Received: 31 October 2023 | Revised: 30 August 2024 | Accepted: 19 October 2024

DOI: 10.1002/asi.24962

RESEARCH ARTICLE

JASIST WILEY

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

Anthony J. Million[1] | **Jeremy York**[2] | **Sara Lafia**[1] | **Libby Hemphill**[1,2]

[1]ICPSR, University of Michigan, Ann Arbor, Michigan, USA

[2]School of Information, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Anthony J. Million, Inter-university Consortium for Political and Social Research, 330 Packard St., Ann Arbor, MI 48104, USA.
Email: millioaj@umich.edu

Present addresses
Jeremy York, University of Michigan Libraries, Ann Arbor, Michigan, USA; and Sara Lafia, NORC at the University of Chicago, Chicago, Illinois, USA.

Funding information
Division of Social and Economic Sciences, Grant/Award Number: 2121789

Abstract

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

1 | 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 Marchionini'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. However, despite this shift, information scientists do not fully understand how researchers discover data and the user-specific discovery considerations that inform secondary data reuse.

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

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2024 The Author(s). *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

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, and its collection is very well structured. Ninety-nine (∼98.7%) percent of ICPSR data is quantitative, so users disproportionately use statistical or computational methods. The archive's user base is academic researchers, so data discovery at ICPSR represents just one type.

We argue that distinguishing between data and documents helps to clarify why emerging research (Gregory, 2021; Gregory & Koesten, 2022; Koesten et al., 2017) finds 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 contained within data (e.g., metadata and contextual documentation) are foundational to data discovery and reuse because the objects that users seek determine their most relevant representative and descriptive attributes. Thus, the type of object that users seek (in particular, whether the object is a dataset or a document) is important to consider when building IR systems. However, the HIB theories we review do not make this point, nor is it emphasized in studies identifying the information researchers use to find and select data (e.g., Friedrich, 2020; Koesten et al., 2020). We conclude by arguing that a mid-level data representation model can support systems design, which we define as a metadata framework intended to surface the representative attributes of data, showing if they are suitable for tasks, uses, and projects.

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) (p. 7)	Borgman et al. (2015), Dourish and Gómez Cruz (2018), Gregory and Koesten (2022), Munzner (2014), as cited in Gregory and 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 (p. 351)	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)

2 | LITERATURE REVIEW

2.1 | 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). This paper operationalizes data as information objects (see Buckland, 1991), focusing on quantitative data used for academic research.

Questions surrounding the identity of texts are important. Documents carry meaning because they organize information 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 used the same, which may hinder the creation of IR tools.

The distinction between data and documents is not necessarily as straightforward as Furner's (2016) paper suggests. Researchers might treat audio recordings, images, physical and biological specimens, or textual documents as data. For instance, primary source documents are central to historians' work. Using machine-learning models, computational social scientists might analyze government reports, white papers, news articles, or other information. Archeologists treat fossils as data to analyze evolutionary patterns. However, we argue that what ultimately matters is not the format or package within which data are contained (e.g., textual documents, images, audio recordings, quantitative files). Rather, what matters is that "data" are a type of information that serves as an object of analysis, which may possess different research affordances.

2.2 | Critiques of information retrieval systems

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

Critiques of IR systems are not new. In 1989, Bates asked us why traditional IR systems required users to represent their needs using 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 IR tool 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.

2.3 | 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" (Wilkinson et al., 2016, para 3). Thus, data discovery is vital to sharing, stewarding, and reusing data. Nevertheless, data discovery is highly challenging to describe.

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 convoluted 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 that 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. In turn, data reuse is discussed in a wide variety of literature, including research on and descriptions of tools and infrastructure.

Because data discovery is often embedded within other 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 norms and practices 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 (Gregory et al., 2020; York, 2022; Zimmerman, 2007), 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).

2.4 | 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 data retrieval systems is challenging. Krämer et al. (2021) find that users often rely on literature searches to identify relevant data, suggesting tighter integration of search systems for datasets and documents might improve 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 for providing software developers and designers with the help 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 (or personas) 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 activities in situ. These purpose-driven tasks include using data for a study; preparing a new project or proposal; teaching; generating new ideas; experimenting with new 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 of these tasks provides "an entry point for design" (p. 46).

Similarly, after reviewing data discovery practices in five disciplines, Gregory (2021) draws another conclusion with implications for building data discovery tools. She argues that "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 that existing user-centric information-seeking theories "are a reasonable starting point" for IR tool development but are inadequate. They note that "to be truly useful, [theory needs...] to consider the search and sense-making activities people carry out when working with structured data" (p. 1279). Koesten and colleagues then provide a taxonomy of data-centric tasks and a five-pillar model of how researchers work with structured data (p. 1284). They also suggest that software developers use their taxonomy and model as touchpoints. But what information do users need to find and evaluate data?

