Exploratory and Directed Search Strategies at a Social Science Data Archive

Sara Lafia[1], A.J. Million[1], Libby Hemphill[1,2]

1. *ICPSR, University of Michigan*; 2. *UMSI, University of Michigan*

Abstract

Researchers need to be able to find, access, and use data to participate in open science. To understand how users search for research data, we analyzed textual queries issued at a large social science data archive, the Inter-university Consortium for Political and Social Research (ICPSR). We collected unique user queries from 988,475 user search sessions over four years (2012-16). Overall, we found that only 30% of site visitors entered search terms into the ICPSR website. We analyzed search strategies within these sessions by extending existing dataset search taxonomies to classify a subset of the 1,554 most popular queries. We identified five categories of commonly-issued queries: keyword-based (e.g., date, place, topic); name (e.g., study, series); identifier (e.g., study, series); author (e.g., institutional, individual); and type (e.g., file, format). While the dominant search strategy used short keywords to explore topics, directed searches for known items using study and series names were also common. We further distinguished exploratory browsing from directed search queries based on their page views, refinements, search depth, duration, and length. Directed queries were longer (i.e., they had more words), while sessions with exploratory queries had more refinements and associated page views. By comparing search interactions at ICPSR to other natural language interactions in similar web search contexts, we conclude that dataset search at ICPSR is underutilized. We envision how alternative search paradigms, such as those enabled by recommender systems, can enhance dataset search.

Keywords

Research data, information search, query log analysis, user behavior, web analytics

Introduction

Data sharing in the social sciences allows researchers to build upon the work of others. Funders require awardees to share their data to increase scientific efficiency, enhance research transparency, and promote fair access, among other benefits (National Research Council et al., 1985). However, data sharing does not guarantee discoverability or reuse by others. Data curation activities, such as creating descriptive metadata and documentation, promote data findability, accessibility, interoperability, and reuse (Levenstein & Lyle, 2018). Large-scale data archives, such as the Inter-university Consortium for Political and Social Research (ICPSR), support long-term data preservation and provide data curation services to enhance the quality of deposited data (Akmon et al., 2020). Data archives also offer search and discovery tools for data retrieval (Pienta et al., 2018). Prior research has studied the impact of curation and archiving decisions on data reuse (He & Han, 2017; Hemphill et al., 2021); however, less is known about intermediate data discovery steps, such as the specific sequences of actions that users take when seeking data (Lafia et al., 2023) and disciplinary search strategies for finding research data (Gregory et al., 2020; Kacprzak et al., 2017).

Search systems facilitate information discovery and retrieval in several ways. In particular, academic search tasks are often exploratory and complex. They emphasize learning and discovery, are often ill-defined, multi-perspective, and require browsing to support learning alongside search

(R. W. White, 2016). Academic search tasks require support for the user as they "learn" a knowledge domain (H. D. White et al., 2004). Ideally, "context-driven discovery" allows users to learn as they search and gain proficiency with a given subject (Solomon, 2002). Approaches, such as the visualizations of scientific terms (e.g., in maps), balance designer-initiated (global) and user-driven (local) conceptualizations (Börner et al., 2003). Other design considerations, such as search facets, can guide users to understand possible kinds of interactions within a system (Hearst, 2009). Well-designed systems overcome human-system communication's "vocabulary problem" (Furnas et al., 1987) by aligning user concepts with system specifications. This is important for supporting interdisciplinary research, where various disciplinary terms may describe similar phenomena (Institute of Medicine et al., 2005) across multiple levels of expertise (Hembrooke et al., 2005). Importantly, search systems must also balance exploratory and directed search tasks by allowing users to retrieve known items (R. W. White, 2016).

Search interfaces are often evaluated based on their support of user search strategies, including monitoring, file structure, search formulation, term, and idea tactics (Wilson et al., 2009). Our prior work found that users follow direct, orienting, and scenic search paths while navigating dataset searches at a large-scale social science data archive (Lafia et al., 2023). Approaches proposed to increase the accessibility of archival collections include introducing novel finding aids that support federated queries across collections and adding context to boost search relevancy within collections (Renspie et al., 2015). Archives and repositories can develop responsive systems that encourage dataset discovery and reuse by studying how users search for data.

To understand how prospective users search for curated social science research data, we analyzed 1,554 unique user queries issued across 988,475 user search sessions spanning four years (2012-16) at ICPSR. We asked: 1) What are the most common features of queries issued at a large-scale social science data archive?; and 2) What strategies do prospective users employ to search for research data? Based on our analysis, we discuss opportunities for improving data discovery and eventual reuse by supporting exploratory and directed search strategies.

