Finding and reusing research data

Agata Bochynska, PhD Open Research and Digital Scholarship Center
Ivana Malovic, PhD Library of Medicine and Science

University of Oslo

27.04.2023

Contact us at research-data@uio.no

Materials developed as a part of the *Skills development for research data* project

UNIVERSITY
OF OSLO

Source: RDMKit, Elixir Europe https://rdmkit.elixir-europe.org/

UNIVERSITY
OF OSLO

Source: RDMKit, Elixir Europe https://rdmkit.elixir-europe.org/

Data discovery is finding and accessing data collected for a different purpose or by a different researcher or institution.

In the process of data discovery and reuse you are working with **secondary data**, as opposed to **primary data** that you would collect yourself.

"**Open Science** has the potential of making the scientific process more **transparent**, **inclusive** and **democratic**. It is (…) a true game changer in bridging the science, technology and innovation gaps and fulfilling the **human right to science**."

UNESCO Recommendation on Open Science

https://youtu.be/I3Wkvx_ZaFo
https://www.unesco.org/en/natural-sciences/open-science

More data sharing – more data to discover!

Increasing need for data reuse

- High **costs** of primary data collection
- **Redundancy** or similarity in different sets of primary data
- High demands for **storage space** by increasing amount of data
- Promoting **transparency** and **reproducibility** in research

Data discovery: how-to

DATA DISCOVERY PROCESS

1. Develop a clear **picture** of the **data** you need
2. Locate appropriate data **resources**
3. Set up a **search query** and search the resource
4. Select data **candidates**
5. Evaluate data **quality**

Based on: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/The-process-of-data-discovery

Develop a clear picture of the data you need

Deciding on what kind of data you need

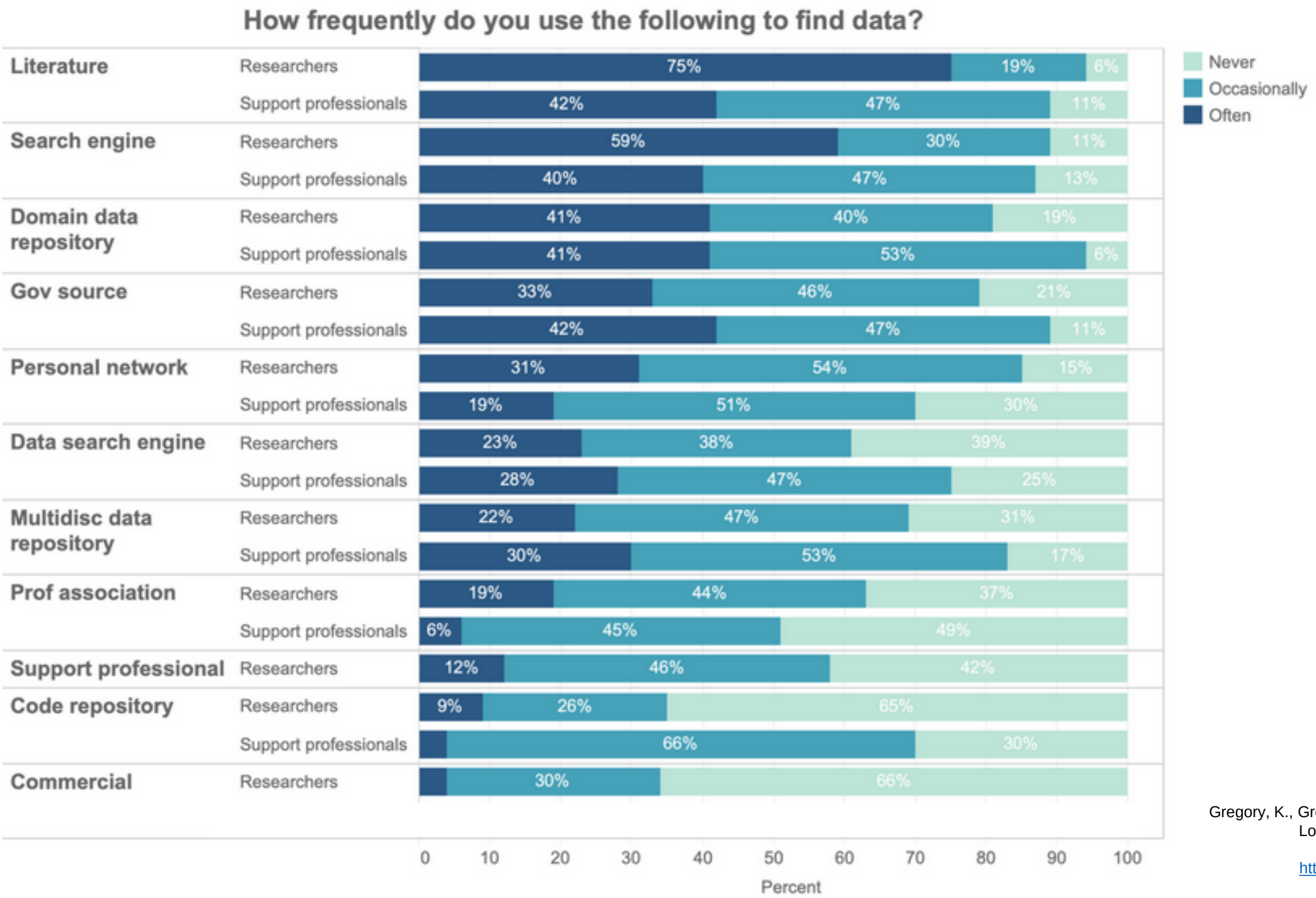
- What is the **theme/domain** you study?
- What is your **research question**?
- What are the **constructs/concepts** and how you will operationalize them?
- What is your **theory**?
- What **study** will you perform?
- What specific characteristics should the data have?

