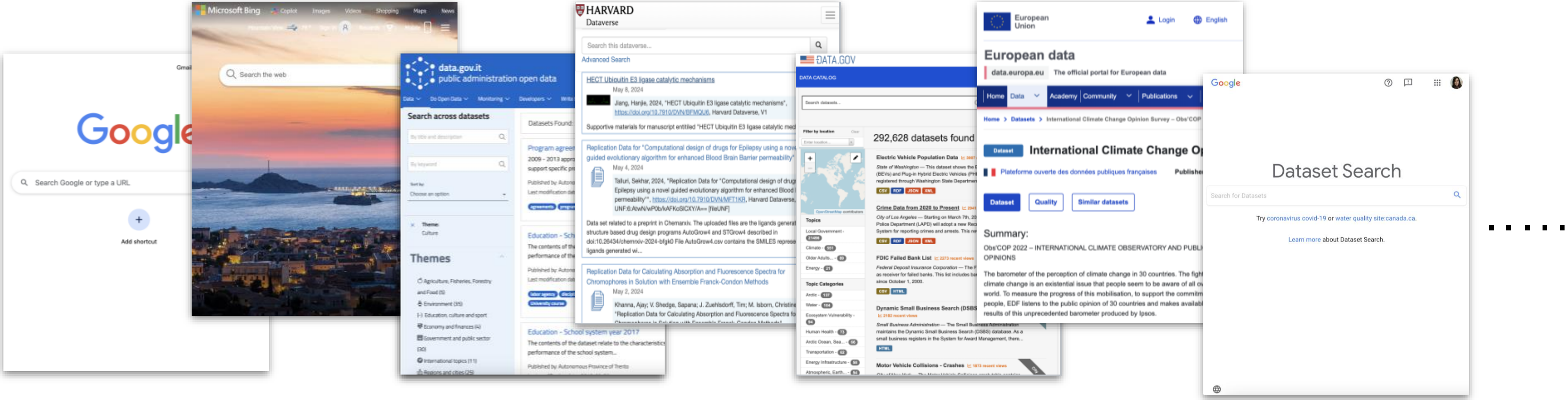
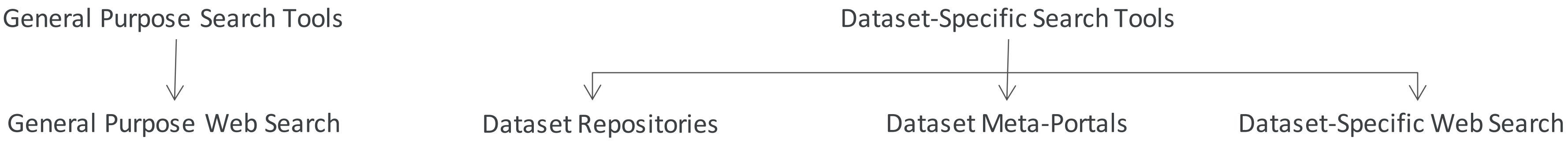
Discovering Datasets on the Web Scale: Challenges and Recommendations for Google Dataset Search

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

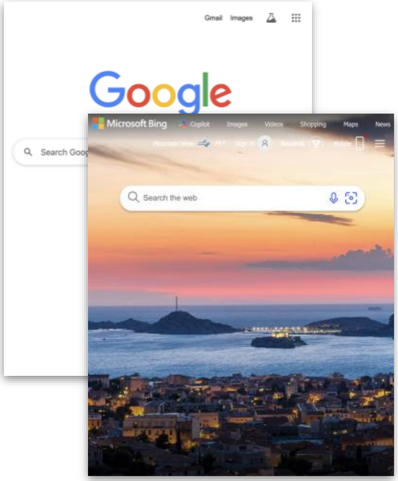
Dataset-Specific Search Tools

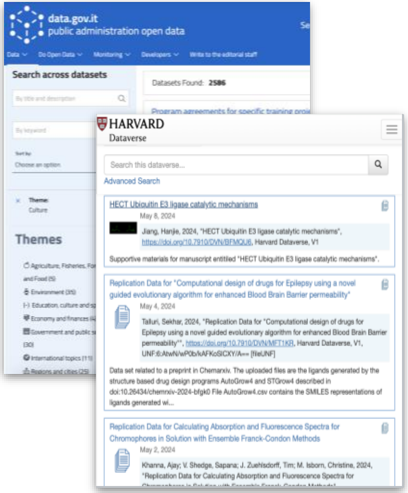
General Purpose Web Search

Dataset Repositories

Dataset Meta-Portals

Dataset-Specific Web Search





Datasets Found: 2586

Program agreements for specific training proje

HECT Ubiquitin E3 ligase catalytic mechanisms

May 8, 2024

Jiang, Hanjie, 2024, "HECT Ubiquitin E3 ligase catalytic mechanisms", https://doi.org/10.7910/DVN/BFMQU6, Harvard Dataverse, V1

Supportive materials for manuscript entitled "HECT Ubiquitin E3 ligase catalytic mechanisms".

Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"

May 4, 2024

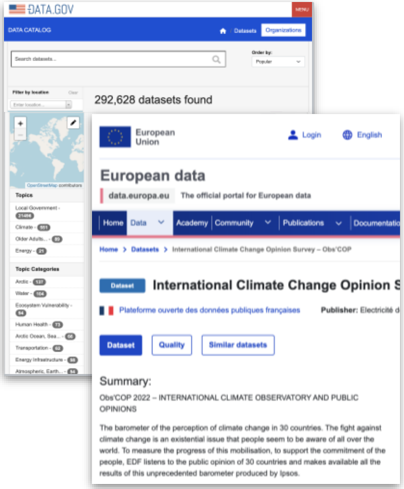
Talluri, Sekhar, 2024, "Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"", https://doi.org/10.7910/DVN/MFT1KR, Harvard Dataverse, V1, UNF:6:AtwN/wP0b/kAFKoSICXY/A== [fileUNF]

Data set related to a preprint in Chemarxiv. The uploaded files are the ligands generated by the structure based drug design programs AutoGrow4 and STGrow4 described in doi:10.26434/chemrxiv-2024-bfgk0 File AutoGrow4.csv contains the SMILES representations of ligands generated wi...

