Enhancing Search and Browse for Scholarly Discovery

Automated Clustering of OAI Metadata

Kat Hagedorn
Suzanne Chapman

University of Michigan
Digital Library Production Service

Introduction

Web search engines trump aggregated bibliographic services in the end-users' minds because they offer searching that is easy to use and results that are easy to understand. Librarians are aware that it is often difficult to create simple interfaces to complex online resources, however a marriage of the two worlds is not impossible. Our research into clustering bibliographic materials provides a test of this marriage.

The metadata aggregator is in a position to add value to the metadata to make it easier to discover. Faced with an ever-expanding corpus of metadata in the OAIster database, and a simple, but increasingly ineffective, method for searching it, we developed a prototype searching and browsing interface that would allow users to access this large corpus using a controlled classification built upon clustered groups of metadata.

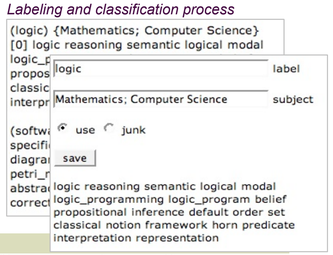
Clustering, in our definition of the term, is taking the words and phrases that make up metadata records and gathering them together into semantically meaningful groupings. We used an automated clustering technique called Topic Modeling, developed at the University of California Irvine. The resulting prototype was part of an Institute of Museum and Library Studies (IMLS) grant to the Digital Library Federation (DLF) on second-generation OAI work.

1. Labeling & Classification

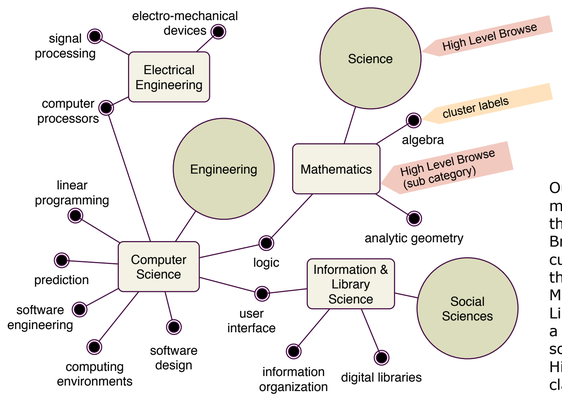
As of September 2006, the OAIster collection included 7.5 million records, a 94,000 word vocabulary, and a total of 290 million word occurrences. This collection was more than sufficient to produce 500 high-fidelity clusters representing the subjects spanned by all the records.

We created a community labeling Web page that would allow colleagues in our department to choose clusters close to their subject expertise and determine labels for those clusters. After the labeling process, there were 352 usable and labeled clusters out of the 500 clusters learned by the Topic Model (junk topics were discarded).

Labeling and classification process



Label	Cluster
tumors	tumor cell human cancer carcinoma normal tumour myc mammary ras expression leukemia growth malignant tpa mouse hpv cell_lines tissue skin
christianity	church churches religious religion cathedral catholic russia_federation moac christian wedding saint chapel methodist photographic_essay bishop christ holy rev mary mission
junk	strong degree weak degrees strength strongly aggregation freedom high weakly stronger depend higher presence large studied highly small exhibit respect
junk	image images motion object segmentation tracking camera shape texture scene contour pixel vision visual stereo algorithm matching detection registration estimation



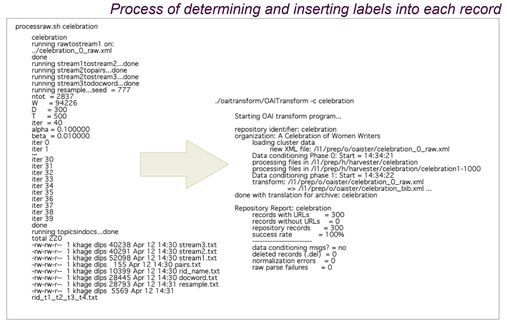
Our next task was to match these labels to the High Level Browse classification currently in use at the University of Michigan University Library. To the left is a graphic depiction of some sections of the High Level Browse classification.

2. Assigning Labels

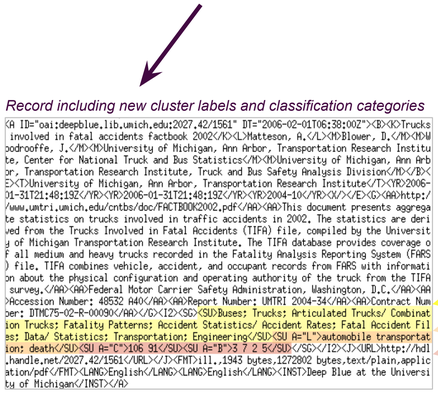
With the classification scheme decided upon, and cluster labels created and mapped to the scheme, we needed to marry the categories and labels to the records. The most effective method for doing so was to include the categories and labels in the records themselves.

