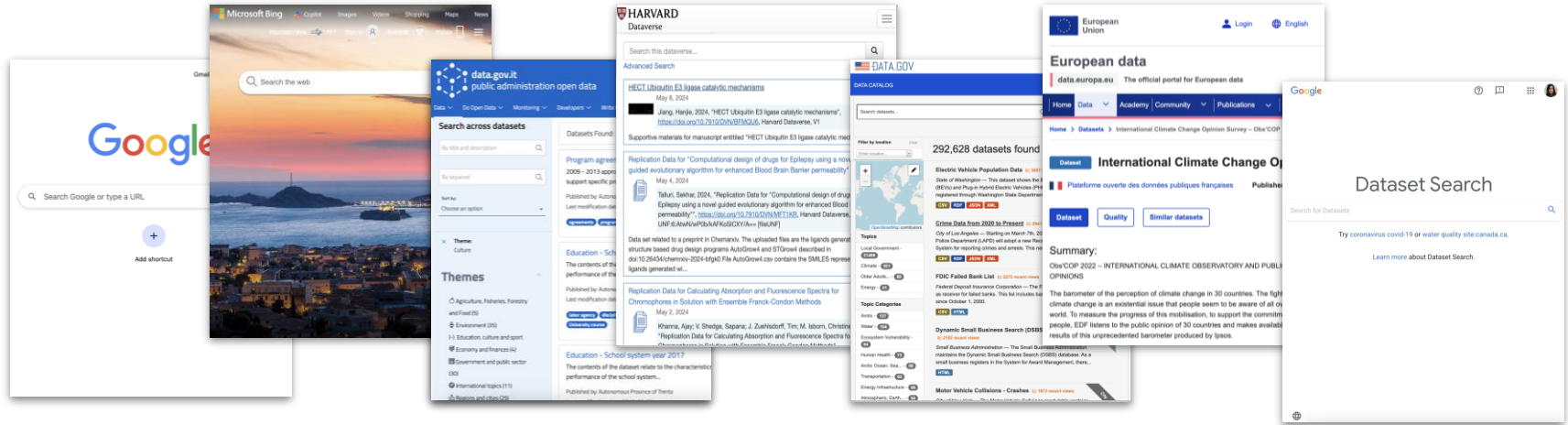


Discovering Datasets on the Web Scale: Challenges and Recommendations for Google Dataset Search



Katrina Sostek, Daniel M. Russell, Nitesh Goyal, Tarfah Alrashed, Stella Dugall, and Natasha Noy

More Data on Web... More Dataset-Search Tools...

