Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery

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

[1.] ICPSR, University of Michigan
[2.] School of Information, University of Michigan

Author Note

A.J. Million, https://orcid.org/0000-0002-8909-153X

Jeremy York is now affiliated with the University of Michigan Libraries, https://orcid.org/0000-0001-8225-9291

Sara Lafia is now affiliated with NORC at the University of Chicago, https://orcid.org/0000-0002-5896-7295

Libby Hemphill, https://orcid.org/0000-0002-3793-7281

Correspondence concerning this article should be addressed to A.J. Million, Inter-university Consortium for Political and Social Research 330 Packard St. Ann Arbor, MI 48104. Email: millioaj@umich.edu

Abstract

Many theories of human information behavior (HIB) assume that information objects are in text document format. This paper argues that four important HIB theories are insufficient for describing users' search strategies for research data because of differences between data and text documents. We first review and compare four HIB theories: Bates' theory of *berrypicking*, Marchionni's theory of *electronic information search*, Dervin's theory of *sense-making*, and Meho and Tibbo's model of *social scientist information-seeking*. All four theories assume that information-seekers search for text documents. We compare these theories to user search behavior by analyzing the Inter-university Consortium for Political and Social Research's (ICPSR) search logs. Users took direct, orienting, and scenic paths when searching for research data. We interviewed ICPSR data users (n=20), and they said they needed dataset documentation and contextual information absent from text documents to find data, which suggested ongoing sense-making. However, sense-making alone does not explain the information-seeking behavior we observed. What mattered most to secondary data discovery were information attributes determined by the type of objects that users sought (i.e., data, not documents). We conclude by suggesting an alternative frame for building data discovery tools.

Keywords

Data, information retrival, systems design

Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery

Introduction

Theories of human-information behavior (HIB) seek to explain user behavior and inform user-centered system development. Influential HIB theories, especially those focused on information-seeking, emerged in response to criticism of a query-centered view of information retrieval (IR) (e.g., Bates, 1989) and the parallel emergence of user-centered design. Seminal theories of HIB that developed in the 1980s and 1990s described information-seekers searching for text that was available online. Theories such as Marchionni's (1997) portrayal of information-seeking in online environments and Ellis's (1989) model of information-seeking by social scientists helped inform IR tool development. Although some of the most influential HIB theories are several decades old, they continue to inform our understanding of information discovery. However, these theories have limits, and we argue their applicability to data discovery and reuse is one such limit.

Scientists are increasingly aware of the importance of data sharing. Despite this shift, information scientists do not fully understand how scientists discover research data and the user-specific discovery considerations that inform secondary data reuse. This knowledge gap hinders the creation of search tools.

To assess the extent to which theory captures how social scientists search for data, we studied behavior at a leading data archive, the Inter-university Consortium for Political and Social Research (ICPSR). ICPSR was founded in 1962 and provides access to over 10,000 social science studies, including over 250,000 data files in more than 17,500 data collections. Researchers, research centers, funders, and governmental agencies contribute datasets to ICPSR's collection. ICPSR maintains a web-based catalog that supports faceted data searches based on metadata standards and controlled vocabularies. ICPSR also maintains a Bibliography of Data-related Literature and a variable database. The consortium curates data to enhance reuse, creates codebooks, and provides restricted data access using one of three modalities. ICPSR provides these resources on a website with a shared, uniform interface.

We argue that distinguishing between data and documents clarifies why emerging research (Gregory & Koesten, 2022; Koesten et al., 2017; Gregory, 2021) finds that information-seeking theories inadequately support data discovery tool design. Building tools for data discovery requires more than considering user behavior. Rather, we find that information about and within data (e.g., metadata and contextual information) is foundational to data discovery and reuse, because the objects users seek determine their most relevant representative and contextual information. Thus, data-specific attributes are important to consider when building IR systems. Still, the HIB theories we review do not make this point, nor is it made in other studies of data discovery. We conclude by arguing that a mid-level data representation model can support future systems design and help create new HIB theories.

Literature review

Definitions of data and documents

The concepts defined in **Table 1**, including "information," "data," "document," and "text," have long been contested in information science (Tredinnick, 2006). Scholars have resolved perennial questions about differences between "data" and "documents" based on the affordances of each class of information object. A comparison of dataset definitions (Renear et al., 2010) found four common functional features: grouping (e.g., a collection), content (e.g., observations), relatedness (e.g., to a subject), and purpose (e.g., a representation). For instance, scientific datasets need persistent, unambiguous identities for researchers to cite them (Wynholds, 2011). In this paper, we operationalize data as information objects (see Buckland, 1991).

Table 1. Key terms and definitions.

Term	Definition	References
Data	Data can include geospatial coordinates, numerical values, and measurements, but also literature corpora, images, or physical samples—any of which may be used to provide evidence of phenomena or to serve as a subject of analysis. Data can be organized in different ways (i.e., in spreadsheets, as networks or graphs, or as collections of related artifacts).	(Borgman et al., 2015; Dourish & Gómez Cruz, 2018; Gregory & Koesten, 2022; Munzner, 2014, as cited in Gregory & Koesten, 2022)
Information	A term used attributively for objects, such as data and documents, that are referred to as such because they are regarded as being informative, imparting knowledge, or communicating information.	(Buckland, 1991)
Quantitative data	"Information-as-a-thing" containing highly structured numerical, categorical, and ordinal variables that comprise files originating from social research methodologies or administrative records.	(Informed by Buckland, 1991; Buckland, 2018; ICPSR, n.d.)
Textual document	An information object containing written text organized into an artifact such as a book, academic journal article, webpage, or gray literature.	(Buckland, 1991; 2018)

Questions surrounding the identity of texts are non-trivial. Documents carry meaning because they organize evidence and convey significance to individuals, sometimes referred to as "relevance" in information retrieval (Buckland, 1997). In his exploration of the distinction between "data" and "documents," Furner (2016) found that "datasets are made up of documents...” and that the dataset is a “species of document” (p. 288). In this view, textual documents consist of texts, while datasets consist of numerical documents. This distinction is important because the representation and organization of textual documents in IR systems impact users’ ability to access and interact with information objects (DeRose et al., 1997). Treating quantitative data and textual documents as identical assumes that each is represented and organized the same—this may hinder the creation of user-centered IR systems.

