Harvard Data Science Review • Special Issue 4: Democratizing Data

Discovering Datasets on the Web Scale: Challenges and Recommendations for Google Dataset Search

Katrina Sostek[1] Daniel M. Russell[2] Nitesh Goyal[3] Tarfah Alrashed[4] Stella Dugall[5] Natasha Noy[4]

[1]Google Research, San Francisco, California, United States of America,
[2]Department of Computer Science, Stanford University, Palo Alto, California, United States of America,
[3]Google Research, New York, New York, United States of America,
[4]Google Research, Mountain View, California, United States of America,
[5]Syllabi, Brooklyn, New York, United States of America

Published on: Apr 02, 2024

DOI: https://doi.org/10.1162/99608f92.4c3e11ca

ABSTRACT

With the rise of open data in the last two decades, more datasets are online and more people are using them for projects and research. But how do people find datasets? We present the first user study of Google Dataset Search, a dataset-discovery tool that uses a web crawl and open ecosystem to find datasets. Google Dataset Search contains a superset of the datasets in other dataset-discovery tools—a total of 45 million datasets from 13,000 sources. We found that the tool addresses a previously identified need: a search engine for datasets across the entire web, including datasets in other tools. However, the tool introduced new challenges due to its open approach: building a mental model of the tool, making sense of heterogeneous datasets, and learning how to search for datasets. We discuss recommendations for dataset-discovery tools and open research questions.

Keywords: data discovery, data search, data reuse, data sharing, open data policy

Media Summary

More and more data are available online, whether from governments, companies, nonprofits, academics, or citizen scientists. However, finding the right dataset for a project can be like searching for a needle in a haystack. In some cases, the desired dataset simply does not exist, but it can be hard to determine that without being able to do an exhaustive search. Most dataset-search tools have too small of a scope, such as dataset repositories limited to datasets in specific disciplines or institutions, or too wide of a scope, such as general search engines where the results cannot easily be filtered to only datasets. Our team at Google developed Dataset Search, which differs from existing dataset search tools because of its scope and openness: *potentially* any dataset on the web is in scope. But how well does this tool work in practice? Can users find datasets they would not have been able to find with other tools? Does the tool help users find datasets more efficiently than they would with other tools? Despite many studies on how well other tools work to find datasets, including general web search tools and dataset-specific search tools, no study has looked at how well a web-scale dataset-specific tool like Google Dataset Search meets user needs. We discuss the implications of our study on how users make sense of new tools, how the tool exacerbates known challenges with the 'messiness' of data on the web, and how data literacy education can be expanded to better prepare people to find, vet, and evaluate real-world datasets and sources.

1. Introduction

In recent years, many disciplines have increasingly relied on open data: machine learning, data science, social science, earth science, health and medicine, data journalism, and countless others (Sambasivan et al., 2021). Just as the web has become an invaluable source of information of any kind, data has also moved online; there

are tens of millions of datasets hosted on thousands of websites covering a dizzying array of topics, geographical locations, and time periods (Benjelloun et al., 2020).

Open-data policies from funders, journals, and governments have helped fuel this increase in online datasets. These policies require that researchers and government agencies share their data (European Commission, 2023; "FAIR Play in Geoscience Data," 2019; Foundations for Evidence-Based Policymaking Act, 2019; "More Research Will Be Publicly Accessible Sooner," 2022). In response to this increase in data sharing, tool developers have created dataset-discovery tools that help dataset creators publish and track their datasets (Chapman et al., 2020). These tools helped incentivize dataset creators to share their data by showing the value of their datasets: how the datasets were originally used and then reused by others (Lane et al., 2022). These changes in the open-data ecosystem—open-data policies and more data discovery tools—have helped tip the scales toward data sharing despite the incentives against sharing (Borgman, 2012; Tenopir et al., 2011; Wallis et al., 2013; Wilkinson et al., 2016). These changes have benefited dataset seekers, who can find more datasets online and find them more easily.

Dataset seekers can pick from a broad set of tools to find and understand online datasets (Chapman et al., 2020; Gregory & Koesten, 2023). These tools range from general purpose web search engines (e.g., Google, Bing) to specialized data portals and search engines for datasets. The specialized tools have dataset-specific features and include only datasets or dataset-related artifacts in their results. Dataset-specific tools include *dataset repositories*, where dataset authors publish datasets, such as academic dataset repositories (e.g., Figshare, Zenodo, or Dataverse). They also include *dataset meta-portals*, where users do not submit datasets directly, but rather the portal aggregates datasets from a set of repositories (Gregory & Koesten, 2023). For example, a government site for a country (e.g., data.gov for the United States) aggregates metadata for datasets from the repositories of multiple agencies and local governments. Finally, at the intersection of general purpose web search tools and dataset-specific tools are *dataset -specific web search tools*. This category of tools provides search capabilities over the subset of pages on the web that describe datasets. The web pages may come from specialized repositories and meta-portals, as well as individual sites from projects and studies that host datasets.

The last category, dataset-specific web search tools, is the newest type of tool with the fewest implementations. A growing body of research on user needs for dataset discovery has evaluated tools of the other three types, but there is no research yet on dataset-specific web search tools (Gregory et al., 2019; Gregory & Koesten, 2023). In addition to the novelty and relative lack of research of this category of tools, we want to study this type of tool because it aims to address a user need identified in previous studies: a tool that combines (1) the reach of general-purpose web search tools that can find any information on the web, and (2) dataset-specific results and features of dataset repositories and meta-portals.

To the best of our knowledge, there is only one current implementation of a dataset-specific and discipline-agnostic web search tool: Dataset Search, which was developed by our team at Google (Noy, 2020). Dataset Search is a specialized search engine that enables users to find datasets on the web by indexing all the pages on

the web that have semantic markup indicating that the page describes a dataset; see Noy, Burgess, & Brickley (2019) for more details on how the tool works. The approach is similar to other vertical search engines for a specific type of information, such as Google Scholar for academic papers. This approach yields a larger and potentially more diverse set of datasets than other dataset-specific tools—a total of 45 million datasets and 13,000 data sources (Benjelloun et al., 2020; Noy & Benjelloun, 2023; Noy, Burgess, & Brickley, 2019).

We sought to better understand the user experience of Dataset Search by conducting semi-structured virtual interviews in March 2022 with 20 participants from a wide range of disciplines, industries, and experience levels; the participants included novices, academics, and industry professionals. During the interviews, participants used Dataset Search to look for datasets of interest to them. Our findings illustrate their successes and challenges with the tool.

This article makes the following contributions:

1. Empirical understanding of user interactions with a dataset-specific web search tool.
2. Understanding of how well Dataset Search met user needs.
3. Understanding of user challenges: building a mental model of the tool, making sense of heterogeneous datasets, and learning the specialized skill of searching for datasets.
4. Recommendations for developers of data discovery tools and open questions for the research community.

2. Background and Related Work

The concept of a dataset is central to the problem of online dataset discovery. Research literature contains many, often vague, definitions of a dataset (Gregory & Koesten, 2023; Renear et al., 2010). What counts as one dataset also varies: one tool may count an entire database as one dataset, while another tool may count many individual slices of the database as different datasets. Alrashed et al. (2021) define a *dataset* as a "collection of data items reflecting the results of such activities as measuring, reporting, collecting, analyzing, or observing" (p. 3). A *dataset page* is a web page that describes a dataset.

With this definition in mind, we now describe different categories of dataset discovery tools to put Dataset Search in context (Section 2.1), discuss the growing body of research on user needs for dataset discovery (Section 2.2), review existing research on building mental models of new tools (Section 2.3), and briefly describe Dataset Search to highlight its uniqueness compared to other tools (Section 2.4).

2.1. Types of Dataset-Search Tools

We can categorize different data-discovery tools based on the scope of what users can find in each tool. Table 1 provides examples for each category.

Table 1. Categories, scope, and examples of dataset-discovery tools.

General versus Dataset-Specific	Tool Type	Scope	Openness	Examples
General Purpose Search Tools	**General Purpose Web Search**	All web pages whether or not they describe a dataset	Open: Any web page	Google Web Search, Bing Web Search
Dataset-Specific Search Tools	**Dataset Repositories**	Datasets submitted by authors or selected by repository maintainers	Curated: Vetted or selected datasets	ICPSR, PANGAEA, Harvard Dataverse, Figshare, Kaggle, GitHub, Zenodo
	Dataset Meta-Portals	Multiple dataset repositories	Curated: Vetted or selected data sources	European Data Portal, data.gov, DataCite Commons, Mendeley Data[a]
	Dataset-Specific Web Search	All dataset pages on the web	Open: Any dataset web page	Dataset Search

[a] Mendeley Data is an example of a meta-portal that allows direct submission of datasets in addition to aggregating across repositories, thus acting as a repository in those cases.

General purpose web search tools, such as Bing or Google, enable users to search for any information on the web, including datasets. These tools do not support dataset-specific features or enable users to limit search results to dataset pages. Bing and Google support filtering results to some file types (e.g., CSV or XLS) (Google Search Central, 2023a; Microsoft, 2022). However, these filters limit the results to specific kinds of dataset download files rather than dataset pages that describe datasets. General purpose tools also can omit datasets in the long tail of the web—pages that receive too little traffic to be indexed at all (Goel et al., 2010; Noy, Burgess, & Brickley, 2019; Search Console Help, 2023).

Dataset repositories allow authors or institutions to publish their datasets.[1] Dataset repositories can also store other materials, such as papers or code. These repositories have many subtypes: discipline-specific (Inter-university Consortium for Political and Social Research [ICPSR] for social sciences, PANGAEA for earth and environmental science, UC Irvine Machine Learning Repository for machine learning); institutional—government, educational, or commercial (NYC Open Data, Harvard Dataverse, OECD Data); repositories with supporting materials for scholarly publishing (Dryad, Figshare, Zenodo); and sites that also host code, competitions, and other features focusing on machine learning (Kaggle, GitHub, Papers With Code, Hugging Face).[2] Repositories have dataset-specific features, such as dataset downloads, previews, or visualizations; number of views or downloads; stars, ratings, or comments; citations or mentions; metadata quality assessments; and search filters based on license, topic, and other metadata fields. Some repositories have

similar features because they use the same underlying software.[3] The scope of a repository is constrained to datasets submitted to the repository or explicitly included by its maintainers. Thus, repositories present a discovery problem of their own: users first need to figure out which repository to use through other means, for example, general web searches, other tools, or recommendations (Gregory & Koesten, 2023; Noy, Burgess, & Brickley, 2019).

Dataset meta-portals aggregate datasets from a defined set of dataset repositories and provide dataset-specific search features across those repositories. Meta-portals determine the set of repositories to include in different ways. For example, data.gov aggregates datasets from U.S. federal agencies and local governments, the European Union (EU) Data Portal collects datasets from individual EU open-government portals, and Mendeley Data curates a list of repositories. Some meta-portals allow users to download datasets from their site, while others link to the repository sites where users can download datasets. Some meta-portals may also allow direct submission in addition to aggregating datasets from other repositories (e.g., Mendeley Data), which means that tools can be both dataset repositories and meta-portals. Thus, meta-portals contain a subset of online datasets, albeit a bigger subset than a single repository. The limitations of meta-portals are similar to those of repositories: users first need to know which meta-portal to use, and thus which repositories (or kinds of repositories) they contain (Gregory & Koesten, 2023; Noy, Burgess, & Brickley, 2019).

Dataset-specific web search tools are similar to general purpose web search tools but provide dataset-specific search features and limit their results to dataset pages. The scope of these tools is dataset pages on the web: the union of dataset pages in repositories, meta-portals, and individual pages on the web that describe datasets (e.g., web pages of a project or a study). Dataset-specific web search tools differ from repositories and meta-portals in the amount of control they have over the datasets they contain. Repositories and meta-portals control the datasets that they include either through user management or, more explicitly, by selecting which repositories to aggregate. However, dataset-specific web search tools do not control or curate the datasets that they contain: any dataset page on the web is potentially in scope. The only tool that currently exists in this category is Dataset Search (Cieslewicz et al., 2018; Gregory & Koesten, 2023; Sansone et al., 2017).[4] Dataset Search indexes dataset pages from all the repository and meta-portal examples listed in Table 1.[5]

2.2. Previous Research on Dataset Seekers' Needs

Starting in 2002 and gaining steam in 2012, a growing body of research on human-computer interaction (HCI) has examined how dataset seekers find, evaluate, and use online datasets (Gregory et al., 2019; Gregory, Cousijn, et al., 2020). These studies focus on different user types: those in specific disciplines, work sectors, or levels of experience.

