

# Enhancing Search and Browse for Scholarly Discovery

## Automated Clustering of OAI Metadata

Kat Hagedorn | University of Michigan  
Suzanne Chapman | 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

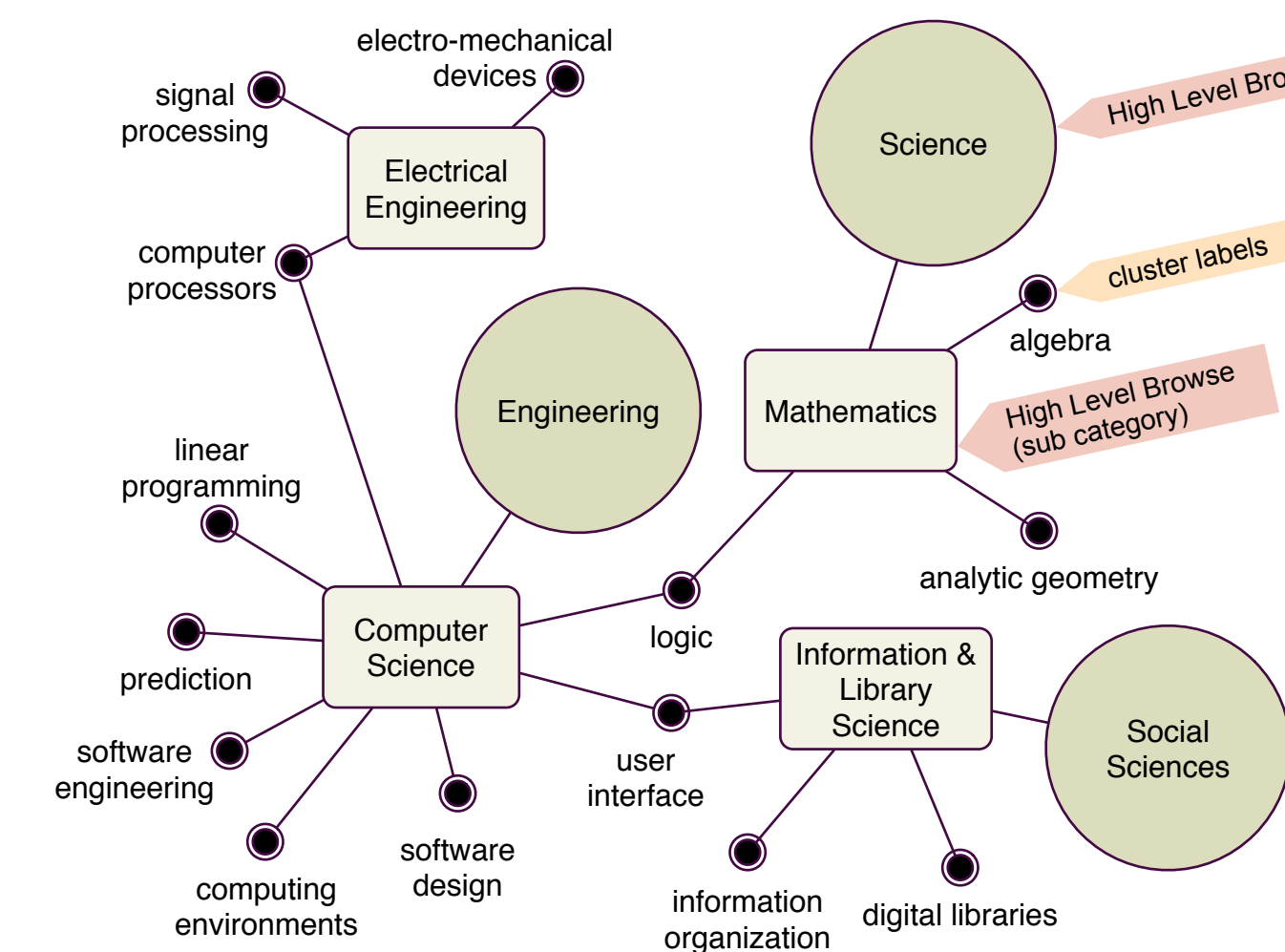
(logic) {Mathematics; Computer Science}  
[0] logic reasoning semantic logical modal

logic reasoning semantic logical modal  
propos classic interpr

use junk  
save

logic reasoning semantic logical modal  
logic\_programming logic\_program belief  
propositional inference default order set  
classical notion framework horn predicate  
interpretation representation

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

```
processraw.sh celebration
celebration
running rawstream1 on:
./celebration_0_raw.xml
done
running stream1stream2_done
running stream2stream3_done
running stream3stream4_done
running stream4stream5_done
running resample_seed = 777
not = 283
W = 94226
D = 500
iter = 40
alpha = 0.100000
beta = 0.010000
iter 0
iter 1
iter 30
iter 31
iter 32
iter 33
iter 34
iter 35
iter 36
iter 37
iter 38
iter 39
iter 40
done
running topicindex_done
total 250
-rw-rw- 1 khage dlp 40238 Apr 12 14:30 stream3.txt
-rw-rw- 1 khage dlp 40291 Apr 12 14:30 stream2.txt
-rw-rw- 1 khage dlp 52589 Apr 12 14:30 stream1.txt
-rw-rw- 1 khage dlp 155 Apr 12 14:30 pairs.txt
-rw-rw- 1 khage dlp 10389 Apr 12 14:31 rd_name.txt
-rw-rw- 1 khage dlp 28445 Apr 12 14:30 docword.txt
-rw-rw- 1 khage dlp 28763 Apr 12 14:31 resample.txt
-rw-rw- 1 khage dlp 5569 Apr 12 14:31 rd_11_12_13_14.txt
```

Starting OAI transform program...