DATA DISCOVERY PROCESS

1. Develop a clear **picture** of the **data** you need
2. Locate appropriate data **resources**
3. Set up a **search query** and search the resource
4. Select data **candidates**
5. Evaluate data **quality**

Based on: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/The-process-of-data-discovery

Locate appropriate data resources



Gregory, K., Groth, P. Scharnhorst, A., Wyatt, S. (2020). Lost or found? Discovering data needed for research. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e38165eb

Where do I look for the data?

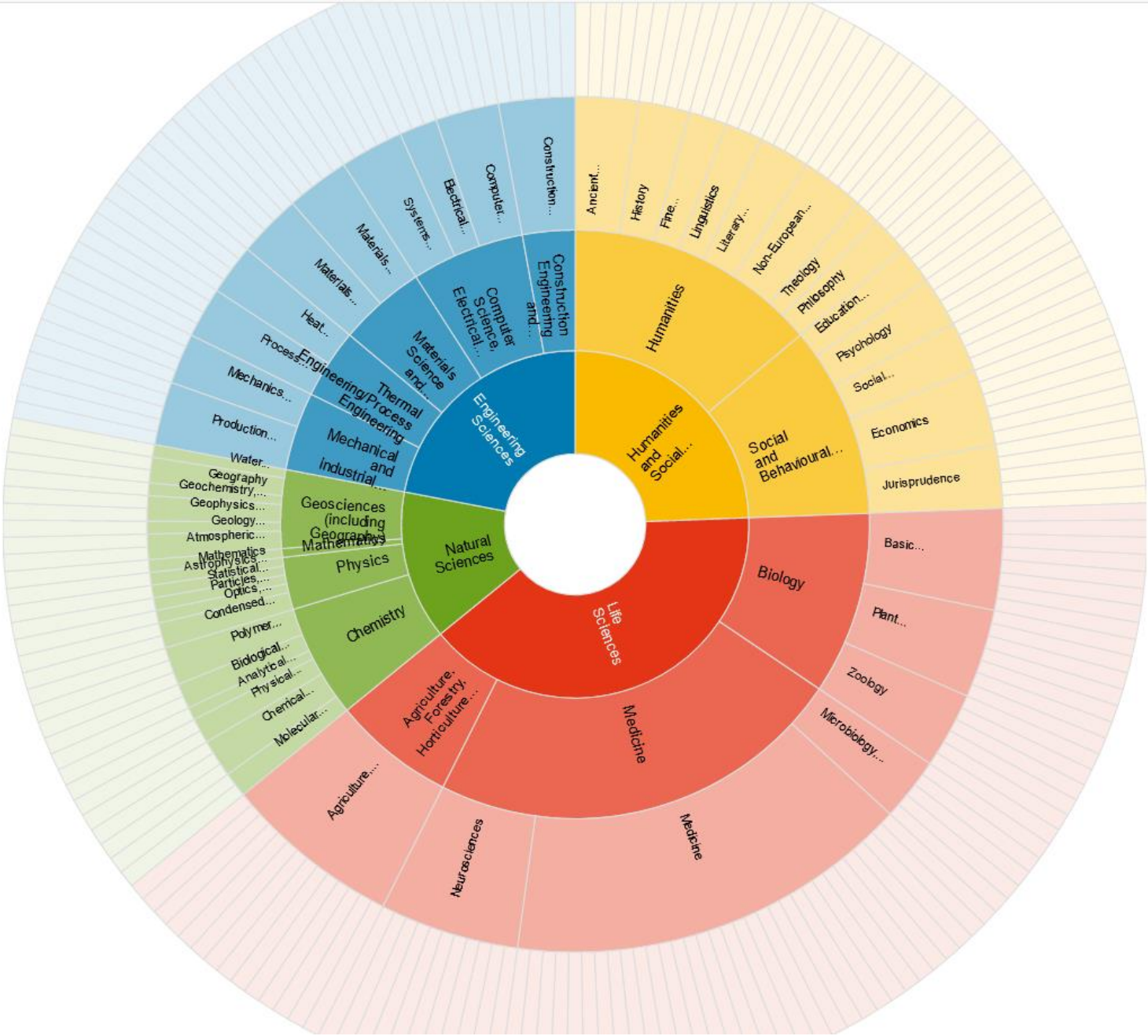
- Discipline-specific repositories
- General-purpose repositories
- A search engine or (meta)data aggregator
- A data journal

