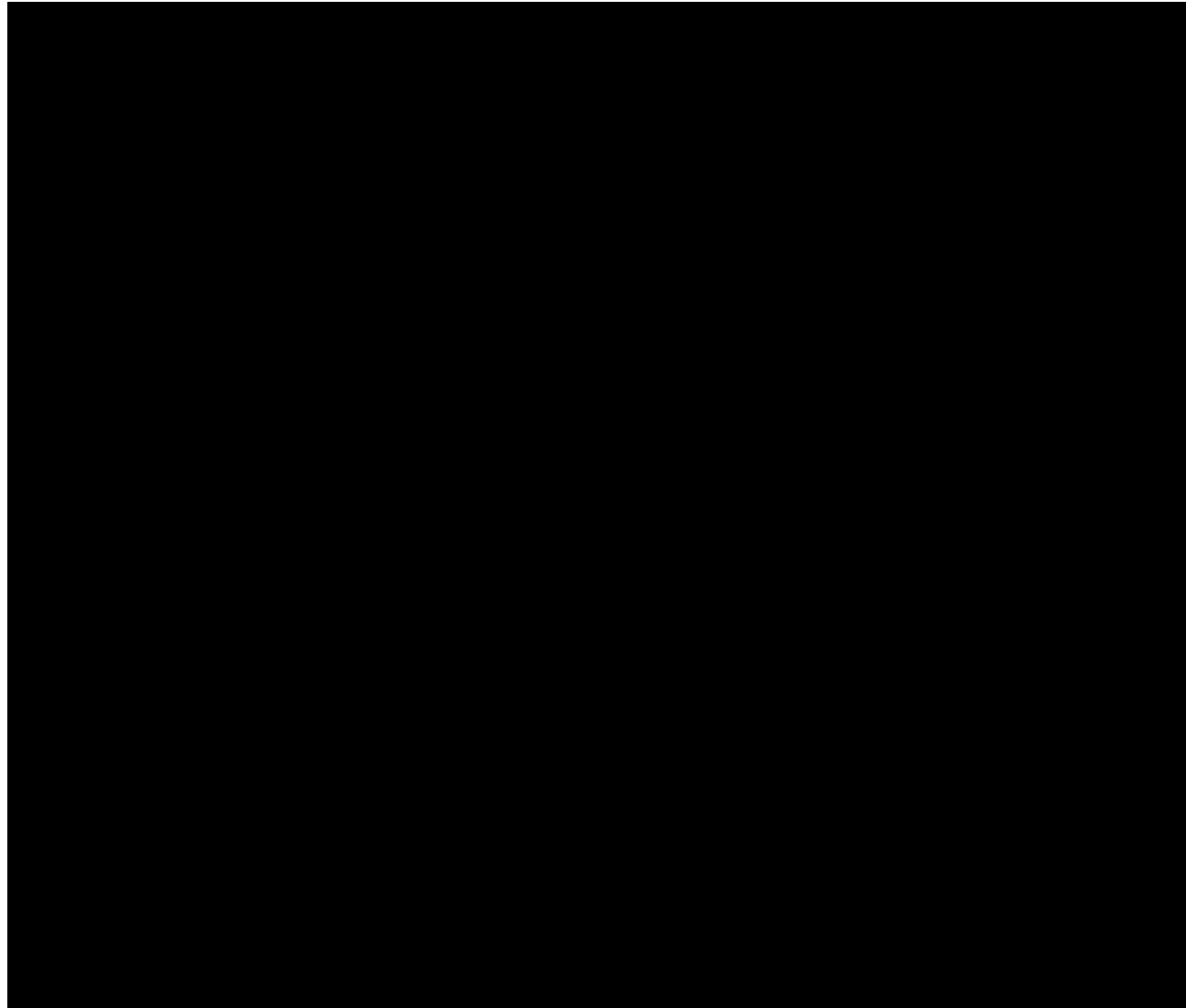




# In academic tag-space



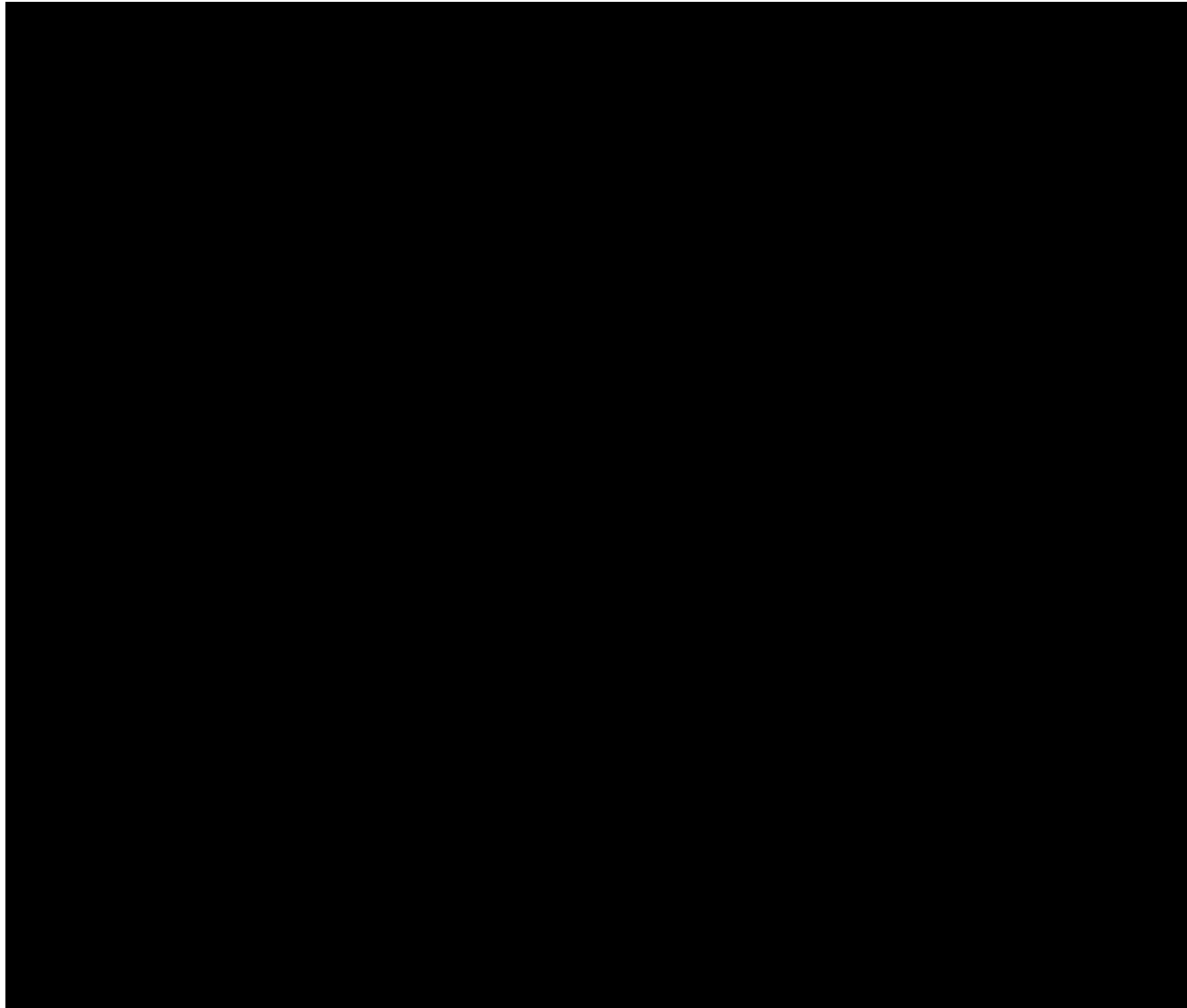
# In our catalog



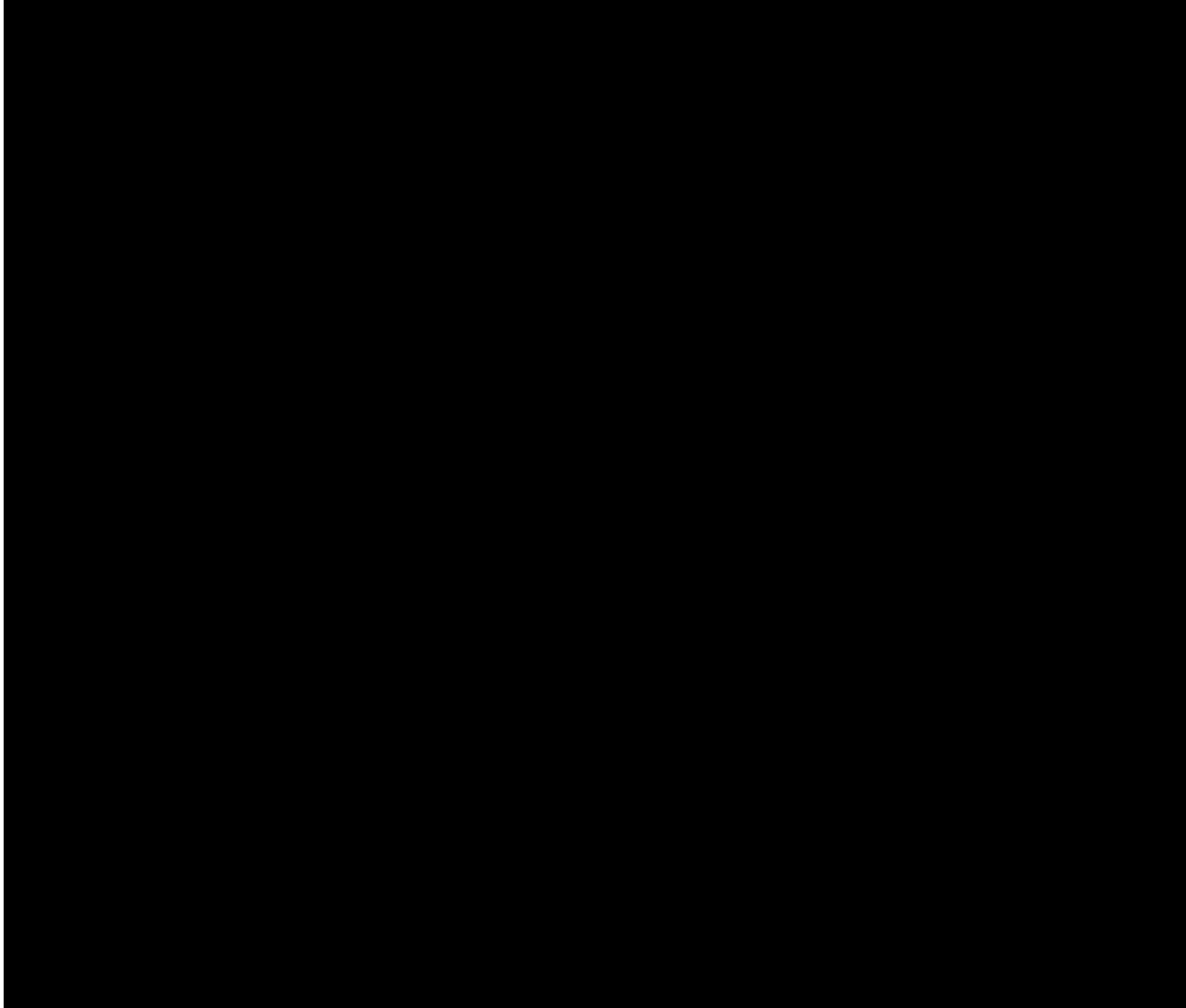
# The strength and the weakness of the subject catalog

- “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

The screenshot shows a web browser window displaying the NCSU Libraries Online Catalog search results for the term 'women'. The browser's address bar shows the URL: <http://www2.lib.ncsu.edu/catalog/?N=0&Nty=1&Ntk=Subject&view=full&Ni>. The search results page features a navigation bar with links for 'Search the Collection', 'Browse Subjects', 'Services', 'Library Information', 'Community', and 'News & Events'. Below the navigation bar, there is a search box with the text 'Catalog Search: women' and a dropdown menu set to 'in Subject Heading'. The search results indicate that 24,729 items were found. A 'Narrow By Call Number Range' section lists various subject categories and their corresponding item counts, such as 'A - General Works (13)', 'M - Music (146)', and 'Z - Bibliography. Library Science. Information resources (ge ... (259)'. A 'Narrow Results By' section is also visible, showing 'Subject: Topic' with sub-categories like 'Women (9989)' and 'History (5829)'. The search results are sorted by 'Relevance' and displayed in a list format. The first result is 'The book as art : artists' books from the National Museum of Women in the arts' by Wasserman, Krystyna, published in 2007. The second result is 'Embracing sisterhood : class, identity, and contemporary Black Women' by McDonald, Katrina Bell, published in 1961. The browser's status bar at the bottom shows 'Done'.