The University of California Irvine created a tool that ranked the top four clusters associated with a record, based on the algorithm's statistical processes. At UM, we then created a modified version of the tool we use to transform harvested metadata for OAIster into our native format (DLXS Bibliographic Class). This tool used the UCI files for each data contributor to insert the cluster labels, and their associated High Level Browse classification categories, into records.

Process of determining and inserting labels into each record

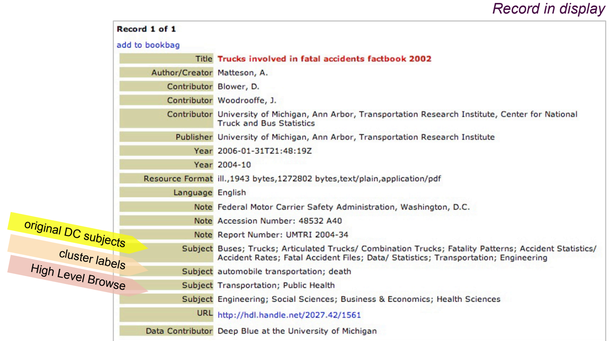


Record including new cluster labels and classification categories



Within the records, the new terms received new subject field attributes (e.g., <SU A="L">) so that our DLXS software could make labels and categories available for searching, browsing and displaying.

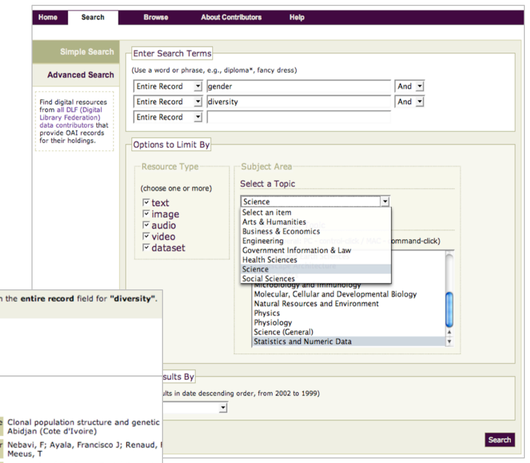
Record in display



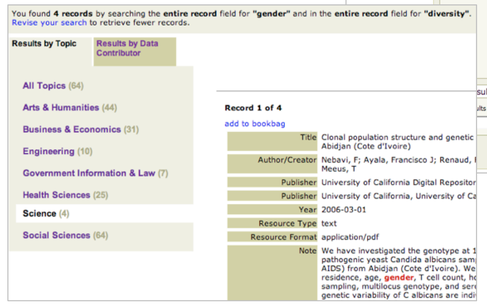
3. Search & Browse Interface

The prototype, or *DLF Portal*, contains both basic and advanced search options and a browse feature.

Only the advanced search interface incorporates the High Level Browse classification. The end-user can choose a top-level category and sub-level category(ies) as a way to limit his search.



http:://quod.lib.umich.edu/i/imls/



The results page enables the end-user to expand or narrow the scope of his search results without needing to perform his search again. The facets ("Browse by Topic" and "Browse by Data Contributor") allow him to view the records using multiple (duple) classifications, which increases the possibility of finding useful materials because he is not limited to a single classification.

Browse showing solely High Level Browse categories (left) and with topic labels (right)

For the browse page, we were able to show sub-level categories so end-users could choose a small enough subject set to browse through. Unfortunately, these sets are often not small enough to browse through in their entirety. The graphic to the right shows a hidden option to view the cluster labels as they relate to the categories.

Lessons Learned

- The Topic Model approach can be time-intensive, e.g., assigning labels and categories to metadata records for the prototype took around 48 hours for 62 repositories of over 2.6 million records.
- Records with a humanities bent fared worse than those describing science resources, e.g., they contain less metadata, often contain metaphors that are lacking in science records.
- The High Level Browse classification scheme had its drawbacks as well, e.g., we were not able to adequately place the clusters that were associated with war (e.g., "world war II") into appropriate sub-categories.
- The real power of including new subject terms was on the search results page, e.g., narrowing/expanding the results, clarification of vague or broad search queries.