Results of YFAD User Query Log Analysis

Presented by M. Matienzo and M. Dotson to FACC on August 22, 2012

---

As outlined in our project update on April 10, 2012, several preliminary conversations had led to the conclusion that any decisions to abolish the stemming function or adjust the search result relevancy rankings of the Yale Finding Aid Database (YFAD) would be advisable only after a thorough analysis of the user query logs. This undertaking would provide a deeper understanding of the search behaviors of YFAD users, i.e., types of search terms, variability of search terms, frequency of basic vs. advanced search terms, etc.

An initial request for YFAD user query logs from Eric James provided us with the following data set, which had already been partially processed:

93031 queries total (July 2011-January 2012)

13548 blank queries

79543 non-blank queries

25137 unique queries

A follow-up request for fully unprocessed query logs yielded a larger sample from an expanded time period:

  151826 queries total (July 2011-May 2012)

   21818 blank queries

  130008 non-blank queries

   42615 unique queries

We approached the data set from two different angles. One was to determine the type of query. We identified the following categories to create a typology for coding the search terms: Persons, Families, Corporate Bodies, Subjects (e.g., events, places, genre terms, etc.), Call #s, and DACS qualifiers (e.g., papers, records, collection, archive, etc.). We also assessed the degree of variance in values entered that mapped to like search term(s).

I coded a subset of over 45,000 queries according to the above categories. Through the use of the “Cluster and edit” function in Google Refine, I was able to determine the degree of variance among seemingly like search terms before data normalization. As might be expected, the variance (i.e., the number of clusters) for known values, such as collection titles or call #s, was lower than for other types of searches, such as personal names or topical subjects.

Next I exported “dc.” (total=12901), “dsm.” (total=4193), and “fgs.” (total = 353) queries to separate spreadsheets in order to pare down the data set to basic searches only (112,201 total queries). I used the customized “Word facet” function in Google Refine to obtain individual search terms listed by frequency in descending order. I then filtered these basic searches by each of the different repositories (MSSA, Beinecke, Arts, etc.), in order to look at YFAD searches within particular collections. I charted the 20 most popular searches for each of the repositories (see MostFrequentSearchTerms Excel spreadsheet).

Some interesting search behaviors or “user stories” emerged from this data set. For example, regarding the stemming functionality of YFAD, a search for “blood root” does not return any hits for the Bloodroot Collective Records. Nor do searches for “o’kiefe” or “o’keffe” return any hits (much less a suggested search for Georgia O’Keeffe). This analysis also raised questions about the content of YFAD and its linkage to other library resources, especially considering that the data set reflects nearly 1400 queries for “voynich,” even though YFAD does not contain a record for the Voynich Manuscript itself.

Further decisions about YFAD stemming and relevancy ranking have been postponed in anticipation of the hiring of a new EAD coordinator at Beinecke and the beginning of next-generation YFAD development in January 2013.