Exploring the role of scale: comparative analysis of digital library user searching

Oksana L. Zavalina
College of Information, University of North Texas, 1155 Union Circle #311068, Denton, Texas, 76203.
Email: Oksana.Zavalina@unt.edu

ABSTRACT
This poster reports preliminary results of a comparative study of user searching in two large-scale digital libraries with history focus in the United States: the federal-level and the state-level digital library. Similarities were observed in search query lengths and average numbers of search categories per search query. At the same time, the study reveals significant differences in the level of use of advanced search options and in search query frequencies, as well as in distribution of most search categories in user queries. The empirical data obtained in this exploratory study will inform further research and will be useful for professionals making decisions on resource description and user interfaces for digital libraries.

Keywords
Search queries, information behavior, federal-level digital libraries, state-level digital libraries, search log analysis.

INTRODUCTION
Cultural heritage materials of historical and educational value, including resources about local and national history, have been a priority in mass digitization, especially in its early stages in 1990s and 2000s. Many of resulting large-scale digital libraries – portals that aggregate millions of digitized or born digital items organized into hundreds of digital collections – function at the state level, some at the regional level, others at the federal level. For example, in the United States, IMLS Digital Collections and Content (IMLSDCC) portal and its spinoff Opening History are well-known federal-scale digital libraries. Opening History, which functioned as a separate entity with a United States history focus for almost 4 years until it was absorbed by its parent broad-scope digital library IMLSDCC in August 2012, was the largest aggregation of digitized content in the nation with over 1500 digital collections and more than a million items. Similarly, the Portal to Texas History is one of the largest state-wide digital libraries and has a history focus. This digital library consists of over 300 local history collections and holds over 600,000 of items that were digitized and contributed by various cultural heritage institutions throughout the state of Texas.

In order to effectively deliver content of large-scale digital libraries to their respective audiences and improve user interaction, the development of large-scale digital libraries should be informed not only by general user tasks such as finding, identifying, selecting, and obtaining information (IFLA, 2008) but also by the specific needs of their user communities. Further, digital library services have to adjust to the evolving information seeking behavior patterns of user groups in the online environment (Horava, 2010; Lavoie et al., 2007; Verheul et al., 2010). To satisfy their information needs, people engage in searching – one of the two major types of interactions between users and information systems such as library catalogs, databases, search engines, or digital libraries (Wilson, 2000). Studies conducted in 1970s-1990s analyzed search queries of the users of library online catalogs (e.g. Bates, 1989; Larson, 1991; Mathews et al., 1983). Such patterns of user searching as the use of Boolean operators, advanced search options and controlled vocabulary were analyzed in many of these research projects (e.g., Borgman, 1996; Bryce, 1991; Fidel, 1992; Hildreth, 1997; Muddamalle, 1998, etc.). Researchers found that information seeking behavior (i.e., selection of search strategy and tactics) and the outcomes of the search depend to a large extent on a searcher’s subject domain. Studies have shown that STEM scholars usually search for information on specific questions or problems that they face when conducting an experiment, reporting results, or verifying the accuracy of information in hand (Ellis, 1993; Palmer, 2005). More specifically, water quality researchers frequently use topical, geographical, and format or genre search terms, and occasionally – chemical formulas, dates, names, and URLs (Nowick & Mering, 2003); medical researchers’ prevailing search query types include laboratory/test results, disease/syndrome, body part/organ/organ component, pharmacological substance, or diagnostic procedure (Natarajan et al., 2010). Humanities and social science researchers, including historians, were found to prefer searching to browsing and use several search types – known author-title search, conceptual/discipline term search, etc.; they most often included in their search queries personal and geographic names, chronological and discipline terms, and expressed the need for search limit by date (Bates, 1996; Buchanan et al., 2005; Harum, 2008; Wu & Chen, 2007, etc.).
Servers of information systems record transaction log data. Transaction log analysis is one of the research methods actively used for unobtrusive observation of user searching, expressed through search queries, in various information retrieval systems (Peters, 1993). Transaction log data is often analyzed quantitatively, with the focus on such measures of search complexity as search query length (i.e., the number of terms included in a search query), Boolean usage rates, and the number of search limits (Jansen, Spink, and Pedersen, 2004; Moulaison, 2008, etc.). Transaction log data is also used in qualitative analysis, for instance, to categorize search queries by content (e.g., Bates, 1996; Beitzel et al., 2007; Jansen et al., 2007, Koshman et al., 2006; Spink et al., 2004; Pu, Chuang, & Yang, 2002, etc.), but few attempts have been made to compare the content of search queries in different information systems (Wolfram, 2008; Yi et al., 2006).