Critiques of information retrieval systems

IR systems allow users to express their information needs as queries that systems subsequently use to return search results. Searches retrieve information including, but not limited to, documents, content in documents, images, metadata, moving images, sound, and databases (Luk, 2022). Web-based search engines are the most visible type of IR system, and they combat information overload by ranking results using measures of relevance.

Critiques of IR systems are not new. In 1989, Bates asked us why traditional IR systems required users to represent their needs in structured queries. “Why cannot the system make it possible for the searcher to express [... their needs] as they would ordinarily, instead of in an artificial query representation for the system’s consumption” (p. 197)? This question underscores an important realization: how humans interact with IR systems shapes performance. Historically, default IR models did not account for user behavior (Robertson, 1977), so IR systems were technically sound but challenging to use. Responding to critiques like the one presented by Bates (1989), researchers began theorizing how users search for and behave using information to improve system usability (Norman & Draper, 1986).

Users search for information differently, and these differences depend on user needs, context, psychology, and other considerations (Fisher et al., 2005). Recognizing this, the Information-seeking in Context Conference (ISIC, n.d.) provides a venue for researchers to present papers about information activities “going beyond a sole focus on [... technology]” (para 2). Researchers also examine information needs, recognizing that these needs motivate information-seeking behavior (Wilson, 1981). Sometimes, information acquisition is unintentional and a product of serendipity (Erdelez, 1999). Scholars also acknowledge that behavior varies by the type of information with which individuals interact (e.g., Albertson, 2015; Lavranos et al., 2015). Discussing modern information retrieval systems, Chapman et al. (2019) note how they are optimized to return tuples (a basic unit of data in a relational database), documents, and web pages but not datasets.

Data discovery and reuse

Data sharing and reuse have a long history (Sherif, 2018), with continued growth in the amount of data available for reuse, the number of reuse studies conducted, and literature examining dataset reuse. Growth during the past decade was caused by an increased awareness among scientists about the importance of data sharing, which facilitates the creation of new knowledge, increases trust in science, and maximizes funders’ return on investment (Borgman, 2012). The well-known FAIR principles (“Findable,” “Accessible,” “Interoperable,” and “Reusable”) note, “The first step in (re)using data is to find them” (para. 3). Thus, data discovery is vital to sharing, stewarding, and reusing data. However, data discovery can be challenging to define.

A substantial number of data discovery systems exist, making it surprising that data discovery is just now receiving close attention from information scientists. National governments operate data portals (e.g., www.data.gov) to promote data reuse. Google maintains a data-specific search tool, while numerous institutional archives and subject-specific repositories exist, such as DataONE (www.dataone.org).

The elusive nature of data discovery could be due to the fact that it encompasses an ongoing cycle of search and evaluation. Koesten et al.'s (2021) study of sense-making in data reuse revealed that researchers engage in multiple patterned individual and collaborative activities to make sense of the data they find. However, we also know from the literature that sense-making happens in stages, and information-seeking and data-related interactions are iterative. Data discovery, thus, involves a sequence of sense-making activities deeply interwoven with considerations about data reuse. Data reuse, in turn, is discussed in a wide variety of literature, including research on and descriptions of tools and infrastructure. Because discovery is often embedded within other data-related activities, Gregory et al. (2019) conclude that "Information documenting data retrieval behaviors are buried across disciplines and data-related literature and is not easy to identify" (p. 420).

Nevertheless, research has uncovered consistent findings. Kriesberg et al. (2013) found that learning to reuse data—including finding data—is an acculturation process for novice researchers who learn the practices and norms of their community. Most importantly, they discovered that novice researchers decide what data to reuse in concert with mentors and more senior researchers. Their finding is supported by others (Zimmerman, 2007; Gregory et al., 2020; York, 2022), who argue that social and other scientists discover data through professional networks or reuse data with which they are already familiar. These studies also found literature is essential for researchers to discover data. Aside from curating and providing access to data, "intermediaries" like libraries promote data discovery by making it findable (Yoon et al., 2018).

User-centered data retrieval

User-centered data retrieval is an emerging area of study, motivated by an effort to create tools that place users at the center of the software development process. However, evidence suggests that building effective, usable data retrieval systems is challenging. Krämer et al. (2021) find that users often rely on literature searches to identify relevant data, suggesting that tighter integration of search systems for datasets and documents might improve user search experiences. However, variation in research practices (see below) makes it challenging to develop user-centered IR tools based on user profiles. Existing HIB theory also appears inadequate to provide software developers and designers with the guidance they need.

In a paper reporting international survey results, Gregory et al. (2020) discuss variations in how researchers use data. They conclude that behavioral differences make creating user-centered IR tools based on profiles difficult because, according to their results, a diversity of data needs, uses, sources, and search practices "appears to be the rule" (p. 40). Instead, Gregory and colleagues call for developers to build IR systems that support discrete data-based tasks *in situ*.

These purpose-driven tasks include using data for a study; preparing a new project or proposal; teaching; generating new ideas; experimenting with methods or techniques; identifying trends or making predictions; comparing datasets to find patterns; modeling algorithms; creating summaries, visualizations, or tools; integrating data to create new datasets; benchmarking; and calibrating models (p. 21). Each task provides "an entry point for design" (p. 46). Partially because data discovery requires attention to such considerations, Gregory and Koesten (2022) argue it is not like searching for literature (p. 3).