Background

Query or transaction logs provide a foundation for analyzing human information behavior (HIB). While HIB models provide a theoretical basis for representing user search behavior (Bates, 1989; Marchionini, 1997; Meho & Tibbo, 2003), query logs offer detailed insights into search strategies that users employ in their everyday lives (Jiang et al., 2013). Taxonomies bridge search log analysis and theoretical models by describing high-level patterns in users' observed search behavior. For instance, log analysis has been used to summarize the intent behind commercial web searches as navigational, informational, and transactional (Broder, 2002).

Query log analyses have been applied to study commercial search engines (Kumar & Tomkins, 2010; Silverstein et al., 1999), digital libraries (Carevic et al., 2020; Jones et al., 2000), and data portals (Degbelo, 2020; Kacprzak et al., 2017). Query log analysis can be used to enhance clickthrough search performance (Joachims, 2002), infer users' information needs by analyzing search topics (Abebe et al., 2018), and appraise gaps in collections by identifying failed searches (Pienta et al., 2018). Analyses can be constrained (e.g., to a single day of searches issued on a given portal) (Herskovic et al., 2007) or cover longer durations to study changing user behaviors (e.g., characterize search as a learning process) (Eickhoff et al., 2014).

Information behavior can also be inferred from users' responses to search systems. For example, during query refinement or reformulation, users modify their search queries to retrieve more relevant results; query modification feedback can be explicitly provided by the user (e.g., clicks within a query) or implicitly derived by the system (e.g., semantic document similarity mining) (Baeza-Yates & Ribeiro-Neto, 2011, Chapter 5). Prior work has identified unique considerations for designing dataset retrieval systems (Wang et al., 2021). For example, while systems index datasets as discrete objects, users may want to perform interactions such as combination and subsetting (Chapman et al., 2019). Leading dataset search systems, such as Google's Dataset Search, rely on original, high-quality metadata for indexing (Brickley et al., 2019). Other dataset search systems, such as government data portals, encourage users to explore and browse for data rather than issue known-item searches (Kacprzak et al., 2017).

Users' dataset search strategies also vary across domains; for example, social scientists tend to trace publication references and explore survey data banks more than earth scientists and astronomers, who follow "bounded" strategies (e.g., searching by journal, location, and time) to find data (Gregory et al., 2019). Social scientists need descriptive metadata to support their search needs; these include contextual information about prior data use (e.g., evidenced in publication citations) (Faniel et al., 2019). However, existing systems do not tend to include explicit, contextual information about how data have been reused by others or curated, for example, in search indexes (Sun & Khoo, 2017). Generally, users' information needs are often far more detailed and expressive than the dataset search queries that they issue (Papenmeier et al., 2021). In this study, we analyze query logs to develop a baseline understanding of users' expressed information needs and search behaviors when seeking social science data.

Methods

We analyzed user search queries at the Inter-university Consortium for Political and Social Research (ICPSR), a large social science data archive. Specifically, we used Google Analytics (GA) to track user queries issued through the ICPSR website's search box (i.e., "site searches") across research metadata, variables, data-related publications, and documentation about ICPSR. ICPSR holdings include over 250,000 data files in 10,000 public use studies and 295 series. GA omits searches performed by ICPSR staff based on IP addresses. We only considered the 30% of sessions (988,475/3,434,937) that included site search interactions. From these sessions, we collected all unique user queries issued across user search sessions from 9/1/2012-9/1/2016. We selected the period for our analysis based on the stability of ICPSR's website design and the consistency of available GA data; site changes to GA since 2016 made more recent data challenging to analyze.

Data processing

We processed website queries using Open Refine, a data-cleaning tool. We removed whitespace, normalized text to lowercase, removed punctuation, transformed plural to singular forms, checked spelling, and clustered similar query strings. This approach matched queries that contained the same words in different orders ("crime mental illness", "mental illness crime") and deduplicated nearly identical queries (e.g.,) by merging them into a single entry. We did not, however, merge name variants or synonyms ("National Longitudinal Study of Adolescent Health", "Add Health", "NLS") since these reflected diverse search strategies. By applying these rules, we merged a total of 900 queries.

Query classification

To classify queries, we first aligned and extended existing categories of data-specific queries proposed by Kacprzak et al. (2017) and Pienta et al. (2018). A summary of the categories and the rules we used to code the ICPSR queries is provided in **Table 1**. The prior analysis by Pienta (2018) found that users relied on exploratory keywords – indicating subjects, locations, and timeframes – along with directed terms corresponding to known items – such as studies, series, and author names – when searching for datasets. We used the intersection of these categories (keyword; name; number; author; and format) to code the 1,554 most popular queries in our sample, which were present in more than 57% of all search sessions (562,723/988,475) and which users searched for more than 100 times across all user sessions in our sample. In cases where queries were ambiguous, we issued the same query to ICPSR's search box and reviewed the search results to decide. We assigned one category to each query. Most queries that contained multiple categories (e.g., "chinese household income 2002") referred to study or series names; however, in ambiguous cases (e.g., "english second language in texas"), we assigned the category that had more words or that appeared first in the query string. We used ICPSR's controlled vocabulary (*ICPSR Thesaurus*, 2023) to disambiguate places, topics, authors, and organizations. To interpret the coded queries, we then grouped them into one of two search task categories: exploratory or directed (R. W. White & Roth, 2009). Exploratory searches facilitate browsing and are not directed to retrieve known items, whereas directed searches indicate a specific item that the user is seeking.