Dataset-discovery studies used four main methods to approach the question of how users find datasets online:

1. User interviews
2. User surveys

3. Logs analysis of specific tools
4. Tool reviews cataloging tools

Table 2 crosses these four methods against the four tool types in Table 1. Note that this table does not contain an exhaustive list of studies about dataset discovery; it includes only the studies most relevant to our research questions.[6]

Table 2. HCI studies of dataset-discovery tools.

Tool Type	User Interviews	User Surveys	Logs Analysis	Tool Reviews
General Purpose Web Search	• Gregory, Cousijn, et al., 2020; • Koesten et al., 2019; • Koesten et al., 2017; • Krämer et al., 2021; • Liu et al., 2022	• Gregory, Groth, et al., 2020[a]		• Chapman et al., 2020; • Gregory & Koesten, 2023
Dataset Repositories	• Gregory, Cousijn, et al., 2020; • Kern & Mathiak, 2015; • Koesten et al., 2017; • Krämer et al., 2021; • Liu et al., 2022	• Gregory, Groth, et al., 2020; • Wu et al., 2019[b]	• Kacprzak et al., 2017; • Koesten et al., 2017; • Xiao et al., 2020	• Chapman et al., 2020; • Gregory & Koesten, 2023; • Koesten et al., 2019
Dataset Meta-Portals	• Dixit et al., 2018; • Gregory, Cousijn, et al., 2020; • Krämer et al., 2021	• Gregory, Groth, et al., 2020; • Wu et al., 2019	• Ibáñez & Simperl, 2022; • Sharifpour et al., 2023	• Chapman et al., 2020; • Gregory & Koesten, 2023; • Koesten et al., 2019
Dataset-Specific Web Search				• Chapman et al., 2020; • Gregory & Koesten, 2023

[a] The survey in Gregory, Groth, et al. (2020) asks about "data search engines" separate from web search engines and dataset repositories. Dataset Search was launched on October 5, 2018, and the survey was fielded from September to October 2018 (Gregory, 2020). It is possible that survey respondents had heard of the tool, but it was not referenced in the survey, so we do not list this study under dataset-specific web search.

[b] The survey in Wu et al. (2019) supplemented a case-study approach that aggregated case studies from five tools, including dataset repositories and meta-portals, and standards working groups.

The only previous user research on dataset-specific web search tools are reviews that describe Dataset Search but are not based on empirical evidence.[7] The dearth of research on dataset-specific web search tools may be

due to their newness or lack of awareness about them (Gregory & Koesten, 2023). Although the user studies are not about dataset-specific web search tools, their findings are relevant and informed our study design.

Despite the wide range of different approaches and types of participants, the studies converged on several findings: Dataset discovery is one step in the complex process of using data for projects, dataset evaluation is a complex activity on its own, trustworthiness and literature search are important parts of evaluating datasets, and recommendations from the literature apply to both tool developers and the open data ecosystem.

Dataset discovery is one step in the complex process of using data for projects. Several studies examined how data discovery fits into a larger context: why participants seek data and what they plan to do with it (Koesten et al., 2017; Krämer et al., 2021; Muller et al., 2019). Koesten et al. (2017) present a model for the process of working with datasets that is highly iterative and can involve repeatedly skipping back one or several steps. These steps include defining the task, searching for datasets, evaluating datasets for use, exploring datasets in depth, and then using the data. Studies showed that users often started with general search tools, investigated specific datasets, then went back to general search tools to refine their searches (Gregory, Groth, et al., 2020; Ibáñez & Simperl, 2022; Koesten et al., 2017; Krämer et al., 2021). Data discovery can also be a collaborative process where different team members contribute their domain or technical knowledge (Choi & Tausczik, 2017; Erete et al., 2016; Passi & Jackson, 2018).

Evaluating datasets for use is a complex activity. Many studies highlight the shortcomings of existing metadata to evaluate datasets for use: it has varying levels of quality (Koesten et al., 2017) and is often incomplete (Davies & Frank, 2013; Gregory, Groth, et al., 2020; Pasquetto et al., 2019), inaccurate, not meaningful (Koesten et al., 2021), and non-standardized (Dixit et al., 2018; Gregory, Cousijn, et al., 2020; Krämer et al., 2021). Additional information about datasets is spread across the underlying data itself and related artifacts, such as articles or documentation (Gregory, Groth, et al., 2020; Koesten et al., 2019; Koesten et al., 2021). Dataset seekers may find multiple versions of a dataset that they may want to evaluate, some of which may be exact replicas (Liu et al., 2022).

Trustworthiness is an important factor in evaluating a dataset. Across many user studies, assessing quality of a dataset was intertwined with establishing trust in the datasets, data authors, and data providers (Gregory, Cousijn, et al., 2020; Gregory, Groth, et al., 2020; Koesten et al., 2017; Krämer et al., 2021). Unlike publishing an article in a peer-reviewed journal or conference, publishing a dataset does not usually require a rigorous vetting process that would confer trust. Previous studies identified several factors that help establish trust in a dataset: the reputation of the data authors, linked articles or other usages that imply others trust a dataset (Färber, & Lamprecht, 2021; Koesten et al., 2021; Krämer et al., 2021; Lowenberg, 2022; Yoon, 2017), recommendations from colleagues, detailed documentation, transparent methods, and the quality of the underlying dataset itself (Faniel et al., 2012; Gregory, Groth, et al., 2020; Passi & Jackson, 2018; Yoon, 2017). Other studies found that users trusted government, educational, and nonprofit institutions more than other types

of institutions (Gregory, Cousijn, et al., 2020; Krämer et al., 2021). Yoon and Lee (2019) described the repositories containing datasets as playing an "intermediary role [that] contributes to data reusers' trust."

Literature search is often an important part of the dataset search process. Dataset seekers use literature search tools, such as Google Scholar and PubMed, to discover, evaluate, and build trust in datasets (Krämer et al., 2021; Zimmerman, 2007). Datasets can be introduced, used, or mentioned in papers; or hosted on journal sites as supplementary material (Walters, 2020). Some repositories, such as Papers With Code, are both literature and dataset search tools that connect papers and datasets to each other. Kern and Mathiak (2015) found that searching for datasets took more time and effort than searching for literature, but empirical social scientists were willing to spend that additional time because "choosing a dataset is a much more important decision for a researcher than choosing a piece of literature" (p. 207). Krämer et al. (2021) found that students were better prepared to search for literature than datasets: "The search for literature is a very common task, but searching for datasets is not. While literature search is often taught at universities, similar courses for handling data are much rarer" (p. 191).

Recommendations apply to dataset-search tools and the open data ecosystem. These previous studies recommended improving existing dataset-search tools or creating new tools to address user needs. Several studies mentioned that new dataset-specific web search tools would make it possible for dataset seekers to efficiently conduct searches over multiple repositories, including repositories they were previously unaware existed (Gregory, Cousijn, et al., 2020; Gregory, Groth, et al., 2020; Krämer et al., 2021; Wu et al., 2019). Liu et al. (2022) suggested building a tool to "Provide federated search of datasets from multiple repositories [...] to reduce the user's burden of searching multiple sources and deal with duplicates of records" (p. 6). Although studies discussed the existence of Dataset Search, we only found one study where participants mentioned that they used it (Liu et al., 2022). The studies also presented recommendations to improve the open data ecosystem: dataset authors and providers could standardize and improve metadata (Fenner et al., 2019; Koesten et al., 2017), and institutions and professional organizations could teach how to search for and evaluate datasets (Krämer et al., 2021). Our findings align with these recommendations (Section 4).

2.3. Previous Research on Building Mental Models

Dataset Search is a new type of tool, which requires users to build a mental model of how it works, whether implicitly or explicitly. Previous studies about building mental models focused on tools that were not dataset specific, such as general web search tools, online library systems, and institutional repositories for research artifacts that include but are not limited to datasets.

Many studies cite Norman's (1983) definition of a mental model applied to HCI, summarized by Holman (2011) as "an internal cognitive representation of a tool or system that helps one master it" (p. 20). Zhang (2013) describes the process of building a mental model as dynamic, starting with existing mental models and updated over time while using a new tool. As users interact with a new tool, they update their mental model by

"assimilating new concepts, phasing out previously perceived concepts, and modifying existing concepts" (p. 169; see also Rieh et al., 2010). Mental models are difficult to study; methods include verbal accounts, drawing, and observing user errors to identify gaps (Zhang, 2008).

Google Web Search loomed large in many studies of information retrieval (IR) systems. Rieh et al. (2010) stated that the tool "dominates people's approaches to new IR systems and decreases their willingness to learn and explore new systems. Rather than trying to understand the nature and scope of a new system, people simply begin entering keywords into the search box without any prior exploration of or orientation to the system" (p. 173). Khoo and Hall (2012) found that users idealized Google Web Search as a model for how a digital library tool should work, but that was often based on an inaccurate "folk model" rather than an accurate "technical model" (p. 10). Zhang (2008) found that undergraduate students' mental models of search engines varied in sophistication, were sometimes naive and incorrect, and "tended to ignore the invisible part of the Web, such as metadata" (p. 2097).

2.4. Dataset Search's Approach to Dataset Discovery

Dataset Search differs from other tools because its scope is any dataset page on the web. As shown in Table 1, the scope of Dataset Search is the union of the scope of meta-portals, repositories, and dataset pages outside of those tools—if they have markup (Figure 1).

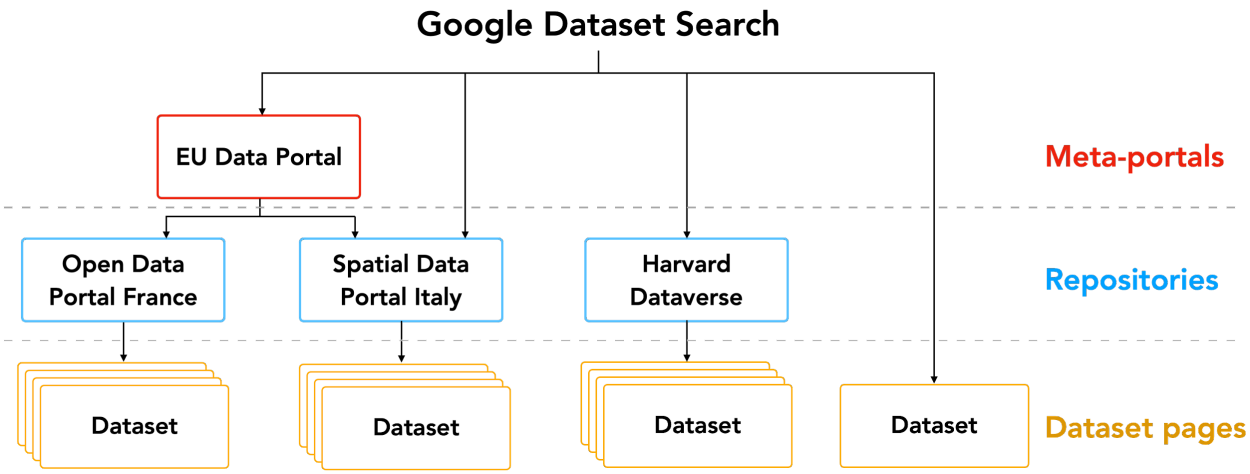

Figure 1. Scope of Dataset Search.

Dataset Search includes individual dataset pages, dataset pages from repositories (e.g., Harvard Dataverse), and meta-portals (e.g., EU Data Portal). Meta-portals, in turn, include datasets from individual repositories, many of which are also indexed directly in Dataset Search.

The large, heterogeneous, and open set of datasets contained in Dataset Search translates into a different user experience than other dataset-search tools. Search results can link to meta-portal pages, repository pages, and project or study pages. A single search result may link to multiple sites containing the same dataset; for

example, the dataset may be available in a repository, on a local government site, and a meta-portal for that country.

Dataset Search's initial search page[8] looks similar to Google Web Search and Google Scholar: a page with a search box, example queries, and a 'learn more' link. After the user types in a query, they see a list of results for their query (Figure 2), select a result, then see details about the dataset and links to dataset pages at the sources.