Similarly, after reviewing data discovery practices in five disciplines, Gregory et al. (2021) draw another conclusion with implications for building data discovery tools. They argue "a theoretical framework based on information retrieval alone is insufficient for deeply understanding practices of data discovery" (p. 249). Koesten et al. (2017) support this finding, reporting on a study to inform systems design. They conclude existing user-centric information-seeking theories "are a reasonable starting point" for IR tool development but are ultimately inadequate. They note "to be truly useful, [information-seeking models…] need to consider the specific search and sense-making activities people carry out when working with structured data" (p. 1,279). Koesten and colleagues then forward a taxonomy of data-centric tasks and a five-pillar iterative model of how researchers work with structured data (p. 1,284). They also suggest that software developers use their taxonomy and model as touchpoints. However, they do not say what information about data users need to complete searches, nor do they differentiate between data types.

Human information behavior models

Here, we review four influential HIB theories:

1. berrypicking (Bates, 1989);
2. electronic information-seeking (Marchionini, 1997);
3. sense-making (Dervin, 1983); and
4. social scientist information-seeking (Meho & Tibbo, 2003).

Despite their age, these theories remain relevant to theory and practice today (especially IR tool design), collectively receiving 3,810 citations since 2013.[i] We present these theories because they characterize the information-seeking literature. Although we do not employ formal inclusion criteria, we exclude theories without a focus on information-seeking. We also exclude redundant or overlapping theories (e.g., Kuhlthau, 1993, because her model is holistic, and the concept of *uncertainty* is explained by Dervin's idea of a *conceptual gap*). Nevertheless, the four we review shed light on the state-of-the-literature with respect to models supporting IR tool development.

Bates' (1989) theory of *berrypicking* first emerged to criticize traditional IR systems. Observing a limitation of simple, query-based IR systems at the time, she found users do not search for information using simple query matching. Rather, she found they:

> [B]egin with just one feature of a broader topic, or just one relevant reference, and move through a variety of sources. Each new piece of information that they encounter gives them new ideas and directions to follow and a new conception of the query. At each

> stage, they are not just modifying the search terms used in order to get a better match for a single query. Rather, the query itself, as well as the search terms used, is continually shifting in part or whole (p. 198).

Bates argued that most searches evolve and that at different stages of the search process, users find information and references to inform the redirection of a larger search process. Thus, information retrieval occurs “bit-at-a-time,” much like picking berries “scattered on the bushes [in the forest]; they do not come in bunches” (p. 198).

Marchionni (1997) also described iterative *electronic information-seeking* in online environments. Specifically, he said the information-seeking process is systematic and opportunistic, composed of several sub-processes that begin with recognizing a problem and end with some stopping point that can make calls to other sub-processes. Marchionni’s model includes identifying information needs, defining problems, selecting information resources, formulating queries, executing queries, examining results, extracting information, and reflecting on the process or ending the search. Like Bates (1989), Marchionni sought to help developers create usable search tools. Uniquely, he developed his model to describe resources increasingly connected by the Internet.

Dervin's (1983) *sense-making* theory portrays search as related to cognition. Conceptually distinct from our other theories, it is a process and a methodology originally used by communications researchers. Dervin’s logic was that:

1. individuals face a *conceptual gap;*
2. in a *situation*;
3. where they bridge this gap using *information*; and
4. bridging the gap is *helpful*.

In information science, researchers use sense-making as a methodological theory to “explicate and study variable analytic measures categorized as information needs, seeking, and use” (Savolainen, 1993, p. 13). Human-computer interaction researchers also began employing Dervin’s sense-making theory around the same time (Russell et al., 1993). The metaphor of sense-making lends itself to understanding cognitive processes that shape IR; however, it is not a theory of HIB *per se*. It is an outlier because Dervin does not focus on user behavior. Rather, sense-making describes how individuals benefit from information when they use it to develop mental models of the world and a methodological frame to ascertain where and when this occurs.

Bates, Marchionni, and Dervin each study people but they do not focus on their attributes. We recognize that different populations may behave differently, and one theory focuses on ICPSR’s user group—Meho and Tibo’s (2003) *social scientist information-seeking*. Meho and Tibo expand on Ellis’ (1989) population-agnostic model to describe how social scientists use IR tools. They found four information-seeking stages: searching, accessing, processing, and ending. Within these stages are 16 sub-behaviors, including four behaviors that Ellis did not find. Unique among the theories we review, they account for on- and offline behaviors. Discussing information resources, Meho and Tibo point out that some are necessary for the research lifecycle to continue (p. 585).

Methodology

Research suggests that HIB theory inadequately describes data discovery, which carries implications for building user-centered IR tools. However, as Gregory and Koesten (2022) remind us, millions of datasets are available online (p. 1), and numerous intermediaries provide these datasets to users. Because data intermediaries provide their datasets using diverse systems, we designed a study mindful of the potential for variation—behavior in one context may be different elsewhere. Data-seeking behavior at ICPSR had not yet been compared to HIB theories, so we tested them in a high-profile context, simultaneously gathering observational data for theory-building to meet the call issued by Gregory et al. (2019).

Search log analysis

We began our study by comparing the abovementioned HIB theories with data from ICPSR search logs (Lafia et al., 2023). Data were collected from ICPSR's website using Google Analytics between 2012 and 2016 and described 98,000 user sessions. ICPSR staff were excluded from our analysis based on IP addresses. The search logs captured users' interactions (i.e., clicks, queries) with ICPSR resources, including the data catalog, study-level pages, dataset variables, and the Bibliography of Data-related Publications.

