12-5-2006

New Maps of the Library: Building Better Subject Discovery Tools Using Library of Congress Subject Headings

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

A white paper accompanying a presentation at the CNI Task Force Meeting, December 5, 2006. (The paper is misdated 2007.) The Powerpoint slides for the presentation are included as a supplementary file.

New Maps of the Library: Building Better Subject Discovery Tools Using Library of Congress Subject Headings

A brief white paper accompanying a CNI presentation, December 2007

John Mark Ockerbloom, University of Pennsylvania Library

Abstract

We describe tools in development at the University of Pennsylvania to generate and display interactive "subject maps" for exploring library collections. Based on the Library of Congress Subject Headings (LCSH), these maps are automatically built from existing authority records, a collection's bibliographic records, and optional local "tweaks" for local interests and search patterns. Users can explore these maps via ordinary text-based web browsing, and browse clusters of related research resources. We now provide these maps for small collections like The Online Books Page, and are experimenting with maps for the entire Penn Library catalog. We hope to enable users to take full advantage of the rich conceptual relationships in LCSH-based library collections, and more effectively browse increasingly diverse and dispersed library collections.

Why subject maps?

Subject-based browsing is an important tool for discovering relevant research materials. Readers often find useful and unexpected resources by browsing open stacks organized by subject, for instance. Libraries likewise have put considerable effort into creating detailed subject descriptions for their materials using complex ontologies, most notably the Library of Congress Subject Headings (LCSH). Increasingly, many resources are available to researchers that are not on local shelves, including resources available via closed or offsite storage, online access, or inter-library loans. Such resources could be browsed online, but the browsing needs to be more informative than the limited alphabetic subject browse and search that are offered by most online catalogs and that give a very limited "worms-eye view" of collection subjects. It should be easier for researchers to find relevant subject areas and items in those and related areas. With appropriate tools, LCSH-based discovery can give more focused results for established subject areas than full text search does, more organized and diverse subject repertoires than folksonomies or simple keyword ontologies provide, and richer and more varied relationships between subjects than purely facet-oriented discovery allows. Subject maps can provide LCSH-based discovery tools that better provide such features.

What are subject maps?

Subject maps are organized networks of well-defined subject terms, and relationships between them, applied to a particular collection of items, and displayed in ways that allow researchers to easily view and navigate among closely related subjects and their associated items. Like geographic maps, they let their readers see details of a particular area (in this case a topical rather than geographical area), attractions of nearby areas, and ways of moving to those areas to explore them in more detail. Subject maps avoid placing unnecessary constraints on how users can move around. For instance, they support many more directions of movement than simple taxonomic hierarchies do. At

any given location, they attempt to present as much detail as readers will find useful, without overwhelming them.

From a user's point of view, subject maps are presented as clusters of subjects and clusters of items described by those subjects. These clusters could be displayed graphically, but in this paper we describe an information-dense text display that supports quick orientation and navigation without requiring special software or high-powered client computers. Technically, the internal representations of subject maps are graphs with annotated nodes for subject terms, and annotated edges for relationships.

Why use LCSH as the basis for subject maps?

The Library of Congress Subject Headings provide a particularly detailed and developed set of subjects and relationships developed and applied to bibliographic items over more than 100 years. LCSH describes a much larger range of subjects than most other controlled vocabularies, it has a large number of relationships defined either explicitly in the ontology or implicitly through facets or other conventions, and its subject terms are based on actual usage in the literature they describe, making them natural searching and browsing targets. Furthermore, millions of items in thousands of libraries are already described using LCSH, and have not been described in any other classification schemes to the same level of detail.

Subject maps can also be generated for other ontologies, but they are especially suitable for ones like LCSH with a wide source vocabulary and repertoire of relationships.

How can the user view and navigate subject maps?

