D-Lib Magazine
July/August 2007

Volume 13 Number 7/8

ISSN 1082-9873

Enhancing Search and Browse Using Automated Clustering of Subject Metadata

Kat Hagedorn
Metadata Harvesting Librarian
Digital Library Production Service, University of Michigan University Library
Ann Arbor, MI
<khage@umich.edu>

Suzanne Chapman
User Testing and Interface Specialist
Digital Library Production Service, University of Michigan University Library
Ann Arbor, MI
<suzchap@umich.edu>

David Newman
Research Scientist
Department of Computer Science, University of California Irvine
Irvine, CA
<newman@uci.edu>

Abstract

The Web puzzle of online information resources often hinders end-users from effective and efficient access to these resources. Clustering resources into appropriate subject-based groupings may help alleviate these difficulties, but will it work with heterogeneous material? The University of Michigan and the University of California Irvine joined forces to test automatically enhancing metadata records using the Topic Modeling algorithm on the varied OAIster corpus. We created labels for the resulting clusters of metadata records, matched the clusters to an in-house classification system, and developed a prototype that would showcase methods for search and retrieval using the enhanced records. Results indicated that while the algorithm was somewhat time-intensive to run and using a local classification scheme had its drawbacks, precise clustering of records was achieved and the prototype interface proved that faceted classification could be powerful in helping end-users find resources.

Introduction

Those interested in discovering academic materials will seek out Web search engines, and more often than not Google, first [1]. The Web search engines tend to trump vendor databases and aggregated bibliographic services in the end-users' minds because they offer searching that is easy to use and provide 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.

During our research we took to heart Peter Morville's truism – "you can't use what you can't find" [2]. Of the methods that lessen the end-user's burden by making information more findable, we were most interested in improving search and implementing faceted classification (providing multiple views, or facets, of the data). Faced with an ever-expanding corpus of metadata in the OAIster database [3], and a simple, but increasingly ineffective, method for searching it, we developed a prototype searching and browsing interface that would allow end-users to access this large corpus in recognizable chunks.

As Shreeves et al. have rightly stated, "shareable metadata should support search interoperability" [4]. Unfortunately, to date with the 750+ repositories that are included in OAIster's collection of shared metadata, variability among repositories severely hinders interoperability [5]. Unlike full-text collections that can be searched on bibliographic metadata and full text, bibliographic collections often have limited metadata (e.g., only title and author). The metadata aggregator is in a position to add value to the aggregation to enhance discoverability. Others have worked at reprocessing aggregated metadata [6,7], and their efforts provided the impetus for our research into improving search and browse for our prototype.

The prototype we developed was part of an Institute of Museum and Library Services (IMLS) grant to the Digital Library Federation (DLF) for second-generation Open Archives Initiative (OAI) research. Our first-generation work created OAIster – this second grant helped us refine issues surrounding OAI as a protocol and its use in the world of digital online resources. To showcase our findings, we developed a prototype portal, called the DLF Portal, modeled on OAIster and using a sub-set of OAIster metadata [8]. This portal contained the searching and browsing interface for the clustered metadata.

Clustering Algorithm

Clustering, in our definition of the term, means taking words and phrases that make up metadata records and gathering them together into semantically meaningful groupings. For instance, a record about the feeding and care of cats can be grouped with a record about the feeding and care of hamsters. While we could cluster records based on a number of different facets (e.g., format, author, publisher), we recognize that clustering by subject area is most useful for our end-users [9].

To achieve this, we used an automated clustering technique called Topic Modeling [10], developed at the University of California Irvine (UCI). The power of this technique is in its ability to cluster effectively using decidedly less text than a full-text indexing algorithm. Additionally, while others have tested similar statistical clustering techniques on a repository-by-repository basis [11,12,13], we chose to test this algorithm on a large, heterogeneous corpus of metadata – one that reflects the problems inherent in effectively searching large databases of information.

The Topic Model is an implementation of Latent Dirichlet Allocation, which is a probabilistic version of Latent Semantic Indexing. To create semantically meaningful clusters, the algorithm automatically "learns" a set of topics that describe a collection of records, i.e., without direct human supervision it can discover meaningful clusters of records. These topics are discovered by the algorithm through finding patterns of words that tend to co-occur throughout the collection. Newman et al. have written a separate paper fully describing the Topic Model algorithm and improvement experiments performed after the current prototype work [14].

The input to the Topic Modeling algorithm was made up of the records from 668 OAIster repositories (the September 2006 update). Of these, 163 primarily non-English repositories and 117 very small repositories were excluded, so that we could limit the scope of the project. For each of the remaining 388 repositories, the contents of the Dublin Core title, subject and description fields were tokenized (made lowercase, punctuation removed, simple stemming). Additionally, we augmented a standard English stopword list (*the*, *and*, *that*, *with*, etc.) with words that had little topic value but occur frequently in metadata records, such as *volume*, *keyword*, *library*, and *copyright*.

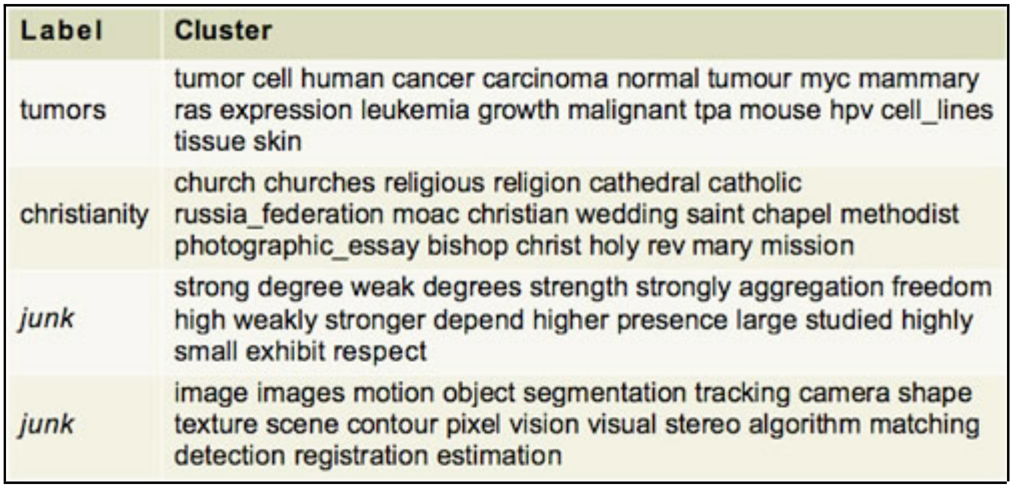
The final representation of our processed OAIster collection included 7.5 million records, a 94,000 word vocabulary, and a total of 290 million word occurrences. Because this collection of records was too large to process on our system, we chose every third record, producing a collection containing 2.5 million records and a total of 96 million word occurrences. This collection was more than sufficient to produce 500 high-fidelity topics representing the subjects spanned by all 7.5 million records.

By using the complete OAIster collection of metadata as input, we were working with a broad and varied subject corpus. Because the quality of the topical clusters would depend upon as comprehensive a vocabulary as possible, this should have resulted in better topical clusters than if we'd used a smaller input. (The prototype itself contained all DLF repositories – collected for the purpose of the grant and numbering 62 repositories and over 2.6 million records.)

Developing Labels and Mapping to Classification

Based on UCI's prior experience, 500 topics (clusters) were sufficient to cover the majority of subject areas. Initially, we thought these clusters would only be used by our system to enable searching and browsing; however, upon review the clusters themselves appeared useful for results display. To that end, we decided to create a unique label for each cluster using a word or phrase associated with each cluster that would define or describe it. These labels would be integrated into the clustered records themselves.