To analyze data, first, we manually classified the 25 most-accessed resources on the ICPSR website. We proposed five categories (i.e., access points, drop-off, lookup, object, and transactional) for these resources based on Broder's (2002) taxonomy of users' information needs and web queries.[ii] We classified these resources according to aspects of the data discovery process that they supported. For example, ICPSR's "Find Data" page was an obvious access point, whereas individual study pages were data objects. Next, we determined behavioral sequences where ICPSR users accessed resources by identifying three paths: *direct*, *scenic*, and *orienting*. These paths led from the main website and concluded at exit pages, where users left the website. Finally, we noted points of failure and recurring behaviors that supported or impeded research data discovery along each path.

Data user interviews

After completing our search log analysis, we conducted a follow-up interview study. The purpose of this study was to 1) validate the behavioral paths we identified and 2) learn more about the information researchers need to find data. Our interviews helped us understand and interpret Google Analytics data, which did not provide granular insights into user interaction with website features (e.g., search facets, metadata fields). Work by Gregory and Koesten (2022) shows that metadata is central to research data discovery. However, as noted earlier, the literature says very little about what types of information are key to sense-making, and we wanted to know more. The literature says sense-making describes data-seeking behavior, but this begs the question—*what is made sense of and how?*

In the summer of 2022, we interviewed 20 social scientists who had published in peer-reviewed journals using ICPSR data. We used ICPSR's Bibliography of Data-related Publications to identify individuals. We recruited participants via email, offering $50 gift cards for their participation. In our emails, we provided a screening survey that collected demographic information and asked about the data-seeking behaviors we uncovered with Google Analytics.

Our interview participants included university faculty (n=12), graduate students (n=5), postdocs (n=3), and government employees (n=1). Most participants were criminologists (n=6), psychologists (n=3), sociologists (n=3), and public health scholars (n=3). Interviewees reported 1 to 5 (n=6), 5 to 10 (n=7), 10 to 20 (n=3), and more than 20 years (n=2) experience as researchers. Sixteen study participants self-identified as quantitative researchers, while four used mixed-methods. We could not recruit qualitative researchers because of the quantitative composition (~98.7%) of ICPSR's collection. All of our interview participants said they wanted data to carry out research. Other reasons that participants sought data included teaching (n=9), policy or program analysis (n=5), and writing class papers (n=3).

After our participants responded to the screening survey, we conducted and recorded our interviews online using Zoom in a semi-structured fashion (Creswell & Creswell, 2017). We began our interviews by asking participants about their research data-seeking behaviors. We also asked about the need for information to enable data discovery on a scale of "1" (not necessary) to "5" (essential). Questions about information pertained to relevance, data usability, a need to conduct cutting-edge research, accessibility, and trust in data resources. Each category provided insight into the information that users need to find datasets.

Findings

The theories we evaluated inadequately describe research data discovery at ICPSR. Two (*berrypicking* and *electronic information-seeking*) fail to explain how users navigate searches to make sense of information rather than acquire information objects. Three (*berrypicking*, *electronic information-seeking*, and *sense-making*) do not encapsulate all of the data discovery paths we identified. We also found that data-specific information attributes are central to data discovery, which suggests that building usable tools requires more than identifying and supporting sense-making behaviors. As a standalone theory, sense-making failed to explain our collective study findings—rather, we find that data attributes and documentation are most relevant.

HIB theory describes documents, not data

The theories we reviewed state that they examine information-seeking, but they actually focus on text documents. **Table 2** compares the theories we evaluated by describing their portrayals of information, whether users conduct multiple queries, whether users can engage in multiple tasks while searching, and user objectives.

Table 2. Comparison of four HIB theories.

Theory	Search Object	Query	Task	User Objective

Bates (1989)	"Documents," "Information" (p. 199)	Multiple possible (p. 199)	Multiple possible (p. 199)	Information retrieval (p. 199)
Marchionni (1997)	"Information" (Online, mostly text)	Multiple possible (p. 50)	Single (But multiple possible sub-processes) (pp. 36-38, 50, 59)	Information retrieval (p. 50, 59)
Dervin (1983)	"Information" (Product of human observing in communication) (p. 4)	Single (Operationalized as a "gap") (p. 9)	Single (A "situation" and "use" of information to fill a "gap") (p. 9)	Filling a "gap" (p. 9)
Meho and Tibo (2003)	"Information" (Books, journal articles, archival materials, fieldwork data, newspapers) (pp. 577–578)	Multiple possible (p. 584)	Multiple possible (p. 584)	Ending a research project (p. 584)

In these HIB theories, descriptions of information objects refer to documents, online texts, books, articles, fieldwork notes, and archival materials. We explain this characterization by noting the date each theory emerged. Their creators developed them when a library catalog or subscription database was the most visible IR system. Analog media was widespread, and system users sought textual information. As analog systems moved online, Bates (1989), Marchionni (1997), and Meho and Tibbo (2003) kept abreast of changes in information-seeking environments to support IR tool development.

The theories we examined describe dynamic behaviors instead of simple query matching, likely because they emerged to critique traditional IR systems. Altogether, these theories portray individuals as conducting multiple queries and engaging in interrelated search tasks. Dervin (1983) is one exception because, to her, information is a product of human communication and not something to obtain like a document. She treats information-seeking as a cognitive process. Bates (1989) and Marchionni (1997) say object retrieval is an objective. However, Dervin (1983) theorizes that users engage in sense-making behavior to fill a cognitive gap, meaning that information retrieval is not their priority. Meho and Tibo (2003) agree, conceptualizing search as a means to an end. In *social science information-seeking*, academic researchers seek information to process it and finish a research project. When they process information, they engage in research tasks and end searches or formulate new queries as needed.

Users look for data along direct, scenic, and orienting paths

Our analysis of trace data confirms that HIB theory provides a starting point for IR tool development. Our analysis also confirms Gregory (2021) and Koesten et al. (2017), who argue that sense-making behaviors characterize research data discovery. As we show in Lafia et al. (2023), users follow one of three paths when searching for ICPSR data (**Figure 1**). Direct paths are linear, narrowly focused, and involve the fewest possible steps to return information. Orienting paths connect access points (e.g., a search page) to resources that provide users with contextual information about data. Scenic paths are where users alternate between access points and objects as they refine their search queries and change their access points.