# Better browsing with facets

## But what are you missing?

The screenshot shows a web browser window displaying the NCSU Libraries Online Catalog search results for the term 'women'. The browser's address bar shows the URL: <http://www2.lib.ncsu.edu/catalog/?N=0&Nty=1&Ntk=Subject&view=full&Ni>. The search results page features a navigation bar with links for 'Search the Collection', 'Browse Subjects', 'Services', 'Library Information', 'Community', and 'News & Events'. Below the navigation bar, there is a search box with the text 'Catalog Search: women' and a dropdown menu set to 'in Subject Heading'. The search results indicate that 24729 matching items were found. A 'Narrow By Call Number Range' section lists various call numbers and their corresponding item counts, such as 'A - General Works (13)' and 'Z - Bibliography. Library Science. Information resources (ge ... (259)'. A 'Narrow Results By' section is also visible, with 'Subject: Topic' and a list of related terms like 'Women (9989)' and 'History (5829)'. The search results are displayed in a list format, with the first result being 'The book as art : artists' books from the National Museum of Women in the arts' by Wasserman, Krystyna. The search interface includes a search bar at the bottom with the text 'Find: zua' and various search options like 'Find Next', 'Find Previous', 'Highlight all', 'Match case', and 'Phrase not found'.

NCSU LIBRARIES

Search the Collection | Browse Subjects | Services | Library Information | Community | News & Events

MY LIBRARY: Library Account | My Course Reserves | My Alerts

Catalog Search: women in Subject Heading Go Start Over Send search to: Go

Search 'women':  
We found 24729 matching items. Limit results to currently available items.

Narrow By Call Number Range:

- A - General Works (13)
- B - Philosophy. Psychology. Religion (1066)
- C - Auxiliary Sciences of History (246)
- D - History (General) and History of Europe (633)
- E - History: America (747)
- F - America: local history (261)
- G - Geography. Anthropology. Recreation (465)
- H - Social sciences (7617)
- J - Political Science (282)
- K - Law in general. Comparative and uniform law. Jurispruden ... (293)
- L - Education (713)
- M - Music (146)
- N - Fine Arts (388)
- P - Language and literature (7450)
- Q - Science (324)
- R - Medicine (671)
- S - Agriculture (70)
- T - Technology. (282)
- U - Military science (General) (63)
- V - Naval science (14)
- Z - Bibliography. Library Science. Information resources (ge ... (259)

Narrow Results By:

Subject: Topic

- Women (9989)
- History (5829)
- History and criticism (2681)
- Women authors (2401)
- Social conditions (2353)

Show More ...

Brief View | Full View Sort By: Relevance

- The book as art : artists' books from the National Museum of Women in the arts  
Author: Wasserman, Krystyna.  
Published: c2007.  
Format: Book  
Design Library  
N7433.35 .U6 W37 2007 Stacks Checked Out
- Embracing sisterhood : class, identity, and contemporary Black Women  
Author: McDonald, Katrina Bell, 1961-

Find: zua Find Next Find Previous Highlight all Match case Phrase not found

Done

# So, what are subject maps?

- **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 (**though they don't look much like them**)**
  - 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 (“Hotel’s just off the Metro from LA”)
    - » Differences in concepts and names (“You mean light rail?”)
- **Work well with detailed ontologies like LCSH**



# LCSH? Why in heaven's name voluntarily use LCSH?

- **Subject headings are expensive to assign, maintain**
  - Though not as expensive as a full MARC record
- **They represent a >100 year old legacy system**
  - With all the backward compatibility issues, mismatches with user expectations, “prejudices and antipathies” that implies
- **Our patrons aren't using subject headings-based discovery much**
  - But with the tools we give them, can you blame them? [see critiques by Karen Schneider and others]
- **LCSH may be on its way out, too....**

# Reasons to use LCSH

## (can we at least say “for now”?)

- **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 described by it, and have no other subject metadata (and might never have)
- **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 (which isn’t as sensitive to main vs. incidental)
  - Much larger array of effectively controlled areas than tags or keywords
- **Subject maps will make it easier for users to navigate sensibly with LCSH**
- **Subject maps and related technologies may also make it easier to make quicker changes to LCSH, transition to better ontologies**

# Aggregated subject map example

- “American Discovery” demo: 40,000 records of various kinds from various OAI-harvested sources
  - 10,000+ books from Making of America (Michigan)
  - 3,000+ books and images from Documenting the American South (UNC)
  - 26,000+ other books selected from American Memory, Early Canadiana Online, Celebration of Women Writers, etc. (The Online Books Page, Penn)
- **That’s not very big, though, is it?**
  - It’s big enough to demonstrate basic aggregation issues
  - We’ll show larger scale in a bit
- **Live online demo:**
  - <http://onlinebooks.library.upenn.edu/adsubjects.html>
- **Did it work?**

