Understanding the Information Needs of Large-Scale Digital Library Users

Comparative Analysis of User Searching

Oksana Zavalina and Elena V. Vassilieva

This paper reports on the results of a comparative study of user search logs in two large-scale, domain-specific digital libraries functioning in the United States: the National Science Digital Library and Opening History. Results demonstrate varying levels of use of advanced search options and substantial differences in the search query lengths, search query frequencies, and distribution of search categories in queries. The empirical data on how the members of the target communities search can be used in deriving important information for domain-specific digital library developers' decisions regarding both the details of information organization and support of various search features.

A growing number of large-scale digital libraries, portals that aggregate millions of digitized or born-digital items of historical, cultural, or educational value that are organized into digital collections, have been developed in recent decades. While many of these large-scale digital libraries have been created for the general public, some serve more specific audiences of scholars and educators in different disciplines or domains, for example, history, science, technology, engineering, and mathematics (STEM), etc.

To improve user interaction with large-scale digital libraries and to make sure they successfully meet their users' information needs, the design and development of large-scale digital libraries' discovery and access systems should be informed by general user tasks such as finding, identifying, selecting, and obtaining information as well as by the needs and information-searching patterns of their specific intended user communities.[1] Various groups of users (e.g., researchers, educators, and enthusiasts) may use digital libraries differently because of their varying information needs; moreover, users' information-searching strategies may differ in the large-scale digital libraries that function in distinct domains, or subject areas. These differences may require specific policies regarding the organization and description of information objects in large-scale digital libraries.

The extensive digitization and organization of large-scale digital libraries require in-depth research of current trends in use of these emerging and rapidly developing resources. However, systematic investigation into the user searching

Dr. Oksana Zavalina (oksana.zavalina@unt.edu) is Associate Director, Interdisciplinary Information Science PhD Program and Assistant Professor, Department of Library and Information Sciences, College of Information, University of North Texas. **Dr. Elena V. Vassilieva** (elena.vassilieva@unt.edu) is an adjunct faculty member in the Department of Library and Information Sciences, College of Information, University of North Texas.

Submitted March 4, 2013; tentatively accepted May 6, 2013 pending modest revision; accepted for publication February 24, 2014.

in the context of large-scale digital libraries is in its infancy. In particular, virtually no research studies have compared user searching in domain-specific large-scale digital libraries. The study reported in this article sought to begin bridging this gap by answering the following research question: What are the differences and similarities in user searching behavior between the two large-scale digital libraries geared toward two different user groups? The results of this research could be useful for professionals applying cataloging practices and procedures for digital materials and addressing resource description and metadata for digital collections.

Literature Review

Large-Scale Digital Libraries Serving US History and STEM Education and Research

In the US, digitization of valuable information resources has been supported with federal and state funding for over fifteen years. The Institute of Museum and Library Services (IMLS) has awarded National Leadership Grants and Library Services and Technology Act grants to more than five hundred digitization projects of various scales since 1998.[2] The National Science Foundation (NSF) has funded more than three hundred STEM digital collections.[3]

To offer easy access to rich pools of information objects that are available in digital format because of efforts of hundreds of digitization projects over the years, large-scale digital libraries aggregate hundreds of separate collections and function as portals to these collections and the individual items contained in them. Large-scale digital libraries provide innovative solutions in transitioning learning and teaching to the digital platform.[4] Many of these large-scale digital libraries were created for the general public (e.g., the European Library, and IMLS Digital Collections and Content). Others serve more specific audiences of educators, students, and scholars in different disciplines or domains.

Cultural heritage materials of historical and educational value, particularly resources about local and national history, have been a priority in mass digitization initiatives, especially in the early stages in 1990s and 2000s. Therefore many large-scale digital libraries were created for users in the domain of history. In the United States, many of them function at the state level (e.g., the Portal to Texas History; http://texashistory.unt.edu), some at the regional level (e.g., Mountain West Digital Library; http://mwdl.org), and several were created at the federal level. The American Memory (AM; http://memory.loc.gov) is without doubt the most well-known large-scale digital library in the US history domain. Although comparatively small in collection size, AM aggregates the most carefully selected information resources of the highest quality. This digital library was created by the Library of Congress (LC), in cooperation with other cultural heritage institutions, in the mid-1990s with financial support from the IMLS.[5] Similarly, IMLS funded creation of the Opening History (OH) digital library (http://imlsdcc.grainger.uiuc.edu/history). This digital library was a spin-off from the IMLS Digital Collections and Content (IMLSDCC; http://imlsdcc.grainger.uiuc.edu) portal to all digital collections supported by IMLS, with the purpose of further developing the strongest content area in the IMLSDCC (US history) and providing access to it. OH's primary user group was broadly defined as history researchers, including both academic and nonacademic history scholars; teachers and students at undergraduate, graduate, and postgraduate levels; and genealogists and "citizen historians."[6] OH functioned as a separate entity from October 2008 to July 2012 and quickly became the largest aggregation of digitized content in the United States with more than 1,500 digital collections and more than a million items. In August 2012, OH was absorbed by its parent digital library, IMLSDCC.

Another important domain served by large-scale digital libraries is STEM. In the past decade, STEM digital libraries became major players in STEM education.[7] The NSDL (http://nsdl.org) was organized by the NSF in 2000 for administrators, educators, general public, learners, parents/guardians, professionals/practitioners, and researchers. As the demand for access to high-quality resources in the area of STEM education for teachers and learners grows, NSDL serves as a starting point to locate and retrieve these discipline-specific resources in a variety of formats for different learning levels.[8] NSDL aggregates educational resources that are available online from a wide variety of providers. The majority of these resources are free, based on Open Educational Resource (OER) access. The resources are organized by educational level (from pre–kindergarten to higher education, including informal education and professional development education). They are grouped by resource type (assessment materials, audiovisual, instructional material, reference material, and other). Subject categories are STEM disciplines, e.g., Education, History/Policy/Law, and others, including a General subject section.[9] The NSDL offers well-defined search, browse, and help instruments, which include keyword searching, browsing collections by subject area, combining search by broad subject area and audience, limiting search to new collections, etc.[10] The NSDL serves not only as a resource repository, but also provides useful services and tools for professional development of educators and for network collaboration among the members of NSDL audiences.[11]

To effectively deliver content of large-scale digital libraries to their respective user groups, improve users' interaction with digital libraries, and facilitate efficient information retrieval, digital library services must adjust to the evolving needs and information-seeking behavior patterns of

user groups in the online environment.[12] The following will provide a historical perspective on research into information searching and review the literature that addresses how users search for information in online environment.

User Searching: From Card Catalogs to Digital Libraries

People engage in searches to satisfy their information needs. Searching is one of the two major types of interactions between users and discovery and access systems such as library catalogs, databases, search engines, or digital libraries.[13] User searching is expressed through queries: sets of one or more symbols (e.g., words or phrases) combined with other syntax and used as commands for the information retrieval system to locate potentially relevant content indexed by that system. The query is a key theoretical construct in both the information-retrieval systems research and the information searching behavior research fields.[14] Figure 1 displays an example of a search query for "digital libraries" in one of the major search engines.

Many library catalog use studies, conducted in the 1930s to early1970s and summarized by Krikelas in 1972, compared how traditional library card catalogs were searched and how different bibliographic description elements (fields) of the catalog card were used by library patrons.[15] It was found that author, title, subject headings, call number, and date of publication were heavily consulted by users while place of publication, publisher, edition, and content notes tended to be consulted less often; size, series note, and illustration statement were of less interest to library users.

The early automated catalogs of 1970s, mostly due to hardware and software limitations, had less functionality than card catalogs and were used by and designed specifically for library staff trained in technology.[16] Only in the early 1980s did end users receive direct access to public access catalogs. Largely because online catalogs brought in new affordances of "search capabilities indexing," which include keyword searching, Boolean searching, truncation, and multi-index searching, users initially expressed much greater satisfaction with online catalogs than with traditional library card catalogs.[17] Such patterns of user searching as the use of Boolean operators and controlled vocabulary in online catalogs were analyzed in many studies. Boolean searching was found to be ineffective, not only because the majority of library users—even highly educated ones—experienced difficulties with Boolean logic concepts, but also because the execution order of Boolean commands was not standardized across different online catalogs.[18] Moreover, some research from the early 1990s found that performance is improved in systems that do not require Boolean operators for complex queries.[19] Many studies demonstrated that users of online catalogs tend to use simple keyword searches more often than any type of advanced search that allows them to search

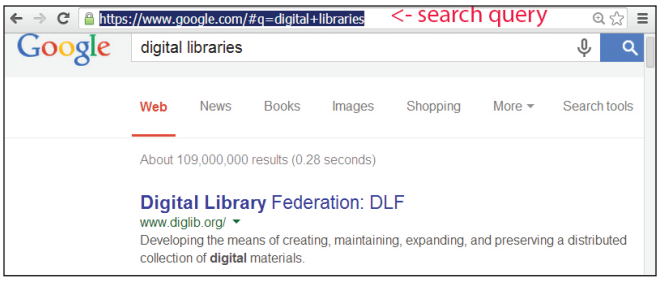

Figure 1. Example of a search query for "digital libraries" in one of the major search engines that automatically limits the number of retrieved results to 109,990

with controlled vocabulary terms.[20]