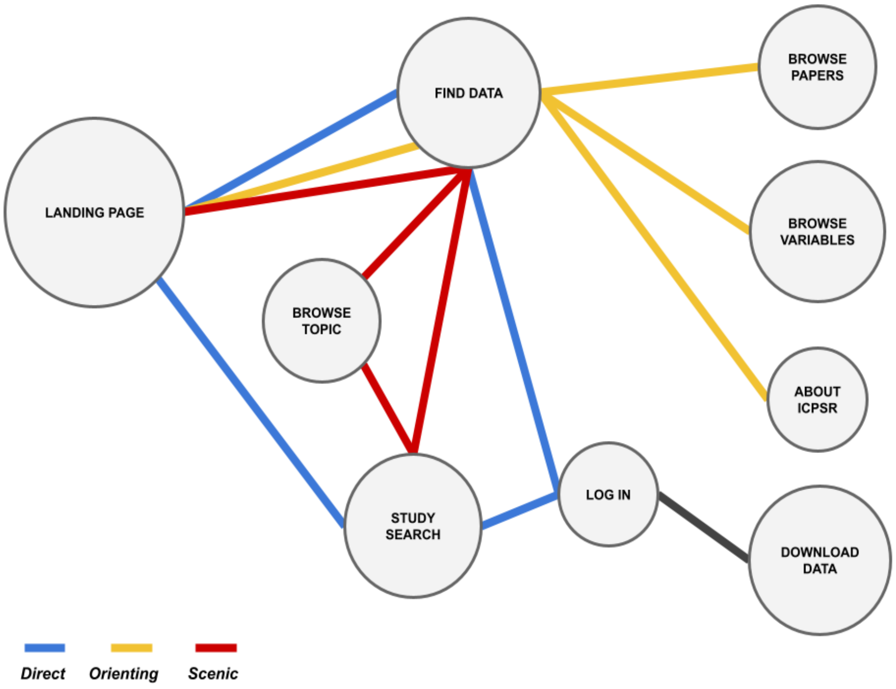

Figure 1. Simplified paths extracted from ICPSR search logs. Pages shown as scaled nodes with overlaid paths: direct (blue), orienting (yellow), and scenic (red).

Direct and scenic paths prioritize data retrieval, which supports Bates (1989) and Marchionni (1997). However, orienting paths emphasize users acquiring information about data or services rather than the data itself. This information was diverse, including codebooks, web pages (e.g., "About ICPSR"), a variable index, and other forms of documentation that researchers used to identify and evaluate data. Users followed paths individually and in combination, which confirms that *sense-making* (Dervin, 1983) is central to research data discovery. Making sense of information about data was necessary for users to meet their information needs, but unlike *document* search (e.g., looking for a journal article), researchers evaluated a wider range of information than tradtional object-level metadata. Specifically, this information included descriptive metadata and study documentation indicating the potential ways data could feasibly be reused.

Because they describe some of what we observed (i.e., direct and scenic paths), *berrypicking* and *electronic information-seeking* are clear starting points for data discovery tool development. However, they fail to specify that users evaluate datasets using diverse information. Meho and Tibo (2003) account for this orienting behavior because they recognize that information-seeking is driven by individuals conducting research, which requires users to work with and analyze disparate information. However, *sense-making* also fails to describe discrete IR tasks associated with the direct and scenic search paths we found. In some cases, users conducted known-item searches for data with which they were already familiar.

Users Rely on Metadata and Documentation to Evaluate Data Relevance

“Doing” research requires sensitivity to the appropriateness of the data used. Answering research questions requires employing appropriate methods. In the case of secondary data, however, data are already collected, so users must evaluate them to ensure their latent affordances support research activities. These activities might range from calculating descriptive statistics to conducting inferential tests. Evaluation might occur before a search for data to reuse in a new study or during the study itself. The varied time and place of evaluation explains why *berrypicking*, *electronic information-seeking*, and *sense-making* were inadequate to describe the combined breadth of paths that we found.

Responding to our screening survey, study participants confirmed engaging in behaviors that led them on direct, scenic, and orienting paths (see **Appendix A**). On a scale of “1” (never) to “5” (always), our 20 participants said that when they downloaded datasets, the data only “sometimes” ($\bar{x}$=3.3) met their needs. Unsuccessful searches incentivized users to follow scenic and orienting paths. However, behaviors associated with these other paths were also unpredictable. Our participants “sometimes” ($\bar{x}$=3.2) said they alternated between searching for and browsing data but “seldom” ($\bar{x}$=2.6) changed their research questions. The variation we observed supports the claim that research data discovery happens in stages as users make sense of datasets and try to meet their information needs. It also confirms that highly varied data-seeking behavior is the norm (Gregory et al., 2020). Indeed, respondents said that data discovery was “almost always” ($\bar{x}$=4.4) harder when they did not know what data contained, so they “almost always” ($\bar{x}$=4.1) used codebooks to evaluate datasets. Beyond confirming that IR and sense-making activities occur, however, none of the theories we examine acknowledged that the attributes of data might shape search behavior (including s*ocial scientist information-seeking*).

In interviews that followed our screening survey, we asked about the information researchers need to find data. We asked about information in traditional digital object metadata and contextual information associated with data discovery (e.g., papers published using data). Eight of the 27 questions we asked about information important to data discovery were “very” or “extremely important” to researchers. The most important information (our subset of 8 questions) also described markers of relevance—social scientists needed information specifically about quantitative data and not textual documents, as is the focus of HIB theory.