Where do I look for the data?

- **Discipline-specific repositories**
- General-purpose repositories
- A search engine or (meta)data aggregator
- A data journal

Search for discipline-specific repositories

https://www.re3data.org/


Engineering Sciences
Construction Engineering and...
Computer Science, Electrical...
Materials Science and...
Thermal Engineering/Process Engineering
Mechanical and Industrial...
Construction...
Computer...
Electrical...
Systems...
Materials...
Materials...
Heat...
Process...
Mechanics...
Production...
Water...
Natural Sciences
Geosciences (including Geography)
Mathematics
Physics
Chemistry
Geography
Geochemistry,...
Geophysics...
Geology...
Atmospheric...
Mathematics
Astrophysics...
Statistical...
Particles,...
Optics,...
Condensed...
Polymer...
Biological...
Analytical...
Physical...
Chemical...
Molecular...
Humanities and Social...
Humanities
Social and Behavioural...
Ancient...
History
Fine...
Linguistics
Literary...
Non-European...
Theology
Philosophy
Education...
Psychology
Social...
Economics
Jurisprudence
Life Sciences
Biology
Medicine
Agriculture, Forestry, Horticulture...
Basic...
Plant...
Zoology
Microbiology...
Medicine
Neurosciences
Agriculture...

Filter

Reset all

Subjects

- **Humanities and Social Sciences (4)**
  - **Humanities (4)**
    - **Linguistics (4)**
      - **General and Applied Linguistics (4)**
      - Individual Linguistics (1)
  - Social and Behavioural Sciences (1)
    - Education Sciences (1)
- Engineering Sciences (1)
  - Computer Science, Electrical and System Engineering (1)
    - Computer Science (1)
      - Artificial Intelligence, Image and Language Processing (1)

Content Types
Countries
API
Data access
Data access restrictions
Database access
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type
Keywords
Metadata standards
PID systems
Provider types
Quality management
Repository languages
Software
Syndications
Repository types
Versioning

Search...

Search

Toogle short help

← Previous | 1 | Next →

Sort by

Found 4 result(s)

Eurac Research CLARIN Centre

ERCC

Subject(s)	Linguistics, Humanities and Social Sciences, Humanities, General and Applied Linguistics, Artificial Intelligence, Image and Language Processing, Computer Science, Computer Science, Electrical and System Engineering, Engineering Sciences
Content type(s)	Databases, Scientific and statistical data formats, Structured text, Software applications
Country	Italy, European Union

The Eurac Research CLARIN Centre (ERCC) is a dedicated repository for language data. It is hosted by the Institute for Applied Linguistics (IAL) at Eurac Research, a private research centre based in Bolzano, South Tyrol. The Centre is part of the Europe-wide CLARIN infrastructure, which means that it follows well-defined international standards for (meta)data and procedures and is well-embedded in the wider European Linguistics infrastructure. The repository hosts data collected at the IAL, but is also open for data deposits from external collaborators.