Although we think that using subject experts would be ideal for the labeling process, for this pilot we did not have appropriate resources to enable this. Instead, we created a "community labeling" Web page [15] that would allow colleagues in the Digital Library Production Service and the Scholarly Publishing Office of the University of Michigan University Library to choose clusters close to their subject expertise and determine labels for them. During the process of labeling, a colleague could decide to "junk" a cluster, i.e., decide that it wasn't a worthwhile subject grouping. This was the case for clusters that contained words that either did not seem to belong together or did not have a strong subject affiliation with each other. The Topic Model created both types of clusters – its statistical nature implies it will create clusters from closely associated terms, even if they are not subject-based.



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

Table 1. Labels for "good" clusters and "junk" clusters.

Once community labeling was finished, one person performed a quality check of the labels to catch inconsistencies in specificity or language. After the labeling process, there were 352 usable and labeled clusters out of the 500 clusters "learned" by the Topic Model.

With labels created, our next task matched these labels to an appropriate classification system in order to provide access to our collection via discrete subject categories. The classification we chose was the High Level Browse classification currently in use at the UM University Library for electronic journals [16] and new books [17], as well as the list of databases in our federated search engine, Search Tools [18]. While a locally-created classification system could theoretically be problematic when mapping to a large, diverse set of labels, the UM system was based directly on the encompassing Library of Congress Classification System. Consequently, we were fairly certain that High Level Browse would provide the subject range we required.

The High Level Browse classification contains fewer than a dozen top-level categories and over 100 sub-level categories. We chose to remove two of the top-level categories that were not subject-based (i.e., "General Reference", "News & Current Events"). Mapping to the classification took place at the same time as labeling, for the sake of efficiency and, hopefully, consistency. Colleagues chose one or more sub-level categories for each label. These sub-level categories were automatically mapped to the top-level categories by our software.

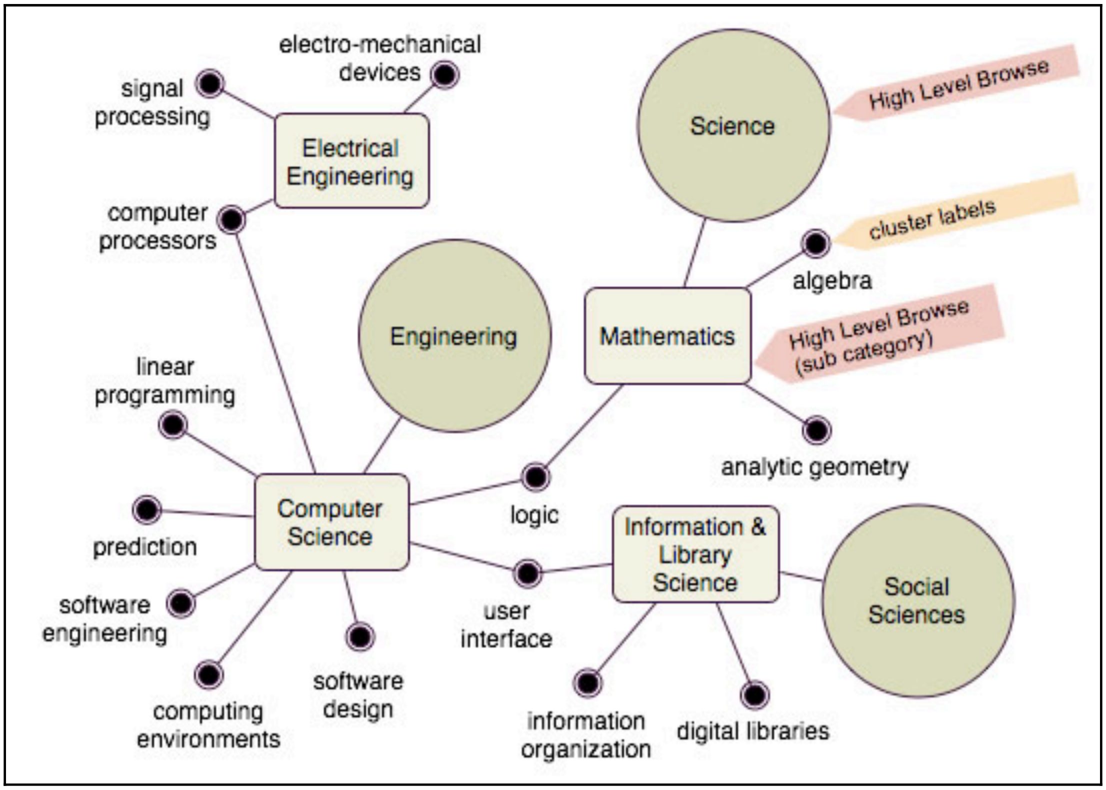


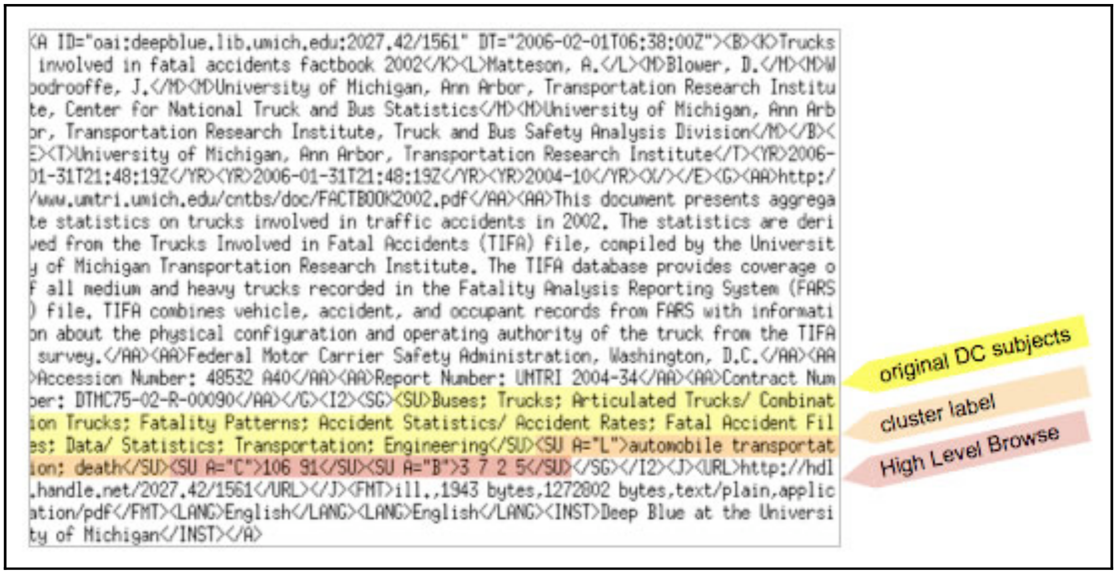
Figure 1. Cluster labels and their assigned High Level Browse top-level and sub-level categories.

Integrating Classified Metadata

With the classification scheme decided upon, and cluster labels created and mapped to the scheme, we needed to marry the classification categories and labels to the metadata. The most effective method for doing so was to include the categories and labels in the metadata records themselves; however, matching to the records during a search would have been too time-intensive and would have prevented them from being searchable (but still browsable). Consequently, inserting labels and categories was a two-step process.