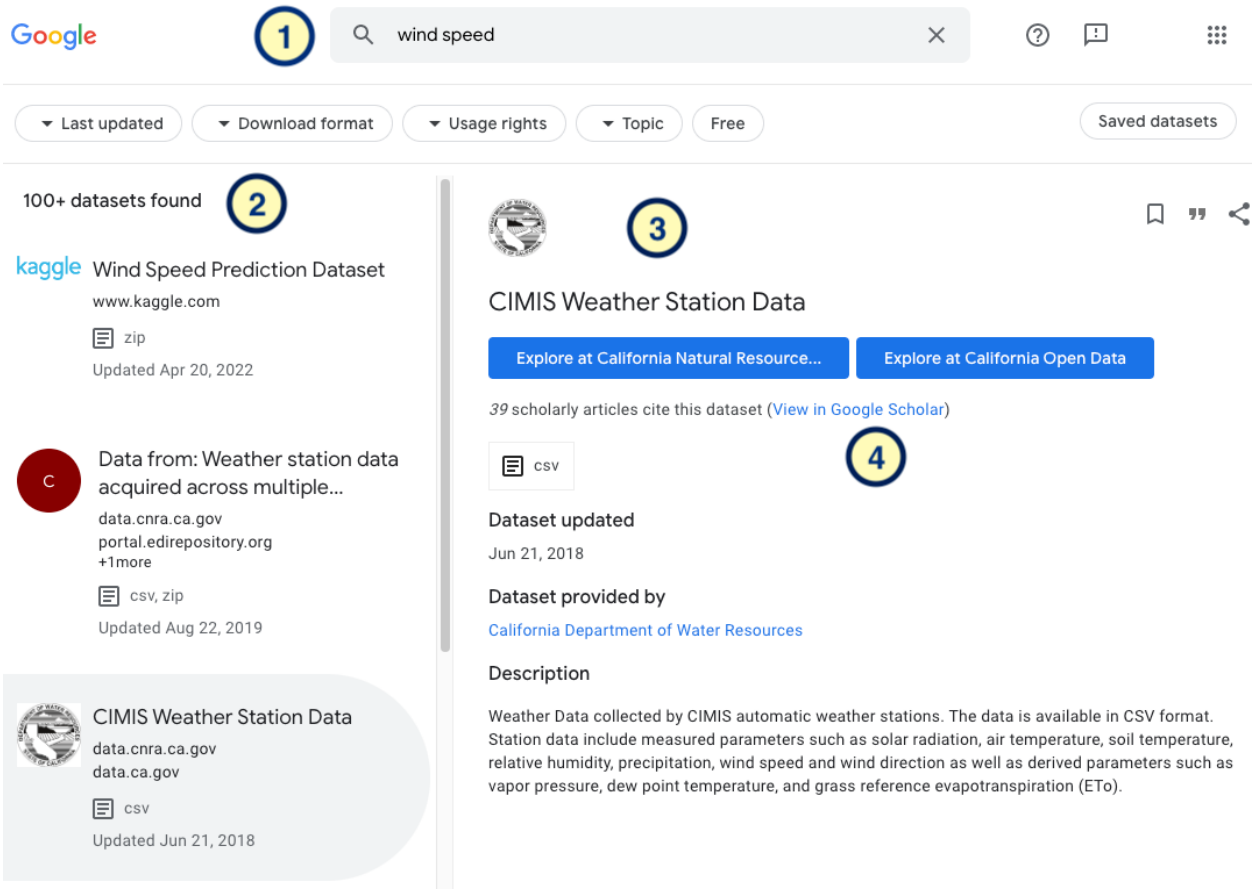

Figure 2. Dataset Search user interface. Screenshot from August 18, 2022.

In Figure 2, (1) the user types a query "wind speed." (2) There are 100-plus results in the list of results (scrolling loads more results). (3) The right-hand side shows a detailed view of the search result that the user selected on the left-hand side, including metadata such as file type and description. The solid blue buttons link to two different sources for this dataset: the California Natural Resources Agency (data.cnra.ca.gov) and the California Open Data Portal (data.ca.gov). (4) Thirty-nine scholarly articles mention this dataset, and the page links to a Google Scholar search with those results.

When a user clicks on a solid blue button, they leave the tool and go to a dataset page on a separate site. On that page, they may find more information about the dataset, including links to additional pages with more

information, such as documentation, methodology, or associated papers.

Dataset Search finds dataset pages by using the Google Web Search crawl to find pages that include the relevant schema.org[9] markup. In some cases, Dataset Search can crawl long-tail pages that are not included in Google Web Search. The metadata for each dataset is derived from the semantic markup that the data provider added to their dataset page. Dataset Search then standardizes the metadata, combines similar fields, and augments the metadata with links to Google Scholar and the Google Knowledge Graph (Noy et al., 2019). The tool also finds dataset duplicates (duplicates on the same site) and replicas (duplicates on different sites). For more details, see Noy, Burgess, & Brickley (2019).

Dataset Search relies on an open ecosystem where website owners add semantic markup to their pages with dataset metadata; there is no submission, validation, or vetting process. The tool encourages data providers to follow best practices (Google Search Central, 2023b), but they have wide latitude within those guidelines. For example, providers can fill in different metadata fields or different levels of detail. The tool displays only the non-empty fields, which vary by dataset. Dataset Search performs only the following checks: (1) the metadata contains the minimal fields, name and description (Benjelloun et al., 2020), and (2) the pages were correctly annotated as datasets and are not spam (Alrashed et al., 2021).

The large and varied scope of Dataset Search addresses a previously identified user need: a dataset-specific web search tool that can potentially find any dataset on the web (Gregory, Cousijn, et al., 2020; Gregory, Groth, et al., 2020; Krämer et al., 2021; Liu et al., 2022). The tool has a wider variety of data sources than other types of tools that control the data sources they contain. This open approach of Dataset Search fuels its large and diverse scope, but may exacerbate the challenge of nonstandard metadata across the ecosystem, as described in the previous literature (Section 2.2).

3. Methods

We focused on two research questions:

- How well does Dataset Search support user needs for dataset discovery as the only dataset-specific web search tool?
- What advantages and challenges do users face with Dataset Search due to its uniquely large scope and open approach?

3.1. Study Design

To address the research questions, we wanted participants to complete a nontrivial task that they were genuinely interested in, had thought of themselves, and had close proximity to real-world usage (Russell & Grimes, 2007). Thus, we asked participants to think of their own searches when using Dataset Search, which meant that the participants did the same type of task but not the same specific task.

We recruited 22 participants from a range of fields, roles, and work sectors. One interviewer conducted 90-minute semi-structured virtual interviews and direct observation with screen shares. In February 2022, we conducted two pilot interviews to test the timing and focus of the questions. In March 2022, we conducted 20 interviews. We analyzed the 20 non-pilot interviews. The interviews had three phases:

- Discussing participants' background and motivations as dataset seekers (~15 minutes).
- Walking through a recent dataset search with the tools the participants used—none of which were Dataset Search—and discussing their challenges (~45 minutes).
- Using Dataset Search to complete the search they discussed in Phase 2 and/or other searches that the participants wanted to try (~30 minutes).

This article discusses the findings from Phase 3 of the interviews (average length = 27 minutes). The interviewer asked participants if they had used the tool before and, if not, introduced the tool as "Google Scholar for datasets" or a "search engine for datasets." While this may have primed the participants to think of the tool in a certain way, we felt that using a metaphor would help the participants quickly get a baseline mental model of the tool before refining their model by using the tool (Brandt & Uden, 2003; Perkins & Salomon, 1992; Zhang, 2013). The tool's visual resemblance to Google Web Search and Google Scholar, in addition to being a Google product, may have possibly had a similar priming effect on its own.

After the participant used the tool, the interviewer asked questions such as "Would it have helped at all in finding the right dataset if you had used it in past research?" "Would it have been more, less, or equally effective as your routine?" and "If you had a magic wand and could build any feature that you like into it, what would it be? Would you remove any existing features?"

3.2. Participants

To address the research questions and fill the gaps from previous studies, we recruited a mix of participants from different disciplines, professional roles and sectors, and at different stages of their careers, including students. We emulated the participant mix of two previous studies that used the same participants (Koesten et al., 2017; Koesten et al., 2019), however, we also included an undergraduate student in addition to graduate students. We recruited a broader mix of participants than previous studies that had a majority from academia or research institutions (Dixit et al., 2018; Gregory, Cousijn, et al., 2020; Gregory, Groth, et al., 2020; Kern & Mathiak, 2015; Krämer et al., 2021; Liu et al., 2022; Wu et al., 2019) or from specific disciplines (Dixit et al., 2018; Kern & Mathiak, 2015; Krämer et al., 2021).

We recruited participants using a database of volunteers for user research studies across Google.[10] We sent a screening survey that asked about demographics, profession, and how frequently the respondents searched for datasets. The screening survey did not ask specifically about Dataset Search. To be included in the study, we required that respondents self-reported that they searched for datasets at least once a month. Out of the 200 respondents, 137 (71%) participants fulfilled that requirement. Of those respondents, we chose participants that

would achieve a mix of sectors, fields, ages, and genders (Table 3). To improve the representation of data journalists and undergraduate students, we supplemented our recruiting with a second round using LinkedIn and the COVID Tracking Project's Slack channel, which yielded three more participants.

Table 3. List of participants in the study.

P	G	Age	Field	Role	Sector	Frequency
1	F	31-40	Epidemiology	Research data analyst	Research	Multiple a day
2	F	24-30	Management consulting	Senior associate	Corporate	Multiple a day
3	F	24-30	Market research	Consultant, firm owner	Independent	Multiple a day
4	F	24-30	Medicine	Research technician	Education	Multiple a day
5	F	24-30	Neurology	Postdoctoral fellow (x)	Research	Multiple a day
6	F	41-50	Technology	Business intelligence analyst	Corporate	Multiple a day
7	M	24-30	Technology	Data analyst (x) (+)	Corporate	Multiple a day
8	M	31-40	Technology	Image science engineer (+)	Corporate	Multiple a day
9	M	18-23	Health Policy	Undergraduate student (+)	Education	Once a day
10	F	24-30	Biotechnology	Associate scientist	Corporate	Multiple a week
11	M	31-40	Biotechnology	Senior scientist	Corporate	Multiple a week
12	M	18-23	Industrial psychology	Graduate student	Education	Multiple a week

13	F	31-40	Insurance	Executive assistant (+)(x)	Corporate	Multiple a week
14	M	18-23	Public Health	Ph.D. student	Education	Multiple a week
15	F	24-30	Psychology	Health science specialist	Government	Multiple a week
16	M	31-40	Technology	Marketing analytics manager	Corporate	Multiple a week
17	M	31-40	Journalism	Data journalist	Corporate	Once a week
18	F	24-30	Journalism	Data journalist (*)	Independent	Once a week
19	F	24-30	Environmental science	Horticulture technician	Education	Once a month
20	F	24-30	Philanthropy	Data analyst (*) (+)	Nonprofit	Once a month

Note.

(*) denotes two participants who had used Dataset Search at least once before the study, but not regularly.

(x) is for three participants who reported that they generally had difficulty concentrating or focusing on tasks.

(+) is for five participants who worked on hobby or volunteer data projects outside of their professional roles.

Table 3 shows the participant number (P), gender (G), age, field, role, sector, and frequency of searching for datasets.

During the screening process, participants were informed that they would receive $100 for their time, redeemable from an online catalog that included cash equivalents, retail items, or donations to charity. All the participants signed informed consent forms and received their incentives.

3.3. Analysis and Coding

To create insights from interview data, we systematically coded the relevant protocol segments using an iterative, inductive, and reflexive qualitative thematic analysis method (Saldaña, 2021). We used a descriptive coding method in a series of thematic analysis coding passes, then tabulated each discrete code (e.g., ‘Building a mental model of the tool’) for each participant. We used an iterative and inductive approach to develop the codes after first reviewing all the interviews.

Three senior researchers coded relevant portions related to the research questions from a sample of the interviews. They discussed discrepancies, revised the codes, and iteratively continued this process until the coding converged with a high level of agreement. One researcher then coded the rest of the interviews. We then identified three themes using an inductive approach from the codes, which we discuss in the next section.

4. Findings

Our findings correspond to three themes:

1. ***Building a mental model of the tool***: Participants based their initial understanding of the tool on their knowledge of more familiar tools. As they used the tool, they updated their mental model in response to their experiences. This approach helped most participants understand the tool, but a few initially confused the tool's purpose with that of other tools (Section 4.1).
2. ***Making sense of heterogeneous datasets***: Participants validated the findings from previous literature that metadata is incomplete and inconsistent, accessing the underlying data is critical but can be challenging, and trustworthiness is an important part of the dataset-search process. Dataset Search exacerbates these issues because it shows results from thousands of dataset repositories and meta-portals, each with different metadata and interfaces (Section 4.2).
3. ***Learning how to search for data***: A few participants were unfamiliar with dataset-specific concepts referenced in the tool, such as data formats and usage rights. Participants recommended educating users on how to search for data through tutorials and in-tool help (Section 4.3).

We describe the findings for each theme in detail below.