2.5 | Information needed to find and evaluate data

Some research has begun documenting the information researchers need to find data, which is closely tied to evaluating and selecting it. For instance, we know that metadata is essential—Phillips and Smit (2021) identify best practices for creating unstructured descriptors (e.g., title, abstract) to enable data sense-making. An important finding is that descriptive metadata may be incapable of surfacing every affordance data might provide. Papenmeier et al. (2021) report on a survey of 72 social scientists, finding that user information needs include dataset *topics*, *metadata*, and *use intentions*. Context also matters for discovery. In their paper on preservation, Faniel et al. (2019) propose 12 types of contexts describing data across three categories: *production*, *repository*, and *data reuse* information (p. 1274). Within these categories, there is variation by academic discipline concerning the information users typically require. Nevertheless, in their 105 interviews, social scientists, archeologists, and zoologists agree (>50%) that *data collection methods* are instrumental for discovery (p. 1281). Information describing the attributes of objects that researchers study also matters. Archeologists and zoologists require context describing *artifacts* or *specimens*. Social scientists need to know about *analytical methods* used in data transformations (p. 1281).

The most relevant information to find data is perhaps best characterized as *selection criteria*. Introduced by

Koesten (2019), selection criteria describe three categories of information (*relevance*, *usability*, and *quality*) that researchers use to decide whether data are "relevant or useful to a task" (p. 120). Seventeen subcategories also describe what individuals consider when evaluating data (Koesten et al., 2020). For example, relevance criteria include the *scope* of a dataset or collection (temporal, geographical, and temporal information), *granularity* (of scope), *context* (the purpose of a research study), and *documentation* (e.g., variables). Usability criteria include data *format*, *documentation* (e.g., labeling), *comparability* (e.g., identifiers, units), *references*, *access* (e.g., license), and *language*. Quality criteria include data *provenance*, *accuracy*, *completeness*, *methodology*, *bias*, *cleanliness*, and *timeliness*. In her presentation of selection criteria, Koesten references data type. However, it bears mentioning that her conceptualization of "type" relates to format (p. 7), which is distinct from data as a type or class of information object—a key focus of this study.

Friedrich's (2020) doctoral thesis begins to generalize Koesten's findings, which apply to only a small sample ($n = 69$) of computer science students. In her study, Friedrich finds that *quality of fit with a research question*, *data accessibility*, *documentation quality*, and *newness* is "important" or "extremely important" for discovery (p. 183). Uniquely, Friedrich's work presents data from a survey of established social scientists ($n = 1475$) affiliated with the Leibniz Institute for Social Sciences (GESIS) in Germany. Her respondents indicate that the most effective data discovery strategy involves consulting study documentation (p. 192).

3 | 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).
4. Social scientist information-seeking (Meho & Tibbo, 2003).

Despite their age, these theories remain relevant to theory and practice (especially IR tool design), collectively receiving 3,810 citations since 2013.[1] Although we did not employ formal selection criteria to choose these theories, we deliberately excluded those not focused on information-seeking and excluded redundant or overlapping theories. For instance, we excluded Kuhlthau (1993) because the concept of *uncertainty* is explained by Dervin's idea of a conceptual gap. More needs to be said regarding our analysis of HIB theory. In 2012, Donald Case identified over 10,000 HIB-related publications (p. 277). There are fewer theories, but the total is considerable—for instance, Fisher et al. (2005) summarize 72 theories, meta-theories, and models (p. xi). Because there are so many studies on the topic, this paper and others (Gregory, 2021; Gregory & Koesten, 2022; Koesten et al., 2017) cannot evaluate every information-seeking theory adjacent to data discovery. Therefore, our review does not intend to paint a comprehensive picture of theory; rather, it clarifies why others find theory inadequate.

3.1 | Selected HIB theories

Bates' (1989) theory of *berrypicking* emerged to criticize traditional IR systems. Observing the limitations of simple, query-based IR systems, she found that users do not search for information using query matching. Rather, 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 [...] to get a better match [...]. Rather, the query itself, as well as the search terms used, is continually shifting (p. 198).

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

Marchionini (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 software developers create usable search tools, but he developed his model to describe resources 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 initially used by communications researchers. Dervin's logic was

that individuals face a conceptual *gap* in *situations* where they bridge this gap using *information*, which is *helpful*.

