Longitudinal Comparative Analysis of Item-Level and Collection-Level User Searching in a Digital Library

Oksana L. Zavalina
Department of Library and Information Sciences
College of Information
University of North Texas
1155 Union Circle #311068
Denton, Texas 76203-5017
1-940-565-3776
Oksana.Zavalina@unt.edu

Elena Vassilieva
Department of Library and Information Sciences
College of Information
University of North Texas
1155 Union Circle #311068
Denton, Texas 76203-5017
1-940-565-3776
Elena.Vassilieva@unt.edu

ABSTRACT
This poster reports the preliminary results of a comparative transaction log analysis study of user searching at collection and item levels in a large-scale cultural heritage digital library that provides a unique option for collection-level search. Substantial changes in user searching have been observed over time, including an important trend of the growing use of collection-level search feature. Results also demonstrate differences in the search query lengths, frequencies, and distribution of search categories between the collection-level and item-level searching. The higher occurrence of place and concept categories in collection-level user searches than in item-level searches has been consistent over time. The analysis also revealed the prevalence of subject searching (including concept, object, event, place, class of persons, and ethnic group search categories) both at the collection level and at the item level.

This study provides empirical data to support digital library developers’ decision making regarding information organization in large-scale aggregations of digital collections.

Keywords
Digital libraries, information search, item-level search, collection-level search, search queries, search categories, transaction log analysis.

INTRODUCTION
Since the 1990s, a large number of digital libraries have been created worldwide that bring together millions of digitized cultural heritage information objects and hundreds of collections of digital objects built according to certain collection development criteria. Many of these digital libraries aggregate curated, thematic research collections (Palmer, 2004). Some of these digital libraries (e.g., The European Library, American Memory, Opening History) are now generating metadata to describe not only individual objects (i.e., item-level metadata) but entire digital collections as integral wholes (i.e., collection-level metadata). Addition of collection-level metadata allows the users to locate entire collections relevant to their research, teaching or other tasks – usually through the basic search option that performs searches against collection-level metadata records as well as item-level metadata records in the database. Some of these large-scale digital libraries also found it important to include a separate option to conduct searches only at the collection level. Inclusion of this feature provides a unique opportunity to compare the information searching patterns at the item level and collection level which can inform the information organization approaches in the digital libraries. Currently, Opening History (http://imsdcc.grainger.uiuc.edu/history), an IMLS-funded University of Illinois-hosted aggregation of over 1,500 digital collections for history researchers, educators, and enthusiasts is the largest cultural heritage digital library in the United States. It is one of the few digital libraries that provide an option for collection-level search in addition to item-level search option.

METHODS
This study used transaction log analysis of the Opening History digital library to answer the following research question: How does collection-level information searching differ from item-level information searching in large-scale digital libraries and are there any changes over time? In particular,

- what (if any) are the differences in the distribution of search categories, such as personal and geographic names, dates, etc?
- how do query length and frequency differ?

The data was collected with the help of the Google Analytics application at two points in time: 12 weeks of search log data systematically sampled in 2008, at the very early stages of the Opening History lifecycle (hereafter referred to as the 1st dataset), and the entire population of
2010 user search queries (hereafter referred to as the 2nd dataset). Both quantitative and qualitative characteristics of search queries were assessed. Quantitative characteristics such as query length and query frequency were measured both for collection and item levels using traditional definitions and approaches (cf., Spink, Wolfram, Jansen, & Saracevic, 2001). Query length was measured as the number of words in a query, query frequency — as the number of times query used in a log.

Collection-level and item-level search queries were grouped with identical queries. This resulted in a total of 3,929 unique search queries: 501 unique collection-level search queries and 713 unique item-level search queries in the 1st dataset and 1,426 unique collection-level search queries and 1,289 unique item-level search queries in the 2nd dataset. Unique search queries were analyzed qualitatively. They were categorized into ten search categories, including seven derived from the Functional Requirements for Bibliographic Records (IFLA, 1998; 2008) (work, person, corporate body, concept, object, event, and place), one from Functional Requirements for Authority Data (IFLA, 2009) (family), and two derived from the earlier study of collection-level searching (Zavalina, 2007): class of persons, and ethnic group. Polysymmetric user search queries and most phrase queries were assigned to multiple categories. Whenever possible, search queries formulated in languages other than English were translated and categorized into appropriate search categories. Qualitative results of search query categorization were quantified and reported here as percentages of search queries in which a certain search category occurs.