Table 1. Query classification scheme and alignment with prior categories

Category	Rules	Related category from Pienta et al. (2018)	Related category from Kacprzak et al. (2017)
Keyword - Place, Date, Topic (*Exploratory*)	Includes a geographic place name, time, or concept.	Keyword or phrase (e.g., "diabetes")	Location (name of city, town, geographical area)
			Time frame (years, months, weekday)
Format - Type (*Exploratory*)	Uses the name of a known file format or analysis method.		File and dataset type (.csv, .pdf, html, table)
Name - Study, Series; Number - Study, Series (*Directed*)	Uses a number in ICPSR's study or series number range (not a year or other identifier).	Study name (e.g., "ICPSR 2896")	Numbers
		Named serial collection (e.g., "NSDUH")	Abbreviations (acronyms - from controlled list or manually verified)
Author - Institutional, Individual (*Directed*)	Includes an author's full or last name, or uses the name of an organization.	Author/principal investigator name (e.g., "Lillard")	

Feature selection

To characterize groups of queries classified in our analysis, we selected features from Google Analytics described in **Table 2**. We chose these query-level features based on prior findings by Kathuria et al. (2010), who defined query intent using query-level features, such as query length and reformulation strategy. We also based our feature selections on work by Sharifpour et al. (2022), who proposed distinct user groups by performing hierarchical clustering on query logs (2022). Both prior studies applied unsupervised clustering to the features they identified to discover user groups across sessions. The features we selected (*Google Analytics*, 2023) were: results page views per search (i.e., the number of items a user looked at after searching); percent search refinements (i.e., the share of sessions where a user adjusted or reformulated their search); average search depth (i.e., number of pages clicked on following a search); time after search (i.e., amount of time spent in the session after a search); and query length (e.g., number of words in the query).

Table 2. Features extracted from Google Analytics to characterize queries

Feature	**Definition from Google Analytics**	**Related category from Kathuria et al.** (Kathuria et al., 2010)	**Related category from Sharifpour et al.** (Sharifpour et al., 2022)
Results page views per search	Views of search result pages divided by total unique searches	Results viewed	Page views
Percent search refinements	Repeated searches using another term divided by views of search result pages	Query reformulation	
Average search depth	Average number of pages viewed after performing a search	Total click-throughs	
Time after search	Amount of time in seconds users spend on site after performing a search		Total time spent
Query length	Number of terms contained in a particular query	Number of query terms	Number of unique query terms

Results

Users searched with short, unique phrases

To characterize the queries, we measured their lengths and checked if they contained interrogative terms (e.g., “who”). All queries were shorter than two words on average, meaning that most users entered a single word or phrase. Exploratory searches, which facilitated browsing and were not directed to retrieve known items (R. W. White & Roth, 2009), were shorter on average than directed searches for known items, such as the names of social science studies (1.5 words versus 2.7 words). In terms of query formulation, only four queries contained one or more interrogative keywords proposed by Bendersky and Croft (2009) suggesting a question (e.g., the word “do” indicates the question in: “'Do children of asian immigrants speak english in the home more often than children of latino immigrants?'”). Users also searched with many distinct query terms, illustrated in a long-tailed distribution (**Figure 1**). For example, the single most popular query in our sample (ICPSR study number "21600") was issued 10,148 times, while many more queries ("surveillance", "infertility", "religious attitudes") were issued 100 times each.