Digital libraries have been developed since the late 1990s. User expectations of digital libraries were shaped by experiences with major easy-to-use search engines (predominantly Google), widely used transactional sites (e.g., Amazon and eBay), the popularity of computer games, and changes in the Western society in general: greater speed of developments, perceived need for immediate gratification, a more information-rich environment, and the popular heuristic of "satisficing" (i.e., an approach when the user is satisfied with "good enough" results that reach the minimum acceptability threshold by meeting some of the criteria and sacrificing other criteria).[21] As a result, users typically expect much more from digital libraries than from conventional library services. These expectations include comprehensiveness, accessibility, immediate gratification, ease of use, and availability of data in multiple formats.[22] Expectations of digital library services are often too high (although this is somewhat context-dependent) and are combined with a lack of appreciation of basic points, such as that digital library collections are created based on the knowledge of user groups' needs.[23]

A study released by British Library researchers in 2008 found that the main characteristics of user behavior include horizontal information seeking (a form of skimming activity, where searchers view just one or two pages from an academic site and then "bounce" out); extended navigation (people spend as much time finding their way around as actually viewing results); horizontal "power browsing" through titles, contents pages and abstracts; squirreling behavior, where the user saves information—particularly free content—in the form of downloads for later use but rarely revisits it; and little time spent in evaluating information.[24]

Domain Knowledge and User Searching

The term *domain knowledge* was first coined in 1991 by Allen, who 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 level of

knowledge both on a specific search topic and the broader subject domain.[25] This observation was confirmed by numerous other studies. Researchers found that domain experts focus on the answers to search questions and have clear expectations for both the answer and the context in which it would appear.[26] For example, digital library user expectations, including "collection expectation"—an expectation that certain kinds of resources and information would be found in library or academic sources and not in search engines—differ by user domain and level of expertise.[27]

Another important finding is that search tactics used by students change over time: as students acquire more domain knowledge on their research topic, they start to use wider and more specific vocabulary in their subject search.[28] Studies show that when the domain knowledge is low, higher numbers of searches per session occur because of an inability to initially choose appropriate terms; more domain-knowledgeable students use advanced search options more but make fewer changes to their searches.[29] It was also found that with an increase in the level of domain knowledge, users tend to use more terms in queries; domain experts use more effective strategy, conduct more complex searches, and incorporate more unique terms.[30]

Many studies considered how representatives of a specific domain or discipline search for information. Studies have shown that scientists' searching is usually aimed at specific questions or problems that they face when conducting an experiment, writing up results, or checking the accuracy of information in hand.[31] Humanities researchers with high domain knowledge were found to use a variety of search types, with known author-title search being the least problematic. Success in more uncertain types of searches (e.g., a conceptual/discipline term search) was found to heavily depend on the level of the searcher's domain knowledge and experience in using a particular digital library. Subject classifications were almost never used by academic searchers because the scholars' conceptual models usually differed from that represented in the classification scheme.[32]

Digital library users—humanities and social science researchers—were found to prefer searching to browsing, and some explained this by the lack of call number browsing capabilities in a digital library environment.[33] The users—both students and scholars—in these domains have been found to actively use timeline and chronological browsing, and interactive map browsing, and express the need for search limit by date.[34]

Researchers also analyzed for what the users of information systems—in particular, scholars (i.e., domain experts)—search. Previous studies of web searching discovered, for example, that humanities scholars most often include in their search queries personal names, geographic names, chronological terms, and discipline terms. As shown by other studies, water quality researchers frequently use topical, geographical, and format or genre search terms, and occasionally, chemical formulas, dates, names, and URLs. Medical researchers' prevailing search query types were found to include laboratory/test results, disease/syndrome, body part/organ/organ component, pharmacological substance, or diagnostic procedure.[35]

The user base for domain-independent information systems that are aimed at a broad user audience tends to include more novice than expert users, while the audience of domain-specific information systems that are aimed at the users in certain domain (e.g., history, science) typically includes a higher proportion of domain experts. The studies discussed above analyzed information seeking behavior of the users of domain-independent and domain-specific information systems of more traditional types: bibliographic databases, including library catalogs, and some web search engines. However, information seeking by users in either domain-specific or domain-independent 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 previously researched and compared.

Transaction Log Analysis Studies of User Searching

Transaction logs recorded by the servers of information retrieval systems provide a wealth of data for analysis of various patterns of user information seeking expressed through queries. Transaction log analysis—"the study of electronically recorded interactions between online information retrieval systems and the persons who search for the information found in those systems"—is one of the methods actively used for unobtrusive observation of user behavior in various information retrieval systems.[36] For example, Markey's summary of research results of the studies into information seeking behavior conducted over the period of 25 years, demonstrates that many of these studies used transaction log analysis method.[37] Transaction log data are often analyzed quantitatively. For example, Jansen, Spink, and Pedersen compared search query length and Boolean usage rates in different digital collections.[38] Moulaison studied transaction logs and analyzed queries by the users of a college online public access catalog and indicated that the number of terms included in a search (a query length) and the number of search limits can serve as "measures of search complexity" and a way of "documenting the sophistication of the queries."[39] Transaction log analysis is frequently used in qualitative analysis. Several studies categorized web search queries into topical categories applying qualitative methods of research.[40]

Despite the popularity of transaction log analysis as a research method for studying user information-seeking behavior, the potential of transaction log analysis has not

been used to its full capacity to benefit large-scale digital libraries' development. Several published earlier studies have analyzed transaction logs of domain-specific, large-scale digital libraries such as the NSDL, AM, OH, or domain-independent, large-scale digital libraries such as IMLS DCC or the European Library.[41] However, only two of these studies examined the content of the user search queries.[42] Moreover, the user search queries in the various types of digital libraries aimed at different user communities have not been previously compared.

The study reported in this paper addresses this gap through a mixed-method comparative analysis of transaction log data for patterns of user searching in two representative domain-specific, large-scale digital libraries: OH and NSDL.

Research Method

While some large-scale digital libraries regularly collect transaction log data, others do not. Moreover, a variety of tools and methods are used to record transaction log data, which significantly complicates comparative analysis. To ensure a meaningful comparison, the decision was made to seek transaction log datasets collected using the same application. Two large-scale digital libraries that had collected transaction log data using the Google Analytics application as of January 1, 2010, and had made that data available to the researchers in late 2011—the NSDL and the OH portal—were selected as the targets of this investigation. All of the user search queries in these two large-scale digital libraries during the one-year period between January 1 and December 31, 2010, were included in the analysis.

The transaction log data collected by the Google Analytics application were imported into Microsoft Excel spreadsheet files for further analysis. A total of 28,531 non-empty search queries (i.e., the search queries that not only contained a search page URL and a search command, as in http://nsdl.org/search/?verb=Search&q=&submitButton=Search—but also contained a string of characters indentifying the specific search term(s) used in the search, as in http://nsdl.org/search/?verb=search&q=sharks=&submitButton=Search) were identified in the log samples. Many of the more popular search terms occurred more than once, so identical queries were grouped together and folded into unique search queries for each of the two domain-specific digital libraries. For example, if the Google Analytics search log file for a digital library contained five instances of user(s) searching with the same search term (e.g., "clouds") at different points in time, for the purposes of further analysis, the authors counted them as a single unique search query "clouds" with five instances, or with a query frequency value of 5. This resulted in 13,965 unique search queries: 2,551 in OH and 11,414 in the NSDL.