Repository identifier: celebration  
organization: A Celebration of Women Writers  
loading cluster data  
new XML file: /1/prep/oai/celebration\_0\_raw.xml  
Data conditioning Phase 0: Start = 14:34:21  
processing files in /1/prep/harvester/celebration/celebration1-1000  
data conditioning phase 1: Start = 14:34:22  
transform: /1/prep/oai/celebration\_0\_raw.xml  
-> /1/prep/oai/celebration\_lab.xml ...  
done with translation for archive: celebration

Repository Report: celebration  
records with URLs = 300  
records without URLs = 0  
repository records = 300  
success rate = 100%

data conditioning msgs? = no  
deleted records (del) = 0  
normalization errors = 0  
raw parse failures = 0

Record including new cluster labels and classification categories

(A ID="oai:deepblue.lib.umich.edu:2027.42/1561" DT="2006-02-01T06:38:00Z"><B>Trucks involved in fatal accidents factbook 2002</B>Mattenes, A.</A><D>Blower, D.</D><D>Woodroffe, J.</D><D>University of Michigan, Ann Arbor, Transportation Research Institute, Center for National Truck and Bus Statistics</D><D>University of Michigan, Ann Arbor, Transportation Research Institute, Truck and Bus Safety Analysis Division</D><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></E><G><A>http://www.umtri.umich.edu/cntrs/doc/FACTBOOK2002.pdf</A></G><A><A>This document presents aggregate statistics on trucks involved in traffic accidents in 2002. The statistics are derived from the Trucks Involved in Fatal Accidents (TIFA) file, compiled by the University of Michigan Transportation Research Institute. The TIFA database provides coverage of all medium and heavy trucks recorded in the Fatality Analysis Reporting System (FARS) file. TIFA combines vehicle, accident, and occupant records from FARS with information about the physical configuration and operating authority of the truck from the TIFA survey.</A></A><A>Federal Motor Carrier Safety Administration, Washington, D.C.</A><A>Accession Number: 48532 A40</A><A>Report Number: UMTRI 2004-34</A><A>Contract Number: DTM75-02-R-00090</A></G><I2><S><D>Buses; Trucks; Articulated Trucks/ Combination Trucks; Fatality Patterns; Accident Statistics/ Accident Rates; Fatal Accident Files; Data/ Statistics; Transportation; Engineering</S></I2></G><I2><S><D>automobile transportation; death</S></I2></G><I2><S><D>Transportation; Public Health</S></I2></G><I2><S><D>Engineering; Social Sciences; Business & Economics; Health Sciences</S></I2></G><I2><S><D>http://hdl.handle.net/2027.42/1561</I2></G><I2><S><D>1943 bytes.1272802 bytes.text/plain,application/pdf</I2></G><I2><S><D>English</I2></G><I2><S><D>Deep Blue at the University of Michigan</I2></G></A>

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.

original DC subjects  
cluster label  
High Level Browse

Record in display

Record 1 of 1  
add to bookbag

Title	Trucks involved in fatal accidents factbook 2002
Author/Creator	Mattenes, A.
Contributor	Blower, D.
Contributor	Woodroffe, J.
Contributor	University of Michigan, Ann Arbor, Transportation Research Institute, Center for National Truck and Bus Statistics
Publisher	University of Michigan, Ann Arbor, Transportation Research Institute
Year	2006-01-31T21:48:19Z
Year	2004-10
Resource Format	ill., 1943 bytes, 1272802 bytes, text/plain, application/pdf
Language	English
Note	Federal Motor Carrier Safety Administration, Washington, D.C.
Note	Accession Number: 48532 A40
Note	Report Number: UMTRI 2004-34
Subject	Buses; Trucks; Articulated Trucks/ Combination Trucks; Fatality Patterns; Accident Statistics/ Accident Rates; Fatal Accident Files; Data/ Statistics; Transportation; Engineering
Subject	automobile transportation; death
Subject	Transportation; Public Health
Subject	Engineering; Social Sciences; Business & Economics; Health Sciences
URL	http://hdl.handle.net/2027.42/1561
Data Contributor	Deep Blue at the University of Michigan

original DC subjects  
cluster labels  
High Level Browse

### 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/ims/

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 (dupe) 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)

High Level Browse

- Philosophy (38383)
- Religious Studies (6936)
- Russian and East European Studies (35551)
- Theatre and Drama (29527)
- West European Studies (168210)
- Business & Economics (334235)
- Business (General) (112709)
- Economics (20071)
- Finance (69150)
- International Business

High Level Browse

- Philosophy (38383)
- Knowledge Representation (8241)
- Epistemology (14957)
- Hypothesis (16244)
- Religious Studies (6936)
- Christianity (6936)
- Russian and East European Studies (35551)
- Soviet Studies (35551)
- Theatre and Drama (29527)
- Artistic Performance
- West European Studies

### 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.