# "It Took Longer than I was Expecting:" Why is Dataset Search Still so Hard?

Madelon Hulsebos, Wenjing Lin, Shreya Shankar, Aditya G. Parameswaran
UC Berkeley

## ABSTRACT

Dataset search is a long-standing problem across both industry and academia. While most industry tools focus on identifying one or more datasets matching a user-specified query, most recent academic papers focus on the subsequent problems of join and union discovery for a given dataset. In this paper, we take a step back and ask: is the task of identifying an initial dataset really a solved problem? Are join and union discovery indeed the most pressing problems to work on? To answer these questions, we survey 89 data professionals and surface the objectives, processes, and tools used to search for structured datasets, along with the challenges faced when using existing systems. We uncover characteristics of data content and metadata that are most important for data professionals during search, such as granularity and data freshness. Informed by our analysis, we argue that dataset search is not yet a solved problem, but is, in fact, difficult to solve. To move the needle in the right direction, we distill four desiderata for future dataset search systems: iterative interfaces, hybrid querying, task-driven search and result diversity.

## 1 THE DATASET SEARCH PROBLEM

Finding relevant datasets through a process of *dataset search* is essential to make sense of and extract value from data, underlying applications in data discovery and preparation, data exploration, data science, and machine learning. Dataset search has a renewed emphasis thanks to the emergence of data cataloging and governance platforms [4, 17] operating alongside data lakes and lakehouses [6]. With the volume of data collected worldwide estimated to reach 180 zettabytes by 2025 [3], dataset search is essential to navigate large data catalogs effectively and efficiently.

Recent work [11] categorizes dataset search into one of two objectives: 1) identifying an initial dataset for a given task, i.e., *basic dataset search*, and 2) *enriching* an already-identified dataset, e.g., via joins or unions. For the former, the input query is a keyword search expression, while for the latter, the input query is a table targeted for enrichment. We find that the majority of recent academic research has been focused on the latter objective, via join and union discovery for a target table [9, 10, 12–14, 19, 21]. Some recent academic papers consider basic dataset search as a component of one-shot question answering (e.g. "*What housing price indexes are available for cities in Pennsylvania?*") [18, 23], targeting lay users rather than data analytics use cases. On the other hand, most industry systems focus on basic dataset search, relying on syntactic keyword matching between queries and tables, as well as filtering over metadata such as date ranges [11]. Recent industry systems continue to focus on basic dataset search, but incorporate semantic search, i.e., retrieving datasets based on similarities between embeddings of queries and datasets (e.g., schemas) [1, 2], question-based search [22], or profiling metrics [7].

The gap between academic research and industry dataset search systems raises the question: is basic dataset search truly "solved"? What are the unmet needs in basic dataset search, and dataset search in general? Furthermore, how important is our community's current research focus on enrichment over basic search? Anecdotally, as well as from our experience, we believe that basic dataset search still very much remains an unsolved problem. However, we lack a deep understanding of user needs—both quantitative and qualitative—with respect to dataset search. A survey of the present-day users of dataset search can help identify best practices, inform challenges, and provide a roadmap for future system development.

We therefore conducted an online survey among 89 data professionals focusing on dataset search, with a focus on structured data and analytical use-cases. We sought to understand why people search for data, how they do it—where do they search, what tools do they use, and what is their process, what aspects of datasets that they are searching for are most important to them, what challenges they face, and what ideal dataset search systems should support. Our survey spanned participants from various industries and organizations, ranging from finance, tech, and healthcare, to energy, real-estate, and government. Our findings illustrate that basic dataset search is still a time-consuming process, relying on iterative workflows and collaboration among coworkers. Moreover, existing systems are not semantically robust and flexible enough to handle the sheer diversity of data characteristics and queries, resulting often in dataset overload. We present four desiderata for future dataset search systems: iterative interfaces, hybrid querying, task-driven search, and result comprehensibility and diversity.