The first author coded all 3,929 unique search queries in the two datasets. Additionally, a subset of 10.54% of the 1st dataset (3.25% of the overall sample) was coded independently by a volunteer coder to avoid coder bias. Another much larger subset of user search queries totaling 52.52% of the 2nd dataset (36% of the overall sample) was independently coded by the second author. Strong intercoder reliability reached 93.36% or Cohen’s Kappa of .710 for the 1st dataset and 99.40% or Cohen’s Kappa of .976 for the 2nd dataset.

FINDINGS
At the initial stages of the Opening History development, the collection-level search was used overall less often than the item-level search. However, collection-level search was used more often than item-level search in the second dataset collected two years later. In the 1st dataset (2008), collection-level search accounted for 32% of all search queries. In the 2nd dataset (2010), collection-level search accounted for 52.5% of all search queries.

Query Length and Query Frequency
In the 1st dataset collection-level search queries had an average of 1.75 and a median of 1 word per query (Table 1). Item-level search queries were longer, with an average of 1.99 and a median of 2 words per query. Two years later, the length of collection-level search queries increased to an average of 2.44 and a median of 2 words per query. The length of item-level search queries also increased to an average of 2.18 words per query but the median remained the same. Thus, collection-level search queries became longer than item-level search queries.

<table>
<thead>
<tr>
<th>Collection-level query:</th>
<th>2008 mean</th>
<th>2008 median</th>
<th>2010 mean</th>
<th>2010 median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>1.75</td>
<td>1</td>
<td>2.44</td>
<td>2</td>
</tr>
<tr>
<td>Frequency</td>
<td>1.54</td>
<td>1</td>
<td>2.06</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item-level query:</th>
<th>2008 mean</th>
<th>2008 median</th>
<th>2010 mean</th>
<th>2010 median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>1.99</td>
<td>2</td>
<td>2.18</td>
<td>2</td>
</tr>
<tr>
<td>Frequency</td>
<td>1.99</td>
<td>1</td>
<td>1.75</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Query length and frequency

Item-level search queries in the 1st dataset had higher frequencies than collection-level search queries (Table 1). On average, the item-level search query was used 1.99 times, while the collection-level search query was used only 1.54 times. The situation changed two years later. In 2010, the average frequency of the item-level search query decreased to 1.75 and became lower than the frequency of collection-level search queries, which slightly increased to 2.06. The median frequency for both item-level and collection-level search queries remained at 1.

Search Categories
Collection-level and item-level search queries exhibited very similar specificity (number of search categories per search query). On average in the 1st dataset, a collection-level search query covered 1.3992 search categories, while an item-level search query covered 1.3955 search categories. These numbers are somewhat higher for the 2nd dataset. In the 2nd dataset, an average collection-level search query covered 1.5374 search categories, while an item-level search query covered 1.5039 search categories. The median number of search categories per search query was the same for collection-level and item-level searches (1) and did not change over time.

The top four categories observed in collection-level searches in the 1st dataset were object (36% of searches), place (26%), concept (22%), and person (19%) (Table 2). Three of these categories are subject search categories (IFLA, 1998; 2008). The corporate body, event, and work search categories were observed somewhat less often: at 13%, 9%, and 8%, respectively. Ethnic group and class of persons search categories were observed in 3% of collection-level searches each. No family search instances were observed in the 1st dataset, either at collection or at the item level.

Somewhat different distribution of search categories was observed in the item-level searches in the 1st dataset (Table 2). While, similarly to collection-level queries, object was the most frequently occurring item-level search category
(28%), the same percentage of searches were observed in the person category (28%). Place and concept search categories were also often observed in item-level search queries. Besides person, three additional search categories occurred more frequently in item-level search queries than in collection-level search queries: event (10%), ethnic group (5%), and class of persons (8%). The item-level search in object, place, concept, and corporate body search categories occurred less often than the collection-level search in these categories, while the same percentage of work (8%) search categories was observed in item-level and collection-level searches.

<table>
<thead>
<tr>
<th>Category</th>
<th>Collection-level search</th>
<th>Item-level search</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2008</td>
<td>2010</td>
</tr>
<tr>
<td>class of persons</td>
<td>3%</td>
<td>8%</td>
</tr>
<tr>
<td>concept</td>
<td>22%</td>
<td>18%</td>
</tr>
<tr>
<td>corporate body</td>
<td>13%</td>
<td>13%</td>
</tr>
<tr>
<td>ethnic group</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>event</td>
<td>9%</td>
<td>11%</td>
</tr>
<tr>
<td>family</td>
<td>0%</td>
<td>0.5%</td>
</tr>
<tr>
<td>object</td>
<td>36%</td>
<td>30%</td>
</tr>
<tr>
<td>person</td>
<td>19%</td>
<td>26%</td>
</tr>
<tr>
<td>place</td>
<td>26%</td>
<td>37%</td>
</tr>
<tr>
<td>work</td>
<td>8%</td>
<td>12%</td>
</tr>
</tbody>
</table>