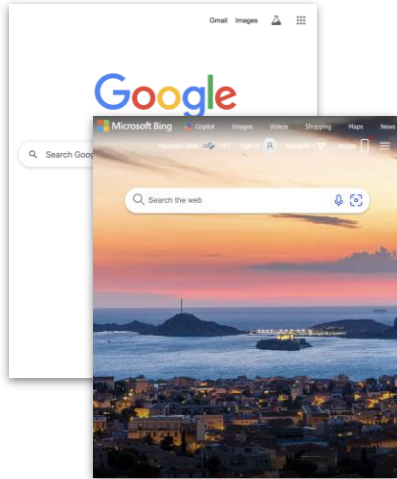


The Scope of Dataset-Search Tools

General Purpose Search Tools



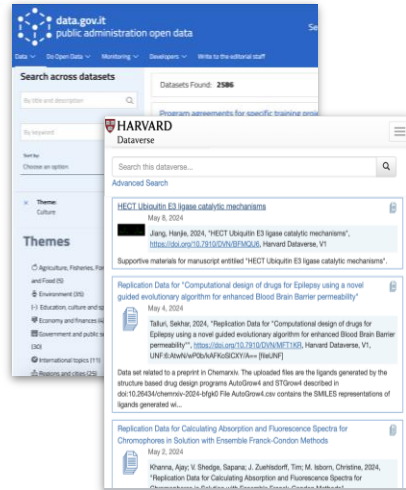
General Purpose Web Search



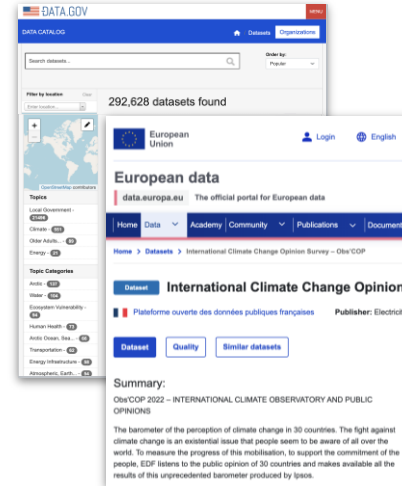
Dataset-Specific Search Tools



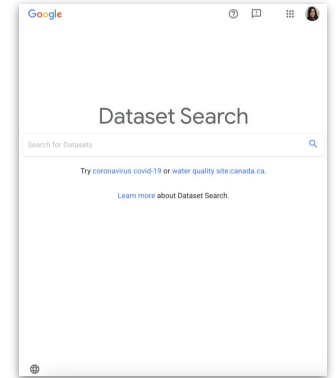
Dataset Repositories



Dataset Meta-Portals



Dataset-Specific Web Search

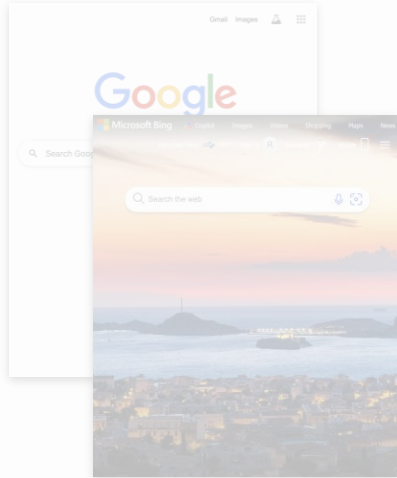


The Scope of Dataset-Search Tools

General Purpose Search Tools



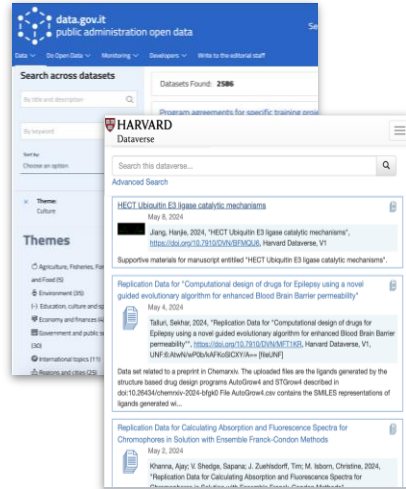
General Purpose Web Search



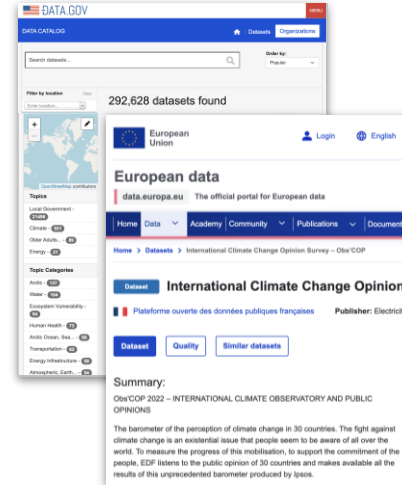
Dataset-Specific Search Tools



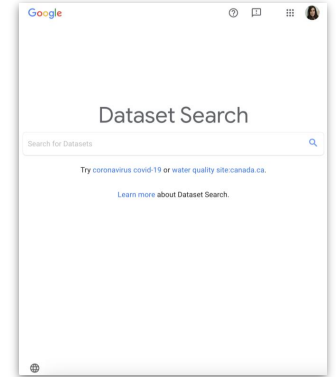
Dataset Repositories



Dataset Meta-Portals



Dataset-Specific Web Search

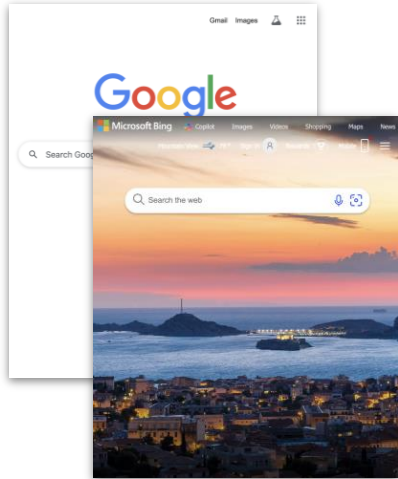


The Scope of Dataset-Search Tools

General Purpose Search Tools



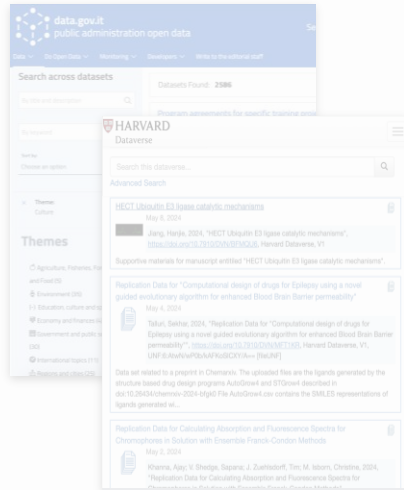
General Purpose Web Search



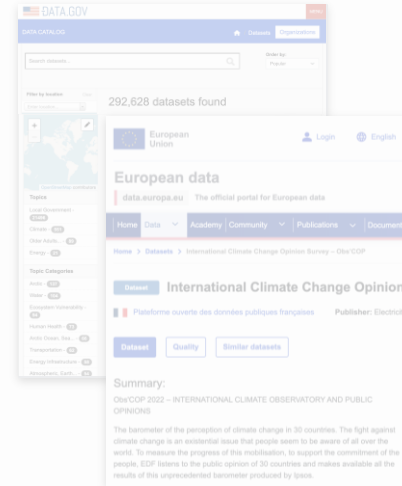
Dataset-Specific Search Tools



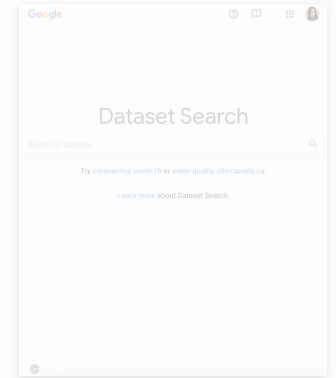