UCI created a tool that ranked the top four clusters associated with a record, as computed by the algorithm. We ran this tool for all the records in each repository included in the prototype (e.g., on all the records from the Library of Congress Digitized Historical Collections repository). The resulting file contained the four cluster labels (numeric representations of these) associated with each record in the repository. In the end, we had 62 files – one per repository.

In order to perform the final step of inclusion of the categories and labels in the records, we modified the UM tool used to transform harvested metadata for OAIster into our native format (DLXS Bibliographic Class). [19] Two global files were created first: one to associate each of the 352 cluster labels with their assigned sub-level categories, as had been determined by our cadre of community labelers during the labeling process, and the second to associate the numeric representations of the clusters with their corresponding labels. These files plus the files from each of the 62 repositories created by UCI's tool were used to insert the labels and categories into metadata records as subject fields with new attributes (e.g., A="L") that could be understood by our DLXS software.



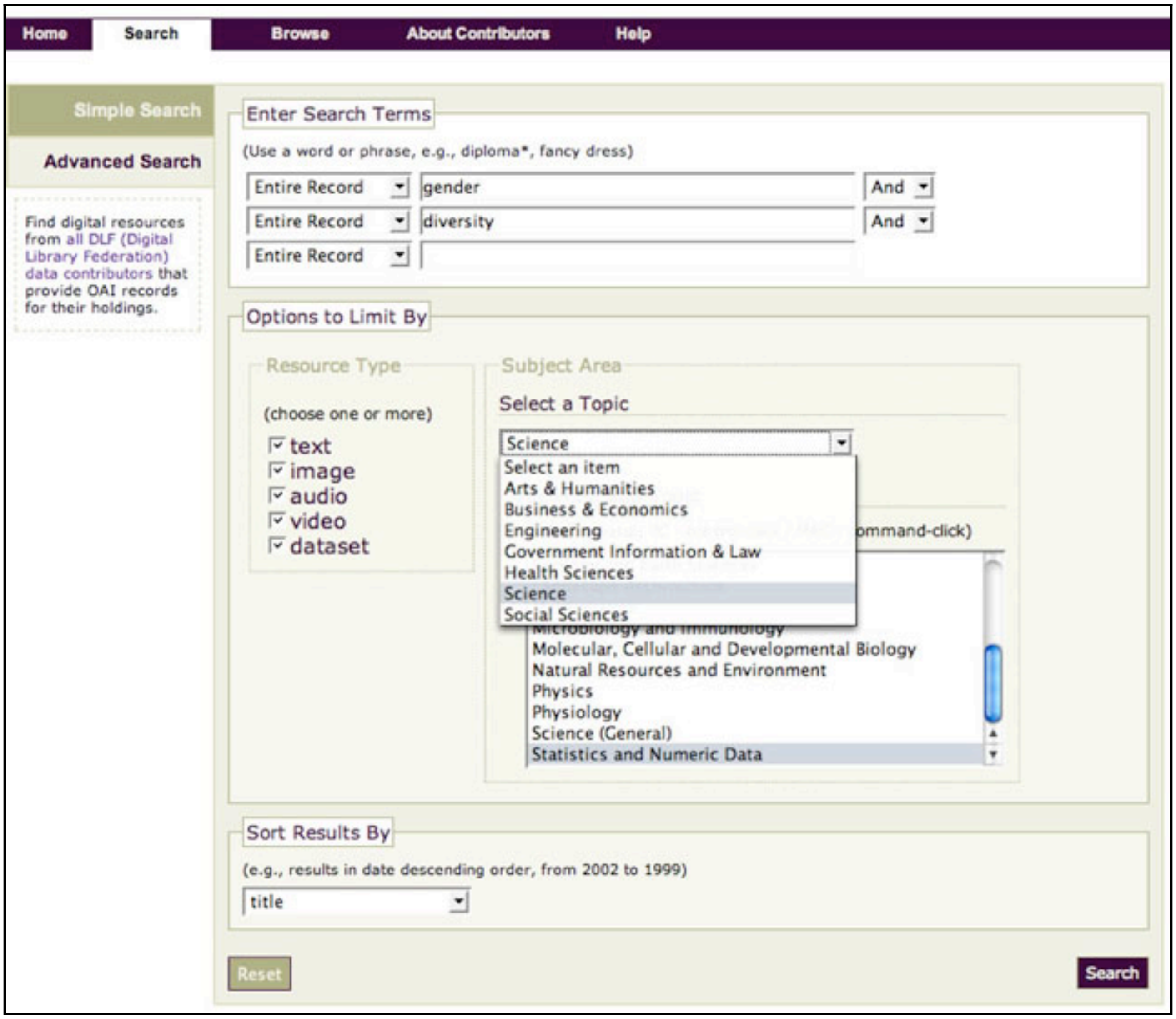
```
<A ID="oai:deepblue.lib.umich.edu:2027.42/1561" DT="2006-02-01T06:38:00Z"><B><K>Trucks
 involved in fatal accidents factbook 2002</K><L>Matteson, A.</L><M>Blower, D.</M><M>W
oodrooffe, J.</M><M>University of Michigan, Ann Arbor, Transportation Research Institu
te, Center for National Truck and Bus Statistics</M><M>University of Michigan, Ann Arb
or, Transportation Research Institute, Truck and Bus Safety Analysis Division</M></B><
E><T>University of Michigan, Ann Arbor, Transportation Research Institute</T><YR>2006-
01-31T21:48:19Z</YR><YR>2006-01-31T21:48:19Z</YR><YR>2004-10</YR><X/></E><G><AA>http:/
/www.umtri.umich.edu/cntbs/doc/FACTBOOK2002.pdf</AA><AA>This document presents aggrega
te statistics on trucks involved in traffic accidents in 2002. The statistics are deri
ved from the Trucks Involved in Fatal Accidents (TIFA) file, compiled by the Universit
y of Michigan Transportation Research Institute. The TIFA database provides coverage o
f all medium and heavy trucks recorded in the Fatality Analysis Reporting System (FARS
) file. TIFA combines vehicle, accident, and occupant records from FARS with informati
on about the physical configuration and operating authority of the truck from the TIFA
 survey.</AA><AA>Federal Motor Carrier Safety Administration, Washington, D.C.</AA><AA
>Accession Number: 48532 A40</AA><AA>Report Number: UMTRI 2004-34</AA><AA>Contract Num
ber: DTMC75-02-R-00090</AA></G><I2><SG><SU>Buses; Trucks; Articulated Trucks/ Combinat
ion Trucks; Fatality Patterns; Accident Statistics/ Accident Rates; Fatal Accident Fil
es; Data/ Statistics; Transportation; Engineering</SU><SU A="L">automobile transportat
ion; death</SU><SU A="C">106 91</SU><SU A="B">3 7 2 5</SU></SG></I2><J><URL>http://hdl
.handle.net/2027.42/1561</URL></J><FMT>ill.,1943 bytes,1272802 bytes,text/plain,applic
ation/pdf</FMT><LANG>English</LANG><LANG>English</LANG><INST>Deep Blue at the Universi
ty of Michigan</INST></A>
```

Figure 2. New subject fields for cluster labels and classification categories.

Designing and Building Prototype

Discussions about the search, browse and results prototype interfaces centered on estimates of the effectiveness of certain designs, and the time and resources needed to build them. While we hoped to briefly test these designs with end-users or staff, we only had limited time to build the prototype and this effectively removed time for testing. We received comments after the prototype launched, which will be discussed in the next section.