4.1. Building a Mental Model of the Tool

Understanding Dataset Search. At the beginning of the study, we introduced the tool in terms of more familiar tools. For the participants who had not used Dataset Search before, we described the tool as "Google Scholar for datasets" if they were familiar with Google Scholar (n = 14) or "a search engine for datasets" (n = 4). The participants built a more complete mental model of the tool through trial and error with 3.2 queries on average per participant (median = 2.5, minimum = 1, maximum = 11).

As they used the tool, nine participants described their evolving mental model of the tool by comparing it to other tools. (P8) "*Basically, it's a specialized Google, right?*" (P4) "*It kind of reminds me of what [you] would come up with a PubMed search with titles of different papers here. But in this case, it's datasets, so it's specifically searching for a dataset.*" P6 asked how Dataset Search differed from Google Web Search and why they were separate tools. (P6) "*Is [Dataset Search] going to be a separate URL instead of Google [Web Search]? Or is [Dataset Search] something that it's going to be part of a widget here? [pointing to the list of categories (News, Images, etc.) at the top of Google Web Search results].*"

Two participants initially did not understand the overall purpose of Dataset Search because they confused the tool's purpose with Google Web Search or literature search tools. P2 used a query she had recently used in Google Web Search, got zero results, refined the query twice and got zero results, then refined it again and got 39 results. When she looked at one of the search results, she realized that she had been searching for a single data point—an average—instead of an entire dataset. (P2) "*This is a Google search. [...] Maybe it's on me. [...] That's probably an adjustment [I would make] when I start using the tool.*" Similarly, P5 adjusted her mental model after trying five queries: "*I'm using it as a Google Scholar rather than a dataset engine.*" P2 and P5 had both self-reported that they searched for multiple datasets a day.

Six participants mentioned that they saw search results that looked like papers. (P11) "*It looks like it's pulling up papers. [...] Is it supposed to be leading to use of papers or to [...] the actual data?*" After seeing papers in the results, three of the six participants questioned their previous mental model of how the tool differed from literature search tools or Google Web Search. (P14) "*I'm seeing reports. Like I'm seeing this type of stuff that I might see when I search on Google.*" In some of these cases, participants were able to find datasets associated with papers after clicking through to the dataset pages.

By the end of the interviews, all participants were able to build a mental model that matched the tool's functionality by trying out queries, seeing the results, and thinking out loud about how the tool worked.

Usefulness of the tool. The participants used queries that they thought of in the moment or had discussed earlier in the interview, such as [Greenhouse gas emissions food system], [Protein amino acid], or [Asian American health]. Six participants with domain expertise used queries where they already 'knew the answer' as a kind of test to make sure they understood how the tool worked. The other 14 participants used queries where they did not yet know the answer to test out the tool. P1 described using both approaches in sequence: "*This is me searching not knowing if the data exists. I knew the DREAMS data exists. [...] But now this is – does this data even exist in the world?*"

The six participants who used queries where they already knew the answer said that the tool passed their test: the results contained at least one of the datasets they expected to see. The other 14 participants found useful datasets but were less definitive about the tool's performance because their expectations were not as well defined as those of the other participants. Six participants said they would use Dataset Search as a secondary resource in parallel with or before using other search tools. (P3) "*I don't think it would replace those initial [Google Web Search] queries that I would do. But it would definitely be in parallel. I would probably do that Google Search first still, to just see, what questions should I be asking, or when I go into this Dataset Search, who's maybe a reliable resource on the topic.*"

Overall, the participants had a positive reaction to the tool, but some said they would want the tool's functionality to be more like other tools if they were to use it in the future. (P11) "*I'm trying to mirror my experience with Google [Web] Search and what I find helpful about that.*" For example, he and two other

participants requested a specific feature from Google Web Search: a visual indication of which query terms match and do not match in each result. Later, P11 referenced more specialized tools: "*Making it a hybrid Google Scholar [and] Google Image [with charts] I think would be valuable.*"

What results are in Dataset Search? Of the 64 queries that participants tried, 52% had 100-plus results, which is the maximum number of results that the tool shows for a query. Of the 20 participants, only three participants got zero results for any queries, and two of those participants were initially confused about the purpose of the tool. Sixteen of the other 17 participants said that the number of results made sense for their query, and they were able to get more results by refining the queries.

One participant said she was surprised because the number of results for her four queries was lower than she expected. (P10) "*Seeing [...] there are only like 30 results, thinking about how many there possibly could be on the internet, it's kind of surprising how few there are.*" Earlier in her interview, P10 said that she would have found Google Web Search results to be more useful than Dataset Search's results because they would contain additional results from Google Scholar, Wikipedia, and PubMed, among others. Her mental model of the tool may have included not only datasets but also contextual information and artifacts about datasets (e.g., papers).

Only three participants asked for details about how results get into the tool: who fills out the metadata and if datasets are submitted and validated, as they are in dataset repositories. P16 speculated that "*maybe the providers*" populated the metadata.

Dataset replication. Dataset Search detects datasets that are replicas (the same dataset on different sites) and collapses them into one search result. A search result with replicas displays multiple links for each source (e.g., the two solid blue buttons in Figure 2 link to two different sources for the same dataset).

Five participants who searched for datasets once a day or once a week mentioned the multiple blue buttons, and four asked about them or said they were confused by them. (P16) "*One thing I'm still not sure is which one [of the four blue buttons] do I use up here? It seems like the first thing I do is I click all four to see what's different, right?*"

P6 asked about the ordering of the blue buttons. One participant wanted to know if one of the blue buttons was the primary source of the data. (P14) "*What would be nice is if I knew who collected, who owns the data, or the entity. Like sometimes I know that it's a CDC data set or an NIH data set, but I mean, this I can guess because of the .gov.*" Meanwhile, P8 saw the multiple blue buttons as a positive signal: "*It has lots of different links to other people using it, so I would bet this is probably the standard [dataset to use].*"

4.2 Making Sense of Heterogeneous Datasets

Metadata shortcomings and inconsistencies. Sixteen participants validated previous findings about dataset metadata's shortcomings (Section 2.2). Metadata helped participants decide whether to click through to dataset

pages, but it was insufficient to make sense of datasets and decide whether to use them.

Metadata in Dataset Search is not consistent across datasets because the results come from a wide range of sources. For example, the description field in some results contained a paper abstract (n = 4), and participants were not sure if the date updated field was when the dataset was originally collected, updated, or when the web page was updated (n = 4). Five participants noticed different amounts of metadata across search results. Dataset Search displays only the non-empty metadata fields, so metadata inconsistency translates into different fields shown in the tool and less effective filtering (Figure 2). Participants had a higher mental load due to inconsistent metadata, which made it harder for them to compare datasets.

Accessing the underlying data. Unlike dataset repositories, Dataset Search does not support downloading the underlying dataset from within the tool. Twelve participants asked if the tool had a download link, dataset preview, or visualization; or they said that those would be helpful. Seven participants said that they wanted to reduce the amount of time, number of clicks, or steps it took to access and explore the underlying data. (P3) "*If I could pull this directly into data visualization so that I'm not like, download the CSV, open Power BI, import into Power BI, and then start building those visualizations—that's just a very time laborious process [...] that's just too in-depth for me to do without knowing I'm going to use it.*"

When participants needed to visit different sites to download the data, they experienced a high mental load, especially with unfamiliar sites (n = 5). P1 described her frustration while trying to find a data download link on a data repository page: "*Metadata is not the data. But it's actually not so clear how to get the data [...] This makes me think that the data is right here, but it's not [...] This is frustrating. I mean, granted, I've given it 30 seconds. But it should be easier than that. [...] They're giving you the data without giving you the data.*"

Search results from meta-portals added more steps and mental load to the process of finding a download link (Figure 1). For example, P19 clicked on a meta-portal result, which linked to another meta-portal, which linked to a journal article with data; she had to click three times from the tool to get to the dataset source. P20 followed a link to a meta-portal and mistakenly thought that the meta-portal was the repository that stored the dataset. Both participants searched for datasets once a month (the least frequent amount).

Building trust in unfamiliar data sources. Dataset Search enables users to discover new sources of data. Ten participants remarked on unfamiliar sites in the search results. (P10) "*I have no idea what this site is.*" Of those 10 participants, four searched for data multiple times a day, one searched once a day, three searched once a week, and two searched once a month. Six participants were curious about the unfamiliar data sources and saw it as an opportunity to discover new sources, especially sites that showed up multiple times in the results. (P3) "*This, dataset guide, all of these look really interesting. [...] Maybe they're not reputable, but they have reputable sounding names.*" However, P3 weighed her curiosity about new data sources against the time and mental load that it would take to vet them.

Other participants were not as interested in exploring or even seeing unfamiliar or untrusted sources in search results. Four participants suggested filtering results to only trusted data sources, such as .gov, .edu, or .org domains; to sources cited in or associated with papers; or to specific vetted sources like scholarly journals. (P13) "*If there was a way to filter for .gov, .org, [...] that would save me lots of time because I lived on .gov and .orgs. [...] Google Scholar has that feature already in it because you're looking for accredited sources.*"

4.3. Learning How to Search for Datasets

Searching for datasets is a skill within data literacy[11] (Krämer et al., 2021) that also requires specialized domain knowledge (Adhikari et al., 2021; Coughlan, 2020; Dixit et al., 2018; Kross & Guo, 2019). Participants were unfamiliar with several concepts referenced in the tool, such as specific data formats (n = 6) and usage rights (n = 3). Other participants shared their ideas about helping users learn how to use the tool. (P9) "*Adjusting the amount of information provided at first glance could be useful, like a novice view versus a pro view [for dataset results to show] data at different levels for people to digest.*" Three participants suggested educational materials to help new users and students (e.g., guides, tutorials, or videos). Two participants suggested making the tool a community platform for people to learn from one another and share how they used datasets.

5. Discussion: Challenges and Recommendations

Our findings suggest that searching for datasets with Dataset Search has benefits but poses new challenges. A significant benefit is that participants can discover potentially useful datasets and data sources that they did not know existed. A challenge is that some participants found it difficult to build a mental model of an open, web-scale, dataset-search tool. Dataset Search exacerbated some of the challenges described in previous research, such as making sense of different datasets, accessing the underlying data, and developing trust in datasets (Section 2.2).

To successfully address the key challenges that we identify in this section, two sides of the ecosystem—developers of dataset-discovery tools and dataset providers—will need to work together to improve the user experience. Previous research has already made many recommendations for dataset providers and the open data ecosystem (Section 2.2), so we will focus on the recommendations for dataset-discovery tools and researchers. While some of these recommendations are specific to Dataset Search, many apply to data-discovery tools in general; we specify the scope of each recommendation below. We summarize these recommendations at the end of each section: Building a mental model of the tool (Table 4), making sense of heterogeneous datasets (Table 5), and dataset discovery is a skill and needs to be a part of data literacy (Table 6). After conducting this study in March 2022, we made several changes to Dataset Search based on the findings; we will note those product changes below in the relevant sections.

5.1. Building a Mental Model of the Tool

Dataset Search is the first implementation of a new type of tool: dataset-specific web search. As with any new tool, users will build their initial mental model of the tool based on familiar tools (Section 2.3). We observed participants' mental models evolve as they tried different queries, explored the results, and clicked through to dataset pages. Participants had a few common questions while developing their mental models:

1. How are Dataset Search and Google Web Search different?
2. Why are some expected results and additional context missing?
3. Why do some results have multiple links (the blue buttons)?

We will discuss each question in turn.

Dataset Search and Google Web Search. We used two methods to observe how the participants built mental models: verbal accounts and observing user errors to identify gaps (Zhang, 2008). One recurring theme was questions about the difference between Dataset Search and Google Web Search and why they were separate tools. This issue is unique for Google products and is not as relevant to non-Google dataset-search tools. However, Google Web Search is often the first starting point for users looking for datasets, especially those who do not already know about specialized search tools (Section 2.2). So, all dataset-search tools have a similar challenge to clarify to users when it makes sense to use general web search tools versus their tools, the pros and cons, and how to use both tools for broad and narrow searches.

After the user study, we partially addressed this issue by integrating results from Dataset Search directly into Google Web Search for some dataset-related queries (Figure 3; Noy & Benjelloun, 2023).