Replication Data for Calculating Absorption and Fluorescence Spectra for Chromophores in Solution with Ensemble Franck-Condon Methods

May 2, 2024

Khanna, Ajay; V. Shedge, Sapana; J. Zuehlsdorff, Tim; M. Isborn, Christine, 2024, "Replication Data for Calculating Absorption and Fluorescence Spectra for



292,628 datasets found

European data

data.europa.eu The official portal for European data

Home › Datasets › International Climate Change Opinion Survey – Obs'COP

International Climate Change Opinion S

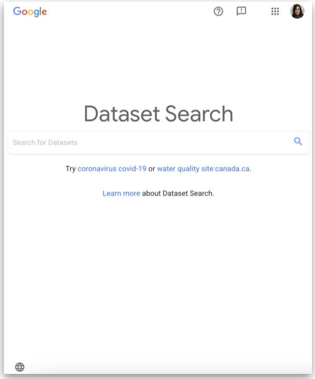
Plateforme ouverte des données publiques françaises

Publisher: Electricité d

Summary:

Obs'COP 2022 – INTERNATIONAL CLIMATE OBSERVATORY AND PUBLIC OPINIONS

The barometer of the perception of climate change in 30 countries. The fight against climate change is an existential issue that people seem to be aware of all over the world. To measure the progress of this mobilisation, to support the commitment of the people, EDF listens to the public opinion of 30 countries and makes available all the results of this unprecedented barometer produced by Ipsos.



Try coronavirus covid-19 or water quality site:canada.ca.

Learn more about Dataset Search.

The Scope of Dataset-Search Tools

General Purpose Search Tools

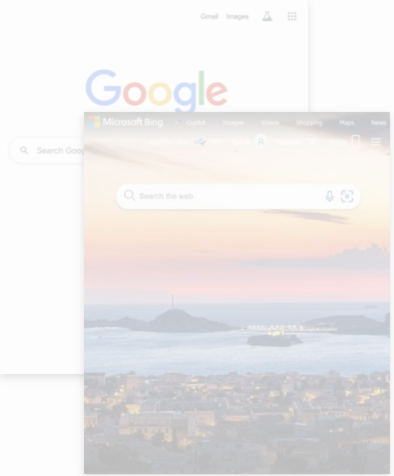
General Purpose Web Search

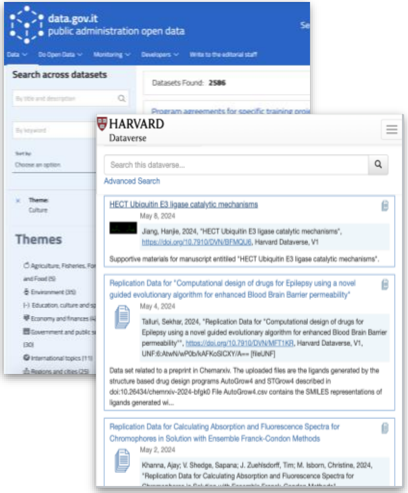
Dataset-Specific Search Tools

Dataset Repositories

Dataset Meta-Portals

Dataset-Specific Web Search





Search across datasets

Datasets Found: 2586

Program agreements for specific training proje

Theme:
Culture

Themes

HECT Ubiquitin E3 ligase catalytic mechanisms

May 8, 2024

Jiang, Hanjie, 2024, "HECT Ubiquitin E3 ligase catalytic mechanisms", https://doi.org/10.7910/DVN/BFMQU6, Harvard Dataverse, V1

Supportive materials for manuscript entitled "HECT Ubiquitin E3 ligase catalytic mechanisms".

Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"

May 4, 2024

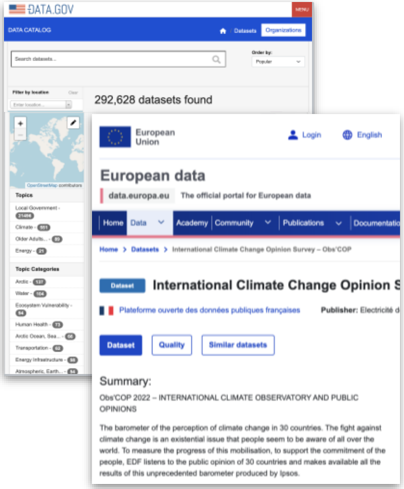
Talluri, Sekhar, 2024, "Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"", https://doi.org/10.7910/DVN/MFT1KR, Harvard Dataverse, V1, UNF:6:AtwN/wP0b/kAFKoSICXY/A== [fileUNF]

Data set related to a preprint in Chemarxiv. The uploaded files are the ligands generated by the structure based drug design programs AutoGrow4 and STGrow4 described in doi:10.26434/chemrxiv-2024-bfgk0 File AutoGrow4.csv contains the SMILES representations of ligands generated wi...

Replication Data for Calculating Absorption and Fluorescence Spectra for Chromophores in Solution with Ensemble Franck-Condon Methods

May 2, 2024

Khanna, Ajay; V. Shedge, Sapana; J. Zuehlsdorff, Tim; M. Isborn, Christine, 2024, "Replication Data for Calculating Absorption and Fluorescence Spectra for



292,628 datasets found

European data

data.europa.eu The official portal for European data

Home > Datasets > International Climate Change Opinion Survey – Obs'COP

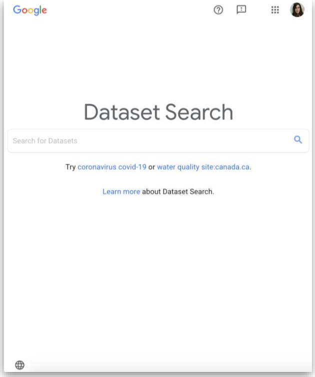
Dataset International Climate Change Opinion S

Plateforme ouverte des données publiques françaises Publisher: Electricité d

Summary:

Obs'COP 2022 – INTERNATIONAL CLIMATE OBSERVATORY AND PUBLIC OPINIONS

The barometer of the perception of climate change in 30 countries. The fight against climate change is an existential issue that people seem to be aware of all over the world. To measure the progress of this mobilisation, to support the commitment of the people, EDF listens to the public opinion of 30 countries and makes available all the results of this unprecedented barometer produced by Ipsos.