Michigan Corpus of Academic Spoken English

MICASE

Subject(s)	Humanities and Social Sciences, Humanities, Linguistics, General and Applied Linguistics, Individual Linguistics
Content type(s)	Structured text, Audiovisual data
Country	United States

MICASE provides a collection of transcripts of academic speech events recorded at the University of Michigan. The original DAT audiotapes are held in the English Language Institute and may be consulted by bona fide researchers under special arrangements. Additional access: https://lsa.umich.edu/eli/language-resources/micase-micusp.html

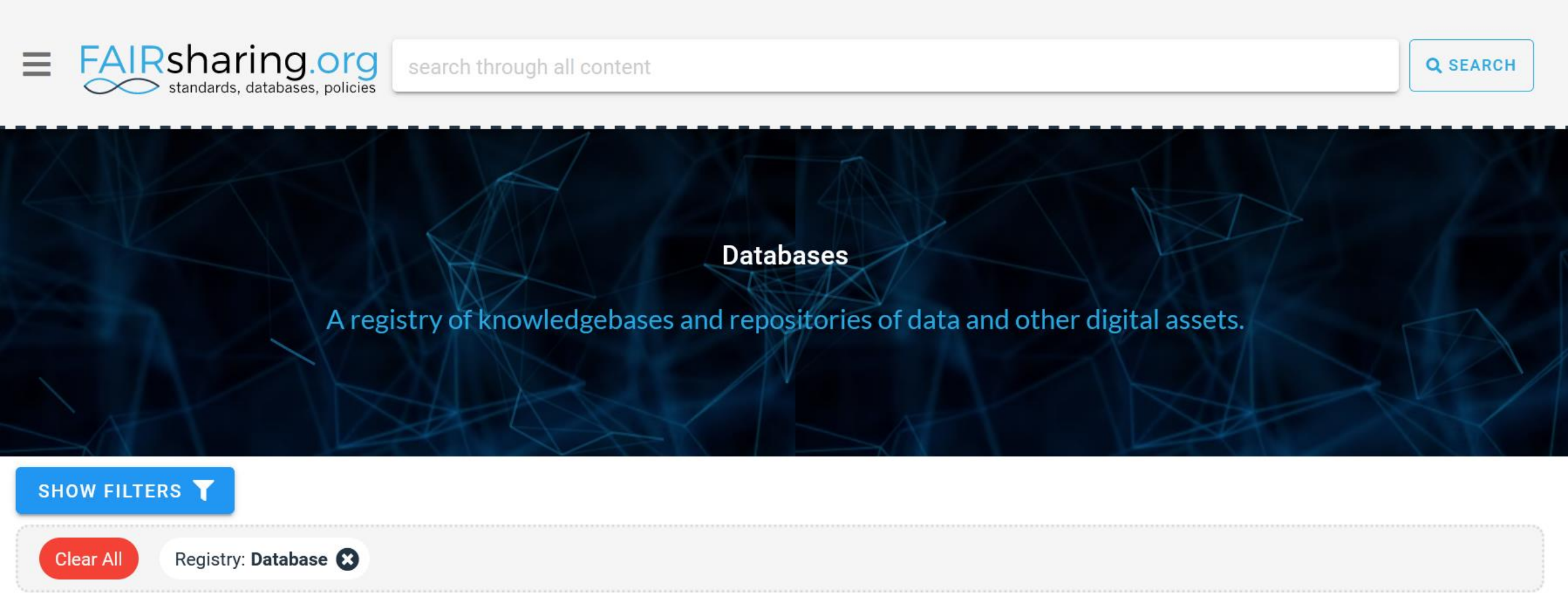
The University of Pittsburgh English Language Institute Corpus

PELIC

Subject(s)	Linguistics, General and Applied Linguistics, Humanities, Humanities and Social Sciences
Content type(s)	Standard office documents, Archived data, Source code, Images
Country	United States

The University of Pittsburgh English Language Institute Corpus (PELIC) is a 4.2-million-word learner corpus of written texts. These texts were collected in an English for Academic Purposes (EAP) context over seven years in the University of Pittsburgh's Intensive English Program, and were produced by over 1100 students with a wide range of linguistic backgrounds and proficiency levels. PELIC is longitudinal, offering greater opportunities for tracking development in a natural classroom setting.

Search for discipline-specific repositories



https://fairsharing.org/

Some examples

Example: NIPH (FHI)

Access to data

Researchers can apply for access to data from health registries and health studies, as well as biological material from the biobanks. Here you will find guidelines and electronic application forms.