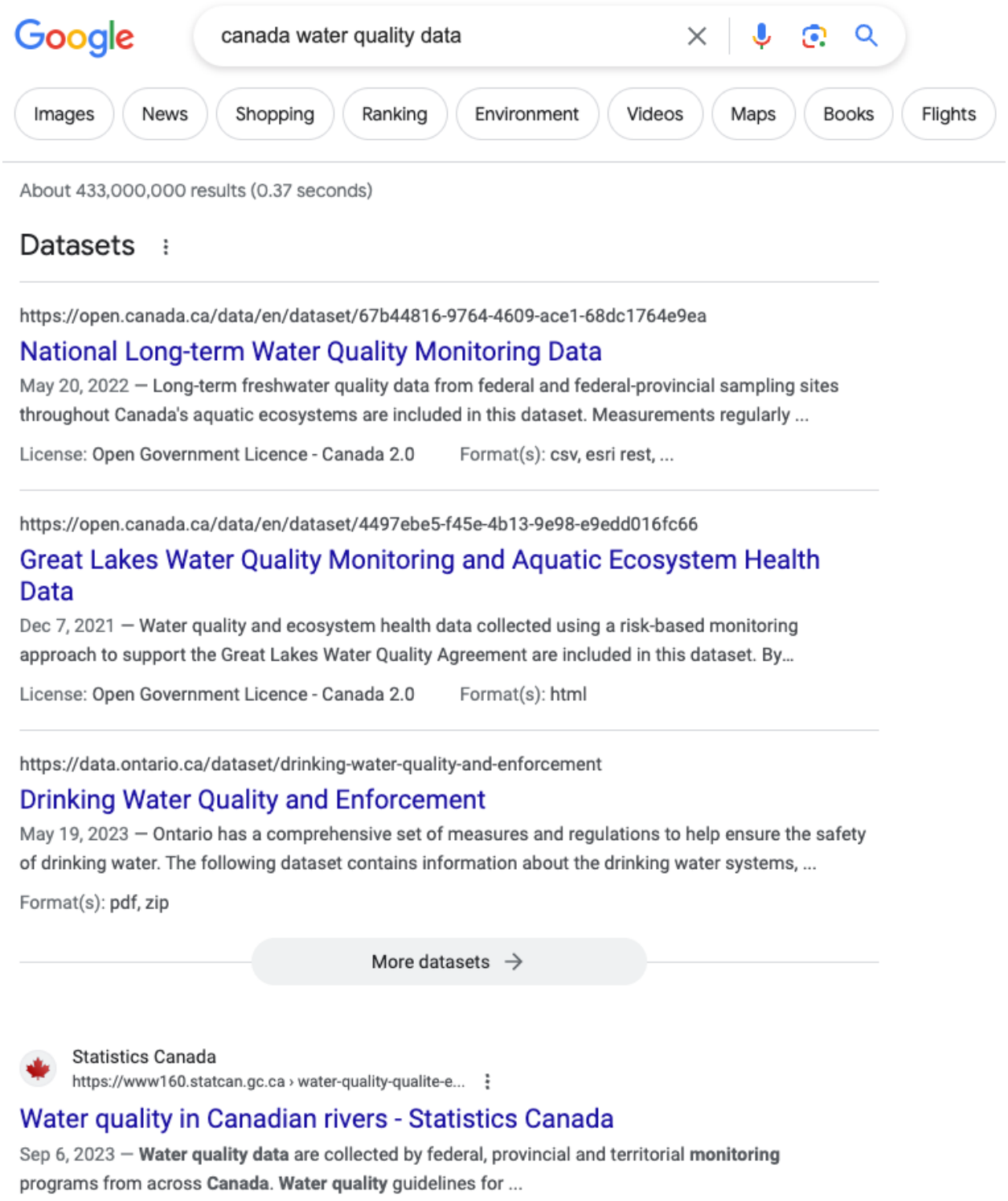

Figure 3. Dataset Search results in Google Web Search. Screenshot from October 30, 2023.

The Google Web Search results show a section for “Datasets” that contains direct links to dataset pages and a link to “More datasets” that goes to a Dataset Search page for that query. The search results may also contain

non-dataset results with more context about the dataset results.

This integration could address several of the issues that users had when building mental models of the tool that were based on Google Web Search. If a user discovers Dataset Search through Google Web Search by clicking on the "More datasets" link, the user may be able to more easily build a mental model of the tool as a specialized version of Google Web Search for datasets, as opposed to an entirely separate tool. Google and Bing Web Search already have a similar user interface for recipes, jobs, and events search. Users can also get 'the best of both worlds' by seeing contextual information about datasets with the dataset results, for example, scholarly articles, related queries, definitions of terms, and Wikipedia entries.

We would need to do further research to validate that this change addresses some of the issues that users had in building mental models of the tool and differentiating it from Google Web Search. We could also compare how successfully users build a mental model of the tool when they discover it through Google Web Search versus using it on its own for the first time.

User expectations about the scope of the tool. The second question that participants had was why some expected results and additional context were missing. For example, some participants saw fewer results than they expected, especially compared to the number of results in Google Web Search, which may have surprised or disappointed them. A previous study on a meta-portal had a similar finding, where participants had difficulty understanding the scope of the tool and "whether it contained data relevant for their research topics" (Dixit et al., 2018). Other participants in our study wanted to see artifacts associated with datasets within the tool, which they would have seen in Google Web Search.

We can look at these questions through the lens of previous research about mental models. Rieh et al. (2010) distinguished between "processing" mental models that focus on how tools store and retrieve information, versus "global-view" models that focus on how tools link to external information (i.e., the rest of the web). Khoo and Hall (2012) distinguished between digital libraries that curate their contents versus search engines that do not, where the 'library' aspect of a tool carries a "responsibility to its users that is not typically expected of commercial search engines" (p. 7). Dataset repositories and meta-portals align with the "processing" and curated mental models, while general-purpose search engines align with "global-view" and open models.

Applying a "global-view" and open mental model to Dataset Search sets unrealistic expectations that all datasets are contained in the tool, and users will be confused when an expected dataset is missing. We could think of a limited version of the "global-view" mental model, where the search results ideally contain all dataset pages, but in reality only contain all the dataset pages that the tool could discover and identify as such. General web search tools are also limited to some degree because they do not contain all web pages, such as long-tail pages that do not get enough traffic to be indexed. Thus, all tools are limited to varying degrees, which users may not realize.

Of course, a tool with limited search results will be less useful to users who are looking for datasets that the tool does not include. We agree with Gregory and Koesten's (2023) recommendation that such tools could better set user expectations: "We cannot assume that data seekers know the constraints of data search, i.e. limited indexing, [...]. Data search systems must translate user behavior with this in mind" (p. 61). When there are zero or only a few search results, tools could fail in a user-centered way to help users understand why results are missing and point them to alternative ways to find those results.

To address the participants who expected to see contextual information about datasets in the results, we must acknowledge that datasets do not exist in a vacuum (Gregory, Groth, et al., 2020; Krämer et al., 2021). Datasets are mentioned and used in articles or blog posts, introduced in scholarly papers, connected with code, interactive notebooks, or machine learning models. There are two approaches to address this user need: (1) dataset-specific search tools could make it easier to find the context of datasets (e.g., articles, blog posts, notebooks, models, and published code) without having to leave the tool; and (2) general or scholarly web search tools could make it easier to find datasets in the context of other search results. Examples of tools that already follow approach (1) include Kaggle, Hugging Face, and Papers With Code, which all focus on machine learning. An example of a tool that could follow approach (1) is Google Scholar, which already includes books, court opinions, and patents alongside academic papers; it could also include datasets and link them to other artifacts. Our work to integrate Dataset Search results into Google Web Search has also brought some datasets into the context of non-dataset search results.

How will these tools connect datasets to their artifacts? One approach is to rely on dataset metadata to link datasets to artifacts. To a limited extent, dataset metadata standards already support linking datasets to artifacts with fields for publications[12] and related works.[13] However, given the existing issues with metadata incompleteness, we are skeptical that we can rely on dataset providers to adopt these fields in a widespread way and frequently update them with new artifacts. An alternative approach is for researchers to infer the links between datasets and their artifacts (Färber, & Lamprecht, 2021; Jain et al., 2020; Lane et al., 2022).

Dataset replication across the web. The third question that users had was about the one-to-many relationship between search results and links to dataset pages (the multiple solid blue buttons in Figure 2). Dataset Search collapses identical replicas of a dataset from different sites into one result with multiple sources. However, participants were not sure which links to click on or if all the different links were identical datasets.

The UI pattern of showing multiple sources for the same search result is different from the familiar tools that participants cited when building their mental model of the tool; for example, Google Web Search and PubMed, which have a one-to-one relationship between search results and source web pages. Google and Bing Job Search both have a one-to-many pattern with one job result showing multiple blue buttons for each site where users can apply (one participant mentioned the visual similarity to Google Job Search). Google Scholar shows one primary link per result along with a link to "All *n* versions" that expands to show more links to the same paper.

Dataset Search does not clarify that all these datasets are identical to each other, and the tool does not show which replica is the primary source versus derived versions. While this feature saved participants time versus showing the datasets as separate search results, participants were confused and, in some cases, still clicked on all the links to compare their results. Dataset Search could make this UI pattern clearer with visual indications in the tool that all the links go to identical versions of the same dataset. If it is possible to identify the primary source, that link could be featured more prominently than the replicas of the dataset.

More generally, we recommend that dataset search tools display the complex relationships between the different versions and data sources in a way that is clear and meaningful to users. To enable tools to determine some of these relationships, one approach is for the standards community to develop and extend metadata to describe richer provenance for datasets, which is already underway and available to a limited extent[14] (Groth & Moreau, 2022; Klump et al., 2021). Data providers could then capture a clear provenance trail in metadata: the primary version and source of the dataset, and how the derived version differs from it. An alternative approach is to infer these relationships automatically. This approach poses challenging research questions: Given imperfect metadata, how do we infer the provenance of a dataset and other relationships between datasets?

Table 4. Recommendations to help users build a mental model of the tool.

Challenge	Audience	Recommendation
Dataset search and web search	Dataset search	Set user expectations that many but not all datasets on the web are in the tool; help users build a 'filtered global-view' mental model of the tool
	Dataset search	Fail gracefully when there are few results for a query
	Researchers	Study how well users build mental models of dataset-specific web search when using it for the first time through Google Web Search versus on its own
Expectations about scope	Dataset discovery tools	Make it easier to find related artifacts in their tools
	Scholarly web tools	Make it easier to find datasets in the context of their search results
	Researchers	Infer the artifacts associated with datasets

Dataset replication	Dataset search	Add indications in the user interface explaining relationships between replicas
	Dataset discovery tools	Display dataset version types and sources in a clear, meaningful way to users
	Researchers	Infer the provenance of datasets
	Researchers	Infer relationships between datasets

5.2. Making Sense of Heterogeneous Datasets

How much metadata do we need for discovery? Most participants wanted more information about datasets from the tool and dataset pages, an observation that validated previous findings (Section 2.2). Metadata quality largely depends on dataset providers and authors. We found that participants got frustrated when they could not rely on the same metadata fields across different search results: not all datasets had a file format, license, authors, last updated date, and so on. Benjelloun et al. (2020) analyzed the field-level incompleteness of metadata in Dataset Search; for example, only 35% of datasets had a license in their metadata. Our findings confirm that this metadata incompleteness negatively affected the user experience.

As mentioned in Dixit et al.'s (2018) study about a meta-portal, inconsistent metadata also makes filtering datasets less effective. In a heterogeneous collection that spans many providers, disciplines, and data types, it is impractical for dataset-search tools to require specific metadata fields (Google Search Central, 2023b). For example, spatial or temporal coverage makes sense for some datasets, while not for others. However, the more consistent and complete the metadata from providers, the less guesswork the discovery tools need to do. Researchers could also attempt to infer and 'fill in' as much of the metadata as possible from other sources, including the underlying data itself.

In the context of responsible reuse of data, metadata helps us understand not only what is in the data, but also how the data were collected, whether or not the sample was representative of the population, and limitations or caveats (Gebru et al., 2021). Arguably, users may not need all this information in the dataset-discovery stage (Koesten et al., 2017). Indeed, two participants in our study commented that they might like to see less information at first. However, the distinction between different stages of data discovery is often hard to make. For example, users may care about usage rights only after deciding that a dataset is useful; or they could instead limit their initial search only to datasets that are available for free. A similar argument could apply to any part of the metadata. The challenge for the discovery tools is to cater to both approaches and any metadata field.

Beyond metadata: Getting underlying data. While metadata was the entry point to discovering a dataset, many users wanted to see the underlying data in the tool, either to explore the dataset in more detail or to fill in gaps

in the metadata. Participants who expected to see data downloads in the tool itself may have been using mental models based on dataset repositories that store datasets and support downloading them directly.[15] When participants clicked on specific search results, some of them had difficulties finding data downloads on dataset pages from unfamiliar providers, including meta-portals that were several clicks away from a repository with a download link. Meta-portals and dataset-specific web search tools could address this issue by adding download links, dataset previews, information about the content, or the ability to easily view datasets in data-analysis tools. If dataset-discovery tools prefer to link to repositories with those features instead of supporting them in their own tools, they can clarify this to users who are confused and disappointed when they cannot find download links.