Dataset Search

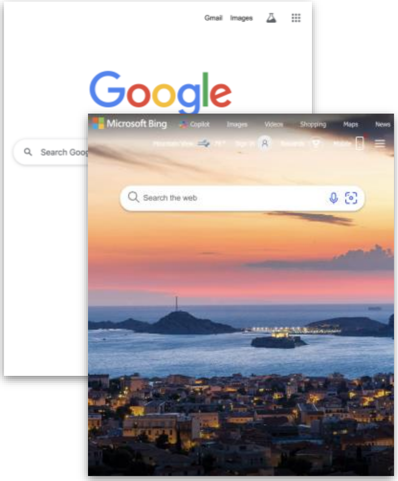
Try coronavirus covid-19 or water quality site:canada.ca.

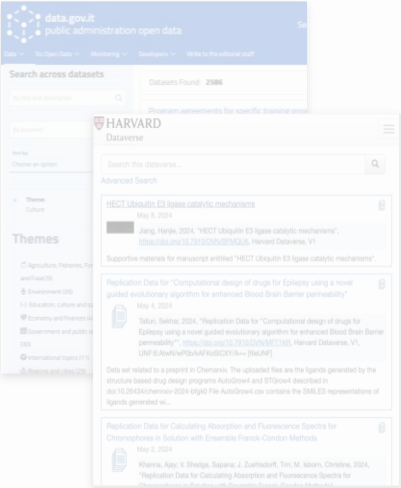
Learn more about Dataset Search.

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





data.gov.it
public administration open data

Search across datasets

Datasets Found: 2586

Program agreements for specific training proje

Theme: Culture

Themes

Agriculture, Fisheries, For and Food (5)
Environment (35)
Education, culture and sp
Economy and finances (4
Government and public s (30)
International topics (11)
Regions and cities (25)

HARVARD
Dataverse

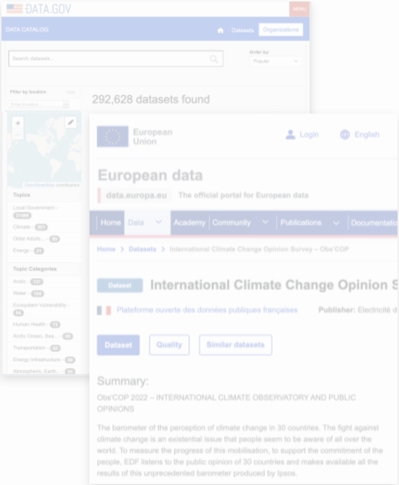
Search this dataverse...

Advanced Search

HECT Ubiquitin E3 ligase catalytic mechanisms
May 8, 2024
Jiang, Hanjie, 2024, "HECT Ubiquitin E3 ligase catalytic mechanisms", https://doi.org/10.7910/DVN/BFMQU6, Harvard Dataverse, V1
Supportive materials for manuscript entitled "HECT Ubiquitin E3 ligase catalytic mechanisms".

Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"
May 4, 2024
Talluri, Sekhar, 2024, "Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"", https://doi.org/10.7910/DVN/MFT1KR, Harvard Dataverse, V1, UNF:6:AtwN/wP0b/kAFKoSICXY/A== [fileUNF]
Data set related to a preprint in Chemarxiv. The uploaded files are the ligands generated by the structure based drug design programs AutoGrow4 and STGrow4 described in doi:10.26434/chemrxiv-2024-bfgk0 File AutoGrow4.csv contains the SMILES representations of ligands generated wi...

Replication Data for Calculating Absorption and Fluorescence Spectra for Chromophores in Solution with Ensemble Franck-Condon Methods
May 2, 2024
Khanna, Ajay; V. Shedge, Sapana; J. Zuehlsdorff, Tim; M. Isborn, Christine, 2024, "Replication Data for Calculating Absorption and Fluorescence Spectra for



DATA.GOV
DATA CATALOG
Datasets
Organizations
Search datasets...
Order by: Popular
Filter by location
292,628 datasets found
Topics
Local Government
Climate
Older Adults...
Energy
Topic Categories
Arctic
Water
Ecosystem Vulnerability
Human Health
Arctic Ocean, Sea...
Transportation
Energy Infrastructure
Atmospheric, Earth...

European Union
Login
English

European data

data.europa.eu The official portal for European data

Home | Data | Academy | Community | Publications | Documentatio

Home > Datasets > International Climate Change Opinion Survey – Obs'COP

Dataset International Climate Change Opinion S

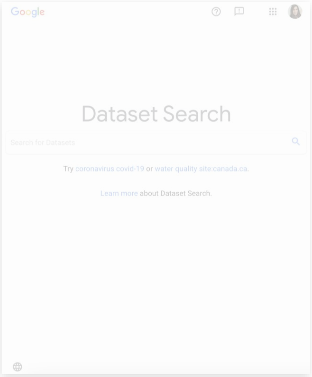
Plateforme ouverte des données publiques françaises Publisher: Electricité d

Dataset | Quality | Similar datasets

Summary:

Obs'COP 2022 – INTERNATIONAL CLIMATE OBSERVATORY AND PUBLIC OPINIONS

The barometer of the perception of climate change in 30 countries. The fight against climate change is an existential issue that people seem to be aware of all over the world. To measure the progress of this mobilisation, to support the commitment of the people, EDF listens to the public opinion of 30 countries and makes available all the results of this unprecedented barometer produced by Ipsos.