Les på norsk Get the latest news

HOW DO I APPLY FOR ACCESS?

ARTICLE

How to apply for access to data

The Norwegian Institute of Public Health (NIPH) can provide access to data from health registries and population-based health surveys once an application for data is approved.

Updated 09.12.2020

ARTICLE

Application form for access to data

For applications for access to data or biological samples from studies or mandatory national health registries at the NIPH.

Updated 09.12.2020

Example: NIPH (FHI) – Big Data

NIPH
Norwegian Institute of Public Health

Content A-Z | Contact us | Norsk nettsted

Search in NIPH | Menu

Norwegian Mother, Father and Child Cohort Study (MoBa)

STATUS: **ACTIVE**

The Norwegian Mother, Father and Child Cohort Study is a unique study where over 90,000 pregnant women were recruited from 1998 to 2008. More than 70,000 fathers have participated.

Les på norsk | Contact | Get the latest news

FOR RESEARCHERS

ARTICLE

What is the Norwegian Mother, Father and Child Cohort Study?

The Norwegian Mother, Father and Child Cohort Study (MoBa) is a study of the causes of disease among mothers and children. MoBa began to recruit pregnant women in 1999. Fathers were also invited.

Updated 05.07.2021

ARTICLE

Access to data and biological material from MoBa

On this page we have gathered relevant information for researchers applying for access to data from the Norwegian Mother, Father and Child Cohort Study (MoBa)-.

Updated 07.07.2021

ARTICLE

Information for MoBa researchers

Here you will find the price list, the variable list, MoBa protocols, admission documents and guidelines for publications.

Updated 18.08.2021

https://www.fhi.no/en/studies/moba/

Example: NIPH (FHI) - Statistics

Norwegian Institute of Public Health

Contact | Health status fact sheets on fhi.no |

Norhealth

Search

- Norhealth (N= Norway, H= health regions, C= county figures)
  - About population
  - Childhood and living conditions
  - Environment
  - Accidents and injuries
  - Living habits
    - Nutrition
      - Fruit consumption, 16-79 years (C)
      - Fruit consumption (N)
      - Fruit consumption daily, by education attainment (NH)
      - Vegetable consumption (16-79 years)
      - Vegetable consumption (N)
      - Vegetable consumption daily, by education attainment (NH)
      - Fruit and vegetable consumption, daily (C)
      - Soft drink consumption (C)
      - Soft drink consumption (N)
      - Soft drink consumption daily, by educational attainment (NH)
    - Physical activity
      - Physically active more than 150 minutes per week, 16-79 years (C)
      - Exercise less than weekly, self-reported at military muster (C)
      - Physically active more than 150 minutes per week (N)
      - Physical activity, by educational attainment (NH)
    - Smoking and snus use
      - Smoking, adults, yearly figures (N)
      - Smoking, by educational attainment, 25-74 years, yearly figures (N)
      - Daily smoking, by educational attainment, 25-74 years (NHC)
      - Snus use, adults (NHC)
      - Snus use, adults, yearly figures (N)
      - Smoking, adults (NHC)
      - Snus use, by educational attainment, 25-74 years, yearly figures (N)
    - Alcohol
  - Health and disease
    - Life expectancy
      - Life expectancy (NC)
      - Life expectancy by age, yearly figures (C)
      - Life expectancy, by educational attainment (NC)
      - Life expectancy, difference between education levels (NC)
      - Life expectancy by age and education, yearly figures (C)
      - Total mortality (NHC)
      - Total mortality, by educational attainment
      - Total mortality, by educational attainment (NHC)
    - **Self-Reported Health**
      - Somatic pain, 16-79 years (C)
      - Somatic pain (N)
      - Somatic pain, by educatonal attainment (NH)
      - Headache or migraine, 16-79 years (C)
      - Headache or migraine (N)
      - Headache or migraine, by educational attainment (NH)

Norhealth

System requirements: Norhealth works best in an updated browser. Especially in older versions of Internet Explorer, you can experience problems, like reduced-size dialog boxes or unresponsive system. If you experience such problems, try to open Norhealth in Firefox or Chrome.

In some indicators, the English version of metadata (definitions) can be lacking and will be added later.

Statistics

Norhealth contains more than 100 indicators. Within each indicator, you can create your own tables, figures and maps by using the menus and icons at the top of the page. Navigate the menu in the left-hand margin or use the free text search to find statistics.