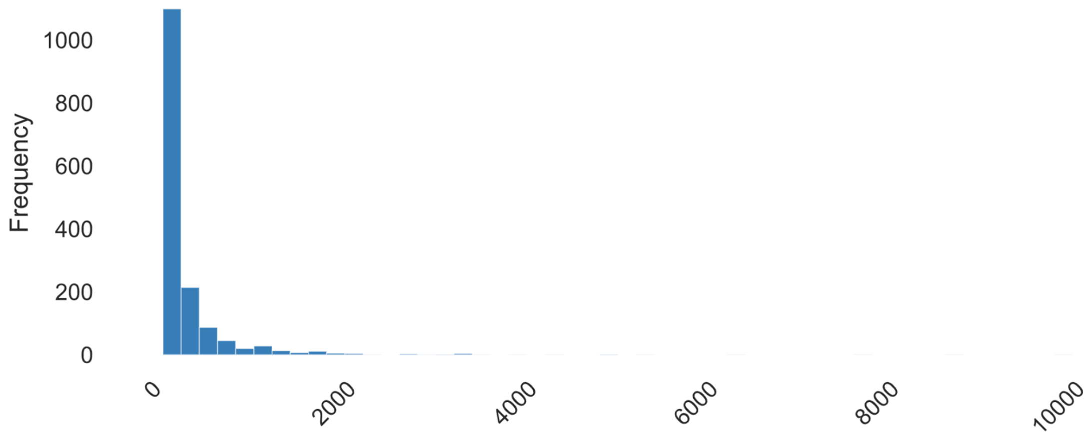

Figure 1. Histogram with fixed size bins (bins=50) of unique query terms (2012-2016)

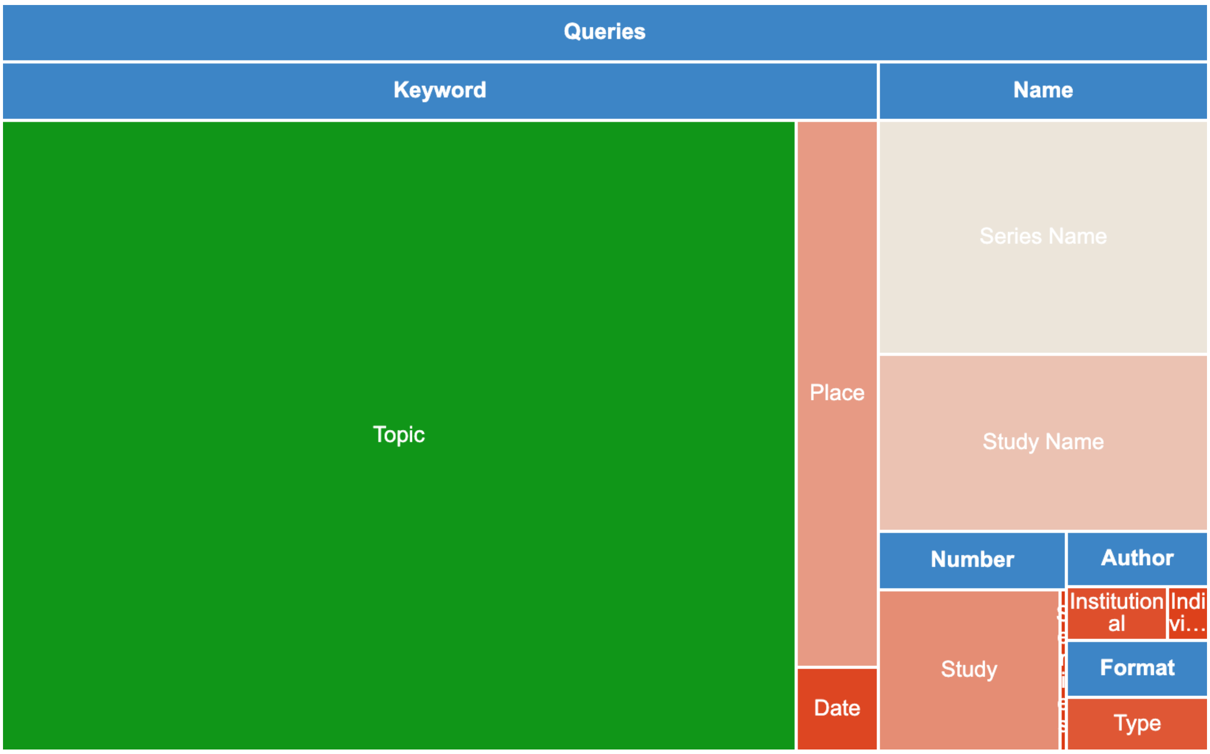

Figure 2. Treemap of labeled queries shows that search by topic and name were most common

Searches were dominated by topics and names

We classified more than 66% (1,030/1,554) of the queries as "Keyword (Topic)", meaning that the user entered one or more social science subject terms into the site search box. The second largest category of queries used part of or the full "Name (Series, Study)" of a social science study or series. Searches by "Number (Series, Study)", "Author (Institutional, Individual)", and "Format (Type)" were the least common kinds s (**Figure 2**).

Exploratory searches included more refinements and page views

Most queries were exploratory (73%), which included keyword and format-based searches, while directed queries (27%) included study and series names, numbers, and authors. We summarized the distributions for each feature across the exploratory and directed query groups (**Figure 3**). We observed that sessions had a similar search duration (in seconds) and search depth (by page views) across query types. However, directed queries tended to be longer than exploratory ones. Sessions with exploratory queries included more refinements, meaning that users edited and re-issued search terms more often; exploratory sessions also included more result page views than their directed search counterparts, suggesting that they enabled more browsing and navigation behaviors.