Since conducting the user study, Dataset Search added previews to search results where the tool can access the underlying data. Users must still click on the dataset page to download the data, but showing a preview helps them determine if they want to explore further. This approach strikes a balance between the needs of users and data providers. Data providers often want users to visit their sites to get datasets to provide access to additional dataset information, track usage metrics, or charge for access.

Accessing the underlying data can also help dataset discovery tools reconcile some of the incomplete or inaccurate fields of dataset metadata. For example, a dataset's metadata may say that it is a tabular file with specific columns, but the underlying data may be completely different. The dataset could have evolved and the metadata is now out of date, or the metadata was not accurate in the first place. Dataset discovery tools showing both metadata and data will need to reconcile these inconsistencies.

Building trust in data sources is key to dataset discovery. Participants in our study discovered new datasets and data sources serendipitously, similar to web search where you do not fully know what results to expect to see. The scope of a web-scale dataset discovery tool can be both a blessing and a curse: a user may find a useful data source they never knew existed, but they cannot assume that all the data sources have already been vetted. Dataset Search removes data sources with spammy or non-dataset pages (Alrashed et al., 2021), which can help to increase trust in its results, but the tool does not do any additional vetting. In line with previous research, we observed that participants developed mechanisms and heuristics to evaluate the reputation of unfamiliar sources, including relying on internet domains (e.g., .gov, .edu, .org) or mentions or usages in scholarly papers (Section 2.2).

Previous research has identified additional signals that build user trust in datasets: related artifacts, quality of metadata and documentation, usage metrics like views and downloads, and user-generated content like ratings or comments (Section 2.2). Many dataset repositories and meta-portals already show some of this information (Section 2.1). For example, Figshare shows citations, views, and downloads; Kaggle shows a usability score, views, downloads, notebooks, comments, and upvotes. Because Dataset Search does not vet datasets like some repositories do, the metadata quality varies widely. Researchers could study how well these signals work in practice when applied to datasets with incomplete metadata and limited artifacts.

Tools could provide signals about the trustworthiness or quality of data sources: background on the organization, the overall theme of the provider's datasets, the number of datasets they maintain, and whether the provider is a primary source for datasets or only contains copies or versions. For example, DataCite Commons' datasets link to a page with more information about their repository, including the number of datasets, citations,, and authors in the repository. Dataset-search tools could support filtering results based on dataset providers and signals about their trustworthiness. For example, data.gov supports filtering search results by data provider, bureau, organization, and organization type (federal, state, county, city, and university). Since conducting the user study, Dataset Search added functionality to filter results by dataset provider.

Table 5. Recommendations to help users make sense of heterogeneous datasets, access underlying data, and build trust in unfamiliar sources.

Challenge	Audience	Recommendation
Metadata quality	Dataset-discovery tools	Allow users to tailor their views and metadata selection based on their specific needs
	Researchers	Infer missing metadata fields
Downloading underlying data	Dataset-discovery tools	Add dataset download links, descriptions of content, previews, visualizations if not already supported
	Researchers	Infer formats and schema from underlying data, regardless of the accuracy of the dataset metadata
Building trust in unfamiliar data sources	Dataset-discovery tools	Provide stronger signals about the trustworthiness or quality of data sources
	Dataset-discovery tools	Show more context about data sources to help users build trust
	Dataset-discovery tools	Allow users to tailor their views and results based on the signals they use to determine trustworthiness
	Researchers	Study how users use trustworthiness signals when datasets have incomplete metadata and artifacts

5.3. Dataset Discovery Is a Skill and Needs to Be Part of Data Literacy

A few study participants suggested ways that Dataset Search could support users through introductory videos, guides, or tutorials, especially for students or novice dataset seekers. Indeed, people have already created dozens of videos on how to use Dataset Search.[16]

We also found that participants were unfamiliar with terminology about licenses and data formats. Dataset-search tools could add educational content or in-tool pointers. Previous research suggested embedding domain knowledge into the tool (Dixit et al., 2018), but we recommend embedding information about the general practice of searching for datasets into tools. For example, Mendeley Data has a "learn more" link next to the license name, and clicking on the link brings up a pop-up window that explains what that particular license means. Dataset-search tools could take a similar approach to specialized terminology like different data formats.

Previous research also found that "Dataset literacy is low" (Krämer et al., 2021, p. 191), which points to a broader opportunity to expand data literacy education. Data science and data literacy educational materials could focus more on web-scale and specialized dataset-search tools; when it makes sense to use one versus another; how to evaluate new data sources; understanding datasets' context, methodology, and limitations; and how to responsibly reuse and cite datasets (Coughlan, 2020; Krämer et al., 2021; Poirier, 2020). Data-search tools could support these educational materials to help novices learn to find, vet, and make sense of datasets.

We can look to information literacy education as a model: when students shifted from using library databases to search engines, information literacy instructors added coursework on how to use search tools and evaluate the trustworthiness of search results and sources (Holman, 2011; Russell, 2019).

The availability of data science courses has increased in recent years, but most of those courses do not teach students how to search for datasets; instead, instructors provide students with datasets (Adhikari et al., 2021; Coughlan, 2020; Mike et al., 2023; Wolff et al., 2016). Previous research has found that even the instructors themselves have trouble finding suitable datasets to use in their courses (Kross & Guo, 2019). Some courses ask students to find their own datasets, but they limit that search to one or a small number of repositories (Atenas & Havemann, 2015; Atenas et al., 2015). Our own participant recruitment process corroborated this hypothesis: We were initially unable to find any undergraduate students who searched for datasets at least once a month and had to supplement our recruitment through more targeted channels.

Table 6. Recommendations to help users learn how to search for datasets.

Challenge	Audience	Recommendation

Learning how to search for datasets	Dataset search	Create more educational materials and in-tool information about licenses, data formats, and so on.
	Dataset-discovery tools	Add educational information to help novice users, for example, specialized terminology, trustworthiness signals
	Open data community	Create more educational materials about web-scale dataset discovery as a part of data literacy

As noted above, many of the challenges and recommendations in this section (Section 5) apply to any dataset-discovery tool with a heterogeneous collection of datasets and other scholarly or technical artifacts.

6. Limitations

Our study has several limitations. First, Dataset Search is the only tool in the dataset-specific web search category (Table 1). Thus, we could not compare it to similar tools to corroborate our findings more broadly. We also explained the tool using metaphors with more familiar tools like Google Scholar and Google Web Search, which may have primed the participants to think about the tool a certain way. If and when more tools in this category exist and users are aware of them, it will likely be easier for the users to build a mental model of dataset-specific web search tools. Future studies could compare how well the same query satisfies user needs in Dataset Search versus a combination of tools of their choice. Future studies could also investigate these questions by using query analysis.

Second, our participants had different experience levels with dataset discovery (Table 3), which translated into different approaches during their interviews. At the same time, we saw commonalities in our findings across a range of participants, which provided additional validity to the findings.

Third, participants used different queries rather than perform the same specific tasks. While this protocol made it harder to compare the findings, participants were more invested in the task and had better context to interpret the results because the query was about something that they cared or knew about.

Fourth, all participants were based in the United States and used English-language queries. Future studies could focus on users searching for datasets from other countries or in non-English languages.

Lastly, we did not have any participants who were regular users of Dataset Search. A separate study with this group of participants could help us understand the features that they find useful or frustrating with repeated use.

7. Conclusion

Along with the rise of online data's scope and significance, the last decade saw an increase in tools for publishing and finding data across different disciplines. In this article, we studied a new type of tool in this landscape: dataset-specific web search tools, specifically Dataset Search. Such a tool allows users to discover datasets and data sources that they did not know existed. However, the openness and heterogeneity of such a tool also bring new complexities to user interactions: (1) building a mental model of the tool that accurately reflects Dataset Search's functionality, (2) making sense of heterogeneous datasets, and (3) learning how to search for datasets that satisfy their needs. To help users navigate this complex space efficiently, researchers, dataset-discovery tool developers, and the open data community must work together to build and support a thriving ecosystem.

Acknowledgments

We would like to thank Jess Holbrook and Gabriella Lanning for their invaluable help with formulating the research questions, developing the interview protocol, and analyzing the interviews. Thank you to Omar Benjelloun for his feedback and Lora Aroyo for her feedback on an earlier draft. We also thank anonymous reviewers for their comments on an earlier draft of this paper.

Disclosure Statement

Katrina Sostek, Daniel M. Russell, Nitesh Goyal, Tarfah Alrashed, Stella Dugall, and Natasha Noy have no financial or non-financial disclosures to share for this article.

Supplementary Files

The codebook for the interviews is available below.

pdf Codebook - Discovering Datasets on the Web Scale.pdf 166 KB

References

Adhikari, A., DeNero, J., & Jordan, M. I. (2021). Interleaving computational and inferential thinking: Data science for undergraduates at Berkeley. *Harvard Data Science Review, 3*(2). https://doi.org/10.1162/99608f92.cb0fa8d2

Alrashed, T., Paparas, D., Benjelloun, O., Sheng, Y., & Noy, N. (2021). Dataset or not? A study on the veracity of semantic markup for dataset pages. In A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P. Barnaghi,

A. Haller, M. Dragoni, & H. Alani (Eds.), *Lecture notes in computer science: Vol. 12922. The Semantic Web – ISWC 2021* (pp. 338–356). Springer, Cham. https://doi.org/10.1007/978-3-030-88361-4_20

Atenas, J., & Havemann, L. (Eds.). (2015). *Open data as open educational resources: Case studies of emerging practice*. London: Open Knowledge, Open Education Working Group. https://dx.doi.org/10.6084/m9.figshare.1590031

Atenas, J., Havemann, L., & Priego, E. (2015). Open data as open educational resources: Towards transversal skills and global citizenship. *Open Praxis*, *7*(4), 377–389. https://doi.org/10.5944/openpraxis.7.4.233

Benjelloun, O., Chen, S., & Noy, N. (2020). Google Dataset Search by the numbers. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicze, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), *Lecture Notes in Computer Science: Vol. 12507. The Semantic Web – ISWC 2020* (pp. 667–682). Springer, Cham. https://doi.org/10.1007/978-3-030-62466-8_41

Brandt, D. S., & Uden, L. (2003). Insight into mental models of novice Internet searchers. *Communications of the ACM*, *46*(7), 133–136. https://doi.org/10.1145/792704.792711

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., & Groth, P. (2020). Dataset search: A survey. *The VLDB Journal*, *29*(1), 251–272. https://link.springer.com/article/10.1007/s00778-019-00564-x

Choi, J., & Tausczik, Y. (2017). Characteristics of collaboration in the emerging practice of open data analysis. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 835–846). ACM. https://doi.org/10.1145/2998181.2998265

Cieslewicz, A., Dutkiewicz, J., & Jedrzejek, C. (2018). Baseline and extensions approach to information retrieval of complex medical data: Poznan's approach to the bioCADDI 2016. *Database, 2018*, Article bax103. https://doi.org/10.1093/database/bax103

Coughlan, T. (2020). The use of open data as a material for learning. *Educational Technology Research and Development*, *68*(1), 383–411. https://doi.org/10.1007/s11423-019-09706-y

Davies, T., & Frank, M. (2013). 'There's no such thing as raw data': Exploring the socio-technical life of a government dataset. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 75–78). ACM. https://doi.org/10.1145/2464464.2464472

Dixit, R., Rogith, D., Narayana, V., Salimi, M., Gururaj, A., Ohno-Machado, L., Xu, H., & Johnson, T. R. (2018). User needs analysis and usability assessment of DataMed–A biomedical data discovery index. *Journal of the American Medical Informatics Association*, *25*(3), 337–344. https://doi.org/10.1093/jamia/ocx134

Erete, S., Ryou, E., Smith, G., Fassett, K. M., & Duda, S. (2016). Storytelling with data: Examining the use of data by non-profit organizations. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1273-1283). ACM. https://doi.org/10.1145/2818048.2820068

European Commission. (2023). *European legislation on open data*. Europa.eu. https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

FAIR play in geoscience data [Editorial]. (2019). *Nature Geoscience*, *12*, Article 961. https://doi.org/10.1038/s41561-019-0506-4

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. https://doi.org/10.1002/meet.14504901068

Färber, M., & Lamprecht, D. (2021). The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies*, *2*(4), 1324–1355. https://doi.org/10.1162/qss_a_00161

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, *6*(1), Article 28. https://doi.org/10.1038/s41597-019-0031-8

Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2019). https://www.congress.gov/bill/115th-congress/house-bill/4174

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010). Anatomy of the long tail: Ordinary people with extraordinary tastes. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 201–210). ACM. https://doi.org/10.1145/1718487.1718513

Google Search Central. (2023a). Dataset. https://developers.google.com/search/docs/appearance/structured-data/dataset

Google Search Central. (2023b). File types indexable by Google. https://developers.google.com/search/docs/advanced/crawling/indexable-file-types

Gorgolewski, C. (2019, July 18). *Making GitHub-hosted datasets discoverable by Google Dataset Search*. Medium. https://chrisgorgo.medium.com/making-github-hosted-datasets-discoverable-by-google-dataset-search-13527f2f657a

Gregory, K. (2020). A dataset describing data discovery and reuse practices in research. *Scientific Data, 7*(1), Article 232. https://doi.org/10.1038/s41597-020-0569-5

Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science, 46*(4), 459–475. https://doi.org/10.1177/0165551519837182

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. (2023). *Human-centered data discovery*. Springer Nature.