Google

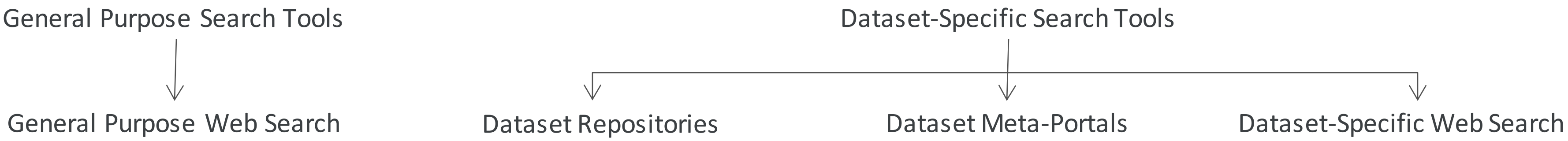
Dataset Search

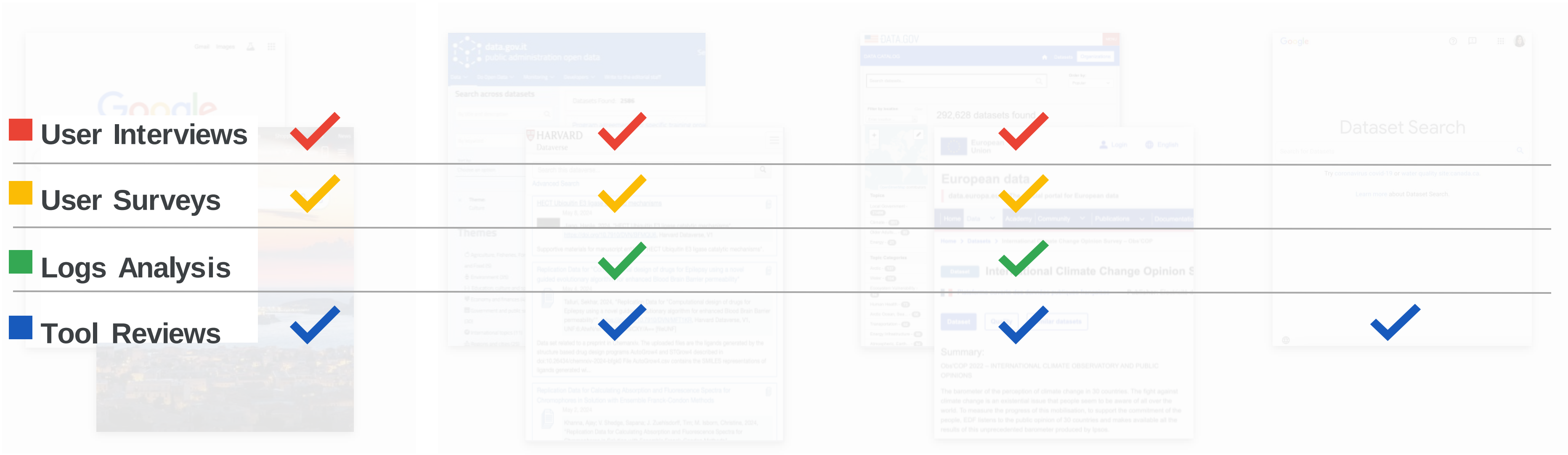
Search for Datasets

Try coronavirus covid-19 or water quality site:canada.ca.

Learn more about Dataset Search.

Previous Studies of Dataset-Discovery Tools

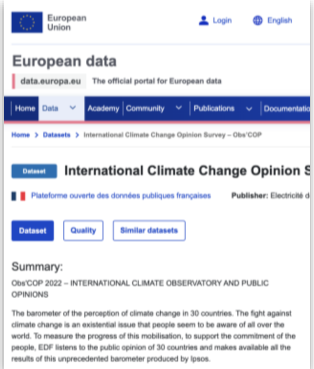

General Purpose Search Tools
Dataset-Specific Search Tools
General Purpose Web Search
Dataset Repositories
Dataset Meta-Portals
Dataset-Specific Web Search



	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



European Union

Login · English

European data

data.europa.eu The official portal for European data

Home | Data | Academy | Community | Publications | Documentatio

Home > Datasets > International Climate Change Opinion Survey – Obs'COP

Dataset **International Climate Change Opinion S**

Plateforme ouverte des données publiques françaises **Publisher:** Electricité d

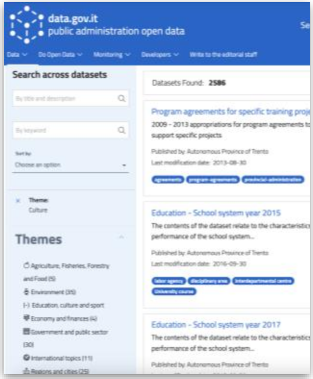
Dataset | Quality | Similar datasets

Summary:

Obs'COP 2022 – INTERNATIONAL CLIMATE OBSERVATORY AND PUBLIC OPINIONS

The barometer of the perception of climate change in 30 countries. The fight against climate change is an existential issue that people seem to be aware of all over the world. To measure the progress of this mobilisation, to support the commitment of the people, EDF listens to the public opinion of 30 countries and makes available all the results of this unprecedented barometer produced by Ipsos.

Dataset Repositories



data.gov.it
public administration open data

Data | Do Open Data | Monitoring | Developers | Write to the editorial staff

Search across datasets

By title and description

By keyword

Sort by:
Choose an option

× Theme:
Culture

Themes

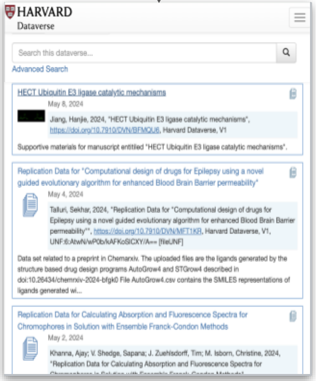
- Agriculture, Fisheries, Forestry and Food (5)
- Environment (35)
- Education, culture and sport
- Economy and finances (4)
- Government and public sector (30)
- International topics (11)
- Regions and cities (25)

Datasets Found: **2586**

Program agreements for specific training proje
2009 - 2013 appropriations for program agreements to support specific projects
Published by: Autonomous Province of Trento
Last modification date: 2013-08-30
agreements · program-agreements · provincial-administration

Education - School system year 2015
The contents of the dataset relate to the characteristics performance of the school system...
Published by: Autonomous Province of Trento
Last modification date: 2016-09-30
labor agency · disciplinary area · interdepartmental centre · University course

Education - School system year 2017
The contents of the dataset relate to the characteristics performance of the school system...
Published by: Autonomous Province of Trento