Dataset Repositories



Dataset Meta-Portals



Dataset-Specific Web Search



Previous Studies of Dataset-Discovery Tools

General Purpose Search Tools



General Purpose Web Search

Dataset-Specific Search Tools



Dataset Repositories



Dataset Meta-Portals



Dataset-Specific Web Search

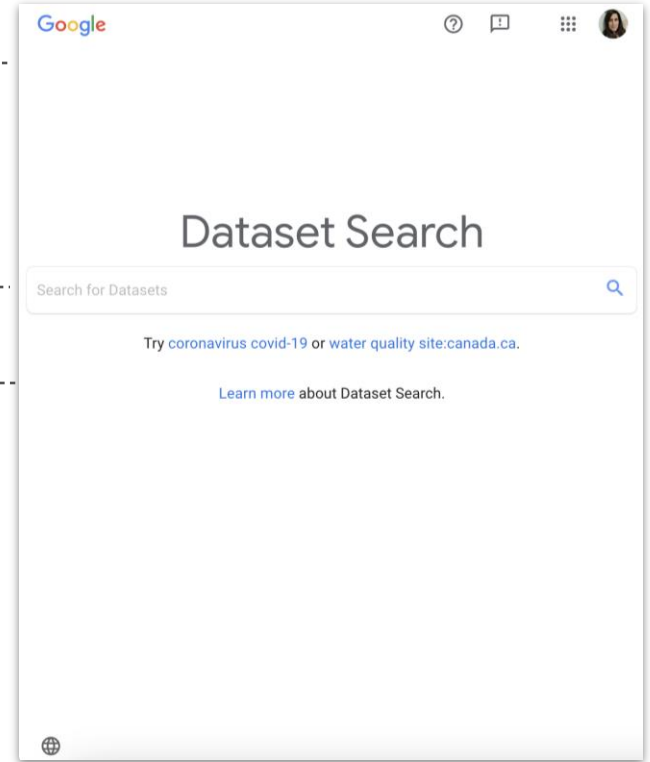
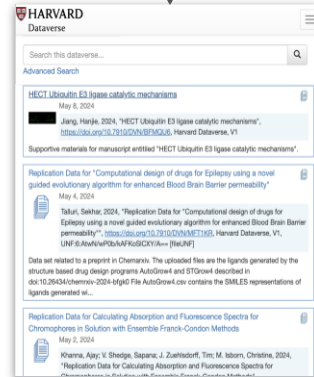
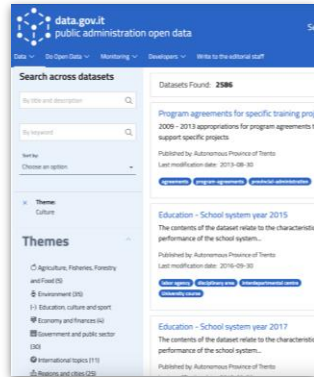
Study Method	General Purpose Web Search	Dataset Repositories	Dataset Meta-Portals	Dataset-Specific Web Search
User Interviews	✓	✓	✓	
User Surveys	✓	✓	✓	
Logs Analysis		✓	✓	
Tool Reviews	✓	✓	✓	✓

The Scope of Google Dataset Search

Dataset Meta-Portals



Dataset Repositories



datasetsearch.research.google.com

Google Dataset Search

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?

The screenshot shows a Google search for 'sitde:dataverse.harvard.edu'. The search results are filtered by 'Provider' (Free) and 'Usage rights' (Croissant). The top result is 'Melbourne's grassland GIS site polygons, and inventory and transect points' from dataverse.harvard.edu, updated May 6, 2015. The result card includes a unique identifier (https://doi.org/10.7910/DVN/SSLIXK), dataset update date, provider name, license (CC0 1.0 Universal Public Domain Dedication), area covered (Melbourne), and a description of the dataset contents.

Google

▼ Last updated ▼ Download format Croissant ▼ Usage rights ▼ Topic Saved datasets

▼ Provider Free

Updated Aug 21, 2021

Melbourne's grassland GIS site polygons, and inventory and transect points

Explore at:

Unique identifier
<https://doi.org/10.7910/DVN/SSLIXK>

Dataset updated
 May 6, 2015

Dataset provided by
 Harvard Dataverse

License
 CC0 1.0 Universal Public Domain Dedication
License information was derived automatically

Area covered
 Melbourne

Description
 A zip file containing three ARCGIS shape files for: 1. Polygons of Melbourne's grassland reserves 2. Points for the start and end of each transect. See the associated inventory and vegetation cover datasets for species information collected over the summer of 2013-14.

