

## Creating Subject Area Groupings

### *Call number class mappings*

General subject area groupings used to narrow recommendation sets were produced by looking at call number class metadata for all circulating items. Groupings were based on Library of Congress subject groupings created by the Columbia University Digital Library Project<sup>1</sup>, adjusted for UCLA records where there were gaps, and supplemented by mappings from National Library of Medicine call number classes to the same scheme of general subject areas.

The hierarchical structure, up to 4 levels deep, was stored in a mySQL table. The structure of the table was defined as:

```
CREATE TABLE class_map (  
  row_num int unsigned NOT NULL auto_increment,  
  classalpha_start varchar(3) NOT NULL default "",  
  classnum_start float(7,3) unsigned NOT NULL default '0.000',  
  classalpha_end varchar(3) NOT NULL default "",  
  classnum_end float(7,3) unsigned NOT NULL default '0.000',  
  category_1 varchar(64) default NULL,  
  category_2 varchar(64) default NULL,  
  category_3 varchar(128) default NULL,  
  category_4 varchar(64) default NULL,  
  PRIMARY KEY (row_num),  
  UNIQUE KEY row_num (row_num)  
)
```

A tab-delimited text file with the data exported from this mySQL database is available for download<sup>2</sup>.

### *Processing*

Processing steps were as follows:

1. Dump the system ID, call number, and call number class from the Lucene index to a text file.
2. Run a perl script that:
  - Reads each line of the text file.
  - Attempts to look up the categories associated with the call number class in the class\_map table.
    - If successful, writes the system ID and categories to a text file.
    - If unsuccessful, skips the record and logs the error.

---

<sup>1</sup> Hierarchical Interface to LC Classification, Arranged by Class Number Range.  
<http://www.columbia.edu/cu/libraries/inside/projects/metadata/hilcc/files/class.html>

<sup>2</sup> Tab-delimited text file containing exported data:  
[http://cdlib.org/inside/projects/melvyl\\_recommender/report\\_docs/category\\_mappings.txt](http://cdlib.org/inside/projects/melvyl_recommender/report_docs/category_mappings.txt)

3. Update the summary table for circulating items in the database (used for generating recommendations) using the new text file.

### ***Final Groupings***

Categories were assigned to the 999,207 records in the database of circulating items as follows:

|                                     |         |
|-------------------------------------|---------|
| No category assigned                | 77,798  |
| Arts, Architecture and Applied Arts | 49,518  |
| Business and Economics              | 66,440  |
| Engineering and Applied Sciences    | 37,002  |
| General                             | 8,196   |
| Health Sciences                     | 50,367  |
| History and Archaeology             | 158,984 |
| Journalism and Communications       | 5,331   |
| Languages and Literatures           | 194,933 |
| Law, Politics and Government        | 53,487  |
| Music, Dance, Drama and Film        | 54,874  |
| Philosophy and Religion             | 64,322  |
| Sciences                            | 53,363  |
| Social Sciences                     | 124,592 |