In information science, researchers use sense-making as a theory to "explicate and study variable analytic measures categorized as information needs, seeking, and use" (Savolainen, 1993, p. 13). The metaphor of sense-making is also highly powerful and lends itself to understanding cognitive processes that shape IR; however, it is not a theory of HIB per se. Sense-making is an outlier among our theories because Dervin does not focus on user behavior. Instead, she describes how individuals benefit from using information to create mental models of the world and provides a methodological framework to explain why and where information-seeking occurs. Throughout this paper, we conceive sense-making as a highly elastic explanation for why individuals iteratively seek and evaluate data, much like Gregory and Koesten, who deliberately align themselves with Dervin's terminology. Although sense-making is not treated as a theoretical framework by the data discovery literature, we believe that it can be misconstrued as such and evaluate the theory accordingly.

Bates, Marchionni, and Dervin study people but 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 Tibbo'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, Meho and Tibo account for both on- and offline behaviors. Discussing information resources, they observe that some are necessary for the research lifecycle to continue (p. 585).

4 | 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, p. 429).

4.1 | Search log analysis

We began our study by comparing the aforementioned 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 their IP address. 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. 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. Although it is beyond the scope of this paper, we analyzed trace data to prototype a data recommender system.

4.2 | 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 found and (2) learn more about the information that researchers need to find data. Our interviews helped us understand and interpret Google Analytics data, which did not provide granular insights describing user interactions with specific website features (e.g., metadata fields).

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 discrete data-seeking behaviors we uncovered with Google Analytics (see Appendix A for the full screening survey). Specifically, we contacted 383 first-authors and received 42 survey responses from individuals willing to participate

in our study. Using information from our survey, we selected a *maximum-variation sample* (Cohen & Crabtree, 2006) of individuals to account for as many professional roles, disciplines, experience levels, and methodologies as possible.

Our interview participants included faculty ($n = 12$), graduate students ($n = 4$), 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 1–5 ($n = 8$), 5–10 ($n = 7$), 10–20 ($n = 3$), and more than 20 years ($n = 2$) experience as professional researchers. Sixteen study participants self-identified as quantitative researchers, and four used mixed methods. Given the composition of ICPSR's collection, we could not recruit qualitative researchers. Every participant said that they wanted data to "do research" and publish scholarly articles. Other reasons that participants looked for data were teaching ($n = 9$), policy or program analysis ($n = 5$), and writing class papers ($n = 3$). Based on our study sample, the individuals we interviewed were data users who had successfully published using secondary data during the first half of 2022.

After our participants responded to the screening survey, we conducted and recorded semi-structured interviews (Creswell & Creswell, 2017). Our protocol asked 27 questions, which we felt broadly described the information researchers need to support data discovery (see Appendix B). We also asked 33 questions about data recommendation. Questions in our interview protocol were informed by an undergraduate information science project that evaluated ICPSR's website, two grant-funded studies examining data curation (www.datacurationimpact.org), and a doctoral dissertation (York, 2022). We further refined our questions during an iterative protocol design process shaped by pilot tests with ICPSR user volunteers. Interviews took place online using Zoom, averaging approximately 1 h and 5 min. To reduce cognitive fatigue, we gave participants a break. Our findings came from a structured section of our semi-structured protocol, but even in the structured section, we encouraged participants to ask clarifying questions and elaborate on their answers.

In our interviews, we framed questions about the information required for data discovery by asking participants to focus on a "typical situation." That is, we limited responses to what researchers believed was their default information need. Next, we asked our study participants to score the information they needed for data discovery on a scale of "1" (not necessary) to "5" (extremely). Questions pertained to data relevance, usability, conducting cutting-edge research, access, and trust in data resources. We did not attempt to differentiate between participant situations, but everyone said they needed quantitative data to publish in their field (e.g., criminology, sociology, public health). To analyze interview data, we tabulated scores indicating the importance of different types of information for discovery. Last, we triangulated (Denzin, 2007) among our collective findings to infer relationships between search behaviors and the information that users needed to succeed in their searches. We did not create a thematic codebook and analyze interview questions other than the 27 we referenced because doing so was beyond the scope of this study.

5 | FINDINGS

The theories we evaluated inadequately describe 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*) fail to encapsulate each of the data discovery paths we identified. We also found that data-specific information attributes are central to discovery, which suggests building usable IR tools requires more than identifying and supporting sense-making behaviors. Rather, we found that data attributes were paramount.

5.1 | HIB theory describes documents, not data

The theories we reviewed state that they examine information-seeking, but they primarily 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. By objectives, we mean the conditions researchers must meet to satisfy their information needs.

In these theories, descriptions of information objects refer to documents, online texts, books, articles, fieldwork data, and archival materials. We explain this characterization by noting the date that 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 documents. As analog systems moved online, Bates (1989), Marchionini (1997), and Meho and Tibbo (2003) kept abreast of changes in information-seeking environments to support IR tool development.

