Large-Scale Digital Library User Searching: What Role Does Domain Play?

Oksana L. Zavalina Oksana.Zavalina@unt.edu
Elena V. Vassilieva Elena.Vassilieva@unt.edu

Department of Library and Information Sciences
College of Information
University of North Texas

Abstract

This poster presents findings of an exploratory comparative analysis of user search queries in three large-scale digital libraries: two domain-specific functioning in the domains of STEM education and US history, and one domain-independent (. This study measured search query lengths and frequencies, and categorized search queries into ten search categories based on content analysis. Results suggest that domain-based differences (i.e., differences in user searching between digital libraries representing different domains) are more substantial than differences in user searching between domain-specific and domain-independent digital libraries. Domain-based differences in distribution of search categories between search query length and search query frequencies are statistically significant. These findings may have implications for design and evaluation of large-scale aggregations of digitized materials.

Keywords: search queries, comparative analysis, domain analysis, transaction log analysis, content analysis

Introduction and Problem Statement

Digitization of information objects of cultural, historical, and educational value has been actively supported by US federal funding agencies such as Institute for Museum and Library Services (IMLS) and National Science Foundation (NSF). In addition, these agencies have supported large-scale digital libraries that aggregate hundreds of digital collections consisting of millions of digitized items, and provide centralized access for this wealth of information. Three of the largest digital libraries of this type in the USA have been:

- IMLS-funded Digital Collections and Content (IMLSDCC)[1] with 650 digital collections
- IMLS-funded Opening History (OH)[2] with over 1500 digital collections focusing on US history[3].
- NSF-funded National Science Digital Library (NSDL)[4] with 120 science, technology, engineering, and mathematics (STEM) education digital collections.

IMLSDCC is a domain-independent digital library aimed at a general audience and covering a wide range of subject areas and disciplines; OH and NSDL are domain-specific, i.e., are created for distinct audiences. User base for domain-independent information systems tends to include more novice

[1] http://imlsdcc.grainger.uiuc.edu
[2] http://imlsdcc.grainger.uiuc.edu/history
[3] On August 1, 2012, the interfaces of IMLSDCC and OH digital libraries were merged into one, under the IMLSDCC name.
[4] http://nsdl.org/.

Acknowledgements: This research was supported by the University of North Texas Office of Research and Economic Development and Texas Center for Digital Knowledge. Authors thank developers of the Opening History and IMLSDCC for providing access to search logs used in the study.
Zavalina, O.L., & Vassilieva, E.V. (2013). Large-scale digital library user searching: What role does domain play? *iConference 2013 Proceedings* (pp. 692-696). doi:10.9776/13322

than expert users while the audience of domain-specific information systems normally includes a higher proportion of domain experts. It was observed in numerous studies (e.g., Allen, 1991; Bates, 1972; Connaway, Johnson, & Searing, 1997; Hembrooke et al., 2005; Hsieh-Yee, 1993; Marchionini et al., 1993; Wildemuth, 2004; Zhang, Anghelescu, & Yuan, 2005) that information seeking behavior, including search query length, frequency, specificity etc., depends to a large extent on searcher's level of knowledge both on a specific search topic and in the broader subject domain.

To ensure seamless intellectual access to rich pools of digital content, that is accumulated in these large-scale digital libraries, by their intended user communities (e.g., broad general audience, historians or STEM educators), the design of their respective information systems, including description of information objects in metadata records, choice of search options etc., should be informed not only by general user tasks of finding, identifying, selecting, and obtaining information (IFLA, 2008) but also by the information needs and search approaches specific to users' respective domains, and levels of users' domain knowledge. Previous studies of web searching discovered, for example, that humanities scholars most often include in their search queries personal and geographic names, chronological and discipline terms (Bates, 1996); water quality researchers frequently use topical, geographical, and format or genre search terms, and occasionally—chemical formulas, dates, names, and URLs (Nowick & Mering, 2003); medical researchers' prevailing search query types include laboratory/test results, disease/syndrome, body part/organ/organ component, pharmacological substance, or diagnostic procedure (Natarajan et al., 2010).