Archaeological Sites and Villages of the Mosul Dam Reservoir...
 dataverse.harvard.edu
 Updated Apr 7, 2020

Melbourne's grassland GIS site polygons, and inventory and...
 dataverse.harvard.edu
 search.dataone.org
 Updated May 6, 2015

Replication data and minimal code examples for: "The Trigger..."
 dataverse.harvard.edu
 bin, nc +2
 Updated Oct 21, 2019

Supplementary data of 40 lunar sampling sites
 dataverse.harvard.edu
 data.mendeley.com

User Interviews

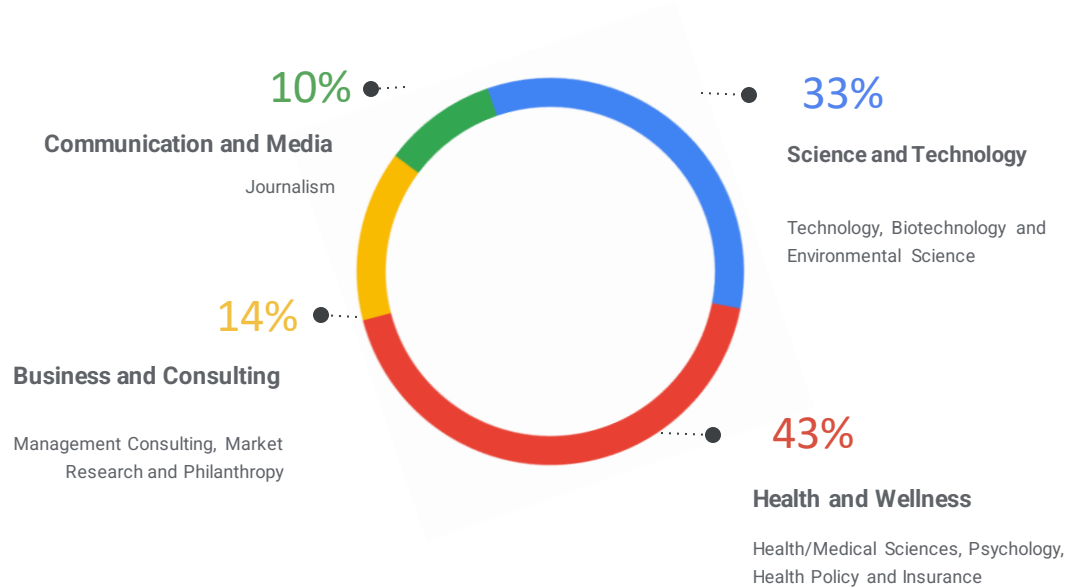
Participants

- We recruited 20 participants: 12 female, 10 male
- Seek datasets at least once a month

Study Design

90-minute semi-structured virtual interviews.

- Background and motivations (~15 min)
- Recent dataset search challenges (~45 min)
- Dataset Search usage (~30 min)



User Interviews

Participants

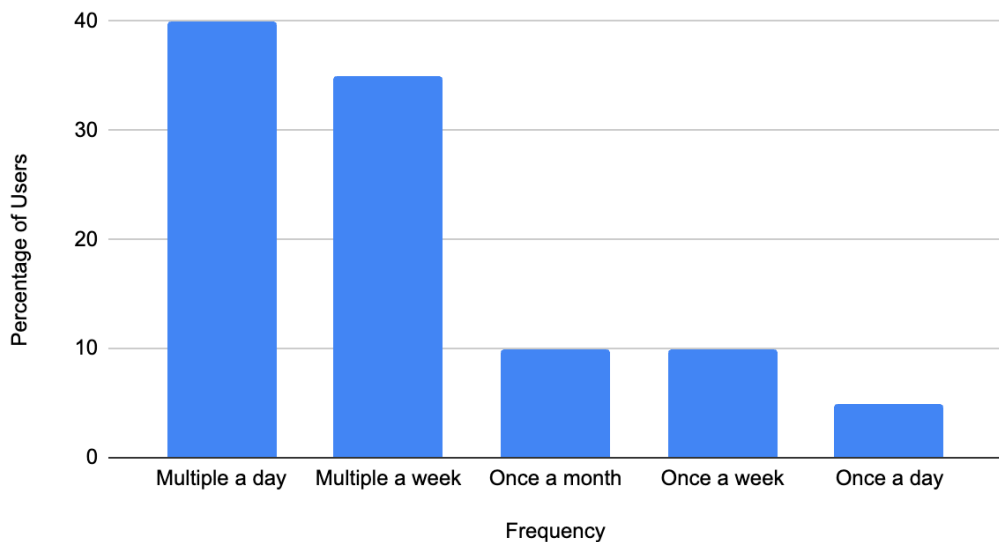
- We recruited 20 participants: 12 female, 10 male
- Seek datasets at least once a month

Study Design

90-minute semi-structured virtual interviews.