Groth, P., & Moreau, L. (2022). *PROV-Overview: An overview of the PROV family of documents* [World Wide Web Consortium]. W3. https://www.w3.org/TR/prov-overview

Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., & Whang, S. E. (2016). Goods: Organizing Google's datasets. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 795–806). ACM. https://doi.org/10.1145/2882903.2903730

Holman, L. (2011). Millennial students' mental models of search: Implications for academic librarians and database developers. *The Journal of Academic Librarianship, 37*(1), 19–27. https://doi.org/10.1016/j.acalib.2010.10.003

Ibáñez, L.-D., & Simperl, E. (2022). A comparison of dataset search behaviour of internal versus search engine referred sessions. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 158–168). ACM. https://doi.org/10.1145/3498366.3505821

Jain, S., van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: A challenge dataset for document-level information extraction. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7506–7516). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.670

Kacprzak, E., Koesten, L. M., Ibáñez, L.-D., Simperl, E., & Tennison, J. (2017). A query log analysis of dataset search. In J. Cabot, R. De Virgilio, R. Torlone (Eds.), *Lecture notes in computer science: Vol. 10360. Web engineering* (pp. 429–436). Springer, Cham. https://doi.org/10.1007/978-3-319-60131-1_29

Kern, D., & Mathiak, B. (2015). Are there any differences in data set retrieval compared to well-known literature retrieval? In S. Kapidakis, C. Mazurek, & M. Werla (Eds.), *Lecture notes in computer science: Vol. 9316. Research and advanced technology for digital libraries* (pp. 197–208). Springer, Cham. https://doi.org/10.1007/978-3-319-24592-8_15

Khoo, M., & Hall, C. (2012). What would 'Google' do? Users' mental models of a digital library search engine. In P. Zaphiris, F. Loizides, G. Buchanan, & E. Rasmussen (Eds.), *Lecture notes in computer science: Vol. 7489. Theory and practice of digital libraries* (pp. 1–12). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33290-6_1

Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R. R., & Asmi, A. (2021). Versioning data is about more than revisions: A conceptual framework and proposed principles. *Data Science Journal*, *20*(12), Article 12. https://doi.org/10.5334/dsj-2021-012

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

Koesten, L., Kacprzak, E., Tennison, J., & Simperl, E. (2019). Collaborative practices with structured data: Do tools support what users need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper no. 100). ACM. https://doi.org/10.1145/3290605.3300330

Koesten, L. M., Kacprzak, E., Tennison, J. F., & 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

Kross, S., & Guo, P. J. (2019). Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper no. 263). ACM. https://doi.org/10.1145/3290605.3300493

Lane, J., Gimeno, E., Levitskaya, E., Zhang, Z., & Zigoni, A. (2022). Data inventories for the modern age? Using data science to open government data. *Harvard Data Science Review*, *4*(2). https://doi.org/10.1162/99608f92.8a3f2336

Liu, Y.-H., Wu, M., Power, M., & Burton, A. (2022). *Elicitation of data discovery contexts: An interview study.* Zenodo. https://doi.org/10.5281/zenodo.7179526

Lowenberg, D. (2022). Recognizing our collective responsibility in the prioritization of open data metrics. *Harvard Data Science Review, 4*(3). https://doi.org/10.1162/99608f92.c71c3479

Microsoft. (2022). *Advanced search keywords*. https://support.microsoft.com/en-us/topic/advanced-search-keywords-ea595928-5d63-4a0b-9c6b-0b769865e78a

Mike, K., Kimelfeld, B., & Hazzan, O. (2023). The birth of a new discipline: Data science education. *Harvard Data Science Review, 5*(4). https://doi.org/10.1162/99608f92.280afe66

More research will be publicly accessible sooner [Editorial]. (2022). *Nature Biomedical Engineering, 6*(9), 1013–1014. https://doi.org/10.1038/s41551-022-00944-9

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How data science workers work with data: Discovery, capture, curation, design, creation, In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper no. 126). ACM. https://doi.org/10.1145/3290605.3300356

Norman, D. A. (1983). Some observations on mental models. In D. Gentner, & A. L. Stevens (Eds.), *Mental Models* (pp. 7–14). Psychology Press. https://doi.org/10.4324/9781315802725

Noy, N. (2020, January 23). Discovering millions of datasets on the web. *The Keyword*. https://blog.google/products/search/discovering-millions-datasets-web

Noy, N. & Benjelloun, O. (2023, February 28). Datasets at your fingertips in Google Search. *Google Research*. https://ai.googleblog.com/2023/02/datasets-at-your-fingertips-in-google.html

Noy, N., Burgess, M., & Brickley, D. (2019). Google Dataset Search: Building a search engine for datasets in an open web ecosystem. In L. Liu & R. White (Eds.), *Proceedings of The World Wide Web Conference WWW 2019* (pp. 1365–1375). ACM. https://doi.org/10.1145/3308558.3313685

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. *Communications of the ACM, 62*(8), 36–43. https://doi.org/10.1145/3331166

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review, 1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Passi, S., & Jackson, S. J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), Article

136. https://doi.org/10.1145/3274405

Perkins, D. N., & Salomon, G. (1992). In T. Husén, & T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education* (2nd ed., pp. 425–441). Oxford.

Poirier, L. (2020, December 10). *Ethnographies of datasets: Teaching critical data analysis through R Notebooks*. CUNY Manifold. https://jitp.commons.gc.cuny.edu/ethnographies-of-datasets-teaching-critical-data-analysis-through-r-notebooks

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

Rieh, S. Y., Yang, J. Y., Yakel, E., & Markey, K. (2010). Conceptualizing institutional repositories: Using co-discovery to uncover mental models. In N. J. Belkin (Ed.), *Proceedings of the Third Symposium on Information Interaction in Context* (pp. 165–174). ACM. https://doi.org/10.1145/1840784.1840809

Russell, D. M. (2019). *The joy of search: A Google insider's guide to going beyond the basics*. MIT Press.

Russell, D. M., & Grimes, C. (2007). Assigned tasks are not the same as self-chosen web search tasks. In R. H. Sprague (Ed.), *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)* (p. 83). IEEE. https://doi.org/10.1109/HICSS.2007.91

Saldaña, J. (2021). The coding manual for qualitative researchers (4th ed.). Sage Publications..

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 39). ACM. https://doi.org/10.1145/3411764.3445518

Sansone, S.-A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J. S., Xu, H., Fore, I. M., Lyle, J., Gururaj, A. E., Chen, X., Kim, H., Zong, N., Li, Y., Liu, R., Ozyurt, I. B., & Ohno-Machado, L. (2017). Dats, the data tag suite to enable discoverability of datasets. *Scientific Data*, *4*(1), Article 170059. https://doi.org/10.1038/sdata.2017.59

Search Console Help. (2023). *Why is my page missing from Google Search?* Google. https://support.google.com/webmasters/answer/7474347

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

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLoS ONE*, *6*(6), Article e21101. https://doi.org/10.1371/journal.pone.0021101

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), Article e67332. https://doi.org/10.1371/journal.pone.0067332

Walters, W. H. (2020). Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights*, *33*(1), Article 18. https://doi.org/10.1629/uksg.510

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. 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*, Article 160018. http://dx.doi.org/10.1038/sdata.2016.18

Wolff, A., Gooch, D., Montaner, J. J. C., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, *12*(3). https://doi.org/10.15353/joci.v12i3.3275

Wu, M., Psomopoulos, F., Khalsa, S. J., & de Waard, A. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, *18*(1), Article 3. https://doi.org/10.5334/dsj-2019-003

Xiao, F., Wang, Z., & He, D. (2020). Understanding users' accessing behaviors to local Open Government Data via transaction log analysis. *Proceedings of the Association for Information Science and Technology*, *57*(1), Article e278. https://doi.org/10.1002/pra2.278

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956. https://doi.org/10.1002/asi.23730

Yoon, A., & Lee, Y. Y. (2019). Factors of trust in data reuse. *Online Information Review*, *43*(7), 1245–1262. https://doi.org/10.1108/OIR-01-2019-0014

Zhang, Y. (2008). Undergraduate students' mental models of the web as an information retrieval system. *Journal of the American Society for Information Science and Technology*, *59*(13), 2087–2098. https://doi.org/10.1002/asi.20915

Zhang, Y. (2013). The development of users' mental models of MedlinePlus in information searching. *Library & Information Science Research*, *35*(2), 159–170. https://doi.org/10.1016/j.lisr.2012.11.004

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), 5–16. https://doi.org/10.1007/s00799-007-0015-8

Footnotes

1. We use the term ‘repository,’ but the term ‘portal’ is also used in the literature, sometimes interchangeably, and sometimes with a distinction between the type and scope of datasets included (Gregory & Koesten, 2023). ↩
2. Another subtype of repositories is those that search nonpublic data internal to a government agency or a corporation, for example, Halevy et al. (2016). Those limited-access repositories are out of the scope of this article. ↩
3. For example, data-focused content management systems such as Socrata or CKAN, which data.gov uses. ↩
4. DataMed planned to rely on a crawl but never implemented that system (datamed.org/about). ↩
5. Dataset Search’s coverage of dataset pages in GitHub is limited; see Gorgolewski (2019) for more details. ↩
6. For a more complete and historical list of studies, see Gregory and Koesten (2023). ↩
7. Participants in Liu et al. (2022) mentioned Dataset Search but only in a list of several tools that they used. ↩
8. datasetsearch.research.google.com ↩
9. schema.org/Dataset ↩
10. userresearch.google.com ↩
11. We use Wolff et al.’s (2016) definition of data literacy: “Data literacy is the ability to ask and answer real-world questions from large and small data sets through an inquiry process, with consideration of ethical use of data.” ↩
12. schema.org/Dataset’s publication field ↩

13. support.datacite.org/docs/connecting-to-works ↩

14. support.datacite.org/docs/versioning ↩

15. Some expert participants may also have used mental models based on general web search tools that have advanced features to filter results to download links of specific file types (Google Search Central, 2023b; Microsoft, 2022). ↩

16. www.google.com/search?q=%22google+dataset+search%22+tutorial ↩

References

- Adhikari, A., DeNero, J., & Jordan, M. I. (2021). Interleaving computational and inferential thinking: Data science for undergraduates at Berkeley. *Harvard Data Science Review*, *3*(2). https://doi.org/10.1162/99608f92.cb0fa8d2 ↩
- Alrashed, T., Paparas, D., Benjelloun, O., Sheng, Y., & Noy, N. (2021). Dataset or not? A study on the veracity of semantic markup for dataset pages. In A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P. Barnaghi, A. Haller, M. Dragoni, & H. Alani (Eds.), *Lecture notes in computer science: Vol. 12922. The Semantic Web – ISWC 2021* (pp. 338–356). Springer, Cham. https://doi.org/10.1007/978-3-030-88361-4_20 ↩
- Atenas, J., & Havemann, L. (Eds.). (2015). *Open data as open educational resources: Case studies of emerging practice*. London: Open Knowledge, Open Education Working Group. https://dx.doi.org/10.6084/m9.figshare.1590031 ↩
- Atenas, J., Havemann, L., & Priego, E. (2015). Open data as open educational resources: Towards transversal skills and global citizenship. *Open Praxis*, *7*(4), 377–389. https://doi.org/10.5944/openpraxis.7.4.233 ↩
- Benjelloun, O., Chen, S., & Noy, N. (2020). Google Dataset Search by the numbers. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicze, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), *Lecture Notes in Computer Science: Vol. 12507. The Semantic Web – ISWC 2020* (pp. 667–682). Springer, Cham. https://doi.org/10.1007/978-3-030-62466-8_41 ↩
- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

↩

- Brandt, D. S., & Uden, L. (2003). Insight into mental models of novice Internet searchers. *Communications of the ACM*, *46*(7), 133–136. https://doi.org/10.1145/792704.792711