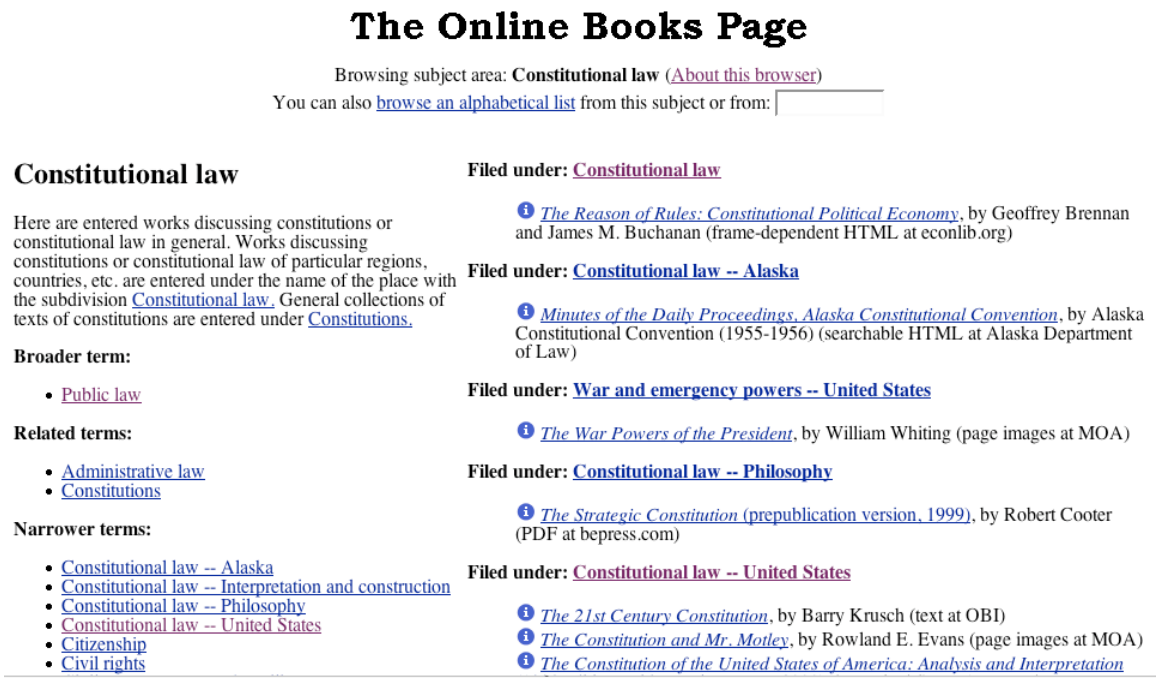

Figure 1: A view of a subject map for The Online Books Page, with focus at "Constitutional law"

In Figure 1, we show a portion of a subject map rendered as a two-column text display in an ordinary web browser. At the top, the display shows the collection being explored, the subject that is the current primary focus of the display, and options to shift to other views, such as alphabetical subject listings, with the same or different primary focuses. If the collection is being filtered through other kinds of descriptive facets (such as media type or date) that information could also be displayed and changed in the upper region.

The larger part of the display is spit into two halves. On the left half appears the subject in primary focus, its usage notes, and clusters of its directly related subjects, organized by the nature of the relationship. This half is generated from data records created when the map is generated. On the right half of the display is a cluster of items described by the subject of primary focus, along with items described by related subjects. This half is generated by real-time subject queries to the collection.

The display has extensive hyperlinks to allow users to switch to different subject focuses, differently oriented subject views, detailed item records, and possibly to full digital content where available.

How can robust subject maps based on LCSH be created?

Subject maps use a variety of relationships for linking and clustering subjects, including broader/narrower term relationships, related term relationships, and "used for/see" relationships, all of which appear in LCSH. These relationships are derived in a variety of ways, including explicit declaration in the ontology, customization by local librarians, usage in the collection being mapped, and the content of the subject terms themselves. In subject maps, relationships are annotated with the method used to derive them, and subjects are annotated with notes on scope and usage within a particular collection. These annotations are used to organize and document item and subject clusters.