Both quantitative and qualitative characteristics of user searching were assessed. Frequencies of occurrence (total and mean) were measured for basic and advanced search approaches. Variability measures—variance and standard deviation—were also assessed. In addition, query length and query frequency indicators were measured using traditional definitions and approaches applied in transaction log analysis studies as suggested by Spink et al., query length was measured as the number of words in a query, and query frequency was measured as the number of times a query appears in the dataset.[43]

Unique search queries were categorized into ten search categories. These coding categories included seven bibliographic entities from three groups of entities in the Functional Requirements for Bibliographic References (FRBR) model—*work* entity from Group 1 (resource entities), *person* and *corporate body* entities from Group 2 (agent entities), and *concept*, *object*, *event*, and *place* entities from Group 3 (subject entities). One entity from the Functional Requirements for Authority Data (FRAD) model—*family*—was also adopted as a coding category. The researchers adopted the following definitions for the search categories from definitions of FRBR and FRAD entities:

- *work*—"a distinct intellectual or artistic creation."[44]
- *person*—"an individual . . . encompasses individuals that are deceased as well as those that are living" and "includes personas established or adopted by an individual through the use of more than one name (e.g., the individual's real name and/or one or more pseudonyms), includes personas established or adopted jointly by two or more individuals (e.g., Ellery Queen—joint pseudonym of Frederic Dannay and Manfred B. Lee). Includes *personas* established or adopted by a group (e.g., Betty Crocker)."[45]
- *family*—"two or more persons related by birth, marriage, adoption, or similar legal status, or otherwise present themselves as a family. Includes royal families, dynasties, houses of nobility, etc. Includes patriarchies and matriarchies. Includes groups of individuals sharing a common ancestral lineage. Includes family units (parents, children, grandchildren, etc.). Includes the successive holders of a title in a house of nobility, viewed collectively (e.g., Dukes of Norfolk)."[46]
- *corporate body*—"an organization or group of individuals and/or organizations acting as a unit, encompasses organizations and groups of individuals and/or organizations that are identified by a particular name."[47]
- *concept*—"an abstract notion or idea . . . encompasses a comprehensive range of abstractions that may be the subject of a *work*: fields of knowledge, disciplines,

schools of thought (philosophies, religions, political ideologies, etc.), theories, processes, techniques, practices, etc. A *concept* may be broad in nature or narrowly defined and precise."[48]

- *object*—"a material thing . . . encompasses a comprehensive range of material things that may be the subject of a *work*: animate and inanimate objects occurring in nature, fixed, movable, and moving objects that are the product of human creation, objects that no longer exist."[49]
- *event*—"an action or occurrence . . . encompasses a comprehensive range of actions and occurrences that may be the subject of a work: historical events, epochs, periods of time, etc."[50]
- *place*—"a location . . . encompasses a comprehensive range of locations: terrestrial and extra-terrestrial, historical and contemporary, geographic features and geo-political jurisdictions."[51]

The coding categories included only one entity—*work*—out of four FRBR family of models' Group 1 (resource entities) categories that include *work*, *expression*, *manifestation*, and *item.* Despite the benefits of search log analysis as an unobtrusive method of observation, one of its limitations is the impossibility to detect from the user's search query what exactly the user is expecting to find any instance of a *work*, its particular *expression* (e.g., Spanish translation), *manifestation* (e.g., 2nd edition), or a specific digital *item*/copy embodying the work. For this reason, only the broadest FRBR Group 1 entity—*work*—was adopted as a search category for this analysis.

In addition to the eight search categories listed above, two more search categories were used in this study: *class of persons* and *ethnic group*; they were derived from the earlier study by the first author of this paper.[52] The following definitions were used for these search categories:

- *ethnic group*—people of the same race or nationality who share a distinctive culture (e.g., Irish Americans, Sioux Indian, Basque).
- *class of persons*—a group of people who shares common attributes, characteristics, qualities, or traits other than race or nationality (e.g., children, graphic designers, prisoners).

Polysemic user search queries (i.e., search queries using the words that can have multiple meanings) were assigned to multiple categories. For example, a polysemic search query "network" was categorized as a *concept* (a network in abstract sense, e.g., social network, Internet, etc.) and an *object* (a physical network, e.g., fish net, spider web, etc.) while another polysemic search query "cologne" was categorized as as an *object* (perfume type) and a place (a city in Germany). Most phrase queries also belonged to multiple categories. For example a "two eagles cherokee" query was categorized as *object*, *person*, and *ethnic group* while an "Atlanta 1864 map" query was categorized as *place*, *event*, and *object*. Whenever possible, search queries formulated in languages other than English were translated and categorized into appropriate search categories, e.g., German-language query "Anti faschisten," which translates as "anti-fascists," was placed into the *class of people* category, the French-language search query "revolution français," which means "French revolution," was categorized as an *event* search.

To assure that the results of the analysis were not "skewed by a single coder's subjective judgment and bias," this research employed two coders.[53] The first author of this paper coded all of the 13,965 unique search queries in this dataset. A subset of 18.8 percent of unique search queries—10.51 percent of the NSDL dataset and 55.89 percent of the OH dataset—was coded by the second author. In the coding process, the authors worked independently of each other and applied the same coding instructions to the same subsets of the units of analysis.[54] A detailed coding manual was developed to support coding activity; it included definitions and examples for each of the coding categories, along with other guidelines. The method had been originally developed and tested by the first author in the study on a sample of 500 search queries and later refined and tested in the study on a sample of 1,200 search queries.[55] For example, the revised coding manual recommended categorizing the names of diseases and medical conditions—which often occurred in NSDL search queries—as *concept* searches if the affected organ was not explicitly indicated in the search query (e.g., "pneumonia"), as *concept* and *object* searches if the organ was named (e.g., "kidney dysfunction") or if the pathogen causing disease was named in the query (e.g., "adenovirus common cold"), and as *concept* and *person* if the disease or condition name included the name of the person who first found or described it (e.g., "Down syndrome").

To establish the reliability of the coding measures, one must examine the similarities and differences in the coders' results and assess the "amount of agreement or correspondence among two or more coders" in coding (i.e., to measure intercoder agreement, also often referred to as intercoder reliability).[56] To ensure that the findings of content analysis are reliable, it is generally recommended to measure intercoder agreement not only as a percentage (as in "two coders agreed with each other in categorizing 83 percent of search terms"), but also using one of the much more complex coefficients (e.g., Cohen's Kappa), which are usually calculated with the help of statistical software packages such as SPSS, SAS, etc. An intercoder agreement coefficient of .90 or greater is considered acceptable to all, and one of .80 or higher is acceptable to most situations.[57]

In the study presented in this paper, a strong intercoder agreement—97.02 percent or Cohen's Kappa of .877 for the NSDL dataset and 99.40 percent or Cohen's Kappa of .976 for the OH dataset—was observed.

At the time of data collection, the NSDL—one of the pioneering digital libraries launched in 2000—was an established and widely used digital library, while OH—created almost a decade later, in 2008—was not yet as widely known to its potential users. A considerable difference in the levels of use of the two digital libraries resulted in an almost fivefold difference in sample sizes between the two digital libraries. To minimize the risk of uneven sample sizes skewing the study results regarding search query frequencies, particularly in relation to the fraction of queries that occurred only once, the authors of this paper decided to analyze and report relative numbers, or percentages (e.g., "30 percent of search queries in the A digital library and 35 percent of search queries in the B digital library had a query frequency of X or more") rather than absolute numbers (e.g., "50 search queries in the A digital library and 320 of search queries in the B digital library had a query frequency of X or more") whenever possible.

In addition, to make sure that differences or similarities observed in user searching between the two digital libraries of different domains did not occur by chance and are therefore real and worth considering in digital library development, the statistical significance of the comparative analysis findings was assessed. The most widely used method of assessment of statistical significance, which is considered appropriate for most types of data, is the t-test. When the t-test is not appropriate—for example, for binary data, where only two values (e.g., yes or 1, and no or 0) are possible, other methods of assessment (e.g., Chi-square test) are used. The authors of this paper used the t-test whenever applicable and substituted it with the Chi-square test as needed. In this study, a t-test was used to assess the statistical significance of the results for search query length, search query frequency, and number of search categories per search query; a Chi-square test was used to assess the statistical significance of the results for the frequencies of selection of advanced and basic search approaches and for the inclusion of particular search categories in users' search queries.

The values of a t-test and Chi-square test are most often calculated with the help of statistical software packages using complex formulae; researchers have to preset the values for probability of error level (p) and degrees of freedom (df) before this calculation. These values are also necessary to interpret the results of the t-test and Chi-square test with the help of special tables developed by statisticians and to determine whether the findings are statistically significant. Usually, a probability of error level of .01 ($p < .01$), which means that a finding has a 99 percent chance of being true, is considered appropriate for establishing the statistical significance of research results with either the t-test or the Chi-square test. In this study, a probability of error level of .01 was used both in t-tests and in Chi-square tests. The degree of freedom for t-test is set based on the number of observations (e.g., search queries) and calculated as the number of observations minus 2. If the data are binary, like in the case with some of the data used in this research, the degree of freedom for a Chi-square test is set as 1. For t-tests with a probability of error level $p < .01$ and large value (i.e., more than 29) for degree of freedom, the value of a t-test over 2.575 indicates that the finding is statistically significant. For Chi-square tests with a probability of error level $p < .01$ and the degrees of freedom value $df = 1$, the value of a Chi-square test over 6.635 indicates that the finding is statistically significant. In the "Findings and Discussion" section of this paper, results of the t-test are reported as the value of t-test ($t = X$) calculated by the SPSS statistical software application, followed by the preset value of degrees of freedom ($df = Y$), and the preset value of probability of error level ($p < Z$), where X, Y, and Z are the specific numbers; for example, as "$t = 8.09$, $df = 13963$, $p < .01$." The results of the Chi-square test are reported the same way, with the value of Chi-square test shown as Chi-square = X (for example, as "Chi-square = 567.5135, $df = 1$, $p < .01$").