Data topics ($\bar{x}$=4.10), geographic coverage ($\bar{x}$=4.20), data collection methods ($\bar{x}$=4.40), collection time period ($\bar{x}$=4.05), and the ways that data operationalized variables ($\bar{x}$=4.20) all mattered to our interviewees. There was also near-universal agreement that descriptive information about study populations ($\bar{x}$=4.50) and variables ($\bar{x}$=4.90) were essential to discovery and the completion of studies. For questions unrelated to relevance, two mattered: dataset documentation (x̄=4.30) and whether data are publicly available (x̄=4.10) were “very important.” Respondents were asked how much each of these factors matters to them, and we provide our complete interview question responses in **Table 3**.

Table 3. Interview Response Averages.

Cutting Edge Research	**Average**
Past papers published with the data	3.6
Information about if data are suitable to conduct a study	3.4
The topic of papers that were recently published using the data	3.3
Where papers were published	2.9
Who has published papers recently reusing the data	2.5
Data Accessibility	**Average**
Whether data are accessible (available for public download or restricted)	4.1
How long it takes to access the data	3.7
Trustworthy Sources	**Average**
Who created the data	3.3
Knowledge about others’ experiences using the data	3.1
Information about the data creator’s affiliations and reputation	3.1
Who funded the original study	2.9
How often others have downloaded the data	2.2
How often others have viewed the data	2.0
General information about who has downloaded the data (assembled from anonymous click patterns)	2.0
Relevance	**Average**
The variables studied (or concepts, if qualitative research)	4.9
The population studied	4.5
The population sample or methods that were used to collect the data	4.4

The geographic coverage of the data	4.2
How variables or concepts were measured	4.2
When data were collected	4.1
The topic of the original study that produced the dataset	4.1
Unique identifiers like grant numbers	1.7
Data Usability	**Average**
What documentation is available (e.g., codebooks)	4.3
Documentation quality (e.g., missing case notes, level of description, distributions)	3.5
How "clean" or well-curated the data are	3.4
The file format(s) of data	3.0
The ability to analyze data online	1.8

Discussion

Distinguishing between quantitative data and textual documents provides a frame that explains our collective findings. Furthermore, making this distinction confirms that HIB theory does not describe data discovery and expands upon work (Koesten et al., 2021; Koesten et al., 2017; Gregory & Koesten, 2022; Gregory et al., 2020; Gregory et al., 2019) characterizing data discovery as a cycle of search and evaluation without treating *sense-making* as an all-encompassing frame. Recall that data include objects used to provide evidence of a phenomenon or serve as the subject of research. Quantitative data are structured, countable "things" (Buckland, 1991) with neatly defined attributes represented in catalog records, variable indices, codebooks, and other lookup resources. Qualitative documents are text objects organized into artifacts like books with "fuzzy" meanings.

Chu (2003) defines *representation* in IR as the extraction of object elements (e.g., keywords or phrases) or assigning terms (e.g., descriptors and subject headings) to an object so systems may characterize its essence. The "essence" of datasets in our study differed from textual documents, which required users to attend to a novel and complex range of metadata and contextual information for evaluation purposes. Differences in object representation and user needs explain the search behaviors we found that our interviewees elaborated upon.

At ICPSR, users sought data, rather than textual documents, to conduct research. However, treating quantitative data and textual documents as identical assumes IR systems represent and organize them the same and researchers use them for equivalent goals. Our four theories made this assumption by operationalizing information as text documents (or products of human observation in the case of *sense-making*). At the same time, the user-centered data retrieval literature neglects to differentiate between the many data types IR systems can represent, and two theories (*berrypicking, electronic information-seeking*) were geared to describing information retrieval alone. However, unlike in these theories, our users goal was not to retrive objects but to do research. *Social scientist information-seeking* acknowledges that information-seeking for academic research is motivated by a need to conduct research, but it focuses on text documents. *Sense-making* fails to account for the latent afordances of data objects that determine if they are suitable for reuse. At ICPSR, the most relevant information for discovery represented quantitative data specifically. Quantative data attributes were central because successful secondary data use (i.e., meeting an information need) depended on users obtaining objects with attributes that are required to successfully publish academic studies.

Theories of HIB first emerged to build user-centered systems and support a broad range of behaviors. Related work demonstrates that context (e.g., Savolainen, 1995), information-seeking tasks (e.g., Li & Belkin, 2008), information needs (e.g., Wilson, 1981), and demographics (e.g., Lorence et al., 2006) all shape user behavior. However, we found that data representation at ICPSR also shaped behavior, because sense-making effectively described users evaluation of a narrow range of relevant data attributes. HIB-related work has already demonstrated that information-seeking may vary by the object type sought (e.g., Albertson, 2015; Lavranos et al., 2015; Beaudoin, 2016). However, no paper we reviewed suggests that research data representation or the type of object sought should be a primary design consideration.

Supporting Researchers

To meet the promise of user-centered data retrieval, we argue that it is crucial for researchers to move beyond a focus on theories of human-information behavior alone. Koesten et al. (2017) portray data-seeking behavior as different than document search, but we go further by providing insight into what information users need. Discussing content-based image retrieval, Beaudoin (2016) found that the individuals who use image-retrieval systems are "primarily concerned with the formal characteristics (i.e., color, shape, composition, and texture) of the images being sought" (p. 350). Although ICPSR users sought quantitative data instead of images, their needs were comparable.

Meho and Tibbo (2003) argue that IR systems should support discrete behaviors accounted for by theory, but without predictable explanations for why and when individuals act, theory is little more than speculation. Software developers and designers require actionable guidance from the theories and models they consult. Sensing that no one theory may be able to provide concrete guidance to developers, Koesten et al. (2017) provide a list of data-centric tasks alongside their model of how researchers work with structured data. However, identifying and predicting when and where search and sense-making behaviors occur with a high degree of precision is difficult. Our interview screening survey revealed that discrete search and sense-making behaviors may be unpredictable, which means even heuristics of the sort provided by Koesten et al. (2017) may be sub-optimal.