The studies discussed above analyzed user searching in information systems of more traditional types: bibliographic databases, library catalogs, and web search engines. However, user searching in information systems of the new type — openly-accessible large-scale digital libraries comprising of digitized and born-digital high-quality content for education and research — has not been actively researched. Despite the popularity of transaction log analysis as a research method for studying information searcher’s behavior, its potential has not been used to its full capacity to benefit large-scale digital libraries’ development. Several studies analyzed transaction logs of large-scale digital libraries such as the National Science Digital Library, American Memory, Opening History, IMLS Digital Collections and Content Collection Registry, and The European Library (Khoo et al., 2008; Pan, 2003; Verberne et al., 2010, etc.). However, virtually no research, with exception of Zavalina and Vassilieva (2013; 2014), compared user search queries in two or more digital libraries, and no study attempted a comparative analysis of user searching in large-scale digital libraries of different scale.

**METHODS**

The exploratory study presented here used a mixed-method comparative analysis of transaction log data on user searching in two representative large-scale cultural heritage digital libraries of state scale and federal scale – Portal to Texas History (PTXH) and Opening History (OH) – and sought answers to the following research question: What are the differences and similarities of user searching in cultural heritage digital libraries of federal scale and state scale? We analyzed and compared transaction log datasets collected using the same application – Google Analytics. To achieve comparable sample sizes, all of the user search queries in OH during the one-year period and a random sample of the PTXH user search queries were included in the analysis. Identical search queries were grouped together for each of the two digital libraries, which resulted in 5,343 unique search queries: 2,551 in OH and 2,792 in PTXH.

The study compared the following quantitative and qualitative metrics: levels of use of basic and advanced search options, search query frequencies (measured as the number of times the identical search query occurs in the sample), search query lengths (measured as the number of words per search query), frequency of occurrence of search categories in user search queries, and co-occurrences of search categories such as concept, object, event, place, class of persons, ethnic group, work, person, and corporate body.

To confirm that differences and similarities observed did not occur by chance, the statistical significance of the findings was assessed. A t-test with probability of error level p<.01 was used to assess statistical significance of the results for search query length, search query frequency, and number of search categories per search query; a Chi-square test with probability of error level p<.01 was used to assess statistical significance of the results for the frequencies of selection of advanced search approaches and for the inclusion of particular search categories in the queries.

**FINDINGS**

In the OH, search queries had an average of 2.32 and a median of 2 words per query. The average length of PTXH search queries was somewhat higher (2.43 words per query) while the median was the same. The median search query frequency was identical (1) in the two digital libraries. However, the mean was substantially higher in OH (1.91) than in the PTXH (1.17).

Most of the search queries in both digital libraries were basic keyword searches. The use of one or more advanced search options was observed in 14.51% of OH search queries and almost twice as often (24.96%) in PTXH. As part of advanced search, both digital libraries provided options to limit search results by the type of information objects (for example, to retrieve only photographs). In addition, OH provided an option to limit search results to a specific digital collection or a group of collections and PTXH provided an option to limit search results by date. A total of 12.58% of search queries in OH and only 4.33% in the PTXH included one or more of these search limits. Fielded search option – an advanced search where the user is allowed to search either by author or by title and subject words – was available for search in the OH but not in the PTXH. Fielded search was observed in 0.61% of OH search queries. Use of another advanced search feature – phrase search with quotes – was observed infrequently (1.32% of search queries) in OH but in every 5th user search query in the PTXH (20.63%).

As shown in Figure 1, the top two user search categories in the federal-level OH were place (36%) and object (32%). The concept category was the fourth most often occurring search category (17%) while the event searching was observed much less (11%). The person, corporate body, and family search categories were observed in 26%, 15%, and 8% of queries respectively, with person being the third top search category. The work category was observed in 9% of queries, while the class of persons and ethnic group search categories occurred in 8% and 5% of queries respectively. The top three user search categories in the
state-level PTXH were the same as in the federal-level OH but appeared in different order: *person* (44% of queries), *place* (32%), and *object* (19%). The *event* was the third most frequently occurring search category in PTXH with 17% of search queries, considerably more than in OH. The *concept* category occurred somewhat less often than in OH (14%). The *corporate body* search category was found almost as often as in OH: 14% of queries. The *family* search category occurred in 14% of PTXH search queries, which represents a substantially higher level than in OH. The three remaining search categories – *work*, *class of persons*, and *ethnic group* – were found in fewer than 10% of search queries each (9%, 5%, and 3%). The *work* search category was observed in the same proportion if queries as in OH, but *class of persons* and *ethnic group* search categories occurred considerably less frequently than in OH.