Findings and Discussion

User Search Approaches

Figure 2 displays search options that were enabled in the OH large-scale digital library at the time of this study. There were two types of item-level search where the user searches for individual information objects. The basic version of *search for items* allowed searching by keyword or phrase anywhere in item-level metadata records that describe individual items. The advanced version of *search for items* allowed searching by keyword or phrase anywhere in item-level metadata records, the faceted searching (i.e., a search in specific field(s) of the metadata record) by author/artist's last name and/or by title/subject word(s), with a possibility to combine search fields and to limit results to specific digital collections selected from the dropdown menu. OH also thoughtfully provided its users with two collection-level search options to accommodate users interested in finding entire collections as opposed to individual items. The basic version of *search collections* allowed for a simple keyword search in all fields in collection-level metadata records that describe the collection of items as whole as opposed to individual items. The advanced version of *search collections* option augmented a simple keyword search in all fields in collection-level metadata records with a possibility to limit

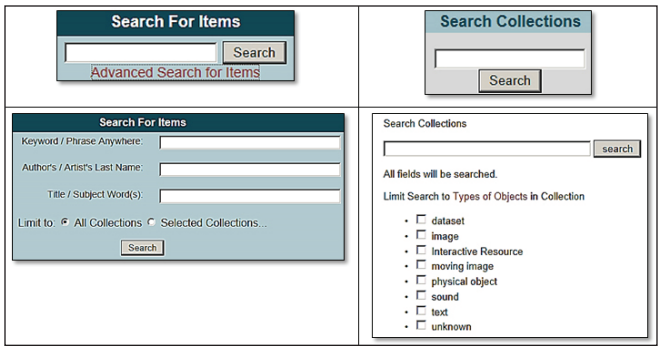

Figure 2. Search options in Opening History

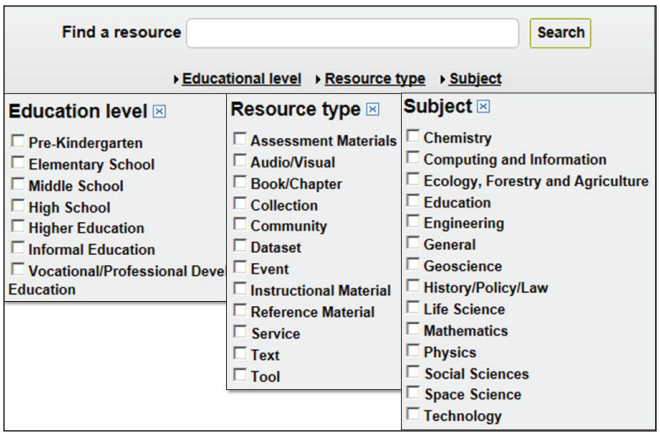

Figure 3. Search options in the National Science Digital Library

search results to collections containing one or more of eight resource types: dataset, interactive resource, physical object, text, image, moving image, and sound.

Figure 3 shows the search options enabled in the NSDL. There is one simple keyword search window, with a possibility to limit search results using one or more of the seven education levels, one or more of the twelve resource types, and one or more of the fourteen broad subject areas. The NSDL does not provide separate search options for advanced item-level search in specific fields of a metadata record (i.e., faceted search) or for collection-level search.

As shown in table 1, most of the search queries in both digital libraries were basic keyword searches, although the percentage of basic keyword searches was significantly higher in OH. The use of various advanced search options was observed in fewer than 15 percent of search queries in OH but in almost 40 percent of search queries in the NSDL. Hence the users of the NSDL engaged in advanced searches much more often than OH users (Chi-square = 1001.602, df = 1, $p < .01$).

Table 2 shows the use of advanced search options in two digital libraries in more detail. Both digital libraries allow the use of various search limits as part of advanced searching (e.g., limit search results to certain collections or by certain object types or audiences). These limits were used in 12.58 percent of search queries in the OH sample. More than a third (38.28 percent) of search queries in the NSDL sample included one or more of search limits. Faceted search queries are the advanced search queries in which the user is allowed to search for matches in specific fields of metadata records. OH users had an option to search by author or by title and subject words. Faceted searching was observed in a small proportion (0.61 percent) of the OH search queries in the sample but not at all in the NSDL sample, which is explained by the absence of faceted search options in this digital library. Use of another advanced search feature—quotes for bound phrases (as in the "'climate change' and water" NSDL query or in the "'John Cobb' Bonneville" OH query)—was also observed infrequently (1.32 percent of search queries in OH and 0.79 percent of search queries in NSDL).

Table 1. Search approaches by digital library (observed frequencies)

Search approach	NSDL	OH	Total
Basic	14,665	3,823	18,488
Advanced	9,406	647	10,053
Total	24,071	4,460	28,531

Chi-square = 1001.602, df = 1, $p < .01$

Table 2. Search approaches: details (N = 28,531)

Search approach	NSDL	OH
Basic keyword search	60.92%	85.49%
Advanced search, including:	39.08%	14.51%
Search limit (audience, collection, genre, etc.)	38.28%	12.58%
Faceted search (e.g., search in author field)	—	0.61%
Bound phrase search	0.79%	1.32%

Search Query Lengths

Search query length is defined as the number of words in the query. For example, the query length of the "national parks" query is two, while the query length of the "John White's narrative of the 1587 Virginia voyage" query is eight. Almost half of user search queries (47 percent) in the OH sample but only 29 percent of users' search queries in the NSDL sample consisted of a single word. As shown in table 3, the search queries of the NSDL users varies more widely in length than those of OH users, which is demonstrated by higher variability—both variance (4.14 as compared to 2.58) and standard deviation (2.03 as compared to 1.61).

Table 3. Search query length (*N* = 13,965)

	NSDL	OH
Number of observations	11,414	2,551
Mean	2.66	2.36
Standard deviation	2.03	1.61
Variance	4.14	2.58

t = 8.09, df = 13963, p < .01

Table 4. Search query frequency (*N* = 13,965)

	NSDL	OH
Number of observations	11,414	2,551
Mean	6.54	1.75
Standard deviation	23.1	1.66
Variance	534	2.76

t = 21.8926, df = 13963, p < .01

For OH, the search query length ranged from 1 to 13 words per query, while the NSDL range was much more considerable: 1 to 53 words. Search queries in NSDL were found, with statistical significance (t = 8.09, df = 13963, p < .01), to NSDL to tend to contain more words than search queries in OH. The average NSDL search query length (2.66 words) was found to be shorter than the average OH search query length (2.36 words).

Search Query Frequencies

Search query frequency is the number of times certain identical queries are found in the transaction log dataset in a digital library. For example, a query "Wisconsin statehood" occurs only once over the period of twelve months in the OH sample, thus its query frequency equals 1; a query "planting" occurs six times in the same sample, thus its query frequency equals 6. Similarly, a query "Ohm's law water" occurs four times over the period of twelve months in the NSDL sample, thus its query frequency equals 4; a query "anticoagulant properties of snake venom" occurs eleven times in the same sample, thus its query frequency equals 11. This study revealed, with statistical significance (t = 21.89, df = 13963, p < .01), that the search queries of the NSDL users occurred more frequently than the search queries of OH users. As shown in table 4, the average NSDL search query frequency (6.54) was found to be considerably higher than the average OH search query frequency (1.75).

Query frequency had a much higher variability in NSDL search queries than in OH search queries—both in variance (534 compared to 2.76) and in standard deviation (23.1 compared to 1.66). In OH, all of the search queries occurred between 1 and 25 times. Similarly, in NSDL, the vast majority of search queries (97.28 percent) occurred between 1 and 25 times. Although some NSDL queries (e.g., "chemistry," "relativity and Einstein," "photosynthesis" etc.) occurred as often as 100 to 885 times their proportion in the total dataset was minuscule, measured in tiny fractions of a percent. Therefore the detailed analysis of search query frequency, results of which are presented in figure 4, focused around search queries that constituted at least 0.01 percent of all search queries in either digital library, i.e., the queries that ranged in query frequency from 1 to 25 (percentage values that are below 0.01 percent are not shown on figure 4). Two-thirds (65.7 percent) of search queries in OH and almost a half (48.55 percent) of search queries in NSDL occurred only once in their respective search log samples. While the number of search queries with frequencies of 1, 11, 12, 13, 14, 16, 19, 20, and 25 were higher in the OH search log sample, search queries with frequencies of 2, 3, 4, 5, 6, 7, 8, 9, and 10 were more common in the NSDL search log sample.

Search Categories

As discussed above, in the categorization of user search queries, the following categories were used: *work*, *person*, *family*, *corporate body*, *concept*, *object*, *event*, *place*, *ethnic group*, and *class of persons.* As discussed above, polysemic and most phrase queries search queries were assigned to multiple categories. Qualitative results of search query categorization were quantified and are reported here as percentages of search queries in which a certain search category occurred.