The theories we examined also describe dynamic behaviors instead of simple query matching, because they emerged to critique traditional IR systems. Altogether,

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)
Marchionini (1997)	"Information" (Online, mostly text)	Multiple possible (p. 50)	Single (But multiple possible sub-processes) (pp. 36–38, 50, 59)	Information retrieval (pp. 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 Tibbo (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)

our four 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 Marchionini (1997) say that object retrieval is a final objective. However, Dervin (1983) theorizes that users engage in sense-making behavior to fill a cognitive gap, meaning that information retrieval is not necessarily their goal. Meho and Tibbo (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.

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

Our analysis of ICPSR website trace data confirms that HIB theory provides a starting point for building user-centered IR tools. Our analysis also confirms Gregory (2021) and Koesten et al. (2017), who argue that sense-making behaviors characterize data discovery. As we report in Lafia et al. (2023), users follow one of three paths when searching for ICPSR data (Figure 1). Direct discovery paths are linear, narrowly focused, and involve the fewest possible steps to return data. 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 information objects as they refine their queries and change data access points. Given the composition of ICPSR's collection and our understanding of the site architecture, we infer that along these paths, users depend on textual documents to find data. The data users sought was distinct from the web pages and documentation they relied upon.

In our findings, direct and scenic paths prioritized data retrieval, which supports Bates (1989) and Marchionini (1997). However, orienting paths emphasized users acquiring information about data or services rather than the data itself. This information was diverse, including codebooks, webpages (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 quantitative 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, including descriptive metadata and study documentation describing how data could be used.

Because they describe some of what we observed (i.e., direct and scenic paths), *berrypicking* and *electronic information-seeking* provide useful starting points for data discovery tool development. However, each fails to note that users evaluate datasets with diverse supplemental resources. Meho and Tibbo (2003) account for orienting (i.e., sense-making) behavior because they recognize that information-seeking is driven by individuals conducting research, which is more complicated than basic IR. However, *sense-making* fails to describe discrete IR tasks associated with the direct and scenic search paths we identified. In some cases, users conducted known-item searches for data with which they were already familiar. We acknowledge that this may result from cases where researchers proceed directly to data from documents (Gregory et al., 2020), treat ICPSR as "online storage" for data with which they are already familiar, or receive instruction from their peers and mentors (e.g., York, 2022). Additionally, we acknowledge that sometimes researchers might transcribe data from documents. However, these qualifiers do not change the fact that *berrypicking*, *electronic information-seeking*, and