- Background and motivations (~15 min)
- Recent dataset search challenges (~45 min)
- Dataset Search usage (~30 min)

How often do you search for datasets?



Findings

Our findings correspond to three themes:

1. Mental model development

Participants initially relied on familiar tool knowledge to understand the tool, updating their understanding as they used it. However, some initially confused its purpose with other tools.

2. Dealing with diverse datasets

Participants confirmed previous findings that incomplete and inconsistent metadata pose challenges, emphasizing the importance of accessing underlying data and ensuring trustworthiness. Dataset Search exacerbates these issues due to its wide range of sources and varying interfaces.

3. Data search proficiency

Some participants lacked familiarity with dataset-specific concepts, suggesting the need for tutorials and in-tool help to improve users' data search skills.

Building a Mental Model of the Tool

Challenge#1: Dataset Search and Google Web Search.

- Set user expectations that not all datasets on the web are in the tool.
- Fail gracefully when there are few results for a query.
- Compare mental model development in dataset web search between Google and standalone use.

Challenge#2: User expectations about the scope of the tool.

- Make it easier to find related artifacts in their tools.
- Make it easier to find datasets in the context of their search results.
- Infer the artifacts associated with datasets.

The screenshot shows a Google search for "canada water quality data". The search bar is at the top with the Google logo and search icons. Below the search bar are tabs for Images, News, Shopping, Ranking, Environment, Videos, Maps, Books, and Flights. The search results show "About 433,000,000 results (0.37 seconds)". A red box highlights the "Datasets" section, which lists three datasets:

- National Long-term Water Quality Monitoring Data**: May 20, 2022 — Long-term freshwater quality data from federal and federal-provincial sampling sites throughout Canada's aquatic ecosystems are included in this dataset. Measurements regularly ... License: Open Government Licence - Canada 2.0 Format(s): csv, esri rest, ...
- Great Lakes Water Quality Monitoring and Aquatic Ecosystem Health Data**: Dec 7, 2021 — Water quality and ecosystem health data collected using a risk-based monitoring approach to support the Great Lakes Water Quality Agreement are included in this dataset. By... License: Open Government Licence - Canada 2.0 Format(s): html
- Drinking Water Quality and Enforcement**: May 19, 2023 — Ontario has a comprehensive set of measures and regulations to help ensure the safety of drinking water. The following dataset contains information about the drinking water systems, ... Format(s): pdf, zip

Below the datasets list is a "More datasets →" button. At the bottom of the search results, there is a result from Statistics Canada: "Water quality in Canadian rivers - Statistics Canada" with a date of Sep 6, 2023.

Results from Dataset Search in Google Web Search

Building a Mental Model of the Tool

Challenge#1: Dataset Search and Google Web Search.

- Set user expectations that not all datasets on the web are in the tool.
- Fail gracefully when there are few results for a query.
- Compare mental model development in dataset web search between Google and standalone use.

Challenge#2: User expectations about the scope of the tool.

- Make it easier to find related artifacts in their tools.
- Make it easier to find datasets in the context of their search results.
- Infer the artifacts associated with datasets.

The screenshot shows a Google search interface with the query 'water quality site:canada.ca'. The search results are filtered to show datasets. The top result is 'National Long-term Water Quality Monitoring Data' from the Government of Canada. Below the main result, there are several 'Explore at' links: 'Government of Canada Open G...', 'catalogue.arctic-sdi.org', 'AmeriGEOSS Community Platfo...', and 'data.wu.ac.at'. A red box highlights a link that says '29 scholarly articles cite this dataset (View in Google Scholar)'. A red arrow points from this link down towards the text at the bottom of the slide.