The presence of particular search categories in user search queries displayed noticeable differences between the two domain-specific digital libraries. As shown in figure 5, the top two categories observed in the OH search queries were *place* (e.g., "Chile") with 34 percent of searches, and *object* (e.g., "drinking vessel") with 31 percent of searches. It is worth noting that both of these categories belong to FRBR Group 3 of entities, or subject entities. Another Group 3 search category—*concept* (e.g., "civil right")—was the fourth most common search category with 17 percent of search queries. However, the fourth FRBR Group 3 subject entity—*event* (e.g., "1935 meat strike")—was observed in the search queries much less often than the other three (10 percent). The FRBR Group 2 (agent) search categories *person* (e.g., "Alfred R. Glancy Jr.") and *corporate body* (e.g. "Dana College," "Kapa Alpha Psi") were observed in 26 percent and 14 percent, respectively, of the search queries. The *work* search category (e.g., "Find It Illinois," "how a colored woman aided john brown") was observed in 9 percent of the search queries, while the *class of persons* (e.g., "fashion designers") and ethnic group (e.g., "Cheyenne") search categories were observed in 8 percent and 5 percent of user

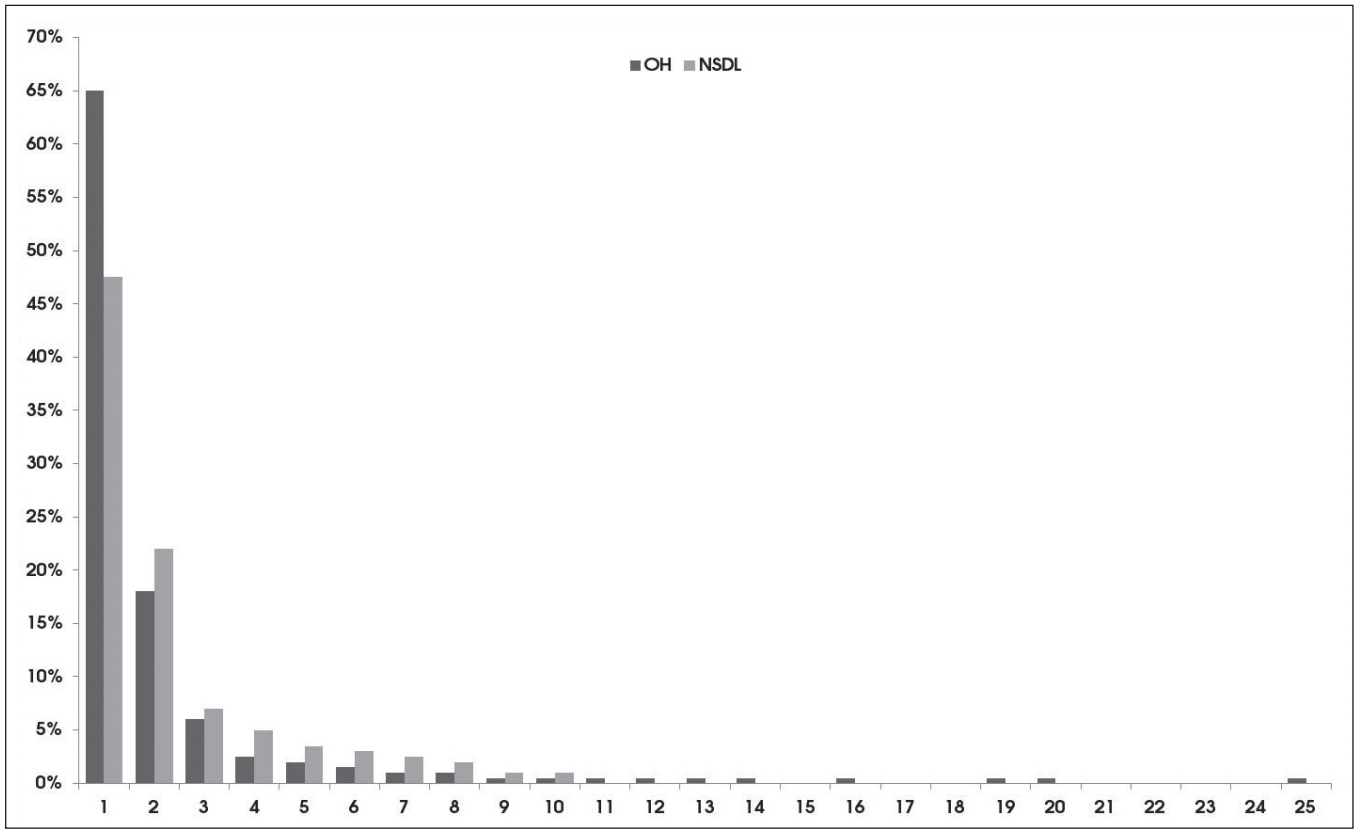

Figure 4. Search query frequency distribution

searches respectively. Finally, the *family* search category (e.g., "Wright brothers") was observed in only 0.43 percent of unique search queries in OH.

In the NSDL search queries (figure 5), the distribution of search categories was different. Two search categories occurred significantly more often in the search queries of the NSDL users as compared to the queries of the OH users. This statistically significant difference was observed for the *object* search category (Chi-square = 292.7135, df = 1, $p < .01$), and the *concept* search category (Chi-square = 1944.959, df = 1, $p < .01$). Similar to the OH sample, *object* (e.g., "starfish student and teacher resource") was one of the most frequently occurring search categories in the NSDL. However, this category occurred in the NSDL sample more often than in the OH sample (51 percent as compared to 31 percent of unique search queries). Unlike in OH search queries, *concept* (e.g., "epigenetics") was the most frequently observed search category in the NSDL sample, with 64 percent of unique search queries containing *concept*. This is a considerably higher proportion of search queries than was found for OH (17 percent). Two search categories that were prevalent in OH search queries but occurred substantially less often in the NSDL search queries, with high statistical significance, included *place* (Chi-square = 2018.511, df = 1, $p < .01$) and *person* (Chi-square = 753.2203, df = 1, $p < .01$).[58] The *Place* category (e.g., "Chesapeake bay") was found in only 5 percent of NSDL search queries as opposed to 34 percent of OH search queries; *person* search category (e.g., "Nikolai Lobachevsky") was found in only 7 percent of NSDL search queries as opposed to 26 percent of OH search queries. It was also found, with statistical significance, that several other search categories had occurred less often in user search queries in the NSDL than in OH: *corporate body* (Chi-square = 567.5135, df = 1, $p < .01$), *ethnic group* (Chi-square = 277.1274, df = 1, $p < .01$), *event* (Chi-square = 276.574, df = 1, $p < .01$), and *class of persons* (Chi-square = 50.4434, df = 1, $p < .01$). The *corporate body* search category (e.g., "NASA") was present in only 3 percent of NSDL search queries as opposed to 14 percent of OH search queries; *ethnic group* category (e.g., "mayans") was present in only 0.7 percent of NSDL search queries as opposed to 5 percent of OH search queries; *event* category (e.g., "middle ages") was present in only 3 percent of NSDL search queries as opposed to 10 percent of OH search queries; and *class of persons* category (e.g., "meteorologist") was present in only 3 percent of NSDL search queries as opposed to 8 percent of OH search