In information science, sense-making is a methodological theory, typically used to describe phenomena like human-information behavior. We found that ICPSR users made sense of metadata, documentation, and varied information sources to conduct research studies; their behavior was an outgrowth of cognitive processes (another view of sense-making) tied to the research enterprise and the latent affordances data provide as objects with with to answer questions. Treating data attributes and representation as a frame for IR tool design enables us to sidestep problems associated with trying to build usable tools while relying on behavioral traces alone. By expanding the frame of IR tool design to include data representation, we can still leverage behavioral traces but understand why these behaviors occur.

We found a predictable set of information needs exist for researchers to find quantitative data. Unlike textual documents, which IR tools describe using metadata or full-text intended to support activities like checking out a book or reading, ICPSR's systems described quantitative data to support object retrieval and evaluation to enable research. Foundational to these processes were information like how variables were operationalized, when and how data were collected, the representativeness of sampling frames, and the topical coverage of datasets. This information, largely unique to quantitative, secondary data, provides an alternative starting point for tool design that, to our knowledge, is unexplored in the user-centered data discovery literature and understudied in the HIB literature.

Missing from the literature is an empirically-validated model of research data documenting information about and within data that users find relevant to identify and select these objects for research. Our findings demonstrate that this information could help create user-centered systems. Making sense of metadata and contextual information is a laborious process—time-consuming enough that researchers prefer to collect their own data or rely on the recommendations of mentors and peers rather than searching for it alone (Zimmerman, 2007; York, 2022; Gregory et al., 2020). Placing relevant data-specific information detailing whether it is suitable to conduct studies at the center of design process would bypass this impediment to secondary data reuse. We, therefore, call for the creation of a mid-level theory of data representation that will provide system designers an account of the information researchers need to evaluate all types of secondary data, not just statistical information. Because secondary data can include documents, and documents are increasingly conceived of and used as data (see, e.g., Park and Pouchard, 2021; Kricka et al., 2020), such a mid-level theory may be beneficial in the design of discovery systems for documents and other research materials more broadly.

Conclusion

In this study, we reviewed four HIB theories and literature focused on user-centered data retrieval. Although the four theories we reviewed do not account for all research studying information-seeking, they are very influential, and we suggest that a data/documents distinction provides insight into how and why social scientists engage in sense-making behaviors when evaluating data for reuse. We provide an alternative frame for building user-centered data discovery tools, and we call for the development of a model of data providing information about and within data that users find relevant.

Our study is limited in that it focused on users at ICPSR, a large data archive, and not all archives use comparable IR systems. Furthermore, because our study population oversampled quantitative and mixed-methods researchers, our findings may not be replicable across populations. Our interview sample focused on successful searches, operationalized as participants publishing papers with data, which also limits the generalizability of our findings.

Nevertheless, we revealed a path of interest to individuals studying human-centered data retrieval. The concept of information is not a monolith, and secondary data reuse depends on researchers finding and using data that enables them to answer questions and conduct research. The attributes of information objects dictate the methods that researchers can use to answer questions, so these attributes ought to be treated as central to research data discovery, as secondary data reuse becomes increasingly widespread.

References

Albertson, D. (2015). Visual information seeking. *Journal of the Association for Information Science and Technology*, *66*(6), 1091–1105. https://doi.org/10.1002/asi.23244

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

Beaudoin, J. E. (2016). Content-based image retrieval methods and professional image users. *Journal of the Association for Information Science and Technology*, *67*(2), 350–365. https://doi.org/10.1002/asi.23387

Borgman, C. L. (2012). The conundrum of sharing research data. *Acta Anaesthesiologica Scandinavica*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries*, *16*(3-4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

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

Buckland, M. (2018). Document theory. *Knowledge Organization*, *45*(5), 425–436. https://doi.org/10.5771/0943-7444-2018-5-425

Buckland, M. K. (1991, June). Information as thing. *Journal of the American Society for Information Science*, *42*(5), 351–360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Buckland, M. K. (1997). What is a “document”? *Journal of the American Society for Information Science and Technology*, *48*(9), 804–809.

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

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

DeRose, S. J., Durand, D. G., Mylonas, E., & Renear, A. H. (1997). What is text, really? *[Asterisk]*, *21*(3), 1–24. https://doi.org/10.1145/264842.264843

Dervin, B. (1983). *An overview of sense-making research: Concepts, methods, and results to date*. Michigan State University.

Dourish, P., & Gómez Cruz, E. (2018). Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society*, *5*(2), 2053951718784083. https://doi.org/10.1177/2053951718784083

Ellis, D. (1989). A behavioural model for information retrieval system design. *Journal of Information Science and Engineering*, *15*(4-5), 237–247. https://doi.org/10.1177/016555158901500406

Erdelez, S. (1999). Information encountering: It's more than just bumping into information. *Bulletin of the American Society for Information Science and Technology*, *25*(3), 26–29. https://onlinelibrary.wiley.com/doi/abs/10.1002/bult.118

Furner, J. (2016). "Data": The data. In M. Kelly & J. Bielby (Eds.), *Information cultures in the digital age: A festschrift in honor of Rafael Capurro* (pp. 287–306). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-14681-8_17

Inter-university Consortium for Political and Social Research (ICPSR). (n.d.). *Glossary of social science terms*. Retrieved February 7, 2023, from https://www.icpsr.umich.edu/web/ICPSR/cms/2042

Gregory, K. (2021). *Findable and reusable? Data discovery practices in research* [Maastricht University]. https://doi.org/10.26481/dis.20210302kg

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*, *2*(2). https://doi.org/10.1162/99608f92.e38165eb