# Building subject maps

## Build them automatically:

- **First, create a collection-independent 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 suitable terms not present, may be inferrable from other info
  - 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**

# Well-ordered ontology? Dream on

- **Making of America Books**
  - LCSH, but sometimes uses obsolete terms
- **Documenting the American South**
  - Images use Thesaurus of Graphic Materials, not LCSH
- **Online Books Page index**
  - Didn't assign LCSH terms to many items
  - But did have LC Call Numbers
- **Many other collections appear to use uncontrolled or collection-specific keywords**
  - Institutional repositories, special image collections, user-tagged items

# How to deal with subjects from more than one ontology?

- **1. Throw them in together, hope for the best**
  - May sometimes be the simplest, cheapest approach
  - Terms not in “standard” ontology may be isolated in map, but still sought by users
  - Automated techniques can often integrate terms that follow general patterns of ontology:
    - » Digital libraries -- Congresses -- California -- Pasadena
    - » Rose Bowl Stadium (Pasadena, Calif.)
  - Automated analysis may also identify “islands” within subject maps that could use integration into map
    - » “ ‘Housing bubble’ is isolated, but is used a lot”

# How to deal with subjects from more than one ontology?

- 1. Throw them in together, hope for the best
- 2. Normalize to a preferred ontology
  - LCCNs -> mapped to most tightly defined LC subject heading
  - Obsolete terms -> resolve “used for” associations; reformat old styles (e.g. geographic headings)
  - Uncontrolled keywords -> Correlations (manual or automated tag correlation); clustering (see Newman and Hagedorn work with OAlster)
  - Appropriate when alternate terms not well controlled, or aren’t very suitable as a basis for browsing
  - But you may want to label auto-assigned terms
  - **And do you really want everything to be LCSH?**



# How to deal with subjects from more than one ontology?

- 1. Throw them in together, hope for the best
- 2. Normalize to a preferred ontology
- 3. Make two subject maps, link them in appropriate places
  - May be appropriate when audience, type of materials is clearly distinct
  - E.g. MeSH for medical researchers and specialist medical literature vs. LCSH for nonmedical researchers and nonspecialist literature on medicine
  - Establish junction points where appropriate (e.g. Cancer (LCSH) <-> Neoplasms (MeSH)). Crosswalks exist.
  - **Not so appropriate where searchers often interested in items in both ontologies (e.g. TGM and LCSH)**

# How to deal with subjects from more than one ontology?

- 1. Throw them in together, hope for the best
- 2. Normalize to a preferred ontology
- 3. Make two subject maps, link them in appropriate places.
- 4. Build a multi-ontology subject map
  - One ontology may dominate (e.g. LCSH's 200,000+ authorized headings vs. TGM's 6,000+)
  - Some terms in one ontology related to ones in other
    - » Tombs & monuments (TGM) -> Two LCSH terms
  - Other subjects may have two equivalent terms
    - » Churches (TGM) = Church buildings (LCSH)
  - **How do you make this clear in the interface?**