## 2 WHERE DO WE STAND IN PRACTICE?

**Survey design.** We conducted a survey to understand the state of dataset search in practice. The survey first inquires about the objectives, workflows and tools used for dataset search, zooms in on challenges faced in the current process, and closes with inquiring about ideal dataset search tools. In total, the survey comprises 15 questions. We begin with a control question to verify that the respondent has experience with data search for work. It continues with 10 questions, of which 6 are multiple-choice questions and 4 are open text answers. The last 4 questions collect information about the participant background (e.g. role and industry). The survey can be viewed at: https://forms.gle/xEHeBuCdiYnXSc2CA. We recruited respondents through mailing lists and social media channels calling specifically for data professionals (e.g., data scientists, analysts, or engineers as well as domain experts who have analyzed data).

, ,
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

**Respondent characteristics.** From the 99 responses we excluded 10 responses of respondents who either explicitly expressed to not have experience with data tasks for work, or were researchers outside of applied fields like retail or finance (and could thus bias our results, for example, if they were CS researchers). In total, our analysis included 89 responses. Most respondents were data scientists, analysts, or engineers, as shown in Figure 1. We had a good spread of professional experience among participants: < 3 years (16%), 3-5 years (37.4%), 5-8 years (17.2%), and >8 years (23.2%).

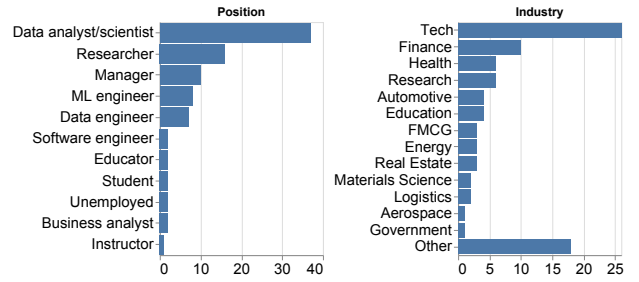


**Figure 1: Roles and organization domains of the survey respondents included in this analysis.**

## 2.1 The Practitioner's Perspective

**Why and where we search.** As discussed, dataset search generally serves one of two goals: 1) identifying an initial dataset for a given task, 2) finding a table to enrich a certain dataset with. The responses show that the search objective is skewed towards the first (79%) while enriching a target dataset is also important (52%). More respondents seek data in internal databases (70%) than external data repositories (61%).

> Takeaway 1. *Search most frequently serves the identification of the initial dataset(s) useful for a given task.*

**Tools used for search.** On inquiring which tools respondents use to search, we observed three frequent terms "SQL" (13), "internal" (14), and "Google" (29). We group the responses into these three categories, and add the category "colleagues" due to recurring comments about discussions with colleagues to inform search. We find that public interfaces (e.g. Google Dataset Search, public data repositories) were mentioned most as seen in Figure 2, aligning with our earlier finding that data workers search both internal and external data stores relatively equally. When searching for internal datasets, databases (e.g. Hive, Databricks or Big Query) and SQL were commonly mentioned (29%), whereas only 9% reported using specialized data search tools such as Domo or tools built in-house.

> Takeaway 2. *For internal search, search features of databases are used most. Only 9% use specialized dataset search tools.*

**Properties of tables searched for.** We asked respondents to identify the properties they consider important when expressing their search needs, specifically relating to content and metadata of structured datasets. With regards to the content, unsurprisingly, the semantics of the table are considered most important (Figure 3).

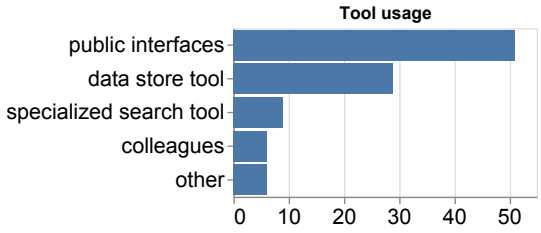


**Figure 2: Tools used in practice to search for datasets.**

However, the survey uncovers that the *granularity* of content (e.g., geographic or temporal) in the table is relevant too, which is rarely an attribute that existing systems enable search over. With regards to metadata, we find that table dimensions, freshness, frequency, and detailed schema are considered most relevant.