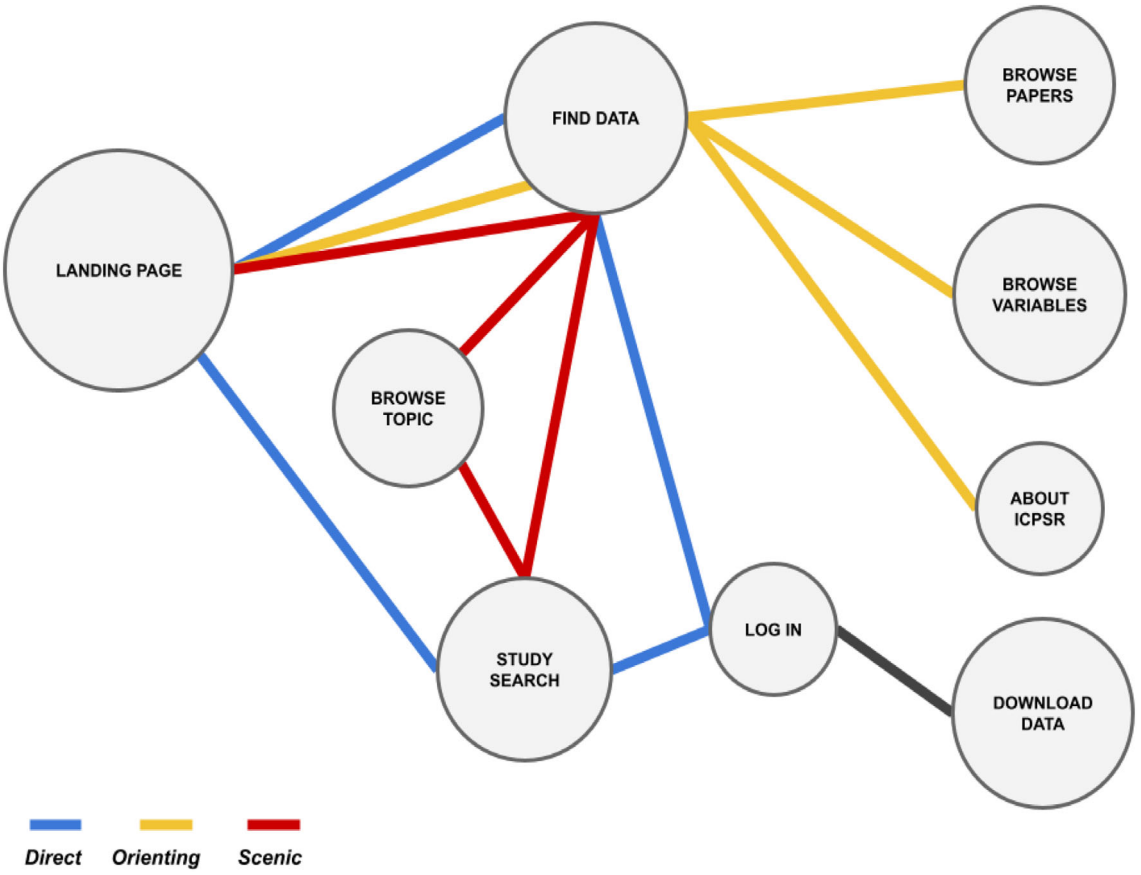


FIGURE 1 Simplified paths extracted from ICPSR search logs. Pages shown as scaled nodes with overlaid paths: direct (blue), orienting (yellow), scenic (red), and unsurfaced transitions to a credentialing/dissemination system where users download data (black).

sense-making inadequately describe the breadth of paths we found.

5.3 | 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. Of our four theories, only Meho and Tibbo (2003) recognize that the objective of data discovery is to complete a research project.

Responding to our screening survey, ICPSR users confirmed engaging in behaviors that led them on direct, scenic, and orienting paths (see Table C1 in Appendix C). On a scale of "1" (never) to "5" (always), our 20 participants said that when they downloaded datasets, the data only "sometimes" ($\overline{x} = 3.3$) 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" ($\overline{x} = 3.2$) said they alternated between searching for and browsing data, but "seldom" ($\overline{x} = 2.6$) changed research questions. The variation we observed supports the claim that research data discovery takes place in stages as users make sense of datasets and try to meet their information needs. Such variation also confirms that highly varied data-seeking behavior is the norm (Gregory et al., 2020). Indeed, respondents said that data discovery was "almost always" ($\overline{x} = 4.4$) harder when they did not know what ICPSR data contained, so they "almost always" ($\overline{x} = 4.1$) used codebooks to evaluate it. Beyond confirming that IR and sense-making activities occur, however, none of the theories we examine acknowledges how the attributes of data might shape search behavior (including *social scientist information-seeking*).

In the interviews following 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). Nine of the 27 questions we asked about information

important to data discovery were "very" or "extremely important" to researchers. The most important information (a subset of seven questions) also described markers of relevance—social scientists specifically needed information about quantitative data and not textual documents. We provide interview question response averages in Table 3.

Data topics ($\overline{x} = 4.10$), geographic coverage ($\overline{x} = 4.20$), data collection methods ($\overline{x} = 4.40$), collection time period ($\overline{x} = 4.05$), and the ways data operationalized variables ($\overline{x} = 4.20$) each mattered to our interviewees. There was also near-universal agreement that descriptive information about study populations ($\overline{x} = 4.50$) and variables ($\overline{x} = 4.90$) were needed for researchers to meet their information needs. For questions unrelated to relevance, two mattered: dataset documentation ($\overline{x} = 4.30$) and if data are publicly available ($\overline{x} = 4.10$) each were "very important." More importantly, however, these responses confirm an important point. Generally, ICPSR users sought quantitative data. The most important information for discovery was related to quantitative data attributes and not general selection criteria.

6 | DISCUSSION

Distinguishing between quantitative data and textual documents provides a frame that explains our collective study findings. Further, making this distinction confirms that HIB theory does not describe data discovery and expands upon prior work (Gregory et al., 2019, 2020; Gregory & Koesten, 2022; Koesten et al., 2017; Koesten et al., 2021) which characterizes data discovery as a cycle of search and evaluation without treating *sense-making* as a theoretical frame explaining why it is different from literature search. 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, p. 353) with neatly defined attributes represented in catalog records, variable indices, codebooks, and other lookup resources. Textual documents are different because they are organized into things like books with "fuzzy" meanings used for other purposes. Although the literature does not explicitly treat sense-making as a theoretical frame that explains data discovery, the metaphor is central in explaining why individuals engage with selection criteria. Nevertheless, we find there is more to the story.

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

TABLE 3 Interview response ($n = 20$) averages, organized by grouping category from the structured section of our protocol (see Appendix B).