HARVARD
Dataverse

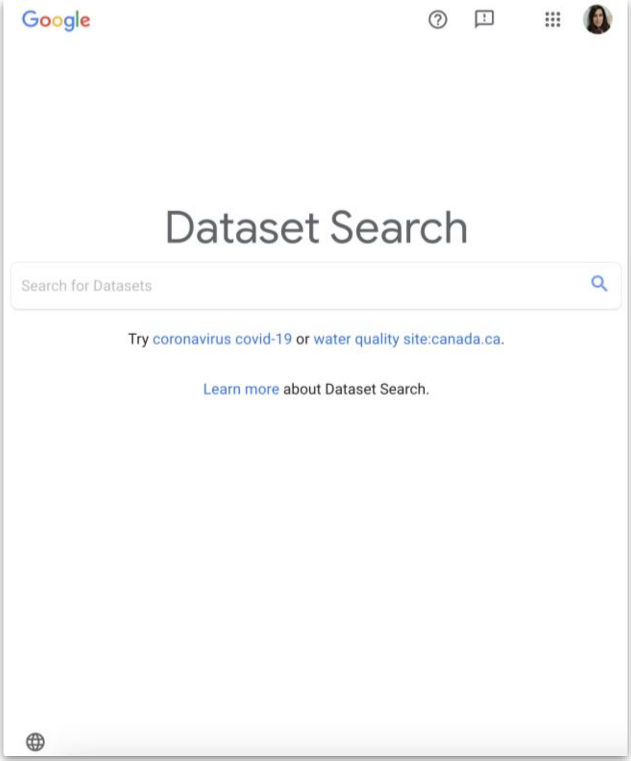
Search this dataverse...

Advanced Search

HECT Ubiquitin E3 ligase catalytic mechanisms
May 8, 2024
Jiang, Hanjie, 2024, "HECT Ubiquitin E3 ligase catalytic mechanisms", https://doi.org/10.7910/DVN/BFMQU6, Harvard Dataverse, V1
Supportive materials for manuscript entitled "HECT Ubiquitin E3 ligase catalytic mechanisms".

Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"
May 4, 2024
Talluri, Sekhar, 2024, "Replication Data for "Computational design of drugs for Epilepsy using a novel guided evolutionary algorithm for enhanced Blood Brain Barrier permeability"", https://doi.org/10.7910/DVN/MFT1KR, Harvard Dataverse, V1, UNF:6:AtwN/wP0b/kAFKoSICXY/A== [fileUNF]
Data set related to a preprint in Chemarxiv. The uploaded files are the ligands generated by the structure based drug design programs AutoGrow4 and STGrow4 described in doi:10.26434/chemrxiv-2024-bfgk0 File AutoGrow4.csv contains the SMILES representations of ligands generated wi...

Replication Data for Calculating Absorption and Fluorescence Spectra for Chromophores in Solution with Ensemble Franck-Condon Methods
May 2, 2024
Khanna, Ajay; V. Shedge, Sapana; J. Zuehlsdorff, Tim; M. Isborn, Christine, 2024, "Replication Data for Calculating Absorption and Fluorescence Spectra for



Google

Dataset Search

Search for Datasets

Try coronavirus covid-19 or water quality site:canada.ca.

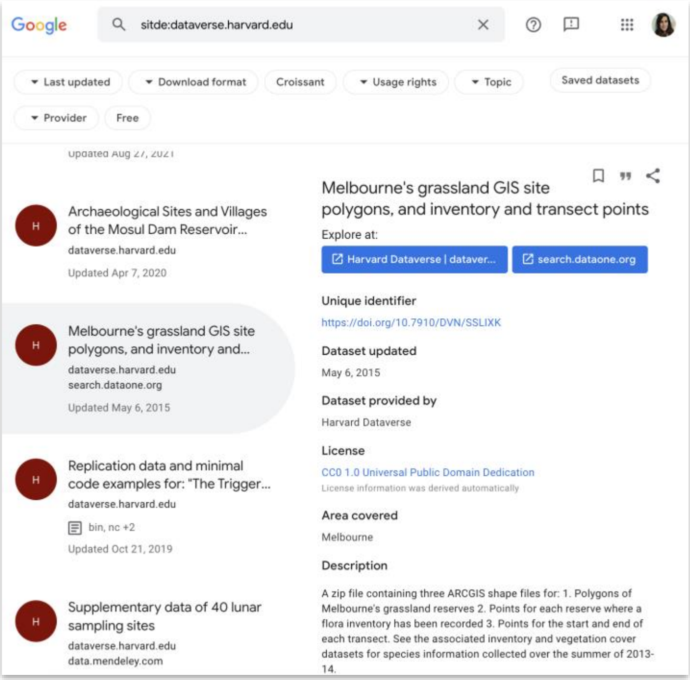
Learn more about Dataset Search.

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?



sitde:dataverse.harvard.edu

Updated Aug 27, 2021

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

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 each reserve where a flora inventory has been recorded 3. 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.