Some datasets are linked to scientific articles in Google Scholar

Building a Mental Model of the Tool

Challenge#3: Dataset replication across the web.

- Add indications in the UI explaining relationships between replicas
- Display dataset version types and sources in a clear, meaningful way.
- Infer the provenance of datasets
- Infer relationships between datasets

It also finds **versions** now!

Dataset Search finds **replicas** of datasets

The screenshot shows a Google search interface with the following elements:

- Search Bar:** "water quality site:canada.ca"
- Filters:** Last updated, Download format, Usage rights, Topic, Provider, Free, Saved datasets
- Results:** 100+ datasets found. The top result is "National Long-term Water Quality Monitoring Data" from Canada, updated May 20, 2022. It lists sources like "open.canada.ca" and "catalogue.arctic-sdi.org". A red box highlights the "See More Versions" link.
- Explore at:** A section with four blue buttons: "Government of Canada Open G...", "catalogue.arctic-sdi.org", "AmeriGEOSS Community Platfo...", and "data.wu.ac.at". A red box highlights these buttons.
- Additional Info:** "29 scholarly articles cite this dataset", "esri rest, csv, wms, html, txt", "Dataset updated: May 20, 2022", "Dataset provided by: Environment And Climate Change Canada", "License: Open Government Licence - Canada 2.0", and "Description: Long-term freshwater quality data from federal and federal-provincial sampling sites throughout Canada's aquatic ecosystems..."

Making Sense of Heterogeneous Datasets

Participants confirmed previous findings on dataset metadata limitations. While helpful for deciding on exploration, metadata lacked clarity for understanding datasets and determining usability. Dataset Search metadata lacks consistency due to diverse sources.

Challenge#1: Metadata shortcomings and inconsistencies

- Allow users to tailor their views and metadata selection based on their specific needs.
- Infer missing metadata fields.

Making Sense of Heterogeneous Datasets

Challenge#2: Downloading underlying data

- Add dataset download links, descriptions of content, previews, visualizations if not already supported.
- Infer formats and schema from underlying data, regardless of the accuracy of the dataset metadata.

Challenge#3: Building trust in an unfamiliar data sources.

- Provide signals about the trustworthiness or quality of data sources.
- Show more context about data sources to help users build trust.
- Enable users to customize views and results based on trust signals.
- Study how users use trustworthiness signals when datasets have incomplete metadata and artifacts.

Google search results for "water quality site:canada.ca". The search results show three datasets from the National Long-term Water Quality Monitoring Data. The first dataset is highlighted with a red arrow pointing to its preview table. The table has columns: VNM_CODE, NATIONAL_VARIABLE_CODE, VARIABLE_COMMON_NAME, and VARIABLE_COMMON_NAME_FR. The table contains 10 rows of data for various elements like ALUMINUM, BARIUM, BERYLLIUM, CADMIUM, COBALT, CHROMIUM, COPPER, IRON, LITHIUM, and MANGANESE.

VNM_CODE	NATIONAL_VARIABLE_CODE	VARIABLE_COMMON_NAME	VARIABLE_COMMON_NAME_FR
100081	371	ALUMINUM EXTRACTABLE	ALUMINIUM EXTRACTIBLE
100082	387	BARIIUM EXTRACTABLE	BIARIUM EXTRACTIBLE
100083	392	BERYLLIUM EXTRACTABLE	BERYLLIUM EXTRACTIBLE
100084	410	CADMIUM EXTRACTABLE	CADMIUM EXTRACTIBLE
100085	454	COBALT EXTRACTABLE	COBALT EXTRACTIBLE
100086	461	CHROMIUM EXTRACTABLE	CHROME EXTRACTIBLE
100087	468	COPPER EXTRACTABLE	CUVRE EXTRACTIBLE
100088	517	IRON EXTRACTABLE	FER EXTRACTIBLE
100089	575	LITHIUM EXTRACTABLE	LITHIUM EXTRACTIBLE
100090	589	MANGANESE EXTRACTABLE	MANGANESE EXTRACTIBLE

Dataset Previews

Learning How to Search for Data

Dataset searching requires data literacy and domain knowledge. Participants were unfamiliar with tool concepts like data formats and usage rights. Some discussed user learning improvements.