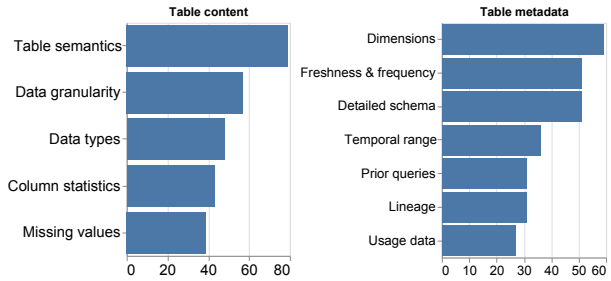


**Figure 3: Importance of table properties for expressing data search intent, along content and metadata axes.**

> Takeaway 3. *Semantics, freshness and frequency, dimensions and granularity, are most important in expressing search intent.*

**The process behind data search.** To better understand the full process behind "basic dataset search", we asked respondents to select the methods they use when searching for datasets. Unsurprisingly, querying data stores is most common. As shown in Figure 4, it appears that consultations with coworkers and experts is the second most common method (61%). Filters and categories are also frequently used when expressing one's search need (43%).

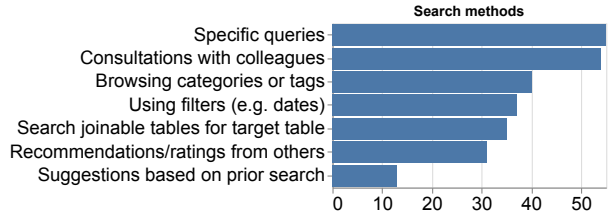


**Figure 4: Methods used to search for datasets.**

Among free-form explanations, we find the below quotes which resemble many similar responses within the respective category. Apart from querying a database, most quotes indicate a collaborative workflow where the search need is informed by coworkers.

> *"Identify the problem and the data for the problem. Then based on the data needed to answer the problem used specific keyword or tag search. Also, identify people who have worked on similar problems and try to contact them to understand data they used."* (Using specific search queries)
>
> *"Having so many tables, I ask more experienced colleagues which ones are most inherent to the analysis I need to do. I then navigate through the categories and tags to looks for others."* (Consultation with coworkers or experts)
>
> *"I use keyword search provided by the repository. Sometimes I browse through categories or use them as filters in addition to keyword search."* (Browsing categories or tags)
>
> *"During consultation with coworkers, I try to determine the use case we are solving and relevant data we'll need to analyse. - Sometimes we have to find the needle in the haystack so we search using common keywords in error logs -Date Time tags are most useful in that respect."* (Utilizing search filters)

Takeaway 4. *Dataset search is not just a query, but an iterative and collaborative process involving many humans-in-the-loop.*

**Why existing systems do not serve search needs.** We inquired about challenges people face in dataset search with existing processes and systems. Respondents' most pressing challenge is the syntactic inconsistencies between table content and metadata, motivating more focus on accurate semantic search methods. Other challenges include an excess of attributes (and datasets), unclear granularity of the data, and more issues listed in Figure 5.

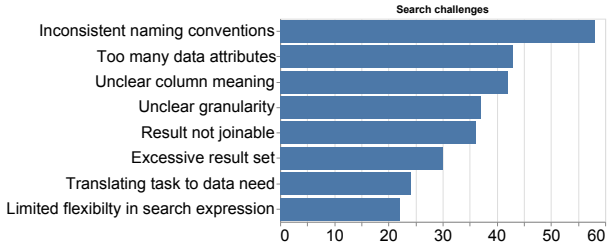


**Figure 5: Challenges in existing dataset search workflows.**

Below, a few exemplary quotes for the most common categories:

> *"Same fields may have different names in two tables, and sometimes the same name may means different in two tables."* (Inconsistent naming conventions)
>
> *"Public real estate datasets often are very wide and contain many columns that are not used for analysis. ... This often confuses junior researchers who don't know which columns are relevant and which are not."* (Too many data attributes)
>
> *"Once I had to do an analysis but it was painful because almost every column had unrecognizable information (like encrypted) it took longer than I was expecting"* (Column values with unclear meanings)
>
> *"Categorical level of detailing is required, which is not possible this days."* (Unclear granularity)

These quotes indicate that the retrieved results are hard to digest and navigate due to unclear semantics and abundance of tables as well as their large sizes. We also identified many quotes emphasizing the limited expressiveness of existing query interfaces.