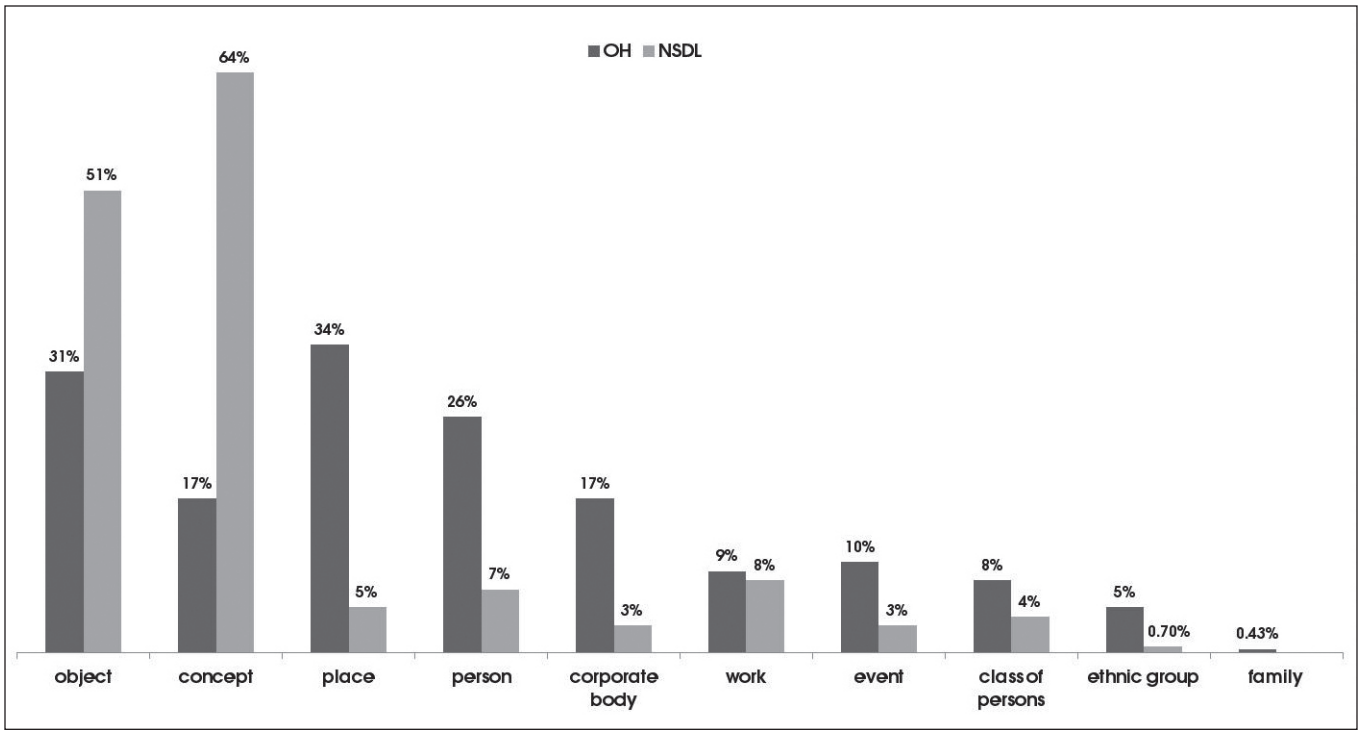

Figure 5. Search categories in OH and NSDL user search queries

queries. The only search category where the difference in occurrence in NSDL and OH search queries was not statistically significant (Chi-square = 2.524082, df = 1, p < .2) was *work* search category. Finally, no *family* searches were observed in the user interactions with the NSDL, as compared to only 0.43 percent of OH search queries that contained the *family* search category.

As discussed above, in the process of analysis, polysemic search queries and most phrase search queries were assigned to multiple categories. The researchers found that although over half of all user search queries in the sample (59 percent in the NSDL and 55 percent in OH) belonged to single search category, a considerable proportion of search queries (45 percent in OH and 41 percent in the NSDL) were multi-category search queries. The number of categories in these multi-category search queries ranged from two to seven. Search queries of the OH users were found, with statistical significance (t = 8.71, df = 13479, p < .01), to contain more search categories than search queries of the NSDL users (see table 5).

As shown in figure 6, roughly a third of search queries (32 percent in the OH sample and 36 percent in the NSDL sample) combined two search categories. Representative examples include "Hisako eagle painting" (*person* and *object*), "mining Arizona Mohave county" (*concept* and *place*), "ADHD in teens" (*concept* and *class of persons*), and "Galapagos tortoise" (*object* and *place*).

The proportion of search queries with three and more search categories was low in both digital libraries. Ten percent of user search queries in the OH sample and 5 percent of queries in the NSDL sample included three search categories. Representative examples of three-category search queries include "California gold rush" (*place*, *object*, and *event*), "mechanics; Newton; acceleration; catapult; force; graphing; laws of motion; mass; motion" (*concept*, *person*, and *object*), and "Japanese culture United States" (*ethnic group*, *concept*, and *place*). Four-category search queries constituted 2 percent of all search queries in the OH transaction log sample, while only 0.4 percent of search queries in the NSDL sample comprised four search categories. Representative examples include "history of oil spills in the United States" (*concept*, *event*, *object*, and *place*) and "the mcgraw-hill dictionary of misspelled and easily confused words" (*concept*, *object*, *corporate body*, and *work*). A small fraction of search queries (0.45 percent in OH and 0.03 percent in the NSDL) included five search categories. For example, "fowler, h. w. 1914. fishes from the rupununi river, british guiana. proceedings of the academy of natural sciences of philadelphia v. 66:229–284" search query from NSDL included the categories: *object*, *place*, *person*, *work*, and *corporate body*. Finally, six category and seven category queries composed only 0.04 percent of the OH user

Table 5. Number of search categories per unique search query (*N* = 13,481)

	OH	NSDL
Number of observations	2,441	11,040
Mean	1.61	1.46
Standard deviation	0.8	0.61
Variance	0.64	0.38

t = 8.71, df = 13,479, p < .01

search queries each, while no multicategory search queries with more than five search categories were observed in the NSDL search logs.

Conclusion

The study shows that in both domain-specific large-scale digital libraries created for different user communities, users preferred a basic keyword search to advanced search options; this observation correlates with results of earlier studies, which found that users of online catalogs and other information retrieval systems tend to prefer simple keyword searches, although evidence from more recent studies shows an upward trend in the use of faceted searches in online catalogs.[59] However, the level of advanced searching observed by the authors of this article in both large-scale digital libraries is high compared with the findings of studies of user searching on the web or in online databases.[60] This may indicate higher a proportion of domain expert users in large-scale digital libraries, as many user studies report that selection of advanced search options increases with increase in user domain knowledge. While NSDL users engaged in advanced searches much more often than OH users, this finding might in part be explained by interface differences between the two large-scale digital libraries, as more search limit options are offered by the NSDL interface. This study also found lower average search query lengths than most transaction log analysis studies of online catalogs and web search engines, as summarized by Markey, with exception of the study of New Zealand Digital Library, the query lengths observed in which were very close with the query lengths observed in the present study.[61]

This study revealed that user searching differed substantially between two digital libraries that are aimed at serving different domains and user populations. For example, the search queries of the NSDL users varied more widely in length, on average were longer, and occurred more frequently than the search queries of OH users. As shown by this study, search queries of the OH users varied more widely in frequency and contained more search categories than search queries of the NSDL users. In addition, the frequency of occurrence of particular search categories in user search queries displayed noticeable differences between the two domain-specific digital libraries. Two search categories—*concept* and *object*—occurred significantly more often in the search queries of the NSDL users while several others—*place*, *person*, *corporate body*, *ethnic group*, *event*, and *class of persons*—occurred significantly more often in OH search queries.

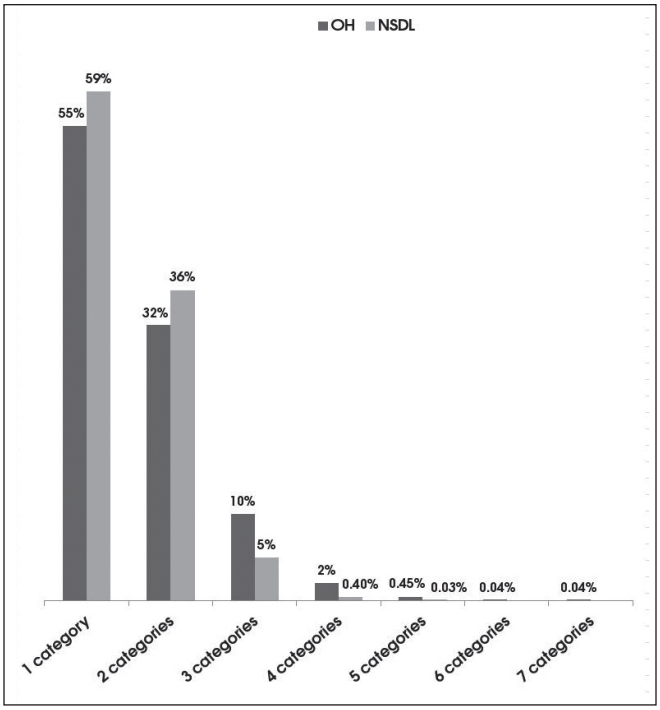

Figure 6. Number of search categories per unique search query

Based on the results of this study, the conclusion can be made that geographical and personal names continue to be prevalent in humanities users' searches, while *object* searches not reported by earlier studies also have prominence. The generally low level of *event* searching in a large-scale digital library in the US history domain (only 10 percent of the OH user search queries contain an *event* search category) observed in this study is somewhat unexpected.

Interface design may have somewhat contributed to user searching differences between OH and the NSDL. Additional investigation into this factor is needed and will be carried out by the authors of this paper. Nevertheless, some of the most statistically significant findings of this exploratory study bear practical implications for digital library developers. For example, the differences revealed by this study suggest that developers of digital libraries serving the STEM population need to give more priority to providing faceted search options and search result limits and to documenting the *concepts* and *objects* (including genres, formats, etc.) in

metadata records. At the same time, as the user-searching data collected by this study suggests, developers of cultural heritage digital libraries that are aimed at serving educators, students, and researchers in the areas of history—and possibly also related social science fields—need to document a wider variety of item attributes in their metadata than need to be documented by their colleagues developing STEM digital libraries, and to pay particular attention to documenting the *persons* and *places* in their metadata. User experience can be improved if large-scale digital libraries—regardless of domain—supply an option to limit search results by geographical area, which is suggested by the high proportion of *place* searching observed in this study.