Both basic and advanced search options are available in the DLF Portal. While arguments can be made that cluster labels and classification categories should be made available in both options, we assumed that simple search should be as Google-like as possible, without added features to distract the end-user. Consequently, only the advanced search interface incorporates the High Level Browse top-level and sub-level categories.



Find digital resources from all DLF (Digital Library Federation) data contributors that provide OAI records for their holdings.

Enter Search Terms

(Use a word or phrase, e.g., diploma*, fancy dress)

Entire Record	gender	And
Entire Record	diversity	And
Entire Record		

Options to Limit By

Resource Type

(choose one or more)

☑ text
☑ image
☑ audio
☑ video
☑ dataset

Subject Area

Select a Topic

Science

- Select an item
- Arts & Humanities
- Business & Economics
- Engineering
- Government Information & Law
- Health Sciences
- Science
- Social Sciences

ommand-click)

- Microbiology and Immunology
- Molecular, Cellular and Developmental Biology
- Natural Resources and Environment
- Physics
- Physiology
- Science (General)
- Statistics and Numeric Data

Sort Results By

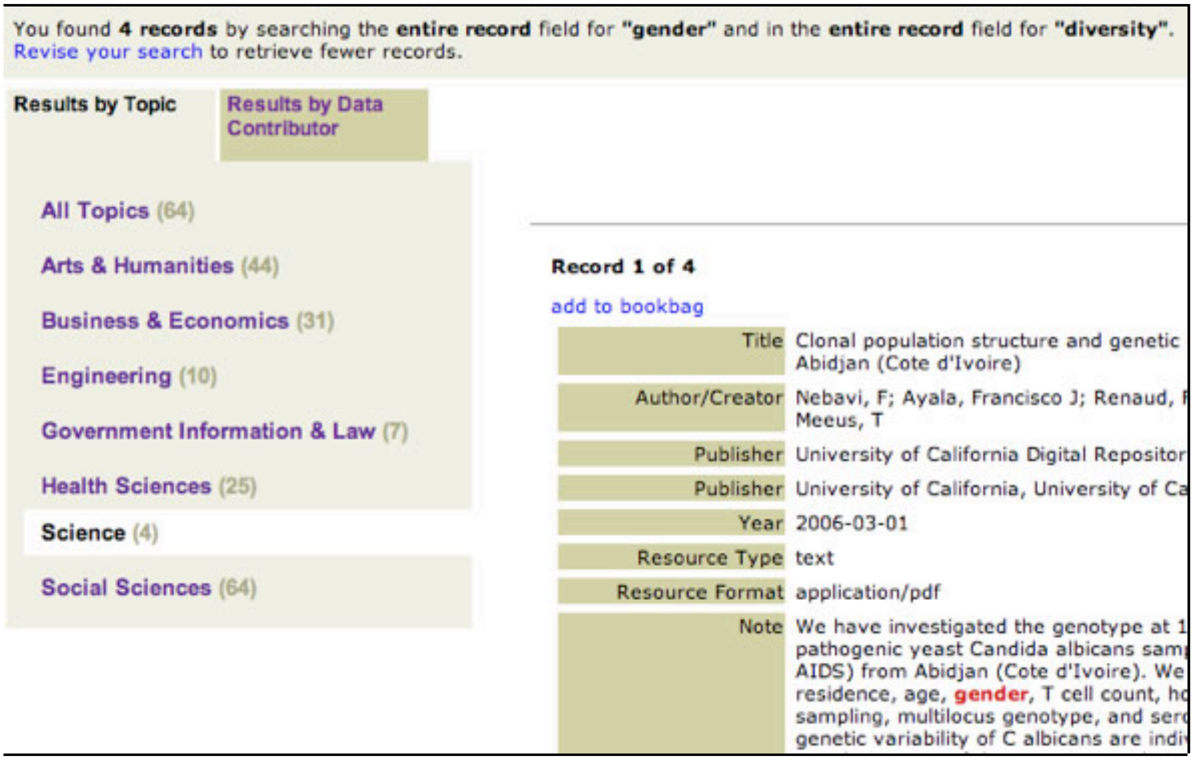
(e.g., results in date descending order, from 2002 to 1999)

title

Figure 3. Advanced search page including High Level Browse categories.

With this new interface, in addition to original options to enter search terms and resource type, the end-user can choose a top-level category and sub-level category(ies) as a way to limit his search. The first option, labeled "Topic," contains the High Level Browse top-level categories. The second option, labeled "Sub-Topic," shows the list of High Level Browse sub-level categories appropriate to the top-level category chosen.

Organizing the interface by this method results in two different ways to perform a search. If, for instance, the end-user enters the search in Figure 3 ("gender" and "diversity"), and chooses "Science" as a top-level category and "Statistics and Numeric Data" as a sub-level category, the end result is Figure 4. By showing the other top-level categories in the results, we allow the end-user to expand his search, e.g., by choosing the 31 hits for "Business & Economics". This enables the end-user to expand the scope of his search results without needing to perform his search again. (Time factors prevented us from implementing a results view of the sub-level categories and their hits.)



You found **4 records** by searching the **entire record** field for **"gender"** and in the **entire record** field for **"diversity"**. Revise your search to retrieve fewer records.

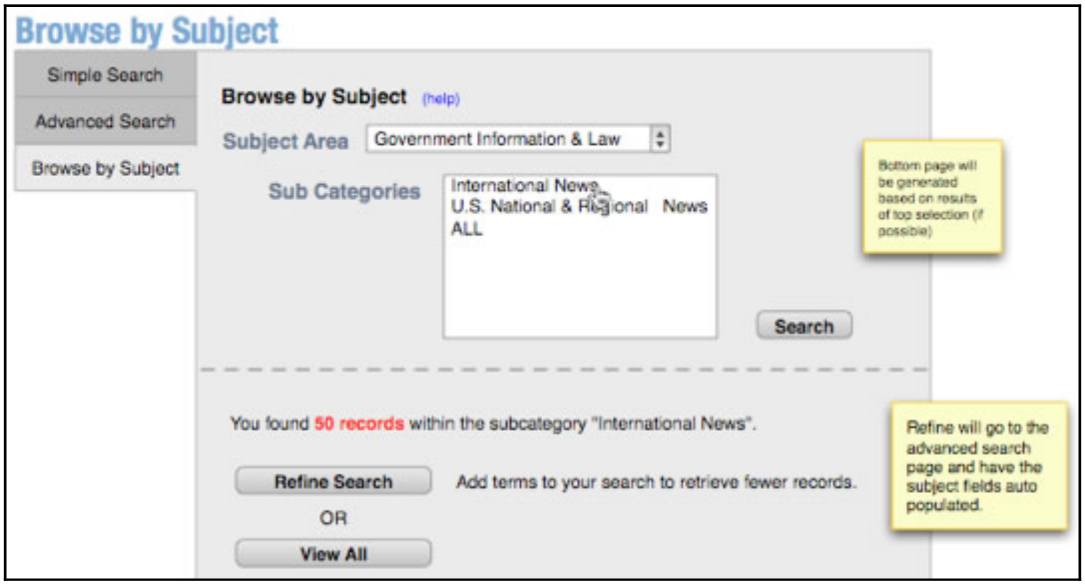
Record 1 of 4

add to bookbag

Title	Clonal population structure and genetic Abidjan (Cote d'Ivoire)
Author/Creator	Nebavi, F; Ayala, Francisco J; Renaud, F Meeus, T
Publisher	University of California Digital Repositor
Publisher	University of California, University of Ca
Year	2006-03-01
Resource Type	text
Resource Format	application/pdf
Note	We have investigated the genotype at 1 pathogenic yeast Candida albicans sam AIDS) from Abidjan (Cote d'Ivoire). We residence, age, gender, T cell count, h sampling, multilocus genotype, and ser genetic variability of C albicans are indi

Figure 4. Results page after choosing High Level Browse categories from search.