Transaction log analysis—"the study of electronically recorded interactions between online information retrieval systems and the persons who search for the information" (Peters, 1993, p. 41)—and its subset, search log analysis (Jansen, Spink, & Taksa, 2008), provide a means for unobtrusive capturing and analyzing user search queries in various information systems, and empirical data to inform information system design. However, the potential of transaction log analysis has not been used to its full capacity to benefit large-scale domain-specific and domain-independent digital libraries' development: although several studies (e.g., Khoo et al., 2008; Pan, 2003; Verberne et al., 2010; Zavalina, 2007; 2012) have analyzed transaction logs of individual large-scale digital libraries—American Memory, The European Library, NSDL, IMLSDCC or OH—only two of them undertook content analysis of search queries; user interactions with different types of digital libraries aimed at different user communities have not been compared.

Methods

Domain analysis approach to information science research suggests comparative empirical studies of users in different fields as one of the important ways to study domains [20]. The study presented in this paper examines the following research questions:

- What are the domain-based differences and similarities between user searching in domain-specific large-scale digital libraries with different subject scope and audience?
- What are the differences and similarities between user searching in domain-specific and domain-independent large-scale digital libraries?

The study compared user search queries in transaction logs of three large-scale digital libraries—one domain-independent (IMLSDCC) and two domain-specific belonging to two different domains of US history and STEM education (OH and NSDL)—recorded over a period of 1 year by Google Analytics. For each of the three digital libraries, search queries were grouped with identical queries. This resulted in a total of over 17,000 unique search queries. The following measures were assessed and compared in the query-level analysis of unique search queries[5]:

- distribution of search options: basic and advanced search
- distribution of search query frequencies[6] and search query lengths[7]

[5] a query that is different from all other queries in the search log, regardless of the searcher; all identical queries are usually collapsed together to give the unique queries

[6] measured as the number of times the identical search query occurs in the search log sample

[7] measured as the number of words per search query

- central tendency measures (mean and median) of search query frequencies and search query lengths.
- frequency distributions of search categories[8] based on the entities of FRBR and FRAD bibliographic models (IFLA, 2008; 2009): *class of persons*, *concept*, *corporate body*, *ethnic group*, *event*, *family*, *object*, *person*, *place*, and *work*.

Findings and Discussion

Overall, a higher proportion of NSDL search queries (39%) used advanced search options and approaches compared to IMLSDCC (20%) and OH (15%). For example, search limits were used in over 38% of NSDL search queries but in only in 13% of IMLSDCC and slightly over 12% of OH queries. Fielded searching, where the user specifies in which metadata field(s) (e.g., author, subject, etc.) the query match should occur, was not used, due to the absence of this option, in domain-specific NSDL occurred very infrequently in domain-specific OH (0.61% of queries) but much more often in domain-independent IMLSDCC (5%). Phrase searching, where the user puts a query or its part in quotation marks, was applied seldom in all three digital libraries, but more often in domain-independent IMLSDCC (1.75 % of search queries) than in two domain-specific digital libraries (1.32% in OH and 0.79% in NSDL).

At the time of analysis, both IMLSDCC and OH were hosted on the same server and, with exception of a color scheme, had identical user interfaces. Thus, any differences in use of search options between these two digital libraries cannot be attributed to the differences in search interface. However, it is possible that interface design may have impacted differences in results between these two digital libraries and NSDL.

Search queries tended to be longer in domain-specific digital libraries (an average of 2.66 words in NSDL, and 2.32 in OH) than in domain-independent IMLSDCC (1.94 words). The median query length however was the same for all three digital libraries: 2 words per query.

The domain-specific digital library with STEM focus substantially differed from two other digital libraries in search query frequency. The average frequency was the distant first in NSDL (6.54), followed by OH (1.91) and IMLSDCC (1.60). The median search query frequency was also substantially higher in NSDL (2) and lower (1) in both OH and IMLSDCC.