The need for advanced search options in the user interface of large-scale digital libraries has been proven in the surveys of digital library users. For instance, in the European Library, the majority (81 percent) of users expressed preference for advanced search.[62] The study reported in this paper provides empirical support that is based on actual user behavior in large-scale digital libraries. By doing this, it makes a substantial contribution to establishing the importance of advanced search options in the digital library user interface—the options that are currently often neglected by digital library developers.[63] In particular, the prevalence of subject searching among the users of both digital libraries, which was observed in this study, suggests that provision of the subject-based advanced search option should be prioritized in the design of large-scale digital libraries, regardless of domain.

As a result of uneven levels of use of the two digital libraries that served targets of this study, sample sizes differed substantially. In future comparative analysis studies of user searching, the difference in the levels of use of the digital libraries should be taken into consideration. For example, to make the sample sizes more comparable while still drawing the samples over the same period, future research could compare a complete sample of the search queries from a less heavily used digital library to a random subset of search queries from a more heavily used digital library.

This exploratory study used a single data collection method—transaction logs—to assess digital library users' searching. To obtain a more complete picture of user searching in large-scale domain-specific digital libraries, more studies that combine both unobtrusive (e.g., in-depth transaction log analysis) and obtrusive (e.g., interview, observation, survey of digital library users) methods and triangulate the results are needed. These future investigations will also need to increase the number of target domain-specific digital libraries to represent wider variety of domains, and to include domain-independent digital libraries that serve a broader general audience.

The results of this exploratory study can also be used in building domain-specific user models for digital library users in the history domain and the STEM meta-disciplinary domain. To build such user models, additional data that were not the focus of this investigation need to be collected and analyzed. One example of such important additional data collection and analysis include session-level transaction log analysis that will go beyond the individual search query and include sequence of search queries, as well as other user interactions with digital libraries, such as browsing and viewing of metadata records. Another example is comparative analysis of traffic in cultural heritage and STEM digital libraries at various time scales: daily, weekly, monthly, and yearly.

References and Notes

1. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report*, accessed June 25, 2013, www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.
2. The numbers are based on IMLS Digital Collections and Content registry of digital collections supported by IMLS (http://imlsdcc.grainger.uiuc.edu).
3. Judith Skog, Richard M. McCourt, and Jessica Gorman, "The NSF Scientific Collections Survey: A Brief Overview of Findings" (white paper, National Science Foundation, 2009), accessed June 25, 2013, http://digital.library.unt.edu/ark:/67531/metadc25978.
4. Marcia A. Mardis et al., "The Digital Lives of U.S. Teachers: A Research Synthesis and Trends to Watch," *School Libraries Worldwide* 18, no. 1 (2012): 70–86.
5. Caroline R. Arms, "Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress: Part 1," *D-Lib Magazine* (April 1996), accessed June 25, 2013, www.dlib.org/dlib/april96/loc/04c-arms.html; Caroline R. Arms, "Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress: Part 2," *D-Lib Magazine* (April 1996), accessed June 25, 2013, www.dlib.org/dlib/may96/loc/05c-arms.html.
6. Carole L. Palmer, Oksana L. Zavalina, and Katrina Fenlon, "Beyond Size and Search: Building Contextual Mass in Digital Aggregations for Scholarly Use," *Proceedings of the Annual Meeting of the American Society for Information Science & Technology* 47, no. 1 (2010): 1–10.
7. Mardis et al.,"The Digital Lives of U.S. Teachers."
8. Daniella Quinones, "Digital Media (Including Video!): Resources for the STEM Classroom and Collection," *Knowledge Quest* 39, no. 2 (2010): 28–32; Eileen McIlvain, "NSDL as a Teacher Empower Point: Expanding Capacity for Classroom Integration of Digital Resources," *Knowledge Quest* 39, no. 2 (2010): 54–63; Anne Marie Perrault, "Making Science Learning Available & Accessible to All Learners: Leveraging Digital Library Resources," *Knowledge Quest* 39, no. 2 (2010): 64–68; Daniel Toomey, "The National Science Digital Library: STEM Resources for the 21st-Century Learner,"

School Library Monthly 27, no. 2 (2010): 54–56.
9. "About NSDL," National Science Digital Library, accessed June 25, 2013, http://nsdl.org/about.
10. Perrault, "Making Science Learning Available & Accessible to All Learners"; Toomey, "The National Science Digital Library."
11. Shonda Brisco, "The Motherlode of STEM," *School Library Journal* 56, no. 2 (2010): 65–66; McIlvain, "NSDL as a Teacher Empower Point."
12. Brian F. Lavoie, Lynn Silipigni Connaway, and Edward T. O'Neill, "Mapping WorldCat's Digital Landscape," *Library Resources & Technical Services* 51, no. 2 (2007): 106–15; Ingeborg Verheul, Anna Maria Tammaro, and Steve Witt, eds., *Digital Library Futures: Users Perspectives & Institutional Strategies* (Berlin; New York: De Gruyter Saur, 2010; Tony Horava, "Challenges and Possibilities for Collection Management in a Digital Age," *Library Resources & Technical Services* 54, no. 3 (2010): 142–52.
13. Tom D. Wilson, "Human Information Behavior," *Informing Science* 3, no. 2 (2000): 49–56.
14. Bernard J. Jansen and Soo Young Rieh, "The Seventeen Theoretical Constructs of Information Searching and Information Retrieval," *Journal of the American Society for Information Science & Technology* 61, no. 8 (2010): 1517–34.
15. James Krikelas, "Catalog Use Studies and Their Implications," *Advances in Librarianship* 3 (1972): 195–220.
16. Micheline Beaulieu and Christine L. Borgman "A New Era for OPAC Research: Introduction to Special Topic Issue on Current Research in Online Public Access Systems," *Journal of the American Society for Information Science* 47, no. 7 (1996): 491–92.
17. Marcia J. Bates, "Rethinking Subject Cataloging in the Online Environment," *Library Resources & Technical Services* 33, no. 4 (1989): 400; Ray Larson, "Between Scylla and Charybdis: Subject Searching in Online Catalogs," *Advances in Librarianship* 15 (1991): 175–236; Joseph A. Matthews, Gary S. Lawrence, and Douglas K. Ferguson, eds., *Using Online Catalogs: A Nationwide Survey: A Report of a Study Sponsored by the Council on Library Resources* (New York: Neal-Schumann, 1983).
18. Christine L. Borgman, "Why are Online Catalogs Still Hard to Use?" *Journal of American Society for Information Science* 47, no. 7 (1996): 493–503.
19. Bryce L. Allen, "Cognitive Research in Information Science: Implications for Design," *Annual Review of Information Science & Technology* 26 (1991): 3–37.
20. Margo Warner Curl, "Enhancing Subject and Keyword Access to Periodical Abstracts and Indexes: Possibilities and Problems," *Cataloging & Classification Quarterly* 20, no. 4 (1996): 45–55; Raya Fidel, "Who Needs Controlled Vocabulary?" *Special Libraries* 83 (January 15, 1992): 1–9; Charles R. Hildreth, "The Use and Understanding of Keyword Searching in a University Online Catalog," *Information Technology & Libraries* 16, no. 2 (1997): 52–62; Manikya Rao Muddamalle, "Natural Language Versus Controlled Vocabulary in Information Retrieval: A Case Study in Soil Mechanics," *Journal of the American Society for Information Science* 49, no. 10 (1998): 881–87.
21. David Bawden and Polona Vilar, "Digital Libraries: To Meet or Manage User Expectations," *Aslib Proceedings* 58, no. 4 (2006): 346–54.
22. Ibid.
23. Ibid.
24. Ian Rowlands and David Nicholas, "Information Behaviour of the Researcher of the Future: A Ciber Briefing Paper" (London: University College, 2008), accessed June 25, 2013, www.jisc.ac.uk/media/documents/programmes/reppres/gg_final_keynote_11012008.pdf.
25. Bryce L. Allen, "Cognitive Research in Information Science: Implications for Design," *Annual Review of Information Science & Technology* 26 (1991): 3–37.
26. Gary Marchionini et al., "Information Seeking in Full-Text End-User-Oriented Search Systems: The Roles of Domain and Search Expertise," *Library & Information Science Research* 15, no.1 (1993): 35–69.
27. Bawden,"Digital Libraries."
28. Pertti Vakkari, Mikko Pennanen, and Sami Serola, "Changes of Search Terms and Tactics While Writing a Research Proposal: A Longitudinal Case Study," *Information Processing & Management* 39, no. 3 (2003): 445–63.
29. Barbara M. Wildemuth, "The Effects of Domain Knowledge on Search Tactic Formulation," *Journal of the American Society for Information Science & Technology* 55, no. 3 (2004): 246–58, doi:10.1002/asi.10367.
30. Xiangmin Zhang, Hermina G.B. Anghelescu, and Xiaojun Yuan, "Domain Knowledge, Search Behaviour, and Search Effectiveness of Engineering and Science Students: An Exploratory Study," *Information Research* 10, no. 2 (2005): paper 217, accessed June 25, 2013, http://informationr.net/ir/10-2/paper217.html; Helen A. Hembrooke et al., "The Effects of Expertise and Feedback on Search Term Selection and Subsequent Learning," *Journal of the American Society for Information Science & Technology* 56, no. 8 (2005): 861–71, doi:10.1002/asi.20180.
31. Carole L. Palmer, "Scholarly Work and the Shaping of Digital Access," *Journal of the American Society for Information Science & Technology* 56, no. 11 (2005): 1140–53; David Ellis, "Modeling the Information-Seeking Patterns of Academic Researchers: A Grounded Theory Approach," *Library Quarterly* 63, no. 4 (1993): 469–86.
32. George Buchanan et al., "Information Seeking by Humanities Scholars," in *Research and Advanced Technology for Digital Libraries: Proceedings of the European Conference on Digital Libraries (ECDL2005)*, ed. Andreas Rauber, Stavros Christodoulakis, and A Min Tjoa (Berlin: Spring-Verlag, 2005): 218–29.