Takeaway 5. *The syntactic data inconsistencies and query flexibility are the most challenging aspects of existing systems.*

**Imagining the ideal dataset search system.** Our final survey questions asked practitioners what they envision for an ideal dataset search system. In their free-form responses, we recognize five main themes, ordered by frequency: semantic search capabilities, more flexible and richer querying, SQL or NL interfaces to data stores, task- or question-driven search, and joinability. Figure 6 shows a distribution of these themes.

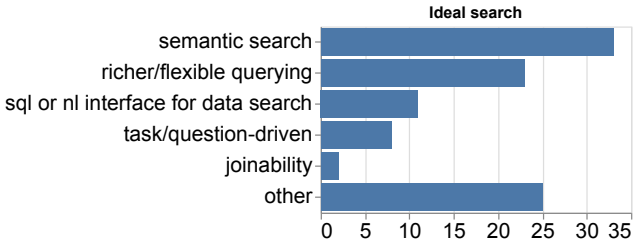


**Figure 6: General themes expressed in inquiring about the ideal search system would be.**

We highlight interesting quotes for the most common themes:

> *"Topic model search results, based on sentence similarity with the dataset description."* (Semantic search)
>
> *"Show me product usage datasets where the main fact table is event-level usage data with hundreds of millions of records and there are dimension tables for user and account."* (Richer/flexible search)
>
> *"Ideally I would have something that could work across all of the various datasources and table and be able to use SQL (or a trustable NLP solution) and pull all the relevant data and metadata."* (SQL or NL interface for data search)
>
> *"Dataset to <solve issue of...> with columns <1,2,3,...> on <granularity desired>"* (Task- or question-driven)

Clearly, semantic search capabilities are key to develop further, whereas a richer set of queryable metadata is strongly desired too. Across many quotes, we also detect a number of conversationally expressed search needs, which we discuss further in Section 3.

Takeaway 6. *Future data search systems are ideally semantic, task-driven, iterative, and coupled with flexible query interfaces.*

## 3 DESIDERATA FOR DATA SEARCH SYSTEMS

Based on our survey analysis, we distill the following four desiderata to inform next-generation dataset search systems.

**D1: Iterative Interfaces.** Respondents highlight two shortcomings regarding the interaction with existing search interfaces. First, existing interfaces require the search intent to be condensed into a single query. However, we find that search needs can be more complex and do not necessarily fit into a one- or two-step procedure. Second, dataset search appears a time-consuming back-and-forth process involving data scientists/analysts, engineers, and domain stakeholders. We wonder: is the search bar still the best suited interface to express complicated search needs? How can we move beyond the search bar? We suggest exploring more iterative interfaces to guide pruning the search space. Conversational interfaces wrapped around generative language models could enable breaking down complicated search queries iteratively [15] while replicating data science and domain expertise.

**D2: Hybrid Querying.** Accurately capturing and flexibly querying the semantics of the tables appears critical. Existing search systems are too sensitive to syntactic inconsistencies, necessitating semantic representations of the table content. Neural table embeddings have been shown to robustly capture the semantics of tables [20]. But our survey analysis also reveals that just semantics won't get the job done. Search intent for data analytics tasks operates over table semantics as well as table metadata (e.g. granularity and freshness), hence should both be treated as first-class citizens in search systems. We need to facilitate querying over hybrid representations seamlessly. The dimensions of metadata should also expand beyond what is supported now, and can arise from curated metadata frameworks like Croissant [5] and Google Dataset Search [8], or build on custom metadata schemas possibly populated using LLMs.

**D3: Task-driven Search.** As many quotes in this paper illustrate, users often don't know what kind of dataset would serve their needs or which datasets are available in the first place. However, existing systems assume that the mapping from use-case to data specification is complete, necessitating consultations with co-workers to inform this specification. We believe that systems should enable searching for datasets for a specific use-case, going beyond question-oriented retrieval as in recent work [18, 23] to serve data professionals who aim to retrieve dataset(s) for, for example, "*developing an ML pipeline for predictive maintenance for motor engines*".