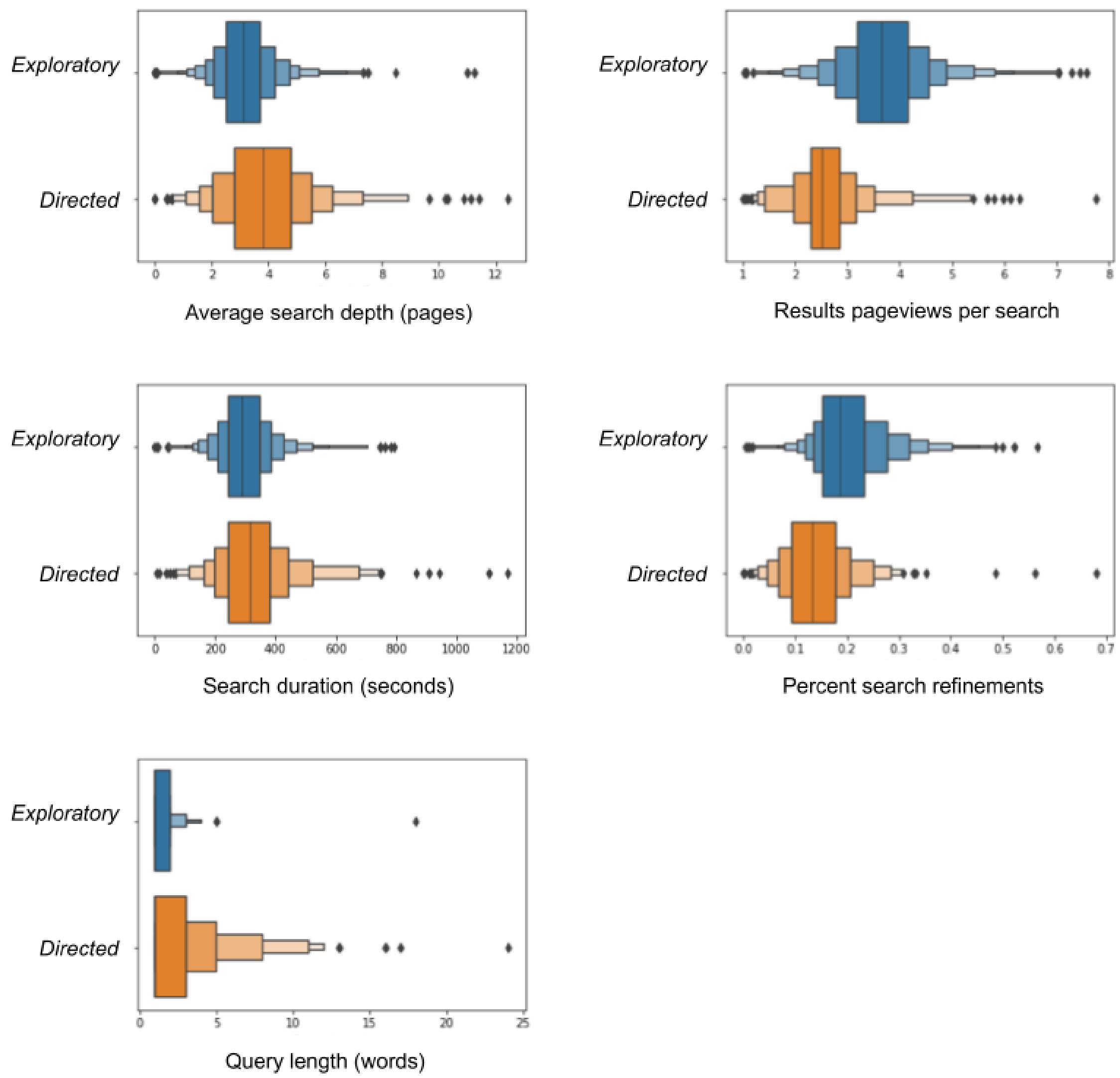

Figure 3. Enhanced boxplots show analytics features of exploratory and directed queries

Discussion

Query log analysis provides insights into users' data discovery behavior. Our findings align with prior studies of information seeking, which differentiate between exploratory and directed search tasks (Bates, 1989; Marchionini, 2006; R. W. White & Roth, 2009). Keyword-based queries that use dates, places, or topics to search suggest that users do not have known items in mind. At the same time, searches for particular study or series names, numbers, and authors are better characterized as “information lookup” tasks, in which users expect to retrieve a specific item (Buckland, 1979). While users who issued exploratory queries were able to navigate ICPSR’s website and take additional actions, such as expanding and refining their searches, users may benefit from more explicit support for query reformation; ICPSR offers search facets, but their integration with users’ queries could be enhanced (Hearst, 2006). In addition, the popularity of known-item searches suggests that there may be a need for additional functions, such as search “bookmarks”, to help users store and take shortcuts to retrieve their previous queries and search results (Aula et al., 2005).