Privacy

Visitor statistics are recorded for Norhealth. The editorial and technical personnel use these statistics to develop the content and navigation. Visitor statistics are based on Google Analytics and we use a feature where the exact IP address of the visitor is not stored at either Google or NIPH.

Copyright

The content and layout on these webpages (www.norgeshelsa.no) are protected by copyright. The Norwegian Insitute of Public Health gives permission to copy text, tables and figures referring to the source, Norwegian Institute of Public Health (www.norgeshelsa.no). Reference to the source must be given for every table and figure. We recommend the use of links by reuse on internet. We are not responsible for misuse of the statistics.

https://norgeshelsa.no/norgeshelsa/

https://www.who.int/data/collections

Example: WHO

World Health Organization

Health Topics | Countries | Newsroom | Emergencies | Data | About WHO

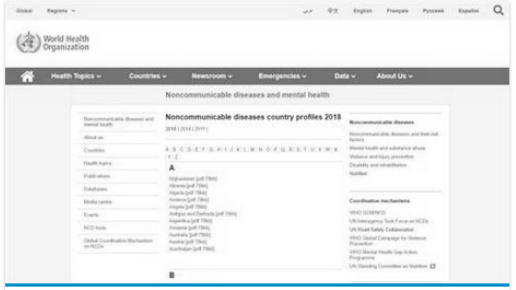
Data / WHO data collections

Data collections

The World Health Organization manages and maintains a wide range of data collections related to global health and well-being as mandated by our Member States.

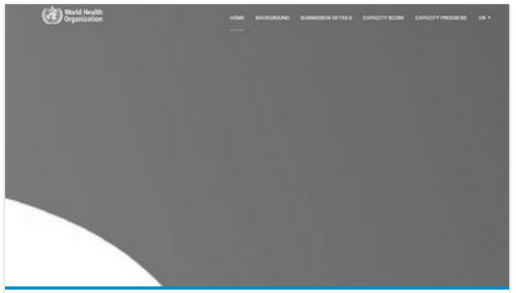
Explore our key health data products and resources from across the organization.

Search



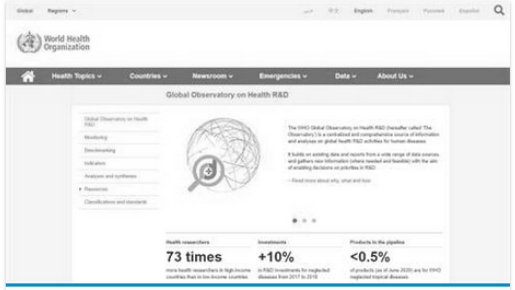
Noncommunicable diseases profiles

Number of lives that can be saved by implementing WHO 'best buys', Risk of premature death, National targets, Risk factors, National Systems Response



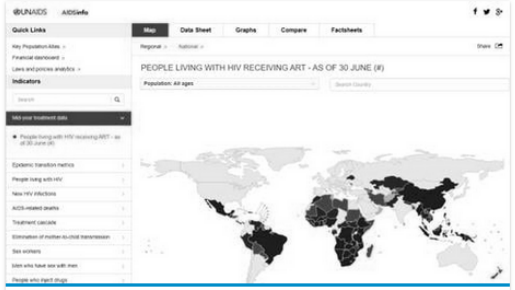
e-SPAR

Electronic State Parties Self-Assessment Annual Reporting Tool (e-SPAR) is a web-based platform proposed to support State Parties of the International Health Regulations (IHR) to fulfil their obligation to



The global observatory for health research and development

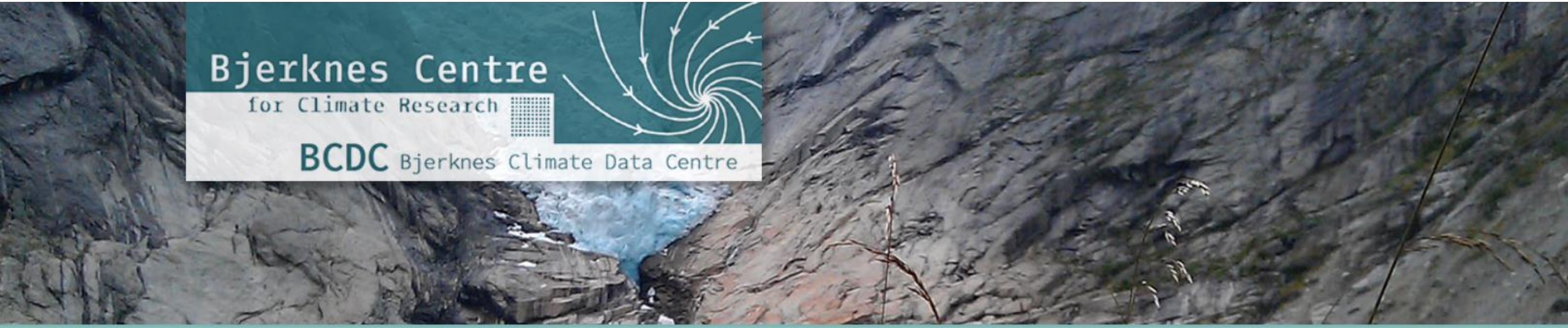
The WHO Global Observatory on Health R&D is a centralized and comprehensive source of information and analyses on global health R&D