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

an object so systems may characterize its essence. The "essence" of datasets in our study differed from textual documents, particularly in how doing quantitative research required users to attend to a novel and complex range of metadata and contextual information for evaluation purposes. Differences in object representation within ICPSR systems and user needs explain the behaviors we found and that our interviewees elaborated upon.

At ICPSR, the most relevant information for discovery were quantitative data attributes. These attributes were key because successful data reuse (i.e., meeting an information need) depends on users obtaining objects with attributes that are required to make valid, empirical truth claims and publish studies. Users might have copied quantitative data from documents or behaved in ways that do not neatly align with our narrative. However, their needs were clear given the composition of ICPSR's collection, our operationalization of data as an object, and the collective profile of our study participants. Further, it bears mentioning that the information researchers needed to find and select data was disproportionately about data (objects) specifically—not "nice to have" context like who had published using the data. To put our findings in Chu's terms, users of ICPSR data sought "quantitative" representations of phenomena (information objects) that were further represented using an array of systems providing documentation potentially relevant to how the data could be analyzed. The content, not merely the format of data, mattered, of course, but because data format provided a connection to research methods, it mattered.

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 data representation at ICPSR also shaped behavior because successful data users said a narrow range of data attributes was the most important. These attributes primarily related to type or format, but also extended to content described by datasets. HIB-related work has already demonstrated that information-seeking may vary by the object type sought (e.g., Albertson, 2015; Beaudoin, 2016; Lavranos et al., 2015). However, no paper we reviewed suggests that research data representation and the type of object sought should be a primary design consideration. It is in light of this finding and our research results that we suggest a mid-level data representation model could support IR tool design. To be sure, many types of information help researchers find and select data, but the formal, structural attributes of the data objects sought, and their content descriptors mattered most to ICPSR users because they indicated how data could be used.

6.1 | 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) and Friedrich (2020) each portray data-seeking behavior as different than document search, but we go further by suggesting the reason it is different has to do with specifics associated with the information objects users seek. Discussing content-based image retrieval, Beaudoin (2016) found that individuals who use image-retrieval systems are "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, their needs were comparable in that the attributes of information objects influenced how users searched.

Below, we summarize areas of alignment between our work and other research documenting the information researchers need for data discovery. Friedrich (2020) demonstrates that data search is a social phenomenon. Findings from her dissertation align with our work because we each conclude that *documentation quality* and *data accessibility* are highly relevant, as is information about a *dataset's fit with a research question*. Friedrich did not ask users if they required information about *data topics*, *geographic coverage*, *data collection methods*, *collection time period*, *study populations*, or the *variables* they contained. However, each attribute is a necessary consideration for quantitative researchers.

We find similar alignment between our study and Koesten's (2019) and Koesten et al.'s (2020) section criteria. Generally, we confirm that data *relevance*, *usability*, and *quality* influence whether users select the data they find. Concerning usability, we also found that dataset *documentation* and *access* are central to making decisions about reuse. We conceptualize *methodology* information as about "relevance" instead of "quality," but our interview participants agreed that this information was important. Concerning data relevance, our results confirm that *scope* (temporal and geographical), *granularity* (of scope), *context* (the purpose of a study), and *documentation* (e.g., variables) all matter. Uniquely, however, because we asked individuals to score criteria, and Koesten did not, our results suggest that relevance considerations stem from the information object sought, both as the thing researchers' study and the format of data representing phenomena.

Meho and Tibbo (2003) argue that IR systems should support behaviors accounted for by theory, but without predictable explanations, theory is little more than speculation. Software developers and designers require actionable guidance. 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, predicting research data discovery behavior is difficult. Recall that our interview screening survey revealed that discrete search behaviors are variable. This finding suggests that even heuristics of the sort provided by Koesten et al. (2017) may fall short in supporting IR tool designers.

In information science, *sense-making* is often cited because it explains human information behavior by appealing to cognitive processes. We found that ICPSR users made sense of metadata, documentation, and varied information sources. Behavior was clearly an outgrowth of cognitive processes tied to the research enterprise and the affordances data provided as objects used to answer study questions. However, treating data object 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 use behavioral traces and better understand why behaviors occur. At ICPSR, we found a predictable set of attributes exists for researchers to find quantitative data. Unlike textual documents, which IR tools traditionally have described using metadata or full-text intended to support activities like checking out a book, ICPSR's systems described data to support data-related tasks in academic research. Foundational to these tasks were information like how variables were operationalized, how data were collected, the representativeness of sampling frames, and the topical coverage of datasets. This information, unique to quantitative, secondary data, provides an alternative and arguably understudied starting point for tool design.