Initially, we thought we could achieve this same functionality by providing browse and search together on the same page, as in Figure 5. However, this method would have prevented the end-user from seeing hits found in other categories.



Subject Area: Government Information & Law

Sub Categories: International News, U.S. National & Regional News, ALL

Bottom page will be generated based on results of top selection (if possible)

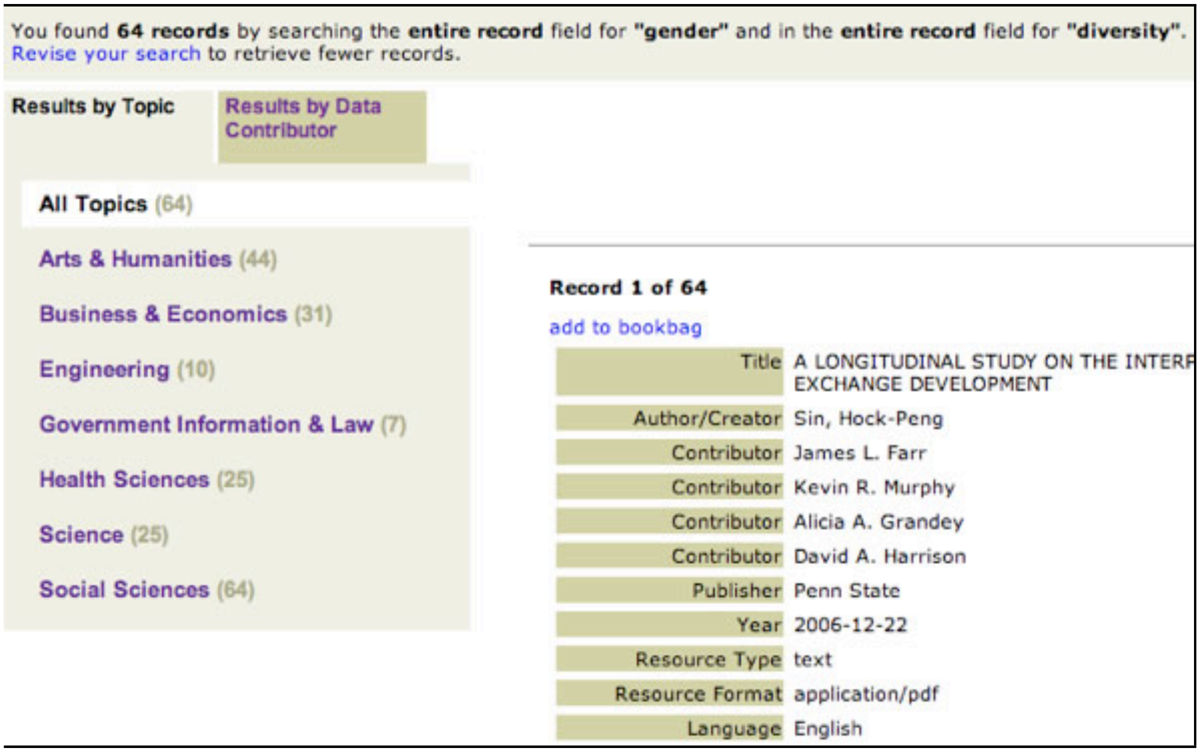
You found 50 records within the subcategory "International News".

Add terms to your search to retrieve fewer records.

Refine will go to the advanced search page and have the subject fields auto populated.

Figure 5. Initial mockup of browse/search page.

The second method of performing a search is to not limit by topic on the advanced search page. If an end-user performs the "gender" and "diversity" search from Figure 3, but does not choose a top-level or sub-level category, his results will look like those in Figure 6. If the end-user is interested in narrowing his search results, he can do that from this page without re-doing his search, e.g., by choosing the 44 hits for "Arts & Humanities". This also allows the end-user to easily change the focus of his original search.



You found **64 records** by searching the **entire record** field for **"gender"** and in the **entire record** field for **"diversity"**. Revise your search to retrieve fewer records.

Title	A LONGITUDINAL STUDY ON THE INTERP EXCHANGE DEVELOPMENT
Author/Creator	Sin, Hock-Peng
Contributor	James L. Farr
Contributor	Kevin R. Murphy
Contributor	Alicia A. Grandey
Contributor	David A. Harrison
Publisher	Penn State
Year	2006-12-22
Resource Type	text
Resource Format	application/pdf
Language	English

Figure 6. Results page without choosing High Level Browse categories from search.

Browse functionality was provided on a separate page. For this page, as shown in Figure 7 (left), 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 very large and end-users may be reluctant to browse through the entire set. We were unable to incorporate the opportunity to revise a search using the browse category selected (similar to that shown in Figure 5). Figure 7 (right) also shows a "hidden" option for viewing the cluster labels associated with their categories that we hope to implement in the future. (To view, place "&displaytids=1" at the end of the browse page URL.)

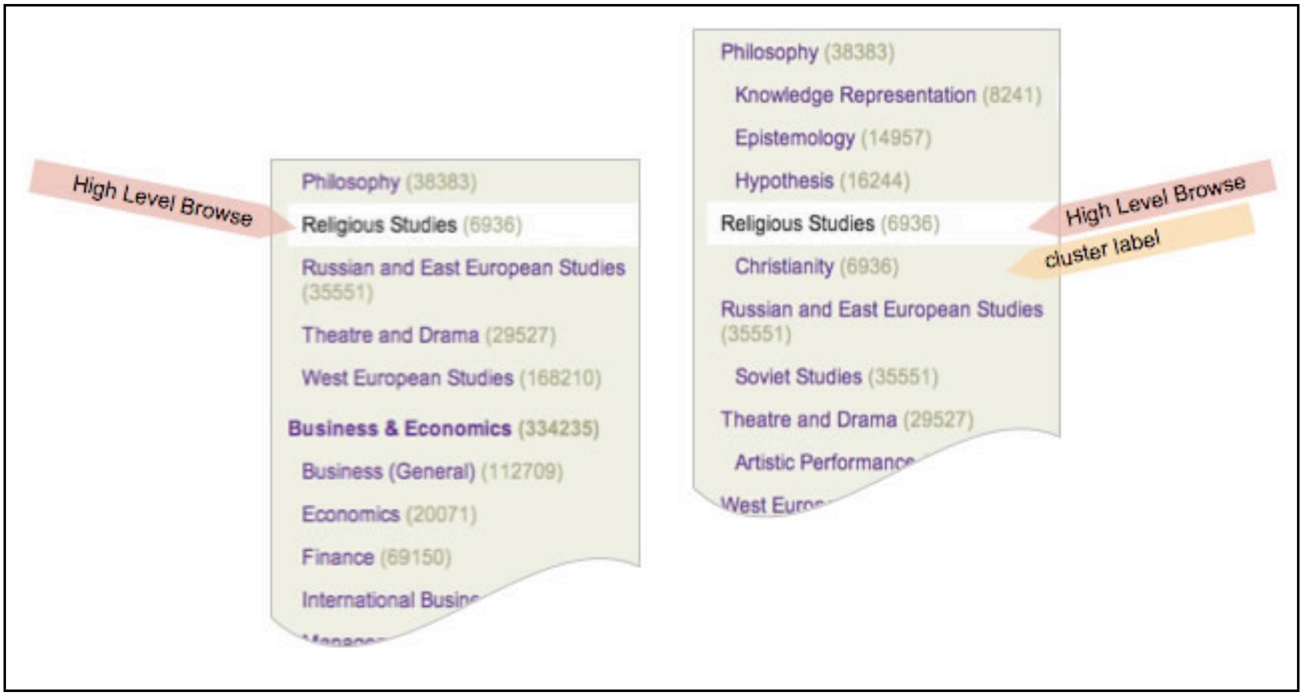


Figure 7. Browse page of categories, and optional expansion of cluster labels.