HIV laws and policies database

Laws and Policies Analytics is a platform to view data on HIV-related laws and policies in countries compiled from official sources and reported by both national authorities and civil society to UNAIDS and the World

Example: BCDC



ABOUT DATA SEARCH PROJECTS SUBMIT PARTNERS NEWS CONTACT

REMOTE SENSING
PALEO
MODEL OUTPUT
OCEANOGRAPHY
ATMOSPHERE
DATA PRODUCTS

>> BCDC Home

DATA PUBLICATION HIGHLIGHTS

12 November 2018

High-Resolution Benthic Mg/Ca Temperature Record of the Intermediate Water in the Denmark Strait Across Dansgaad-Oeschger Stadial-Interstadial Cycles

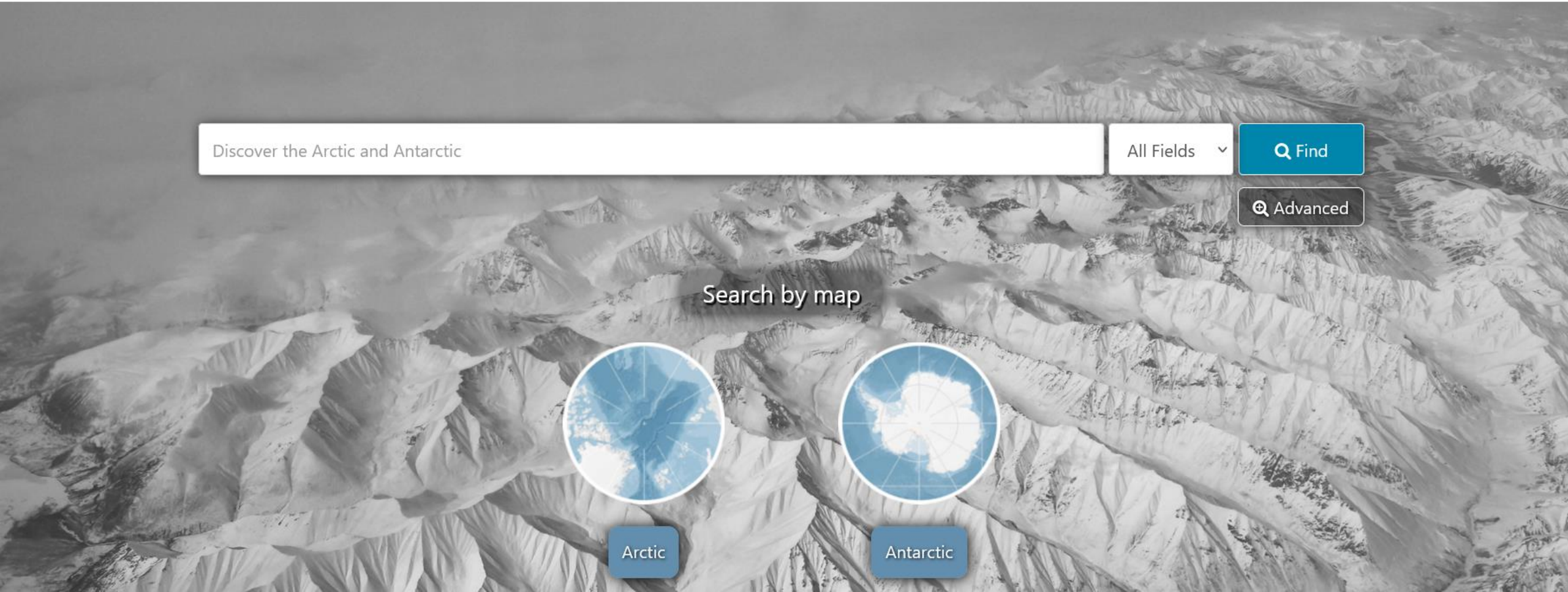
https://www.bcdc.no/

https://openpolar.no/

Example: OPEN POLAR

OPEN POLAR

The Global Open Access Portal for Research Data and Publications on the Arctic and Antarctic