Missing from the literature is an empirically validated model of research data representation documenting information about and within data that users find relevant to identify and select these objects. 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 (Gregory et al., 2020; York, 2022; Zimmerman, 2007). Placing agreed-upon relevant, data-specific information at the center of the design process could bypass this impediment to secondary data reuse. We, therefore, call for the creation of a model that extends to all secondary data, not just quantitative data. Because secondary data can include text documents, and documents are increasingly conceived of as data (e.g., Park & Pouchard, 2021), such a mid-level theory may be beneficial in the design of discovery systems for other types of information objects.

What might our proposed model look like? Minimally, the model will center on Koesten's (2019) argument that selection criteria pertain to whether data are relevant or useful to a task. However, because different types of data afford themselves to different tasks (e.g., quantitative data cannot provide rich accounts of lived human experiences), our proposed model should try to describe the most useful attributes of these data to researchers. Hypothetical categories include textual documents, images, sound, and moving images. However, as we noted earlier, what matters most is not the categories but rather that systems can leverage them as descriptors to surface information about how data might be used. The model might also differentiate between contexts in which different data are needed and the phenomena documented using data.

7 | 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 most 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. Because our study population oversampled quantitative researchers, our findings likely do not describe all social scientists. Our interview sample focused on successful searches operationalized as participants publishing papers with data, which also limits the generalizability of our findings. Additionally, the responses to our analyzed interview questions were structured, limiting the richness of data and imposing our understanding of what makes them appropriate for reuse. While it makes sense to treat quantitative datasets as objective "things" with neatly defined use cases, other data types are less easy to categorize. Nevertheless, we revealed a path of interest to those studying human-centered data retrieval. The concept of information is certainly not a monolith, and secondary data reuse depends on researchers finding and evaluating data that enables them to do research. The attributes of information objects influence the methods researchers can use, so these attributes should be treated as central to data discovery tool design, especially as secondary data reuse becomes more widespread and lowers barriers to participation in science.

ORCID

Anthony J. Million https://orcid.org/0000-0002-8909-153X
Jeremy York https://orcid.org/0000-0001-8225-9291
Sara Lafia https://orcid.org/0000-0002-5896-7295
Libby Hemphill https://orcid.org/0000-0002-3793-7281

ENDNOTE

[1] Number current in Google Scholar as of September 21, 2023.

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). 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. *VLDB Journal: Very Large Data Bases*, *29*, 251–272. https://doi.org/10.1007/s00778-019-00564-x

Chu, H. (2003). *Information representation and retrieval in the digital age*. Information Today: Medford, NJ.

Cohen, D., & Crabtree, B. (2006). "Maximum Variation Sampling." *Qualitative Research Guidelines Project*. Retrieved from https://www.qualres.org/HomeMaxi-3803.html

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

Denzin, N. K. (2007). Triangulation. In *The Blackwell encyclopedia of sociology*. Wiley.

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://doi.org/10.1002/bult.118

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

Fisher, K. E., Erdelez, S., & McKechnie, L. (2005). *Theories of information behavior*. Information Today: Medford, NJ.

Friedrich, T. (2020). *Looking for data: Information seeking behaviour of survey data users* (Doctoral dissertation), Humboldt-Universität zu Berlin.

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. https://doi.org/10.1007/978-3-658-14681-8_17

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), 1–51. 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

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

ISIC Conference. (n.d.). *ISIC conference*. Retrieved from https://www.isic-conference.org/

Koesten, L. (2019). *A user centred perspective on structured data discovery* (Doctoral dissertation), University of Southampton. Retrieved from https://eprints.soton.ac.uk/438583/1/Final_thesis.pdf

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

Koesten, L., Simperl, E., Blount, T., Kacprzak, E., & Tennison, J. (2020). Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies*, *135*, 102367. https://doi.org/10.1016/j.ijhcs.2019.10.004

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. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 1277–1289). ACM. 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

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

Kuhlthau, C. C. (1993). A principle of uncertainty for information seeking. *Journal of Documentation*, *49*(4), 339–355. https://doi.org/10.1108/eb026918

Lafia, S., Million, A. J., & Hemphill, L. (2023). Direct, orienting, and scenic paths: How users navigate search in a research data archive. In *Proceedings of the ACM on human information interaction and retrieval (CHIIR)* (pp. 128–136). ACM. https://doi.org/10.1145/3576840.3578275

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 and Technology*, *54*(6), 570–587. https://doi.org/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

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

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

Phillips, D., & Smit, M. (2021). Toward best practices for unstructured descriptions of research data. *Proceedings of the Association for Information Science and Technology*, *58*(1), 303–314. https://doi.org/10.1002/pra2.458