**D4: Result Comprehensibility and Diversity.** With an increasing growth in data volumes, hence large search spaces, the set of tables resulting from a single query appears too excessive for users to navigate through. While more iterative interfaces help pruning this space, practitioners find a high semantic overlap across retrieved results due to generic schema semantics and duplication of tables and columns. Including more metadata attributes such as popularity, granularity, and freshness may be a strong component in the solution, it might not solve semantic overlap in the result set entirely. To further address this pain point, we can take inspiration from the communities working on information retrieval and recommendation systems and enforce result diversity and drill-down techniques [15, 16].

## 4 CONCLUSION

While systems and algorithmic papers on dataset search abound, we lack a human-centered perspective on the processes and open challenges in dataset search. We presented an analysis of a survey across data professionals for data analytics use-cases. Our insights reveal a need for systems with more flexible retrieval engines and richer search interfaces. We propose four concrete desiderata for next-generation search systems: 1) iterative interfaces, 2) hybrid querying, 3) task-driven search, and 4) result diversity.

## REFERENCES


[1] [n. d.]. Adding Intelligence to Databricks Search. https://www.databricks.com/blog/adding-intelligence-to-databricks-search. Accessed: 2024-03-29.

[2] [n. d.]. A peek inside how Snowflake's new Universal Search feature was built. https://medium.com/snowflake/a-peek-inside-how-snowflakes-new-universal-search-feature-was-built-dfd1188176d0. Accessed: 2024-03-29.

[3] [n. d.]. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/. Accessed: 2024-03-29.

[4] 2023. Datahub: A Modern Data Catalog. https://datahubproject.io/docs/next.

[5] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, et al. 2024. Croissant: A Metadata Format for ML-Ready Datasets. *arXiv preprint arXiv:2403.19546* (2024).

[6] Michael Armbrust, Ali Ghodsi, Reynold Xin, and Matei Zaharia. 2021. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR*, Vol. 8. 28.

[7] Alex Bogatu, Norman W Paton, Mark Douthwaite, and Andre Freitas. 2022. Voyager: Data discovery and integration for data science. *JDIQ* (2022).

[8] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *WWW*. 1365–1375.

[9] Riccardo Cappuzzo, Gael Varoquaux, Aimee Coelho, and Paolo Papotti. 2024. Retrieve, Merge, Predict: Augmenting Tables with Data Lakes. *arXiv preprint arXiv:2402.06282* (2024).

[10] Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2791–2794.

[11] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. *The VLDB Journal* 29, 1 (2020), 251–272.

[12] Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Renée J Miller. 2022. Semanticsaware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. CoRR abs/2210.01922 (2022). *arXiv preprint arXiv.2210.01922* (2022).

[13] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 1001–1012.

[14] Javier de Jesús Flores Herrera, Sergi Nadal Francesch, and Óscar Romero Moral. 2021. Towards scalable data discovery. In *EDBT'21*. 433–438.

[15] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural approaches to conversational information retrieval*. Vol. 44. Springer Nature.

[16] Hector Garcia-Molina, Georgia Koutrika, and Aditya Parameswaran. 2011. Information seeking: convergence of search, recommendations, and advertising. *Commun. ACM* 54, 11 (2011), 121–130.

[17] Mark Grover. 2019. Amundsen — Lyft's data discovery & metadata engine. https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine-62d27254fbb9. Accessed: 2024-03-29.

[18] Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. *arXiv preprint arXiv:2103.12011* (2021).

[19] Zezhou Huang, Jiaxiang Liu, Haonan Wang, and Eugene Wu. 2023. The Fast and the Private: Task-based Dataset Search. *arXiv preprint arXiv:2308.05637* (2023).

[20] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çagatay Demiralp, and César Hidalgo. 2019. Sherlock: A deep learning approach to semantic data type detection. In *KDD Conference (KDD'19)*.

[21] Moe Kayali, Anton Lykov, Ilias Fountalis, Nikolaos Vasiloglou, Dan Olteanu, and Dan Suciu. 2023. CHORUS: foundation models for unified data discovery and exploration. *arXiv preprint arXiv:2306.09610* (2023).

[22] Vidya Setlur, Andriy Kanyuka, and Arjun Srinivasan. 2023. Olio: A Semantic Search Interface for Data Repositories. In *UIST'23*. 1–16.

[23] Qiming Wang and Raul Castro Fernandez. 2023. Solo: Data Discovery Using Natural Language Questions Via A Self-Supervised Approach. In *SIGMOD*.