User Interviews

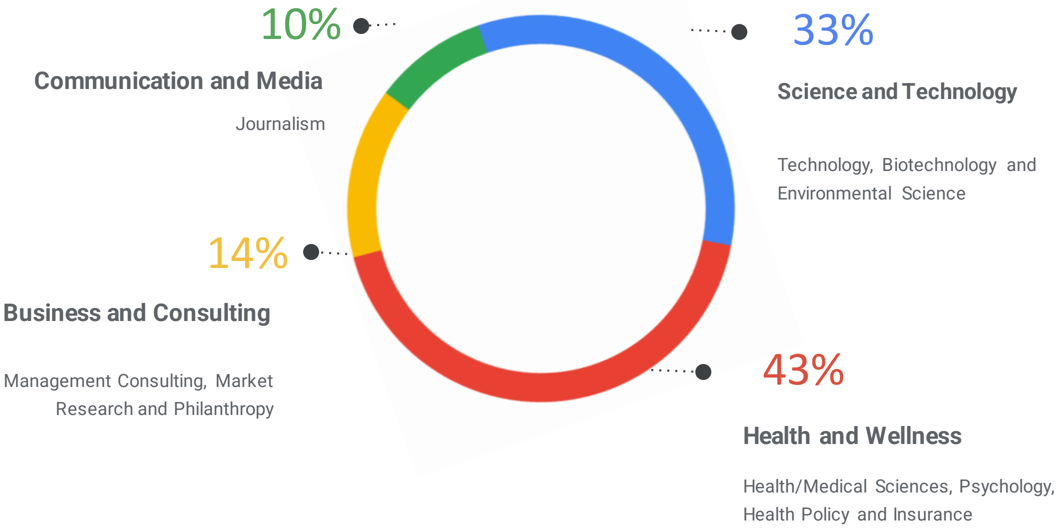
Participants

- 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

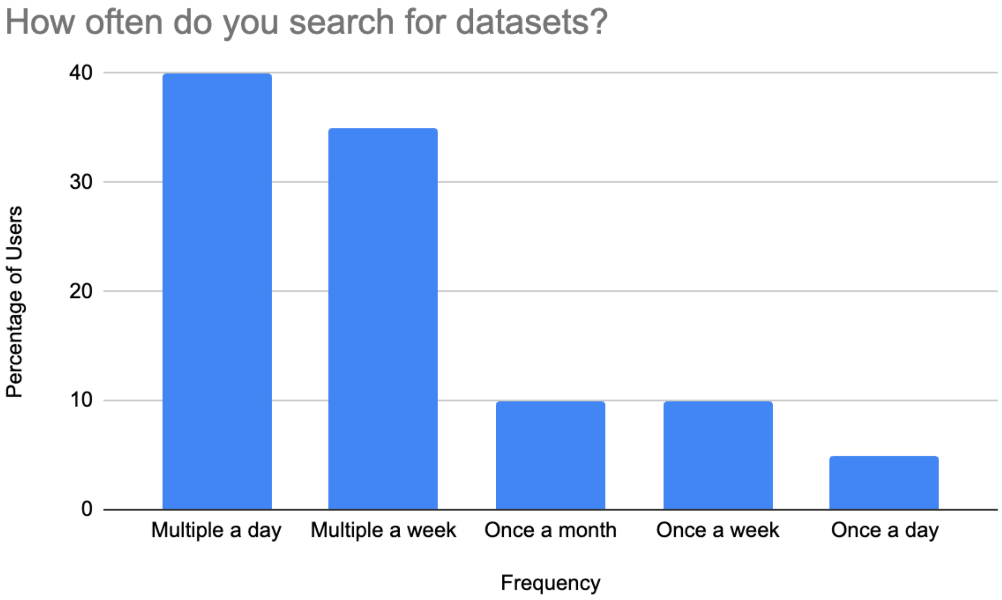
Participants

- 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)



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.

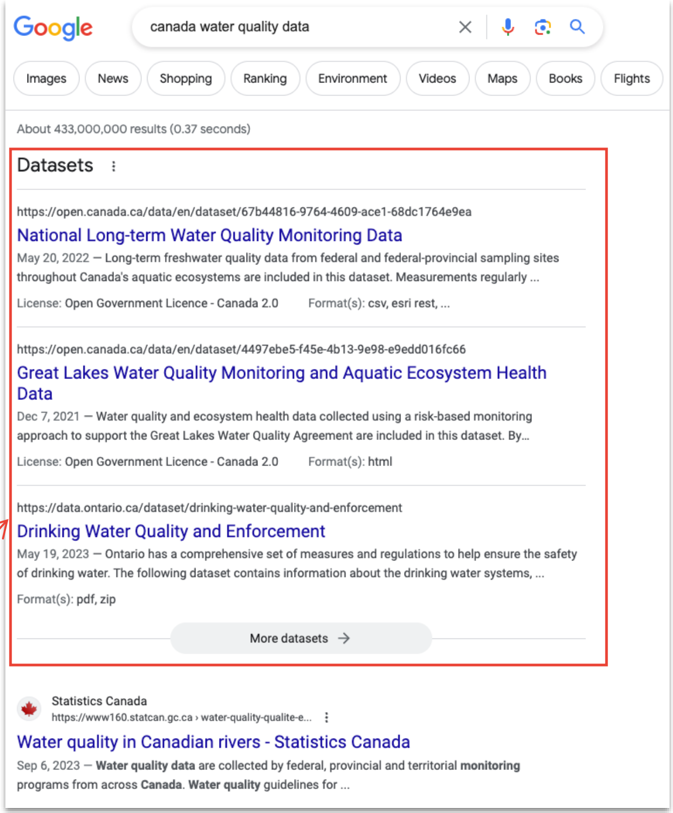
Building a Mental Model of the Tool

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.



About 433,000,000 results (0.37 seconds)

Datasets

https://open.canada.ca/data/en/dataset/67b44816-9764-4609-ace1-68dc1764e9ea

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, ...

https://open.canada.ca/data/en/dataset/4497ebe5-f45e-4b13-9e98-e9edd016fc66

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

https://data.ontario.ca/dataset/drinking-water-quality-and-enforcement

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

Statistics Canada
https://www160.statcan.gc.ca › water-quality-qualite-e...

Water quality in Canadian rivers - Statistics Canada

Sep 6, 2023 — **Water quality data** are collected by federal, provincial and territorial **monitoring** programs from across Canada. **Water quality** guidelines for ...