Robertson, S. E. (1977). Theories and models in information retrieval. *Journal of Documentation*, *33*(2), 126–148. https://doi.org/10.1108/eb026639

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). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. 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

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

How to cite this article: Million, A. J., York, J., Lafia, S., & Hemphill, L. (2025). Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery. *Journal of the Association for Information Science and Technology*, *76*(4), 649–664. https://doi.org/10.1002/asi.24962

APPENDIX A: SCREENING SURVEY QUESTIONS A

Background information

- What best describes your current job/role?
  – Faculty member – Government employee – Journalist – Librarian – Graduate student (e.g., Ph.D., J.D., M.A.) – Undergraduate (e.g., B.A.) – Teacher or lecturer – Other
- With which academic discipline are you most closely affiliated?
  – Anthropology – Area studies – Business studies – Civics – Communication Studies – Criminology – Demography – Development studies – Economics – Education – Environmental studies – Folkloristics – Gender studies – Geography – History – Industrial relations – Information science – International relations – Law – Library science – Linguistics – Media studies – Political science – Psychology – Public administration – Sociology – Sustainable development – None of the above
- How long have you worked in your field?
  – Less than 1 year – 1 to 5 years – 5 to 10 years – 10 to 20 years – More than 20 years
- What research methods do you typically use? For example, you may collect your own data or use data others have collected.
  – Quantitative methods – Qualitative methods – Mixed methods – I'm not a researcher
- How do you usually obtain data for your research? For example, you may collect your own data or use data others have collected.
  – Open-ended text
- How do you use data other than for academic research?
  – Teaching – Writing class papers – Policy or program analysis – Journalism – Other
- What else should I know about you (e.g., your age or education) that relates to the study?
  – Open-ended text

Search experience

- Have you ever searched for data at ICPSR before?
  – Yes – No

Direct paths

- I know what I am trying to find (e.g., a specific dataset)
- When I open datasets, they have what I need
- I get what I want, typically on the first page of my search results
- I stop searching if I do not quickly get what I want
  – Never – Rarely – Sometimes – Often – Always

Orienting paths

- Discovery is harder when I do not know what datasets contain (e.g., variables)
- I use codebooks to decide if a dataset meets my needs
- I switch between searching for data and evaluating datasets I find
- I revise queries based on my search results
- I download data to evaluate it
- I alternate between searching for and browsing data
  – Never – Rarely – Sometimes – Often – Always

Scenic paths

- I look for data to see what is out there
- Data I found earlier is useful later
- I am willing to change my research question(s) when searching
- The data discovery process is more important to me than the result
  – Never – Rarely – Sometimes – Often – Always

APPENDIX B: INTERVIEW PROTOCOL QUESTIONS AND PROMPT TEXT B

[We asked 33 recommender system questions before this point]

We're almost done! To conclude we need to determine which of the following types of information you need to discover datasets. Your answer will probably vary based on context so answer for a typical situation.

- What is this typical situation?
  – Open-ended text

Great. Now, when I list each of the following, provide a score about how important it is. Feel free to explain why you provided your score.

Conduct cutting edge research

- Past papers published with the data
- Information about if data are suitable to conduct a study
- The topic of papers that were recently published using the data
- Where papers were published
- Who has published papers recently reusing the data
  – Not at all – Rarely – Sometimes – Very – Extremely

Accessible data

- Whether data are accessible (available for public download or restricted)
- How long it takes to access the data
  – Not at all – Rarely – Sometimes – Very – Extremely

Trustworthy sources

- Who created the data
- Knowledge about others' experiences using the data
- Information about the data creator's affiliations and reputation
- Who funded the original study
- How often others have downloaded the data
- How often others have viewed the data
- General information about who has downloaded the data (assembled from anonymous click patterns)
  – Not at all – Rarely – Sometimes – Very – Extremely

Relevance

- The variables studied (or concepts, if qualitative research)
- The population studied
- The population sample or methods that were used to collect the data
- The geographic coverage of the data
- How variables or concepts were measured
- When data were collected
- The topic of the original study that produced the dataset
- Unique identifiers like grant numbers
  – Not at all – Rarely – Sometimes – Very – Extremely

Data usability

- What documentation is available (e.g., codebooks)
- Documentation quality (e.g., missing case notes, level of description, distributions)
- How "clean" or well-curated the data are
- The file format(s) of data
- The ability to analyze data online
  – Not at all – Rarely – Sometimes – Very – Extremely

APPENDIX C: FREQUENCY OF RESPONSES ACROSS SURVEY POPULATION WITH QUESTIONS ORGANIZED BY SEARCH PATH TYPE C

TABLE C1 Frequency of responses.

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