For each record on a results page, the subjects are displayed in the same order as in the raw data – Figure 8 in the next section provides a visual description. For the prototype, we neglected to differentiate among the different subject types in display, which should be rectified in a future version of the prototype.

Additionally, results pages (Figures 4 and 6) provide both a "Results by Data Contributor" facet and a "Results by Topic" facet, containing the High Level Browse categories. These facets allow end-users to view the records using multiple (duple) classifications, increasing the possibility of finding useful materials because the end-user is not limited to a single taxonomy. This is discussed further in a later section.

Lessons Learned: Topic Model

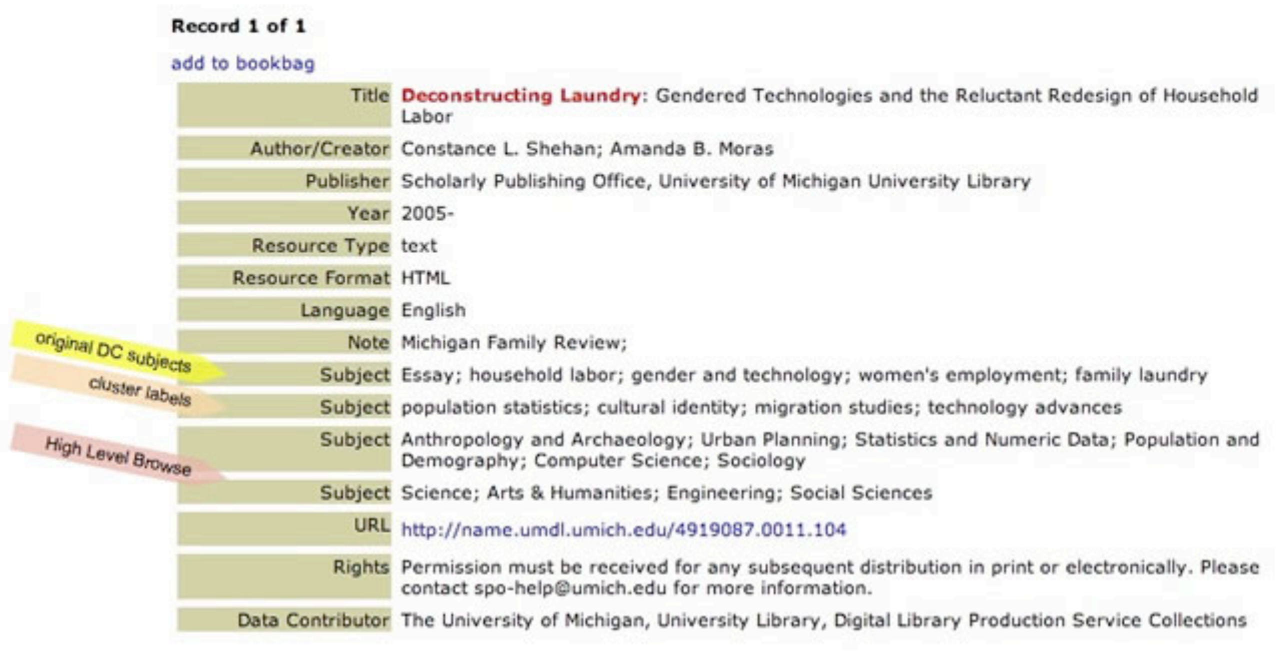
The Topic Model approach can be time-intensive. In particular, assigning labels and categories to metadata records for the DLF Portal took around 48 hours for the 62 repositories of over 2.6 million records. However, the biggest expenditure was performing the initial Topic Model clustering run on the 668 repositories – 6 GB of memory and 12 days of computing time on a 3GHz processor were necessary.

We expect that the clustering run would only need to be performed every couple of years to pick up new words and phrases that have entered the vocabulary. Since we clustered millions of records from hundreds of diverse repositories, it is likely that new records added to the prototype portal would be appropriately labeled by the existing topics. Our proposed approach is to "freeze" these learned topics for a several-year period, and use these topics to classify new records. After this several-year period, if we believe that the existing set of topics does not adequately describe new records, we can re-cluster and "learn" an entirely new set of topics.

The Topic Model process was specifically intended to classify very large sets of records, and often records with quite minimal metadata, with little manual intervention. In our concurrent paper [14], we discuss in depth the algorithm and different experiments undertaken after this project to improve the accuracy of the creation of the clusters. In two experiments, the initial vocabulary is combed over to remove "useless" words, one performed manually and the other via an automated process. Initial results showed that an automated process performed well enough that it could replace the time-intensive manual process.

Lessons Learned: Accuracy of Labeling

Some qualitative analysis was undertaken on the accuracy of the cluster labeling process – an in-depth analysis still needs to be undertaken, although a few quantitative measures are described in the paper noted above. As examples of both accuracy and inaccuracy of labeling, Figures 8 and 9 describe these, respectively.