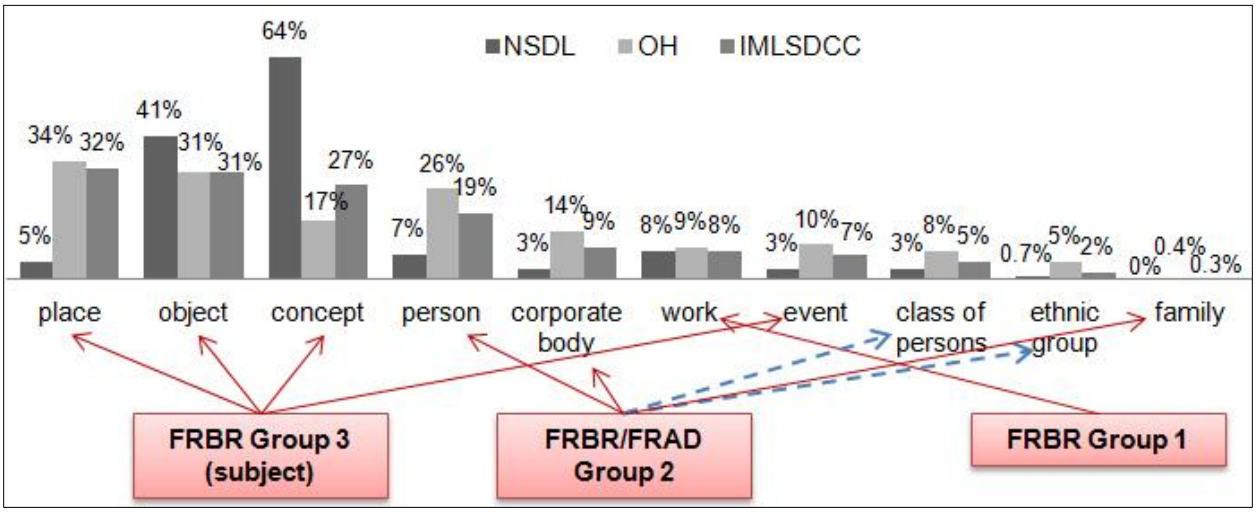

Figure 1. Distribution of search categories

[8] Intercoder reliability test was performed on 55% of unique search queries in the sample independently coded by the author and another coder; high intercoder reliability (98.44 % or Cohen's Kappa of .936) was observed.

As shown on Figure 1, there were more similarities in search category distribution between domain-specific OH and domain-independent IMLSDCC than between two domain-specific digital libraries. This, at least in part, can be explained by overlap in content of IMLSDCC and OH which likely causes some overlap in the user base. However, in all three digital libraries the top two search categories were FRBR Group 3, subject, entities: *concept* (64% of queries) and *object* (41%) in NSDL, *place* (34% in OH and 32% in IMLSDCC) and *object* (31% in both OH and IMLSDCC) in two other digital libraries. *Concept* was the 3rd most often occurring search category in IMLSDCC (27%) and the 4th in OH (17%). However, the 4th FRBR Group 3 subject entity—*event*—was observed much less than the other three (10% in OH, 7% in IMLSDCC, and 3% in NSDL). The FRBR Group 2, agent, entities—*person* and *corporate body* search categories—were observed in 26% and 14% of queries respectively in OH, but only in 19% and 9% of queries in IMLSDCC and even less (7% and 3% respectively) in NSDL. The *work* search category displayed the most similar frequencies across the three digital libraries; it was observed in 9% of OH, and 8% of IMLSDCC and NSDL queries. The *class of persons* search category occurred in 8% of OH queries, 5% on IMLSDCC queries, and 3% of NSDL queries. The *ethnic group* search category occurred in 5% of OH queries but substantially less often in the other two digital libraries: 2% in IMLSDCC and 0.7% in NSDL. Finally, *family* search category was observed very infrequently in all three digital libraries: in only 0.4% of OH queries, 0.3% of IMLS search queries, and in none of NSDL queries.

Conclusion

Preliminary findings of this exploratory comparative study of user searching in three US-based large-scale digital libraries—two domain-specific and one domain-independent—reveal some similarities as well as several notable differences. Regardless of domain, in all three digital libraries users often search by *object* and rarely initiate a phrase search. Overall, levels of use of advanced search options vary and cannot be safely attributed to either a specific domain or domain-independence. Distribution of search categories overall differs more drastically between large-scale digital libraries representing different domains—US history and STEM education—than between domain-specific and domain-independent digital libraries. Domain-specific library search queries exhibit longer query lengths and higher query frequencies than domain-independent library queries, although domain-based difference is statistically significant.

Interface design may have contributed to user searching differences between IMLSDCC and OH, which have very similar user interfaces, and NSDL. Additional investigation into this factor is needed and will be carried out by the authors of this poster.