Table 2. Search categories

The study revealed lower average search query lengths than most other transaction log analysis studies, as summarized by Markey (2007) in the 1st earlier dataset. This might be explained by the fact that none of the previous studies analyzed user search queries in the context of large-scale digital libraries aggregating digital collections. The information retrieval systems targeted by these studies included mainly online catalogs and web search engines. The only study that measured user search query lengths in a comparable environment — New Zealand Digital Library (Jones et al., 2000) — observed end-user search queries with a length between 2.2 and 2.6 words per query, which is somewhat higher than 1.54 to 1.99 words per query observed in this study in the 1st dataset, collected in 2008. However, the search queries in the 2nd dataset, collected in 2010, exhibit the lengths consistent with Jones’ et al. observations (between 2.18 and 2.44 words).

It is worth noting that collection-level query lengths have substantially increased over time. Likewise, collection-level queries have become longer than item-level queries; the same trend is observed for query frequencies.

Results of this study demonstrate the significant increase in proportion of collection-level search queries over time. Given that there were no changes to the Opening History search interface between 2008 and 2010, this observation indicates that the users of this digital library have discovered the benefits of collection-level search and use it more and more often.

Results of this study confirm relevance of Bates (1996) taxonomy of key search query term types — names of individuals, geographical names, chronological terms, and discipline terms — in the context of the large-scale digital library by showing that both item-level and collection-level search queries in the Opening History digital library often include person, place, event, and concept search categories. This study’s observation about the prominence of geographic search agrees with Bates’ results. However, this study has also revealed additional categories of searches— object, corporate body, work, ethnic group, and class of persons — that were not observed in Bates’ earlier research.

This study revealed certain differences in the distribution of search categories between collection-level and item-level user searches. Some of these differences have decreased over time. For example, the striking almost two-fold difference in occurrence of corporate body search category between collection-level and item-level search queries in 2008 has become a much more modest 23% difference in 2010. However, the higher occurrence of place and concept categories in collection-level searches than in item-level searches has been consistent over time. Since both of these search categories belong to subject categories (IFLA, 1998; 2008), this finding indicates higher prominence of subject searching at collection level.
Due to the fact that very few large-scale digital libraries provided collection-level search option at the time of analysis, this exploratory study was limited to studying user search queries in only one large-scale digital library. More generalizable results could be produced by a comparative analysis of item-level and collection-level search queries in multiple large-scale digital libraries of different subject scope. However, authors believe that since Opening History is the largest aggregation of cultural heritage digital collections in the United States, patterns of user searching observed in it can be generalized to other similar large-scale digital libraries that share the overlapping client base with Opening History (e.g., American Memory, California Digital Library, Mountain West Digital Library, Texas Heritage Online, etc.).

This study provides empirical data to support effective information organization solutions in large-scale cultural heritage digital libraries. First, its results indicate high usage of the collection-level search option and thus point to the benefit of including this search option, which is not currently an accepted practice. Second, findings of this study with respect to user search categories provide insights into the kinds of information that should be present in collection-level and item-level metadata records to facilitate information access to digital collections. For example, in the item-level metadata creation, special attention should be given to documenting corporate and personal names and dates related to digital objects. Furthermore, in the collection-level metadata, creators need to ensure recording concepts, places, and ethnic groups related to collections of digital objects. Third, user experience can be improved if large-scale digital libraries supply – especially for collection-level information discovery – an option to limit search results by place, which is suggested by the high proportion of place searching observed in this study. Finally, this study revealed the overall prevalence of subject searching (including concept, object, event, place, class of persons, and ethnic group search categories) in the large-scale digital library, both at the collection level and at the item level. This suggests that provision of the subject-based advanced search option should be prioritized.

While accurately and unobtrusively capturing objective data on the actions of digital library users, transaction log analysis is unable to provide data on users’ motivations and reasoning behind these actions. Triangulation of transaction log analysis results along with results obtained through other methods, for example through think-aloud protocol observations of the users, would overcome this limitation and help develop a fuller picture of digital library user searching at both collection and item levels.

**ACKNOWLEDGMENTS**

This research was partially supported by the University of North Texas Office for Research and Development grant. The authors would like to acknowledge research and implementation team of the IMLS-funded Digital Collections and Content project at the University of Illinois at Urbana-Champaign for making the Opening History transaction log data available for this research as well as for the valuable feedback on the 1st dataset data analysis.

**REFERENCES**