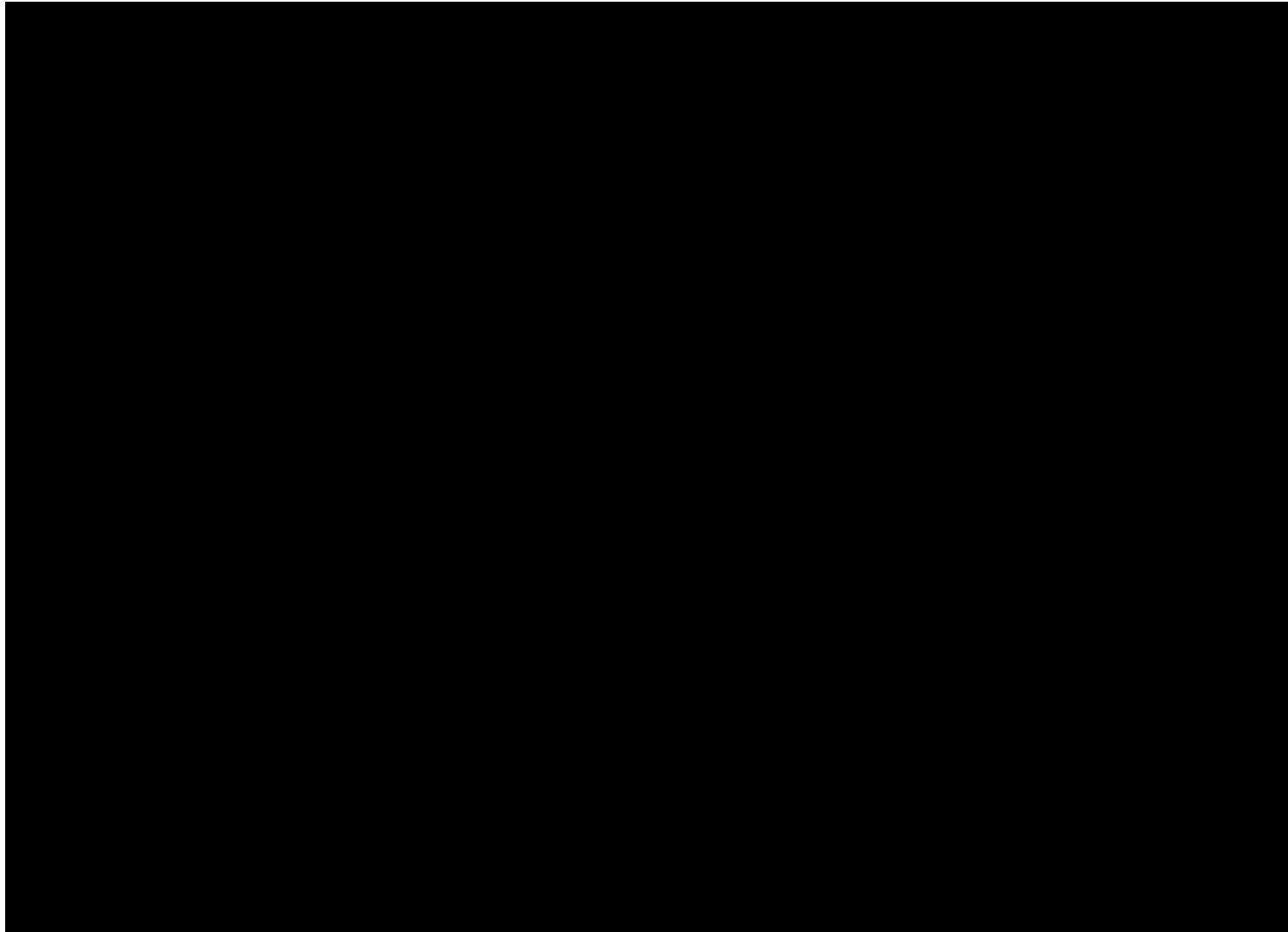
# Bringing a new collection into an aggregation

- **Examine the metadata used**
  - How are subjects represented?
  - What ontologies are used, if any?
  - How reliable or precise are they?
  - Should you normalize or suppress terms?
- **Survey items in the collection, items already present**
  - Do you need to deduplicate?
  - For duplicates (at manifestation or work level), should one collection's metadata override another's?
  - If cataloging is good, union of subject terms may be best
- **Some expert judgment required for best fit**
  - But may be feasible to do this collection-by-collection (**please, not item-by item!**)

# But why stop with just digital collections?

- **We have lots of Americana and other useful content in our print holdings too**
  - Franklin lists over 3 million print items
  - Over 1 million subject terms
  - We can build subject maps for it too
    - » Can be newly generated from harvested data in ~40 minutes on up-to-date hardware with 8GB memory
- **Live online demo:**
  - <http://onlinebooks.library.upenn.edu/adsubjects.html>
  - (Not yet aggregated with earlier digital collections, but can be done with a little more programming)
- **Did it work?**

# Subject maps on Franklin



# Can I contemplate doing this with the OPAC I'm stuck with?

- We can use some tools to do it with our ILS (Voyager)
  - **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 Franklin catalog
- Subject map-aware catalogs might be even nicer, but the tools above suffice to get subject maps going

# But can subject maps play with other cool 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**

- Tags can be correlated with formal ontology terms, as discussed earlier

# 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 next?

- **Try putting it in front of our university's users**
  - Perhaps as add-on to Franklin, or interface to E-resources
- **Try aggregating more collections**
- **Develop smarter map building and display techniques**
- **User studies?**
  
- **Think about how our users will discover the best resources (ours or others') in a networked, highly digital world**

# Conclusions (again?)

- Subject browsing is an important mechanism for resource discovery across aggregated collections
- 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](mailto:ockerblo@pobox.upenn.edu)
- **Thanks!**