Gregory, K., & Koesten, L. (2022). *Human-centered data discovery*. Springer. https://doi.org/10.1007/978-3-031-18223-5

ISIC Conference. (n.d.). ISIC Conference. Retrieved February 8, 2023, from https://www.isic-conference.org/

Koesten, L., Gregory, K., Groth, P., & Simperl, E. (2021). Talking datasets – Understanding data sensemaking behaviours. *International Journal of Human-Computer Studies*, *146*(102562), 102562. https://doi.org/10.1016/j.ijhcs.2020.102562

Koesten, L. M., Kacprzak, E., Tennison, J. F. A., & Simperl, E. (2017). The trials and tribulations of working with structured data: A study on information seeking behaviour. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1277–1289. https://doi.org/10.1145/3025453.3025838

Krämer, T., Papenmeier, A., Carevic, Z., Kern, D., & Mathiak, B. (2021). Data-seeking behaviour in the social sciences. *International Journal on Digital Libraries*, *22*(2), 175–195. https://doi.org/10.1007/s00799-021-00303-0

Kricka, L. J., Polevikov, S., Park, J. Y., Fortina, P., Bernardini, S., Satchkov, D., Kolesov, V., & Grishkov, M. (2020). Artificial Intelligence-Powered Search Tools and Resources in the

Fight Against COVID-19. EJIFCC, 31(2), 106–116. https://doi.org/10.1101/2020.05.23.112284

Kriesberg, A., Frank, R. D., Faniel, I. M., & Yakel, E. (2013). The role of data reuse in the apprenticeship process. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–10. https://doi.org/10.1002/meet.14505001051

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)*.

Lavranos, C., Kostagiolas, P. A., Martzoukou, K., & Papadatos, J. (2015). Music information seeking behaviour as motivator for musical creativity: Conceptual analysis and literature review. *Journal of Documentation*, *71*(5), 1070–1093. https://doi.org/10.1108/JD-10-2014-0139

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, *44*(6), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005

Lorence, D. P., Park, H., & Fox, S. (2006). Assessing health consumerism on the web: A demographic profile of information-seeking behaviors. *Journal of Medical Systems*, *30*(4), 251–258. https://doi.org/10.1007/s10916-005-9004-x

Luk, R. W. P. (2022). Why is information retrieval a scientific discipline? *Foundations of Science*, *27*(2), 427–453. https://doi.org/10.1007/s10699-020-09685-x

Marchionini, G. (1997). *Information seeking in electronic environments*. Cambridge University Press.

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 & Technology*, *54*(6), 570-587. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.10244

Munzner, T. (2014). *Visualization analysis and design*. CRC Press.

Norman, D. A., & Draper, S. W. (1986). *User centered system design*. Lawrence Erlbaum Associates. https://doi.org/10.1201/b15703

Park, G., & Pouchard, L. (2021). Advances in scientific literature mining for interpreting materials characterization. Machine Learning: Science and Technology, 2(4), 045007. https://doi.org/10.1088/2632-2153/abf751

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4. https://doi.org/10.1002/meet.14504701240

Robertson, S. E. (1977). Theories and models in information retrieval. *Journal of*

Documentation, *33*(2), 126–148. https://doi.org/10.1108/eb026639

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 269–276. https://doi.org/10.1145/169059.169209

Savolainen, R. (1993). The sense-making theory: Reviewing the interests of a user-centered approach to information seeking and use. *Information Processing & Management*, *29*(1), 13–28. https://doi.org/10.1016/0306-4573(93)90020-E

Savolainen, R. (1995). Everyday life information seeking: Approaching information seeking in the context of 'way of life.' *Library & Information Science Research*, *17*(3), 259–294. https://doi.org/10.1016/0740-8188(95)90048-9

Sherif, V. (2018). Evaluating preexisting qualitative research data for secondary analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *19*(2). https://doi.org/10.17169/fqs-19.2.2821

Tredinnick, L. (2006). *Digital information contexts: Theoretical approaches to understanding digital information*. Elsevier.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016, December). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation*, *37*(1), 3–15. https://doi.org/10.1108/eb026702

Wynholds, L. (2011). Linking to scientific data: Identity problems of unruly and poorly bounded digital objects. *International Journal of Digital Curation*, *6*(1), 214–225. https://doi.org/10.2218/ijdc.v6i1.183

Yoon, A., Copeland, A., & McNally, P. J. (2018). Empowering communities with data: Role of data intermediaries for communities' data utilization. *Proceedings of the Association for Information Science and Technology*, *55*(1), 583–592. https://doi.org/10.1002/pra2.2018.14505501063

York, J. (2022). *Seeking equilibrium in data reuse: A study of knowledge satisficing*. University of Michigan. https://doi.org/10.7302/6170

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1-2), 5–16. https://doi.org/10.1007/s00799-007-0015-8

Appendix A.
Frequency of responses across survey population with questions organized by search path type.

Direct	**Average**
I know what I am trying to find (e.g., a specific dataset)	4.0
When I open datasets, they have what I need	3.3
I get what I want, typically on the first page of my search results	3.1
I stop searching if I do not quickly get what I want	2.6
Orienting	**Average**
Discovery is harder when I do not know what datasets contain (e.g., variables)	4.4
I use codebooks to decide if a dataset meets my needs	4.1
I switch between searching for data and evaluating datasets I find	3.6
I revise queries based on my search results	3.5
I download data to evaluate it	3.3
I alternate between searching for and browsing data	3.2
Scenic	**Average**
I look for data to see what is out there	3.5
Data I found earlier is useful later	3.4
I am willing to change my research question(s) when searching	2.6
The data discovery process is more important to me than the result	2.4

[i] *Number current in Google Scholar as of September 21, 2023.*
[ii] *See the methods section of Lafia et al. (2023) for detail.*