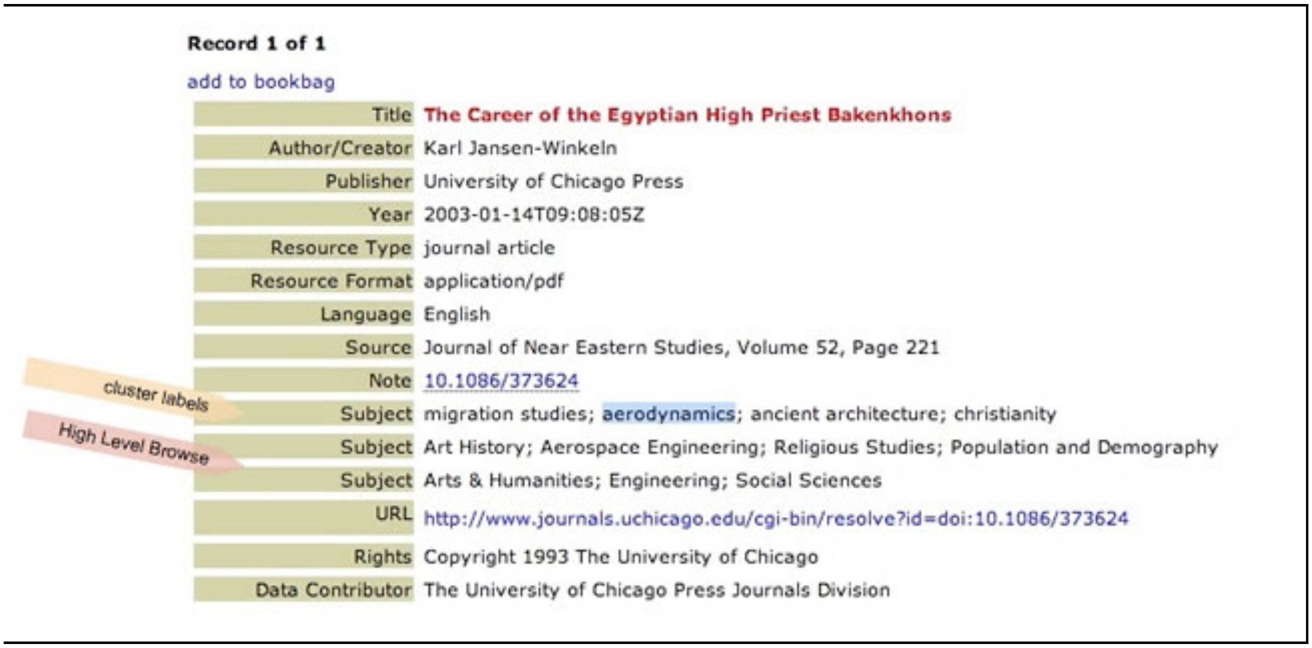


Record 1 of 1

add to bookbag

Title	Deconstructing Laundry: Gendered Technologies and the Reluctant Redesign of Household Labor
Author/Creator	Constance L. Shehan; Amanda B. Moras
Publisher	Scholarly Publishing Office, University of Michigan University Library
Year	2005-
Resource Type	text
Resource Format	HTML
Language	English
Note	Michigan Family Review;
Subject	Essay; household labor; gender and technology; women's employment; family laundry
Subject	population statistics; cultural identity; migration studies; technology advances
Subject	Anthropology and Archaeology; Urban Planning; Statistics and Numeric Data; Population and Demography; Computer Science; Sociology
Subject	Science; Arts & Humanities; Engineering; Social Sciences
URL	http://name.umdl.umich.edu/4919087.0011.104
Rights	Permission must be received for any subsequent distribution in print or electronically. Please contact spo-help@umich.edu for more information.
Data Contributor	The University of Michigan, University Library, Digital Library Production Service Collections

Figure 8. Accuracy of assigned, ranked topic clusters to a record.



Record 1 of 1

add to bookbag

Title	The Career of the Egyptian High Priest Bakenkhons
Author/Creator	Karl Jansen-Winkeln
Publisher	University of Chicago Press
Year	2003-01-14T09:08:05Z
Resource Type	journal article
Resource Format	application/pdf
Language	English
Source	Journal of Near Eastern Studies, Volume 52, Page 221
Note	10.1086/373624
Subject	migration studies; aerodynamics; ancient architecture; christianity
Subject	Art History; Aerospace Engineering; Religious Studies; Population and Demography
Subject	Arts & Humanities; Engineering; Social Sciences
URL	http://www.journals.uchicago.edu/cgi-bin/resolve?id=doi:10.1086/373624
Rights	Copyright 1993 The University of Chicago
Data Contributor	The University of Chicago Press Journals Division

Figure 9. Inaccuracy of assigned, ranked topic clusters to a record.

In Figure 8, the second subject line contains the four ranked cluster labels. Each of these labels describes an aspect of the digital resource well. In Figure 9, the same line of cluster labels is decidedly inaccurate – in particular, we point out the inappropriate "aerodynamics" cluster label.

It became clear as we viewed the results of the clustering that records with a humanities bent fared worse than those describing science resources. Humanities records typically contain very little metadata – author, title, some description, but very often no subjects or other fields. This provides little for the topic model to work with, and as a result often no labels or mis-leading labels are assigned (Figure 9). It is possible that humanities records were more difficult to process because they have been grouped with the science-based records. Certainly, there were more science clusters created than humanities clusters [15], and this could have skewed rankings. A potential method for getting around this problem would be to divide the 352 clusters into science and humanities sets and classify the humanities repositories only with the humanities topics. However, often repositories do not solely contain humanities records, therefore this could be difficult to achieve, but still worth testing.

On the other hand, humanities records often contain metaphors that are lacking in science records. For instance, one humanities record contained the phrase "from the lighthouses" in its title, which has nothing to do with brick-and-mortar lighthouses *per se*. In the case of this record, one of the labels assigned was "energy efficiency," clearly an inaccurate label. Other problems noted were multi-volume works that, while accurately labeled with 1-2 cluster labels consistently, varied widely in the assigning of the remaining 2-3 labels. These records should be assigned the same cluster labels since they contain the same metadata except for volume number. We expect that for end-users these inaccuracies can confuse and potentially irritate.

Lessons Learned: Classification

The classification scheme used when mapping the cluster labels had its drawbacks as well. For instance, we were not able to adequately place the clusters that were associated with war (e.g., "world war II") into appropriate sub-level categories. This is because the High Level Browse classification, being built upon the call numbers of the UM Library collection, lacks the history of war category because the University lacks a specific history of war major. In our case, we made do with the "History" sub-level category, clearly too broad but not completely inappropriate.

Our understanding coming into the project was that we would not have the resources to develop an excellent subject hierarchy for searching and browsing. In retrospect, we might have avoided the labeling of topics and used only the High Level Browse classification. While labeling of topics inserted useful information into the metadata records (and made them more findable), the complications of maintaining consistency of specificity and accurately determining "junk" topics can easily produce a sub-standard set of labels. When this is mirrored in the records, and in the discovery process, the quality of the endeavor depreciates.

Our underlying intent for choosing a classification scheme for end-users was to enable a controlled vocabulary that could provide end-users with standardized access to our content. We feel we achieved this, even if the development of labels and mapping to the classification scheme was not an error-free process. As Käki notes, categories used in real-life situations find more relevant results for end-users, are helpful when ranking of results fails, and perhaps most heartening to us, provides a work-around for mistakes in search queries by offering an alternate avenue for discovery [20].

Lessons Learned: Community Labeling

While our "community labeling" effort was not ideal, it was remarkably efficient and produced decent results. In general, colleagues chose "subject sets" of clusters to label, which helped them become familiar with both the process and the scope of the clusters. The process of creating labels and mapping them to sub-level categories was easier when the following was taken into account:

- The original community labeling page contained no sample records associated with clusters, so labelers resorted to performing keyword searches to get a better sense of the clusters in context.
- Either viewing all the clusters on the same subject before beginning and/or returning to tweak labeling after a first run-through resulted in better labels.
- Knowing the entire High Level Browse classification was necessary before mapping, otherwise it was easy to miss a label that could be classified under more than one subject category.
- Both Google and Wikipedia were useful when faced with a cluster word that was unfamiliar, even though it is understood these are not definitive sources.

Involving subject specialists could alleviate these problems, but could also compound them – subject domains can be difficult to discretely define and effectively training domain experts to use the same approach across the board could prove impracticable. However, with enough iteration and the involvement of a critical mass of experts we could develop an expert community, and potentially an open-source archive of cluster labels for use by anyone in the library community (or beyond).

Lessons Learned: Interface

The clustering methods we used enabled us to improve the findability of appropriate materials by providing additional subject categories for use in searching and browsing. These new terms were then used as selections within the search interface, on a new browse page, and as a way to filter the search results. We quickly realized that the real power of including new subject terms was on the search results page. To reiterate the benefits, we provided:

- narrowing and/or expanding the results using the "Results by Topic" facet,
- the freedom and context of viewing records in individual categories,
- the discovery of different results by choosing a different category, and

- clarification of vague or broad search queries.