The brevity of queries, and the lack of questions entered in ICPSR’s site search, may indicate room for improvement to the user experience. Kacprzak et al. (2017) for example, found that most queries entered into open government data portals were a single word in length. Shorter queries may indicate that users are not confident in the capabilities of search engines to interpret intent in complex prompts and return relevant results (Jansen & Spink, 2006). User behavior reflected in ICPSR’s search logs suggests that site search is generally treated as an entry point for data exploration. Dataset search is still nascent, but as the technology matures, the complexity and length of natural language queries may resemble other web-based search interactions (Taghavi et al., 2012).

Given that keyword-based exploration by topic is the dominant category of search interaction at ICPSR (i.e., 66% of the queries we coded), we plan to extend our analysis by investigating the relationships between the topics of user queries and search results. We are interested in exploring the potential to support query expansion with word embeddings. Developing a finer set of categories such as social science methods (“factor analysis”), populations of interest (“homeless youth”), and historical events (“hurricane katrina”) would allow for detailed search refinements. **Figure 4** illustrates the prevalence of specific words and phrases, which may help predict query intent. Prior studies of search behavior at ICPSR found evidence of the stability of search topics across time (Pienta et al., 2018). Thus, detecting significant shifts in topic popularity may also be informative for supporting data search and discovery.

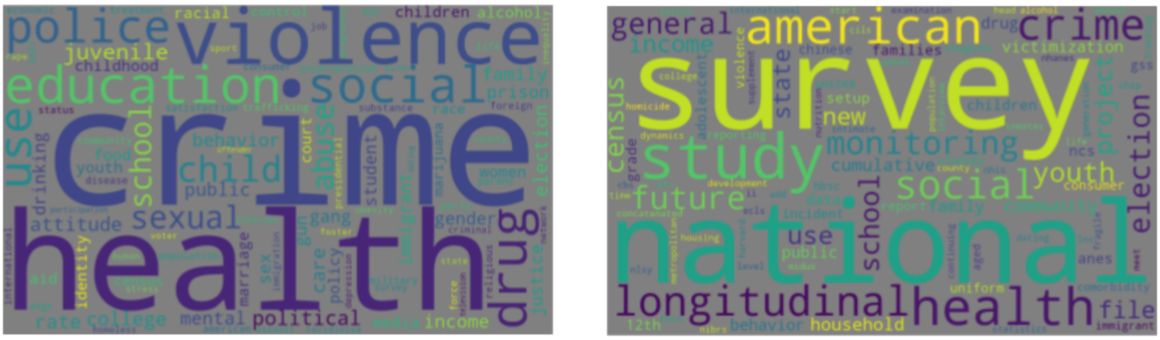

Figure 4. Wordclouds of queries coded as (a) topics (keywords) and (b) names (study, series)

In terms of the limitations of our study, we were restricted to using queries issued between 2012 and 2016. We also focused on the relationship between the most popular queries and features identified in prior studies. This meant that less popular queries, which may not be well-supported by ICPSR’s search system, were omitted from our analysis. To code the queries, we developed a scheme where a single label neatly described most queries; however, we encountered queries that would be better represented by multiple labels (e.g., “chicago homicide” includes place and topic keywords). In some cases, it was also unclear whether a search (e.g., “are you happy”) referred to a known variable label or was an exploratory topic. We note that we are also limited in the inferences we can draw from queries alone, which indicate how users approach search, but do not describe what exactly users are evaluating or their internal cognitive states.

Conclusion

By charting the sequences of actions users take to discover research data (Lafia et al., 2023), and describing directed and exploratory data search strategies, we are better positioned to propose responsive search tools that support research data discovery and encourage data reuse. While current search methods support exploratory browsing and known-item retrieval for research data, the ability to explore semantically related datasets still needs to be improved. In addition, site search at ICPSR is underutilized, and the ways that users query the system are limited. Directed searches were longer and more descriptive than exploratory searches. Compared to directed searches, exploratory searches required users to expend more effort to refine their queries and review results. Future work will explore approaches that balance search efficiency with data exploration to support the serendipitous discovery of research data available in archives, such as ICPSR.

Acknowledgments

We thank Aalap Doshi, Lara Cooper, Sai Sandeep Reddy Bedadala, and the User Experience Engineering team at ICPSR.

Award Information

This material is based upon work supported by the National Science Foundation under grant 2121789.