Dataset discovery is a skill and needs to be part of data literacy

- Create more educational materials and in-tool information about licenses, data formats, etc.
- Add educational information to help novice users, for example, specialized terminology, trustworthiness signals
- Develop educational content on web-scale dataset discovery for data literacy.

The screenshot shows the Google Dataset Search User Support Center page. At the top, there is a navigation bar with links for 'About', 'What's New', 'For Users', 'For Providers', and 'Contact'. Below this is the 'User Support Center' header. The main content area is divided into two columns. The left column features a 'Dataset Search' section with a search bar and the text '25 million datasets at your fingertips' followed by a link to 'g.co/datasetsearch'. The right column contains three sections: 'What is Dataset Search?' which explains that it is a search engine for datasets; 'Our Mission' which states the goal of making datasets accessible and useful, and lists two bullet points: 'Foster a data sharing ecosystem that will encourage data publishers to follow best practices for data storage and publication' and 'Give scientists a way to show the impact of their work through citation of datasets that they have produced'; and 'The Keyword Blog Post' which includes a link to 'Read more about Dataset Search in The Keyword blog post.' The page also features two images: one showing a computer screen with data charts and another showing a newspaper printing press.

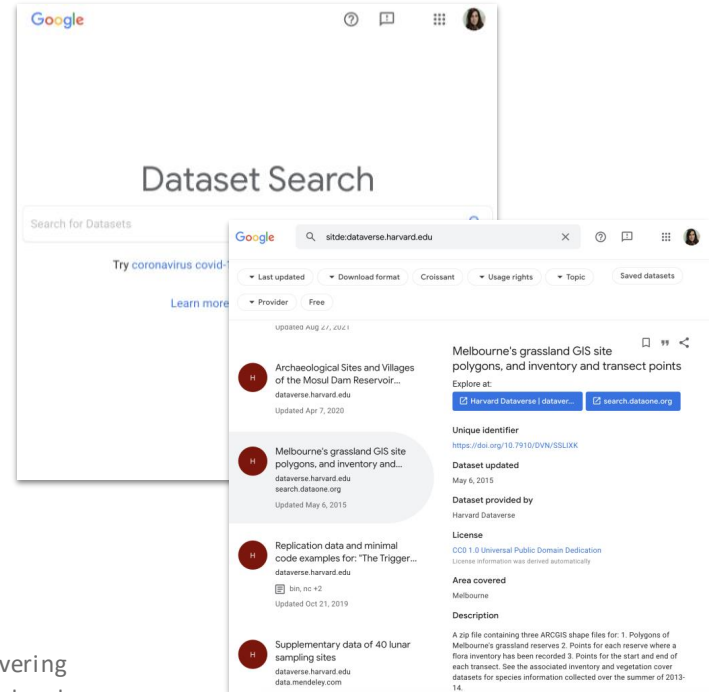
Limitations

- Dataset Search is the sole tool in its category, hindering broader comparison with similar tools.
- Participants varied in dataset discovery experience, shaping their interview approaches. However, common findings among participants enhanced the validity of our results.
- Participants used different queries rather than perform the same specific tasks.
- All participants were based in the US and used English-language queries.
- We did not have any participants who were regular users of Dataset Search.

Thank You

Any Questions?

Sostek, K., Russell, D. M., Goyal, N., Alrashed, T., Dugall, S., & Noy, N. (2024). Discovering Datasets on the Web Scale: Challenges and Recommendations for Google Dataset Search. *Harvard Data Science Review*, (Special Issue 4). <https://doi.org/10.1162/99608f92.4c3e11ca>



datasetsearch.research.google.com