Unfortunately, due to time constraints and technical limitations, we were unable to implement the interface exactly as we wanted. The browse page includes both High Level Browse top-level categories and sub-level categories but the results page only includes the top-level categories. Because each top-level and sub-level category is populated dynamically based on the search query, and with over 2.6 million records to search, this proved too computationally intensive and would have required an even lengthier delay in retrieving the results. This is not an issue for the browse page because the record counts are static between index updates, and can be generated and cached.

We are also concerned that allowing end-users to choose top-level and sub-level categories from the advanced search interface may inadvertently narrow the search query too radically. Since there is no record count indication in the search interface for categories, an end-user could easily enter a query that resulted in too few records or too few suitable records. This could necessitate multiple searches, which more often than not leads to frustration.

Hearst claims that "HFC (hierarchical faceted categories) enabled interfaces are overwhelmingly preferred over the standard keyword-and-results listing interfaces used in Web search engines." [21] In the future, we hope to include facets in addition to topic and contributor, such as resource type, author, and language. We also hope to create a more user-friendly display for the "Results by Topic" facet on the results page and the "Browse by Topic" facet on the browse page that will allow end-users to expand and contract top-level categories to reveal and/or hide the related sub-level categories.

Lessons Learned: End-User Testing

Once the prototype was available in production, we asked our grant colleagues to review it and provide comments. Unfortunately, this was the closest we came to end-user testing. Nevertheless, we received some useful information.

- Reviewers were not certain if selecting a different topic within the search results would retrieve all records in that topic or just the records in that topic limited by the search query. (It retrieves just the records limited by the search query.)
- Reviewers who had performed a search limited by a sub-level category were confused because the sub-level categories were not listed on the results page. (As mentioned, this was not implemented due to time constraints.)
- Because of technical limitations, the left column on the search results page was much slower to load than the rest of the page, frustrating several of our reviewers.

This latter point has also been reiterated in unsolicited comments about the OAIster service. Speed is a concern not to be taken lightly in the Web environment, and is always a priority. In this case, time restraints and software issues hindered us from implementing a faster interface.

Future testing should include quantitative analysis of accuracy in assigning labels to records, search log analysis, and usability studies with end-users. If we hope to use elements of the prototype design in OAIster, this further testing is required before we can be comfortable implementing it in a non-prototype system.

Conclusion

The importance of this research is in its relevance to the fields of metadata enrichment and automated classification techniques, and in particular, the melding of these two worlds. Our efforts specifically approached the difficulties of classifying metadata in a heterogeneous environment. Prior studies have not showcased how discovery can be enhanced using classification techniques – our prototype displays the benefits and limitations of the technology we used. We feel strongly that the lessons learned during this research can only advance these fields: making algorithms tighter, interfaces more user-friendly, and retrieval more accurate.

Acknowledgements

We would like to thank the Andrew W. Mellon Foundation, the Institute for Museum and Library Services (grant no. LG-02-04-0006-04), the Digital Library Federation, and the entire Digital Library Production Service and Scholarly Publishing Office departments at the UM University Library for their support. In particular, we would like to thank Phil Farber, Perry Willett, Christina Powell and Kevin Hawkins for assisting us in our design and analyses.

References

[1] Griffiths, J., Brophy, P. (Spring 2005) "Student Searching Behavior and the Web: Use of Academic Resources and Google." *Library Trends*, v. 53, no. 4, pp. 539-554. [http://findarticles.com/p/articles/mi_m1387/is_4_53/ai_n14732768].

[2] Morville, P. Findability website. Accessed April 9, 2007. [http://www.findability.org/archives/cat_findability.php].

[3] OAIster website. Accessed April 9, 2007. [http://www.oaister.org/].

[4] Shreeves, S.L., Riley, J., Milewicz, L. (August 7, 2006) "Moving Towards Shareable Metadata." *First Monday*, v. 11, no. 8. [http://www.firstmonday.org/issues/issue11_8/shreeves/index.html].

[5] Hagedorn, K. (October 3, 2003) "OAIster: A 'No Dead Ends' Digital Object Service." *Library and Information Technology Association (LITA) National Forum 2003*. [http://www.oaister.org/pres/LITA03_Hagedorn.ppt].

[6] Foulonneau, M., Cole, T.W. (2005) "Strategies for Reprocessing Aggregated Metadata." In *Research and Advanced Technology for Digital Libraries, Proceedings of the 9th European Conference, ECDL 2005, Lecture Notes in Computer Science Series*, pp. 290-301. [http://dx.doi.org/10.1007/11551362_26].

[7] Hillmann, D., Dushay, N., Phipps, J. (2004) "Improving Metadata Quality: Augmentation and Recombination." *International Conference on Dublin Core and Metadata Applications 2004*. [http://purl.org/metadataresearch/dcconf2004/papers/Paper_21.pdf].

[8] DLF Portal website. Accessed April 9, 2007. [http://quod.lib.umich.edu/i/imls/].

[9] Seaman, D. (June 20-21, 2005) "The Distributed Library: OAI for Digital Library Aggregation." *OAI Scholars Advisory Panel Meeting*. [http://www.diglib.org/architectures/oai/imls2004/OAISAP05.pdf].

[10] "UCI Researchers 'Text Mine' the New York Times, Demonstrating Evolution of Potent New Technology." (July 26, 2006) *UC Irvine Donald Bren School of Information and Computer Sciences Press Release*. [http://www.ics.uci.edu/community/news/press/view_press?id=51].

[11] Krowne, A., Halbert, M. (June 2005) "An Initial Evaluation of Automated Organization for Digital Library Browsing." In *JCDL'05, Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 246-255. [http://doi.acm.org/10.1145/1065385.1065442].

[12] Paynter, G.W. (June 2005) "Developing Practical Automatic Metadata Assignment And Evaluation Tools For Internet Resources." In *JCDL'05, Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 291-300. [http://doi.acm.org/10.1145/1065385.1065454].

[13] Zhang, B. (2006) "Intelligent Fusion of Evidence from Multiple Sources for Text Classification." Dissertation, Virginia Polytechnic Institute and State University. [http://scholar.lib.vt.edu/theses/available/etd-07032006-152103/].

[14] Newman, D., Hagedorn, K., Chemudugunta, C., Smyth, P. (2007) "Subject Metadata Enrichment using Statistical Topic Models." In *JCDL '07, Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 366-375. [http://doi.acm.org/10.1145/1255175.1255248].

[15] Topic Labeler website. Accessed April 10, 2007. [http:// yarra.ics.uci.edu/umich2/gettopic.php].

[16] UM Library Electronic Journals & Newspapers List website. Accessed April 10, 2007. [http://www.lib.umich.edu/ejournals/].

[17] UM Library Newly Cataloged Items website. Accessed April 10, 2007. [http://www.lib.umich.edu/newbooks/].

[18] Search Tools website. Accessed April 10, 2007. [http://searchtools.lib.umich.edu/].

[19] Digital Library eXtension Service Bibliographic Class Documentation website. Accessed May 7, 2007. [http://dlxs.org/docs/13/class/bib/index.html].

[20] Käki, M. (April 2005) "Findex: Search Result Categories Help Users When Document Ranking Fails." In *CHI 2005*, pp. 131-140. [http://doi.acm.org/10.1145/1054972.1054991].

[21] Hearst, M.A. (April 2006) "Clustering Versus Faceted Categories for Information Exploration." In *Communications of the ACM*, v. 49, no. 4, pp. 59-61. [http://doi.acm.org/10.1145/1121949.1121983].

Top | **Contents**
Search | **Author Index** | **Title Index** | **Back Issues**
Commentary | **Next Article**
Home | **E-mail the Editor**

doi:10.1045/july2007-hagedorn