These preliminary results suggest a hypothesis that domain-based differences in user searching are more substantial than the differences between user searching in domain-specific and domain-independent large-scale digital libraries. Studies that will test this hypothesis will need to extend the set of targets to include several domain-independent and a variety of domain-specific digital libraries representing different domains.

References

Allen, B. (1991). Cognitive research in information science: implications for design. *Annual Review of Information Science and Technology*, *26*, 3-37.

Bates, M. (1977). Factors affecting subject catalog search success. *Journal of the American Society for Information Science*, 28(3), 161-169. doi:10.1002/asi.4630280304

Bates, M. (1996). The Getty end-user online searching project in the humanities: Report no. 6: Overview and conclusions. *College and Research Libraries, 57*(6), 514-523.

Connaway, L., Johnson, D., & Searing, S. (1997). Online catalogs from the user's perspective: the use of focus group interviews. *College & Research Libraries, 58*, 403-420.

Hembrooke, H., Granka, L., Gay, G., & Liddy, E. (2005). The effects of expertise and feedback on search term selection and subsequent learning. *Journal of the American Society for Information Science and Technology, 56*(8), 861-871. doi:10.1002/asi.20180

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: domain analysis. *Journal of the American Society for Information Science, 46*(6), 400-425. doi:10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44*(3), 161-174. doi:10.1002/(SICI)1097-4571(199304)44:3<161::AID-ASI5>3.0.CO;2-8

IFLA Study Group on the Functional Requirements for Bibliographic Records. (2008). Functional Requirements for Bibliographic Records: Final report: As amended and corrected through February 2008. Retrieved from http://www.ifla.org/files/cataloguing/ frbr/frbr_2008.pdf

IFLA Working Group on Functional Requirements and Numbering of Authority Records. (2009). Functional Requirements for Authority Data - A Conceptual Model. Ed. by Glenn E. Patton. Munchen: K.G. Saur.

Jansen, B.J., Spink, A., & Taksa, I., Eds. (2008). *Handbook of research on web log analysis*. Hershey, NY: IGI Global.

Khoo, M., Pagano, J., Washington, A.L., Recker, M., Palmer, B., & Donahue, R.A. (2008). Using web metrics to analyze digital libraries. In Larsen, R. et al. (Eds.), *Proceedings of the Joint Conference on Digital Libraries* (pp. 375-384). New York: ACM. doi:10.1145/1378889.1378956

Marchionini, G., Dwiggins, S., Katz, A., & Lin, X. (1993). Information seeking in full-text end-user-oriented search systems: the roles of domain and search expertise. *Library and Information Science Research, 15*(1), 35-69.

Natarajan, K., Stein, D., Jain, S., & Elhadad, N. (2010). An analysis of clinical queries in an electronic health record search utility. *International Journal of Medical Informatics, 79*(7), 515-522. doi:10.1016/j.ijmedinf.2010.03.004

Nowick, E., & Mering, M. (2003). Comparisons between Internet users' free-text queries and controlled vocabularies: A case study in water quality. *Technical Services Quarterly, 21*(2), 15-32. doi:10.1300/J124v21n02_02 Pan, B. (2003). Capturing users' behavior in the National Science Digital Library (NSDL). Unpublished report. Ithaca, NY: Cornell University Human Computer Interaction Research Group.

Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech, 11*(2), 41-66. doi:10.1108/eb047884

Verberne, S. et al. (2010). How does the library searcher behave? A contrastive study of library search against ad-hoc search. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation*. Retrieved from http://clef2010.org/resources/proceedings/clef2010labs_submission_42.pdf

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3), 246-258. doi:10.1002/asi.10367

Zavalina, O. L. (2007). Collection-level user searches in federated digital resource environment. *Proceedings of the American Society for Information Science and Technology, 44*(1), 1-16. doi:10.1002/meet.1450440225

Zavalina, O.L. (2012). Subject access: Conceptual models, functional requirements, and empirical data. *Journal of Library Metadata, 12*(2/3), 140-163. doi:10.1080/19386389.2012.699829

Zhang, X., Anghelescu, H., & Yuan, X. (2005). Domain knowledge, search behavior, and search effectiveness of engineering and science students: An exploratory study. *Information Research, 10*(2), paper 217. Retrieved from http://informationr.net/ir/10-2/paper217.html