Approximately half of user search queries (49.6% in OH and 50.7% in PTXH) included a single search category. The number of categories in the remaining, multi-category search queries (50.4% in OH and 49.3% in PTXH) ranged from 2 to 7 for OH and from 2 to 5 for PTXH. Search queries of PTXH users were found on average to contain more search categories (1.66) than search queries of OH users (1.61). Almost a third of queries (36% in OH and 32% in PTXH) combined two search categories. The proportion of queries with three and more search categories was quite low in both digital libraries. Eleven percent of user search queries in OH and 15% of queries in PTXH included three search categories. Four-category search queries constituted 2% in OH and 3% in PTXH. A small fraction of search queries (0.45% in OH and 0.22% in PTXH) included five search categories. Finally, six category and seven category queries comprised only 0.04% of the OH queries each, while no multi-category search queries with more than five categories were observed in PTXH. The top four pairs of co-occurring search categories in OH, found in 5% or more of the search queries, were *place* and *object* (10.04%), *person* and *family* (7.49%), *concept* and *object* (5.72%), and *place* and *corporate body* (5.10%). In PTXH, where only three pairs of search categories occurred in more than 5% of search queries, two of these pairs were the same as in OH but ranked in different order: *person* and *family* (12.75%), and *place* and *object* (6.91%). The third pair of co-occurring search categories in PTXH was *event* and *corporate body* (5.01%).

**DISCUSSION**

The findings of this study reveal several notable differences and some similarities in user searching in digital libraries of federal scale and state scale with the focus on historical materials. A statistically significant difference was observed in the rate of use of advanced search option: state-level digital library users initiated advanced search substantially more often than federal-level digital library users (Chi square=354.74, df=1, p<.01). While the use of phrase search was substantially higher in the state-level digital library, the rate of use of such search limits was much higher in the federal-level digital library. Another statistically significant difference was observed in search query frequencies, with average frequency of search queries substantially higher in federal-level digital library (t=-15.63, df=5341, p<.01). However, search query lengths were similar, with somewhat higher average query length observed in the state-level digital library (t=1.62, df=5341, p<.01). This study’s findings for digital library query lengths are consistent with results of the only previously published study (Jones et al., 2000) that measured search query lengths in a comparable, large-scale digital library environment.

The numbers of search categories per search query were also similar in the two digital libraries, and no statistically significant differences were observed (t=2.17, df=5154, p<.01). However, when it comes to the specific search categories, similar rates of occurrence – with statistically insignificant difference between the state-level and federal-level digital libraries – were observed for only two search categories: *corporate body* (Chi square=0.15, df=1, p<.01) and *work* (Chi square=0.15, df=1, p<.01). The rate of occurrence for remaining search categories differed substantially, with the most statistically significant difference observed for *person* and *family* categories that occur more often in the state-level digital library (Chi square=165.15, df=1, p<.01 and Chi square=42.83, df=1, p<.01 respectively), and *object* search category that occurred more frequently in the federal-level digital library (Chi square=111.85, df=1, p<.01).

**CONCLUSION**

This exploratory comparative study of user searching in digital libraries with history focus at state scale and federal scale provides empirical data to support digital library developers’ decisions regarding information organization in large-scale digital libraries in the cultural heritage domain. For example, the users of both federal-level and state-level cultural heritage large-scale digital libraries would benefit from prioritizing recording *places*, *objects*, *persons*, *corporate bodies* and *concepts* into metadata records in development of these digital libraries. In addition, results of this study suggest that in metadata creation for state-level cultural heritage digital libraries, special attention should be given to documenting *events* and *families*. As suggested by
the high proportion of place searching observed in this study, user experience can be improved if large-scale digital libraries with local and national history focus – both state-level and federal-level – supply an option to limit search results by location.

REFERENCES