Results from Dataset Search in Google Web Search

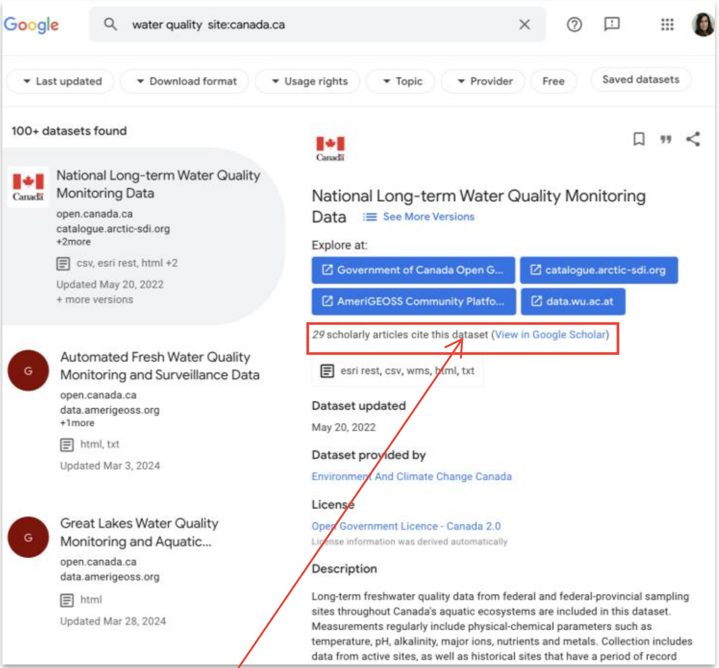
Building a Mental Model of the Tool

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.



100+ datasets found

National Long-term Water Quality Monitoring Data
open.canada.ca
catalogue.arctic-sdi.org
+2more
csv, esri rest, html +2
Updated May 20, 2022
+ more versions

Automated Fresh Water Quality Monitoring and Surveillance Data
open.canada.ca
data.amerigeoss.org
+1more
html, txt
Updated Mar 3, 2024

Great Lakes Water Quality Monitoring and Aquatic...
open.canada.ca
data.amerigeoss.org
html
Updated Mar 28, 2024

National Long-term Water Quality Monitoring Data See More Versions

Explore at:

Government of Canada Open G...
catalogue.arctic-sdi.org
AmeriGEOSS Community Platfo...
data.wu.ac.at

29 scholarly articles cite this dataset (View in Google Scholar)

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
License information was derived automatically

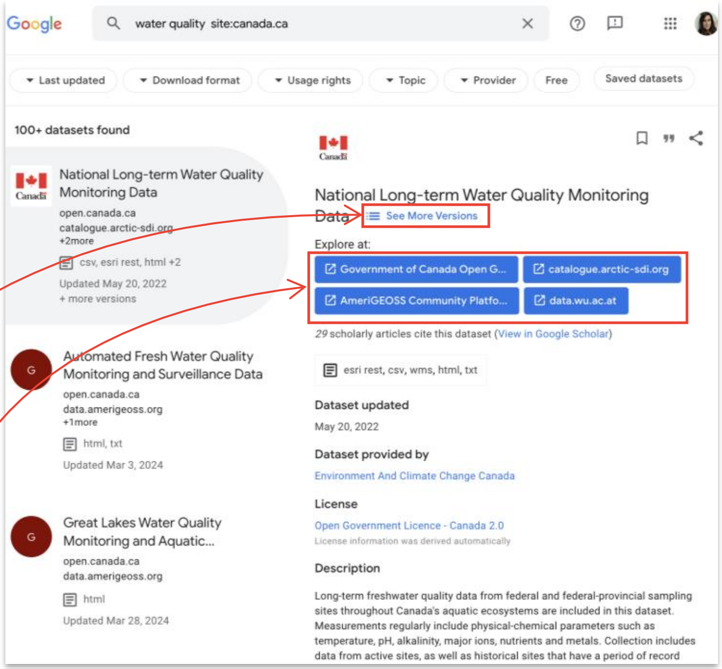
Description
Long-term freshwater quality data from federal and federal-provincial sampling sites throughout Canada's aquatic ecosystems are included in this dataset. Measurements regularly include physical-chemical parameters such as temperature, pH, alkalinity, major ions, nutrients and metals. Collection includes data from active sites, as well as historical sites that have a period of record

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



100+ datasets found

National Long-term Water Quality Monitoring Data
open.canada.ca
catalogue.arctic-sdi.org
+2more
csv, esri rest, html +2
Updated May 20, 2022
+ more versions

Automated Fresh Water Quality Monitoring and Surveillance Data
open.canada.ca
data.amerigeoss.org
+1more
html, txt
Updated Mar 3, 2024

Great Lakes Water Quality Monitoring and Aquatic...
open.canada.ca
data.amerigeoss.org
html
Updated Mar 28, 2024

National Long-term Water Quality Monitoring Data

Explore at:

29 scholarly articles cite this dataset (View in Google Scholar)

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

License information was derived automatically

Description

Long-term freshwater quality data from federal and federal-provincial sampling sites throughout Canada's aquatic ecosystems are included in this dataset. Measurements regularly include physical-chemical parameters such as temperature, pH, alkalinity, major ions, nutrients and metals. Collection includes data from active sites, as well as historical sites that have a period of record

It also finds **versions** now!

Dataset Search finds **replicas** of datasets

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.

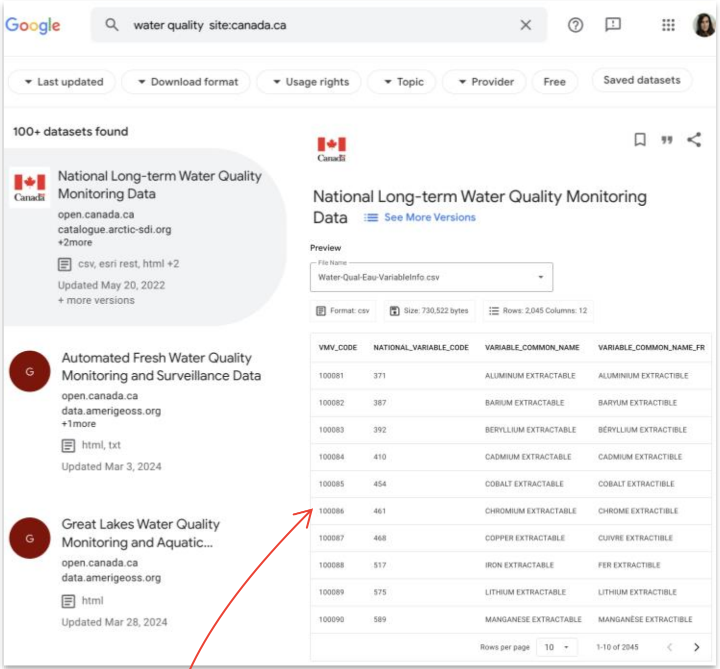
Making Sense of Heterogeneous Datasets

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.