References

Abebe, R., Hill, S., Vaughan, J. W., Small, P. M., & Andrew Schwartz, H. (2018). Using Search Queries to Understand Health Information Needs in Africa. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1806.05740

Akmon, D., Lafia, S., Thomer, A., Hemphill, L., Pienta, A., Yakel, E., Bleckley, D., & Tyler, A. (2020). *Measuring and Improving the Efficacy of Curation Activities in Data Archives*. https://deepblue.lib.umich.edu/handle/2027.42/163501

Aula, A., Jhaveri, N., & Käki, M. (2005). Information search and re-access strategies of experienced web users. *Proceedings of the 14th International Conference on World Wide Web*, 583–592. https://doi.org/10.1145/1060745.1060831

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, *13*(5), 407–424. https://doi.org/10.1108/eb024320

Bendersky, M., & Croft, W. B. (2009). Analysis of long queries in a large scale search log. *Proceedings of the 2009 Workshop on Web Search Click Data*, 8–14. https://doi.org/10.1145/1507509.1507511

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255. https://doi.org/10.1002/aris.1440370106

Brickley, D., Burgess, M., & Noy, N. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. *The World Wide Web Conference on - WWW '19*, 1365–1375. https://doi.org/10.1145/3308558.3313685

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552

Buckland, M. K. (1979). On types of search and the allocation of library resources. *Journal of the American Society for Information Science. American Society for Information Science*, *30*(3), 143–147. https://doi.org/10.1002/asi.4630300305

Carevic, Z., Roy, D., & Mayr, P. (2020). Characteristics of Dataset Retrieval Sessions: Experiences from a Real-Life Digital Library. *Digital Libraries for Open Knowledge*, 185–193. https://doi.org/10.1007/978-3-030-54956-5_14

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., & Groth, P. (2019). Dataset search: a survey. *The VLDB Journal: Very Large Data Bases: A Publication of the VLDB Endowment*. https://doi.org/10.1007/s00778-019-00564-x

Degbelo, A. (2020). Open Data User Needs: A Preliminary Synthesis. *Companion Proceedings of the Web Conference 2020*, 834–839. https://doi.org/10.1145/3366424.3386586

Eickhoff, C., Teevan, J., White, R., & Dumais, S. (2014). Lessons from the journey: a query log analysis of within-session learning. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 223–232. https://doi.org/10.1145/2556195.2556217

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of*

Documentation, *75*(6), 1274–1297. https://doi.org/10.1108/JD-08-2018-0133

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971. https://doi.org/10.1145/32206.32212

Google Analytics. (2023). https://support.google.com/analytics/

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *Journal of the Association for Information Science and Technology*, *70*(5), 419–432. https://doi.org/10.1002/asi.24165

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*.

Hearst, M. (2006). Design recommendations for hierarchical faceted search interfaces. *ACM SIGIR Workshop on Faceted Search*, 1–5. https://flamenco.berkeley.edu/papers/faceted-workshop06.pdf

Hearst, M. (2009). *Search User Interfaces*. Cambridge University Press.

He, L., & Han, Z. (2017). Do usage counts of scientific data make sense? An investigation of the Dryad repository. *Library Hi Tech*, *35*(2), 332–342. https://doi.org/10.1108/LHT-12-2016-0158

Hembrooke, H. A., Granka, L. A., & Gay, G. K. (2005). The effects of expertise and feedback on search term selection and subsequent learning. *Journal of the American Society for Information Science and Technology*. https://doi.org/10.1002/asi.20180

Hemphill, L., Pienta, A., Lafia, S., Akmon, D., & Bleckley, D. (2021). How do properties of data, their curation, and their funding relate to reuse? *Journal of the American Society for Information Science and Technology*, *73*(10), 1432–1444. https://doi.org/10.1002/asi.24646

Herskovic, J. R., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A day in the life of PubMed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association: JAMIA*, *14*(2), 212–220. https://doi.org/10.1197/jamia.M2191

ICPSR Thesaurus. (2023). https://www.icpsr.umich.edu/web/ICPSR/thesaurus

Institute of Medicine, National Academy of Engineering, National Academy of Sciences, Committee on Science, Engineering, and Public Policy, & Committee on Facilitating Interdisciplinary Research. (2005). *Facilitating Interdisciplinary Research*. National Academies Press.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248–263. https://doi.org/10.1016/j.ipm.2004.10.007

Jiang, D., Pei, J., & Li, H. (2013). Mining search and browse logs for web search: A Survey. *ACM Trans. Intell. Syst. Technol.*, *4*(4), 1–37. https://doi.org/10.1145/2508037.2508038

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142. https://doi.org/10.1145/775047.775067

Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. (2000). A transaction log analysis of a digital library. *International Journal on Digital Libraries*, *3*(2), 152–169. https://doi.org/10.1007/s007999900022

Kacprzak, E., Koesten, L. M., Ibáñez, L.-D., Simperl, E., & Tennison, J. (2017). A query log analysis of dataset search. In *Lecture Notes in Computer Science* (pp. 429–436). Springer International Publishing. https://doi.org/10.1007/978-3-319-60131-1_29

Kathuria, A., Jansen, B. J., Hafernik, C., & Spink, A. (2010). Classifying the user intent of web queries using k-means clustering. *Internet Research*, *20*(5), 563–581. https://doi.org/10.1108/10662241011084112

Kumar, R., & Tomkins, A. (2010). A characterization of online browsing behavior. *Proceedings of the 19th International Conference on World Wide Web*, 561–570. https://doi.org/10.1145/1772690.1772748

Lafia, S., Million, A. J., & Hemphill, L. (2023). Direct, Orienting, and Scenic Paths: How Users Navigate Search in a Research Data Archive. *Proceedings of the ACM on Human Information Interaction and Retrieval (CHIIR)*.

Levenstein, M. C., & Lyle, J. A. (2018). Data: Sharing Is Caring. *Advances in Methods and Practices in Psychological Science*, *1*(1), 95–103.

Marchionini, G. (1997). *Information Seeking in Electronic Environments*. Cambridge University Press.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41. https://doi.org/10.1145/1121949.1121979

Meho, L. I., & Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society for Information Science and Technology*. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.10244

National Research Council, Division of Behavioral and Social Sciences and Education, Commission on Behavioral and Social Sciences and Education, & Committee on National Statistics. (1985). *Sharing Research Data*. National Academies Press.

Papenmeier, A., Krämer, T., Friedrich, T., Hienert, D., & Kern, D. (2021). Genuine information needs of social scientists looking for data. *Proceedings of the Association for Information Science and Technology*, *58*(1), 292–302. https://doi.org/10.1002/pra2.457

Pienta, A. M., Akmon, D., Noble, J., Hoelter, L., & Jekielek, S. (2018). A Data-Driven Approach to Appraisal and Selection at a Domain Data Repository. *International Journal of Digital Curation*, *12*(2). https://doi.org/10.2218/ijdc.v12i2.500

Renspie, M., Shepard, L., & Childress, E. (2015). *Making Archival and Special Collections More Accessible*. OCLC Research.

Sharifpour, R., Wu, M., & Zhang, X. (2022). Large-scale analysis of query logs to profile users for dataset search. *Journal of Documentation*, *79*(1), 66–85. https://doi.org/10.1108/JD-12-2021-0245

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, *33*(1), 6–12. https://doi.org/10.1145/331403.331405

Solomon, P. (2002). Discovering information in context. *Annual Review of Information Science and Technology*, *36*(1), 229–264. https://doi.org/10.1002/aris.1440360106

Sun, G., & Khoo, C. S. G. (2017). Social science research data curation: issues of reuse. *Libellarium: Journal for the Research of Writing, Books, and Cultural Heritage Institutions*, *9*(2). https://doi.org/10.15291/libellarium.v9i2.291

Taghavi, M., Patel, A., Schmidt, N., Wills, C., & Tew, Y. (2012). An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces*, *34*(1), 162–170. https://doi.org/10.1016/j.csi.2011.07.001

Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, *72*(9), 1161–1182. https://doi.org/10.1002/asi.24483

White, H. D., Lin, X., Buzydlowski, J. W., & Chen, C. (2004). User-controlled mapping of significant literatures. *Proceedings of the National Academy of Sciences*, *101*(Supplement 1), 5297–5302. https://doi.org/10.1073/pnas.0307630100

White, R. W. (2016). Exploration, Complexity, and Discovery. In *Interactions with Search Systems* (pp. 201–230). Cambridge University Press. https://doi.org/10.1017/CBO9781139525305.009

White, R. W., & Roth, R. A. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts Retrieval and Services*, *1*(1), 1–98. https://doi.org/10.2200/s00174ed1v01y200901icr003

Wilson, M. L., Schraefel, M. C., & White, R. W. (2009). Evaluating advanced search interfaces using established information-seeking models. *Journal of the American Society for Information Science and Technology*, *60*(7), 1407–1422. https://doi.org/10.1002/asi.21080

Wu, M., Psomopoulos, F., Khalsa, S. J., & de Waard, A. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, *18*. https://doi.org/10.5334/dsj-2019-003

Zhang, G., Wang, J., Liu, J., & Pan, Y. (2021). Relationship between the metadata and relevance criteria of scientific data. *Data Science Journal*, *20*(1), 5. https://doi.org/10.5334/dsj-2021-005