↩

- Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., & Groth, P. (2020). Dataset search: A survey. *The VLDB Journal*, *29*(1), 251–272. https://link.springer.com/article/10.1007/s00778-019-00564-x

↩

- Choi, J., & Tausczik, Y. (2017). Characteristics of collaboration in the emerging practice of open data analysis. In C. P. Lee et al. (Eds.), *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 835–846). ACM. https://doi.org/10.1145/2998181.2998265

↩

- Cieslewicz, A., Dutkiewicz, J., & Jedrzejek, C. (2018). Baseline and extensions approach to information retrieval of complex medical data: Poznan's approach to the bioCADDI 2016. *Database*, *2018*, Article bax103. https://doi.org/10.1093/database/bax103

↩

- Coughlan, T. (2020). The use of open data as a material for learning. *Educational Technology Research and Development*, *68*(1), 383–411. https://doi.org/10.1007/s11423-019-09706-y

↩

- Davies, T., & Frank, M. (2013). 'There's no such thing as raw data': Exploring the socio-technical life of a government dataset. In H. Davis et al. (Eds.), *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 75–78). ACM. https://doi.org/10.1145/2464464.2464472

↩

- Dixit, R., Rogith, D., Narayana, V., Salimi, M., Gururaj, A., Ohno-Machado, L., Xu, H., & Johnson, T. R. (2018). User needs analysis and usability assessment of DataMed–A biomedical data discovery index. *Journal of the American Medical Informatics Association*, *25*(3), 337–344. https://doi.org/10.1093/jamia/ocx134

↩

- Erete, S., Ryou, E., Smith, G., Fassett, K. M., & Duda, S. (2016). Storytelling with data: Examining the use of data by non-profit organizations. In D. Gergle et al. (Eds.), *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. (pp. 1273-1283). ACM. https://doi.org/10.1145/2818048. 2820068

↩

- European Commission. (2023). *European legislation on open data*. Europa.eu. https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

↩

- FAIR play in geoscience data [Editorial]. (2019). *Nature Geoscience*, *12*, Article 961. https://doi.org/10.1038/s41561-019-0506-4

↩

- Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. https://doi.org/10.1002/meet.14504901068

↩

- Färber, M., & Lamprecht, D. (2021). The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies*, *2*(4), 1324–1355. https://doi.org/10.1162/qss_a_00161

↩

- Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, *6*(1), Article 28. https://doi.org/10.1038/s41597-019-0031-8

↩

- Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2019). https://www.congress.gov/bill/115th-congress/house-bill/4174

↩

- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

↩

- Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010). Anatomy of the long tail: Ordinary people with extraordinary tastes. In B. D. Davison et al. (Eds.), *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 201–210). ACM. https://doi.org/10.1145/1718487.1718513

↩

- Google Search Central. (2023a). Dataset. https://developers.google.com/search/docs/appearance/structured-data/dataset

↩

-

↩

- Gregory, K. (2020). A dataset describing data discovery and reuse practices in research. *Scientific Data*, *7*(1), Article 232. https://doi.org/10.1038/s41597-020-0569-5

↩

- Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science*, *46*(4), 459–475. https://doi.org/10.1177/0165551519837182

↩

- Gregory, K., & Koesten, L. (2023). *Human-centered data discovery*. Springer Nature.

↩

- 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

↩

- Groth, P., & Moreau, L. (2022). *PROV-Overview: An overview of the PROV family of documents* [World Wide Web Consortium]. W3. https://www.w3.org/TR/prov-overview

↩

- Holman, L. (2011). Millennial students' mental models of search: Implications for academic librarians and database developers. *The Journal of Academic Librarianship*, *37*(1), 19–27. https://doi.org/10.1016/j.acalib.2010.10.003

↩

- Ibáñez, L.-D., & Simperl, E. (2022). A comparison of dataset search behaviour of internal versus search engine referred sessions. In D. Esweiler et al. (Eds.), *ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 158–168). ACM. https://doi.org/10.1145/3498366.3505821

↩

- Jain, S., van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: A challenge dataset for document-level information extraction. In D. Jurafsky (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7506–7516). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.670

↩

- Kacprzak, E., Koesten, L. M., Ibáñez, L.-D., Simperl, E., & Tennison, J. (2017). A query log analysis of dataset search. In J. Cabot, R. De Virgilio, R. Torlone (Eds.), *Lecture notes in computer science: Vol. 10360. Web engineering* (pp. 429–436). Springer, Cham. https://doi.org/10.1007/978-3-319-60131-1_29

↩

- Kern, D., & Mathiak, B. (2015). Are there any differences in data set retrieval compared to well-known literature retrieval? In S. Kapidakis, C. Mazurek, & M. Werla (Eds.), *Lecture notes in computer science: Vol. 9316. Research and advanced technology for digital libraries* (pp. 197–208). Springer, Cham. https://doi.org/10.1007/978-3-319-24592-8_15

↩

- Khoo, M., & Hall, C. (2012). What would 'Google' do? Users' mental models of a digital library search engine. In P. Zaphiris, F. Loizides, G. Buchanan, & E. Rasmussen (Eds.), *Lecture notes in computer science: Vol. 7489. Theory and practice of digital libraries* (pp. 1–12). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33290-6_1

↩

- Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R. R., & Asmi, A. (2021). Versioning data is about more than revisions: A conceptual framework and proposed principles. *Data Science Journal*, *20*(12), Article 12. https://doi.org/10.5334/dsj-2021-012

↩

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

↩

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

↩

- Koesten, L., Kacprzak, E., Tennison, J., & Simperl, E. (2019). Collaborative practices with structured data: Do tools support what users need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper no. 100). ACM. https://doi.org/10.1145/3290605.3300330

↩

-

↩

- 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

↩

- Kross, S., & Guo, P. J. (2019). Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper no. 263). ACM. https://doi.org/10.1145/3290605.3300493

↩

- Lane, J., Gimeno, E., Levitskaya, E., Zhang, Z., & Zigoni, A. (2022). Data inventories for the modern age? Using data science to open government data. *Harvard Data Science Review*, *4*(2). https://doi.org/10.1162/99608f92.8a3f2336

↩

- Liu, Y.-H., Wu, M., Power, M., & Burton, A. (2022). *Elicitation of data discovery contexts: An interview study.* Zenodo. https://doi.org/10.5281/zenodo.7179526

↩

- Lowenberg, D. (2022). Recognizing our collective responsibility in the prioritization of open data metrics. *Harvard Data Science Review, 4*(3). https://doi.org/10.1162/99608f92.c71c3479

↩

- Microsoft. (2022). *Advanced search keywords*. https://support.microsoft.com/en-us/topic/advanced-search-keywords-ea595928-5d63-4a0b-9c6b-0b769865e78a

↩

- Mike, K., Kimelfeld, B., & Hazzan, O. (2023). The birth of a new discipline: Data science education. *Harvard Data Science Review*, *5*(4). https://doi.org/10.1162/99608f92.280afe66

↩

- More research will be publicly accessible sooner [Editorial]. (2022). *Nature Biomedical Engineering*, *6*(9), 1013–1014. https://doi.org/10.1038/s41551-022-00944-9

↩

- Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How data science workers work with data: Discovery, capture, curation, design, creation, In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper no. 126). ACM. https://doi.org/10.1145/3290605.3300356

↩

- Norman, D. A. (1983). Some observations on mental models. In D. Gentner, & A. L. Stevens (Eds.), *Mental Models* (pp. 7–14). Psychology Press. https://doi.org/10.4324/9781315802725

↩

- Noy, N. (2020, January 23). Discovering millions of datasets on the web. *The Keyword*. https://blog.google/products/search/discovering-millions-datasets-web

↩

- Noy, N. & Benjelloun, O. (2023, February 28). Datasets at your fingertips in Google Search. *Google Research*. https://ai.googleblog.com/2023/02/datasets-at-your-fingertips-in-google.html

↩

- Noy, N., Burgess, M., & Brickley, D. (2019). Google Dataset Search: Building a search engine for datasets in an open web ecosystem. In L. Liu & R. White (Eds.), *Proceedings of The World Wide Web Conference WWW 2019* (pp. 1365–1375). ACM. https://doi.org/10.1145/3308558.3313685

↩

- Noy, N., Burgess, M., & Brickley, D. (2019). Google Dataset Search: Building a search engine for datasets in an open web ecosystem. In L. Liu et al. & R. White (Eds.), *Proceedings of The World Wide Web Conference WWW 2019* (pp. 1365–1375). ACM. https://doi.org/10.1145/3308558.3313685

↩

- Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review, 1*(2). https://doi.org/10. 1162/99608f92.fc14bf2d

↩

- Passi, S., & Jackson, S. J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), Article 136. https://doi.org/10.1145/3274405

↩

- Perkins, D. N., & Salomon, G. (1992). In T. Husén, & T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education* (2nd ed., pp. 425-441). Oxford.

↩

- Poirier, L. (2020, December 10). *Ethnographies of datasets: Teaching critical data analysis through R Notebooks*. CUNY Manifold. https://jitp.commons.gc.cuny.edu/ethnographies-of-datasets-teaching-critical-data-analysis-through-r-notebooks

↩

- 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 ↩
- Rieh, S. Y., Yang, J. Y., Yakel, E., & Markey, K. (2010). Conceptualizing institutional repositories: Using co-discovery to uncover mental models. In N. J. Belkin (Ed.), *Proceedings of the Third Symposium on Information Interaction in Context* (pp. 165–174). ACM. https://doi.org/10.1145/1840784.1840809 ↩
- Russell, D. M. (2019). *The joy of search: A Google insider's guide to going beyond the basics*. MIT Press. ↩
- Russell, D. M., & Grimes, C. (2007). Assigned tasks are not the same as self-chosen web search tasks. In R. H. Sprague (Ed.), *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)* (p. 83). IEEE. https://doi.org/10.1109/HICSS.2007.91 ↩
- Saldaña, J. (2021). The coding manual for qualitative researchers (4th ed.). Sage Publications.. ↩
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Y. Kitamura et al. (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 39). ACM. https://doi.org/10.1145/3411764.3445518 ↩
- Sansone, S.-A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J. S., Xu, H., Fore, I. M., Lyle, J., Gururaj, A. E., Chen, X., Kim, H., Zong, N., Li, Y., Liu, R., Ozyurt, I. B., & Ohno-Machado, L. (2017). Dats, the data tag suite to enable discoverability of datasets. *Scientific Data, 4*(1), Article 170059. https://doi.org/10.1038/sdata.2017.59 ↩
- Search Console Help. (2023). *Why is my page missing from Google Search?* Google. https://support.google.com/webmasters/answer/7474347 ↩
- Sharifpour, R., Wu, M., & Zhang, X. (2022). Large-scale analysis of query logs to profile users for dataset search. *Journal of Documentation, 79*(1), 66–85. https://doi.org/10.1108/JD-12-2021-0245 ↩

- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLoS ONE*, *6*(6), Article e21101. https://doi.org/10.1371/journal.pone.0021101 ↩
- Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), Article e67332. https://doi.org/10.1371/journal.pone.0067332 ↩
- Walters, W. H. (2020). Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights*, *33*(1), Article 18. https://doi.org/10.1629/uksg.510 ↩
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. 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*, Article 160018. http://dx.doi.org/10.1038/sdata.2016.18 ↩
- Wolff, A., Gooch, D., Montaner, J. J. C., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, *12*(3). https://doi.org/10.15353/joci.v12i3.3275 ↩
- Wu, M., Psomopoulos, F., Khalsa, S. J., & de Waard, A. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, *18*(1), Article 3. https://doi.org/10.5334/dsj-2019-003 ↩
- Xiao, F., Wang, Z., & He, D. (2020). Understanding users' accessing behaviors to local Open Government Data via transaction log analysis. *Proceedings of the Association for Information Science and Technology*, *57*(1), Article e278. https://doi.org/10.1002/pra2.278 ↩
- Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956. https://doi.org/10.1002/asi.23730 ↩

- Yoon, A., & Lee, Y. Y. (2019). Factors of trust in data reuse. *Online Information Review, 43*(7), 1245–1262. https://doi.org/10.1108/OIR-01-2019-0014

  ↩

- Zhang, Y. (2008). Undergraduate students' mental models of the web as an information retrieval system. *Journal of the American Society for Information Science and Technology, 59*(13), 2087–2098. https://doi.org/10.1002/asi.20915

  ↩

- Zhang, Y. (2013). The development of users' mental models of MedlinePlus in information searching. *Library & Information Science Research*, *35*(2), 159–170. https://doi.org/10.1016/j.lisr.2012.11.004

  ↩

- 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), 5–16. https://doi.org/10. 1007/s00799-007-0015-8

  ↩