100+ datasets found

National Long-term Water Quality Monitoring Data
open.canada.ca
catalogue.arctic-sdi.org
+2more
csv, esri rest, html +2
Updated May 20, 2022
+ more versions

Automated Fresh Water Quality Monitoring and Surveillance Data
open.canada.ca
data.amerigeoss.org
+1more
html, txt
Updated Mar 3, 2024

Great Lakes Water Quality Monitoring and Aquatic...
open.canada.ca
data.amerigeoss.org
html
Updated Mar 28, 2024

National Long-term Water Quality Monitoring Data

Preview

File Name: Water-Qual-Eau-VariableInfo.csv

Format: csv | Size: 730,522 bytes | Rows: 2,045 Columns: 12

VMV_CODE	NATIONAL_VARIABLE_CODE	VARIABLE_COMMON_NAME	VARIABLE_COMMON_NAME_FR
100081	371	ALUMINUM EXTRACTABLE	ALUMINIUM EXTRACTIBLE
100082	387	BARIUM EXTRACTABLE	BARYUM EXTRACTIBLE
100083	392	BERYLLIUM EXTRACTABLE	BÉRYLLIUM EXTRACTIBLE
100084	410	CADMIUM EXTRACTABLE	CADMIUM EXTRACTIBLE
100085	454	COBALT EXTRACTABLE	COBALT EXTRACTIBLE
100086	461	CHROMIUM EXTRACTABLE	CHROME EXTRACTIBLE
100087	468	COPPER EXTRACTABLE	CUIVRE EXTRACTIBLE
100088	517	IRON EXTRACTABLE	FER EXTRACTIBLE
100089	575	LITHIUM EXTRACTABLE	LITHIUM EXTRACTIBLE
100090	589	MANGANESE EXTRACTABLE	MANGANÈSE EXTRACTIBLE

Rows per page 10 1-10 of 2045

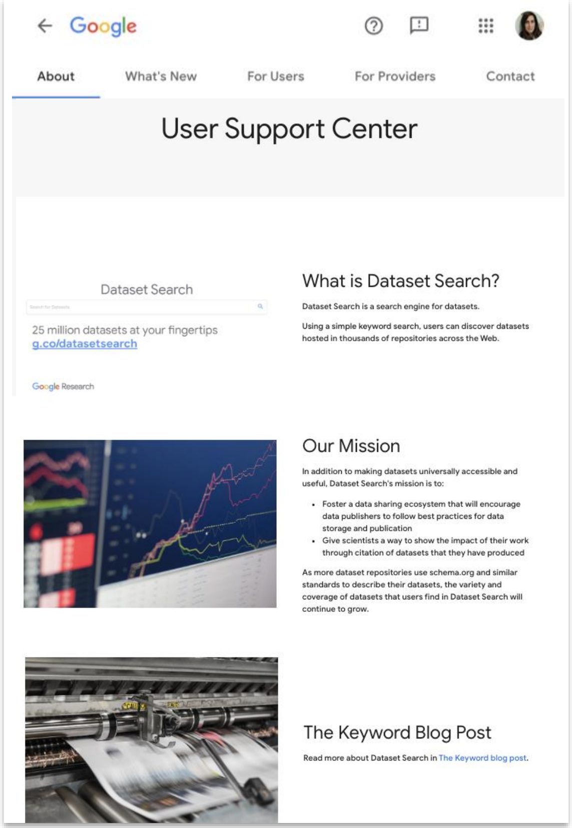
Dataset Previews

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.



User Support Center

25 million datasets at your fingertips
g.co/datasetsearch

What is Dataset Search?

Dataset Search is a search engine for datasets.

Using a simple keyword search, users can discover datasets hosted in thousands of repositories across the Web.

Our Mission

In addition to making datasets universally accessible and useful, Dataset Search's mission is to:

- Foster a data sharing ecosystem that will encourage data publishers to follow best practices for data storage and publication
- Give scientists a way to show the impact of their work through citation of datasets that they have produced

As more dataset repositories use schema.org and similar standards to describe their datasets, the variety and coverage of datasets that users find in Dataset Search will continue to grow.

The Keyword Blog Post

Read more about Dataset Search in The Keyword blog post.

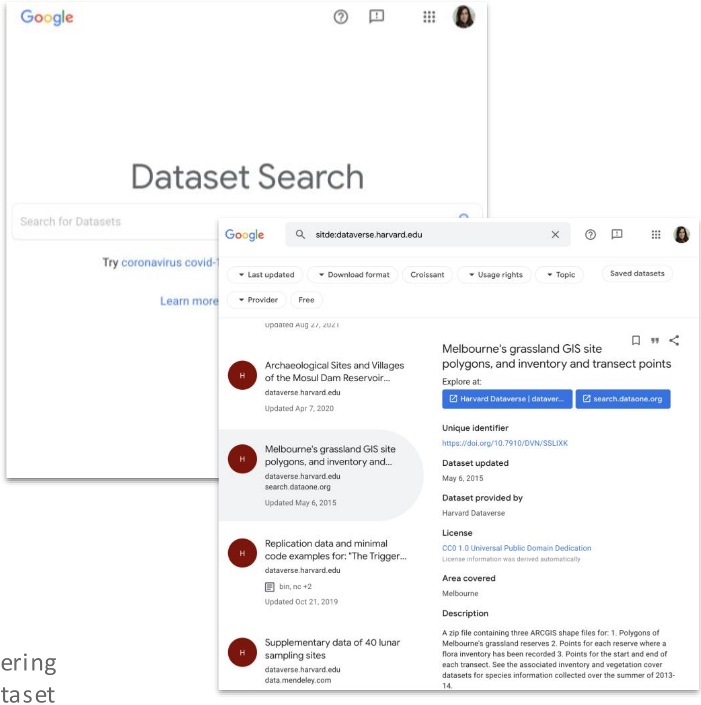
Limitations

- 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



Updated Aug 27, 2021

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

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 each reserve where a flora inventory has been recorded 3. 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.

datasetsearch.research.google.com