Specific sources of subject relationships include:

- **Authority files:** The Library of Congress Authorities define a large set of subject terms, accompanied with scope notes, related terms, and other information. This set can be mined to create an initial, collection-independent subject map. This step is only the start, however, of creating a subject map tuned to a particular collection.
- **Local customizations:** Subject maps need not be limited to authorized headings. Our implementation of subject maps can also read auxiliary data (which we call "tweaks") to add new subject terms and relationships useful to a local community, without the overhead of changing an official authority database. One potentially useful fruitful source of customizations is analysis of local search logs. Terms that are repeatedly searched without success might be automatically detected, so that librarians can review them and add appropriate cross-references to authorized terms.
- **Terms used in collection items:** The collection to be mapped is scanned for subject terms in its metadata. A new map is then drawn (or overlaid) on the original collection-independent map. New subject terms used in the collection are added,

and terms from the original map not referenced in the collection (either directly, or indirectly via narrower terms) are dropped. This results in a different collection-tailored map, possibly very different in scale from the one originally generated. Subject terms are also annotated with data on usage in the collection, most notably the number of items described by each term. This information helps us build appropriately sized and organized cluster views.

- **Facet analysis:** LCSH is partially based on facets that not always independent of each other. It is not difficult to automatically derive new terms and relationships from them. The term "Hospitals – Arizona – History", for instance, is inferred to be related to a broader term "Hospitals – Arizona". To find additional relationships, we can also check for existing subject terms that might remove facets not at the end of the term (such as "Hospitals – History") or that use permutations of the facets (which in LCSH may reflect subtly different meanings, or simply inconsistency in original cataloging).
- **Lexical and domain analysis:** Many subject relationships are implicit in LCSH. For example, it is not uncommon for terms to be related through alphabetical arrangement (e.g. "Charities" and "Charities, Medical") without this relationship explicitly declared in LC authorities. Simple lexical analysis may derive many relationships like these with high degrees of confidence. (An occasionally spurious relationship inference is not a big problem, if it means that many relevant items, and only a few irrelevant items, get added to a cluster.) A small amount of domain information can aid in the lexical inference of other relationships. For example, using a list of states and their abbreviations, and tracking subject assignments in a particular collection, we might infer that "San Francisco (Calif.)" is related to the broader term "California".
- **Co-location of subjects in item records:** In some cases, we can infer new relationships between subjects from the way they are used in a given collection. Some relationships can be inferred from known metadata patterns. For instance, if an item is given a primary subject that is a personal name (such as "Stanton, Elizabeth Cady, 1815-1902") and a secondary subject that ends with a biography facet (such as "Suffragists -- United States -- Biography") it is highly likely that the person named in the first subject is an instance of the term qualified by "Biography" in the second.) Some other systems also infer relationships through simple co-location. If a significant number of items are classified both under topical term A and topical term B, A and B might be usefully related and clustered together. The New York Public Library, for instance, has used this technique in its expanded subject search for its 500,000-item digital image gallery.

Many of the relationship inferences above can be calculated by independent computational components. Depending on an institution's ingenuity, interest in extensive clustering, and computing resources, different inference components can be added to or removed from an institution's map-generation process. Also, if some of the inferences occasionally produce highly conspicuous incorrect relationships, "tweak" data could also be introduced to suppress those relationships.

What scales of collections can subject maps be used for?

Because subject maps adapt their source ontology to the collection in which they are used, subject maps can be applied at a variety of scales. In small collections, irrelevant subject terms are deleted from the maps, preventing navigational dead ends. Large collections can also be supported; while the maps themselves take longer to build than for smaller collections, interactive map display simply requires the system to query a particular subject and its nearby neighbors. The time required for this does not increase with the overall size of the map. (It does increase somewhat with the size of the local cluster, but in practice, response tends to fast even for densely clustered areas.)