33. Ibid.; Ming-der Wu and Shih-chuan Chen, "Humanities Graduate Students' Use Behavior on Full-Text Databases for Ancient Chinese Books," in *Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers*, ed. Dion Hoe-Lian Goh et al. (Berlin: Springer-Verlag, 2007), 141–49.
34. Susan Harum, personal conversation with the authors, January 15, 2008; Wu, "Humanities Graduate Students' Use Behavior on Full-Text Databases."
35. Marcia J. Bates, "The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions," *College & Research Libraries* 57 (1996): 514–23; Wu, "Humanities Graduate Students' Use Behavior on Full-Text Databases," 141–49; Elaine A. Nowick and Margaret Mering, "Comparisons between Internet Users' Free-Text Queries and Controlled Vocabularies: A Case Study in Water Quality," *Technical Services Quarterly* 21, no. 2 (2003): 15–32, accessed June 25, 2013, www.tandfonline.com/doi/pdf/10.1300/J124v21n02_02; Kathik Natarajan et al., "An Analysis of Clinical Queries in an Electronic Health Record Search Utility," *International Journal of Medical Informatics* 79, no. 7 (2010): 515–22, doi:10.1016/j.ijmedinf.2010.03.004.
36. Thomas A. Peters, "The History and Development of Transaction Log Analysis," *Library Hi Tech* 11, no. 2 (1993): 41–66.
37. Karen Markey, "Twenty-Five Years of End-User Searching, Part 1: Research Findings," *Journal of the American Society for Information Science & Technology* 58, no. 8 (2007): 1071–81.
38. Bernard Jansen, Amanda Spink, and Jan O. Pedersen, "The Effect of Specialized Multimedia Collections on Web Searching," *Journal of Web Engineering* 3 no. 3-4 (2004): 182-99.
39. Heather L. Moulaison, "OPAC Queries at a Medium-Sized Academic Library: A Transaction Log Analysis," *Library Resources & Technical Services* 52, no. 4 (2008): 230–37.
40. Bates, "The Getty End-User Online Searching Project in the Humanities"; Steven M. Beitzel, Eric C. Jensen, and Abdur Chowdhury, "Temporal Analysis of a Very Large Topically Categorized Web Query Log," *Journal of the American Society for Information Science & Technology* 58, no. 2 (2007): 166–78; Bernard J. Jansen, Amanda Spink, and Sherry Koshman, "Web Searcher Interaction with the Dogpile.com Metasearch Engine," *Journal of the American Society for Information Science & Technology* 58, no. 5 (March 2007): 744–55; Sherry Koshman, Amanda Spink, and Bernard J. Jansen, "Web Searching on the Vivisimo Search Engine," *Journal of the American Society for Information Science & Technology* 57, no. 14 (2006): 1875–87; Amanda Spink and Bernard J. Jansen, "A Study of Web Search Trends," *Webology* 1, no. 2 (2004), accessed June 25, 2013, http://webology.ir/2004/v1n2/a4.html.
41. Michael Khooet et al., "Using Web Metrics to Analyze Digital Libraries," in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM, 2008), 375–84; Bin Pang, "Capturing Users' Behavior in the National Science Digital Library (NSDL)" (unpublished report, Cornell University Human Computer Interaction Research Group, 2003), accessed June 23, 2013, http://arizona.openrepository.com/arizona/bitstream/10150/106332/1/nsdl-user-report.pdf; Suzan Verberne et al., "How does the Library Searcher Behave? A Contrastive Study of Library Search against Adhoc Search," Paper presented at the meeting of the CLEF (Notebook Papers/LABs/Workshops), 2010, accessed June 25, 2013; www.clef-initiative.eu/documents/71612/86374/CLEF2010wn-LogCLEF-VerberneEt2010.pdf; Oksana L. Zavalina, "Collection-Level User Searches in Federated Digital Resource Environment," in *Proceedings of the American Society for Information Science & Technology* (2007), 1–16.; Oksana L. Zavalina, "Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval," *Journal of Library Metadata* 11, no. 3–4 (2011): 104–28.
42. Zavalina, "Collection-Level User Searches in Federated Digital Resource Environment"; Zavalina, "Contextual Metadata in Digital Aggregations."
43. Amanda Spink et al., "From E-Sex to E-Commerce: Web Search Changes," *IEEE Computer* 35, no. 3 (2002): 107–9.
44. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report* (IFLA: September 1997, amended February 2009), accessed June 25, 2013, www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.
45. Ibid.; Glenn E. Patton, ed., *Functional Requirements for Authority Data: A Conceptual Model*, (Munich: K.G. Saur, 2009), 13.
46. Patton, *Functional Requirements for Authority Data*, 13.
47. IFLA, *Functional Requirements for Bibliographic Records: Final Report* www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.
48. Ibid., 26.
49. Ibid., 27.
50. Ibid., 28.
51. Ibid., 28–29.
52. Zavalina, "Collection-Level User Searches in Federated Digital Resource Environment."
53. Kristina M. Spurgin, and Barbara M. Wildemuth, "Content Analysis," in *Applications of Social Research Methods to Questions in Information and Library Science*, ed. Barbara M. Wildemuth, 129–37 (Westport, CT: Libraries Unlimited, 2009).
54. Klaus Krippendorff, "Reliability in Content Analysis Some Common Misconceptions and Recommendations," *Human Communication Research* 30, no. 3 (2004): 411–33.
55. Zavalina, "Collection-Level User Searches in Federated Digital Resource Environment"; Zavalina, "Contextual Metadata in Digital Aggregations."
56. Kimberly A. Neuendorf, *Content Analysis Guidebook* (Thousand Oaks, CA: Sage, 2002).

57. Ibid.
58. This observation is in line with earlier studies of humanities scholars' information-seeking behavior conducted in the 1980s, 1990s, and early 2000s: Bates, "The Getty End-User Online Searching Project in the Humanities"; Buchanan et al., "Information Seeking by Humanities Scholars"; Stephen Wiberley and William Goodrich Jones, "Patterns of Information Seeking in the Humanities," *College & Research Libraries* 50, no. 6 (1989): 638–45.
59. Curl, "Enhancing Subject and Keyword Access to Periodical Abstracts and Indexes: Possibilities and Problems"; Fidel, "Who Needs Controlled Vocabulary?"; Hildreth, "The Use and Understanding of Keyword Searching in a University Online Catalog"; Muddamalle, "Natural Language Versus Controlled Vocabulary in Information Retrieval"; Spurgin and Wildemuth, "Content Analysis"; Xi Niu and Bradley M. Hemminger, "Tactics for Information Search in a Public and an Academic Library Catalog with Faceted Interfaces," in *Proceedings of the 4th Workshop on Human-Computer Interaction & Information Retrieval* (2010): 83–86, accessed June 25, 2013, http://research.microsoft.com/en-us/um/people/ryenw/hcir2010/docs/HCIR2010Proceedings.pdf.
60. Spink and Jansen, "A Study of Web Search Trends"; David Nicholas et al., "Online Use and Information Seeking Behavior: Institutional and Subject Comparisons of UK Researchers," *Journal of Information Science* 35, no .6 (2009): 660–76, doi:10.1177/0165551509338341.
61. Markey, "Twenty-Five Years of End-User Searching"; Steve Jones et al., "A Transaction Log Analysis of a Digital Library," *International Journal on Digital Libraries* 3, no. 2 (2000): 152–69.
62. Maristella Agosti et al., "Analysing HTTP Logs of a European DL Initiative to Maximize Usage and Usability," in *Proceedings of the10th International Conference on Asian Digital Libraries (ICADL 2007)*, ed. Dion Hoe-Lian Goh et al., 35–44 (Berlin: Springer, 2007).
63. Ibid.; Moulaison, "OPAC Queries at a Medium-Sized Academic Library."