At Penn, we use subject maps in production for our Online Books Page collection of digital books, which consists of over 26,000 items and over 13,000 subject terms. The server, which runs on Sun hardware now several years old, tends to display map views quickly, and generates new maps once a day in about five minutes. We also have a prototype subject map interface to our full Franklin catalog, with over 2 million items and 1 million subject terms. The map takes about 90 minutes to generate on newer hardware. In both cases, map generation is only necessary to reflect new subject terms and relationships; new items appear in subject maps immediately if they are indexed under existing subjects. Hence, it would be reasonable to generate maps infrequently, such as once a week, if necessary. (In fact, both the Online Books and the Franklin maps build from an initial collection-independent map from the Library of Congress authorities that takes a couple of hours to generate on fast hardware. But once that map is generated, we can reuse it repeatedly, only re-generating it when our subject authority database changes substantially.)

Heavily used subjects and relationships, which tend to appear in larger collections, may in some cases call for alternative display strategies, often to avoid overwhelming the user with too much information rather than to avoid system performance problems. We are now investigating techniques for improved display and navigation of such subjects.

How can subject maps be used to improve LCSH?

LCSH has drawn criticism over the years for mismatches between the terms and relationships it prescribes and those used by present-day library users. Also, like other very large subject ontologies, it is difficult to maintain, users have difficulty finding the right subject terms to use, and items may be cataloged inconsistently over time. These problems are to be expected in any large-scale, widely-scoped subject ontology. Our subject map system is designed to alleviate many of these problems. Simultaneous display of subjects, related terms, and applicable items make it easier for users to learn the use of their library's subjects as they view and navigate through the map. It also may make it easier for librarians familiar with the ontology to identify missing relationships. The customized "tweaks" on top of the native ontology, informed by log analysis of failed searches, can allow librarians to quickly fill in missing relationships and subject terms encountered by users. Clustered display of related subjects and their items allows users to find the items of interest to them in multiple related subjects displayed together.

Subject maps therefore should make LCSH and other ontologies like it easier both to use and to maintain.

How can subject maps be integrated into a broader searching and browsing experience?

Subject maps represent just one of a variety of tools that researchers can use to find items, including free-text search, alphabetic browsing, citation analysis, tagging, and facet-based filtering. Subject maps can coexist with many of these techniques and systems. For instance, in the Online Books Page we allow users to switch between traditional alphabetic subject listings and clustered subject map views. We are also considering combining subject maps with filtering on non-subject facets, such as date, language, and media type, common search filters in many online catalogs. This can be supported by annotating subject nodes at map-generation time with the set of non-subject facets or numeric ranges that apply to items described by that subject or its narrower relations. (This annotation can often be concisely noted for facets that have limited controlled vocabularies, or integer values.) Then, when displaying subject maps under facet filtering, we simply suppress subjects that have not been annotated with the appropriate facets in the local clusters, while adding the appropriate facet conditions to the real-time queries that retrieve clustered items.

We can also integrate subject maps with our existing vendor-supplied OPAC. We have already integrated our OPAC with PennTags, a locally-developed academic shared annotation syetem, by adding custom JavaScript to our library catalog record displays to show annotations and links for bibliographic items that have been tagged by our users, and by adding computed links from PennTags postings back into the OPAC display for tagged books. We could use similar techniques to add links to appropriately focused subject maps from subject listings and item records in our OPAC, and vice versa.

What is the future of subject maps, and where can one learn more?

Our work on subject maps is still in its early stages. We have developed them sufficiently to use them in production on The Online Books Page, and to create proof of concept demonstrations for Penn's full library catalog. In future work, we hope to improve our construction, clustering, and display of subject maps, test integrating them with our library catalog, evaluate their performance and effectiveness with users, and investigate further how they compare and relate to other discovery mechanisms.

We believe our work to date is sufficient to demonstrate the promise of subject maps with LCSH and similar ontologies. We hope that presenting out work thus far inspires further ideas and collaboration on subject maps, and promotes an appreciation of what can be done with LCSH-based cataloging and appropriate tools.

More information about the subject maps project, and links to demonstrations of subject maps in action, can be found at http://labs.library.upenn.edu/subjectmaps . The author can be contacted at ockerblo@pobox.upenn.edu.