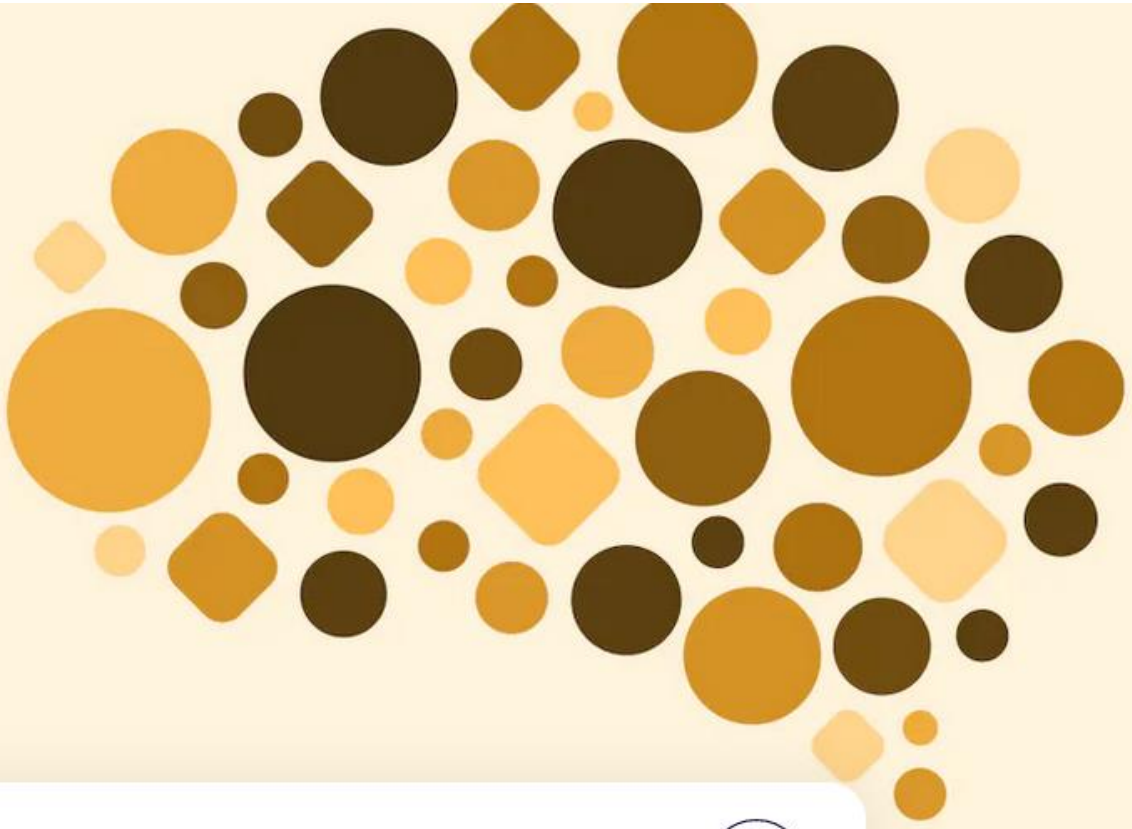
https://sikt.no/surveybanken

Example: SURVEY BANK (Sikt)

Surveybanken

Her kan du finne, analysere og laste ned data fra spørreundersøkelser. Surveybanken inneholder 750 000 spørsmål fra mer enn 3000 spørreundersøkelser tilbake til 1957.

Se hvordan folks holdninger og meninger har endret seg siden midten av forrige århundre.



Søk i surveydata. F.eks. tillit politikere, valgundersøkelse

Where do I look for the data?

- Discipline-specific repositories
- General-purpose repositories
- A search engine or (meta)data aggregator
- A data journal

Where do I look for the data?

- Discipline-specific repositories
- **General-purpose repositories**
- A search engine or (meta)data aggregator
- A data journal

General purpose repositories

zenodo

DataverseNO
Dataverse Network Norway

NTNU Open Research Data
NTNU

OSF

figshare

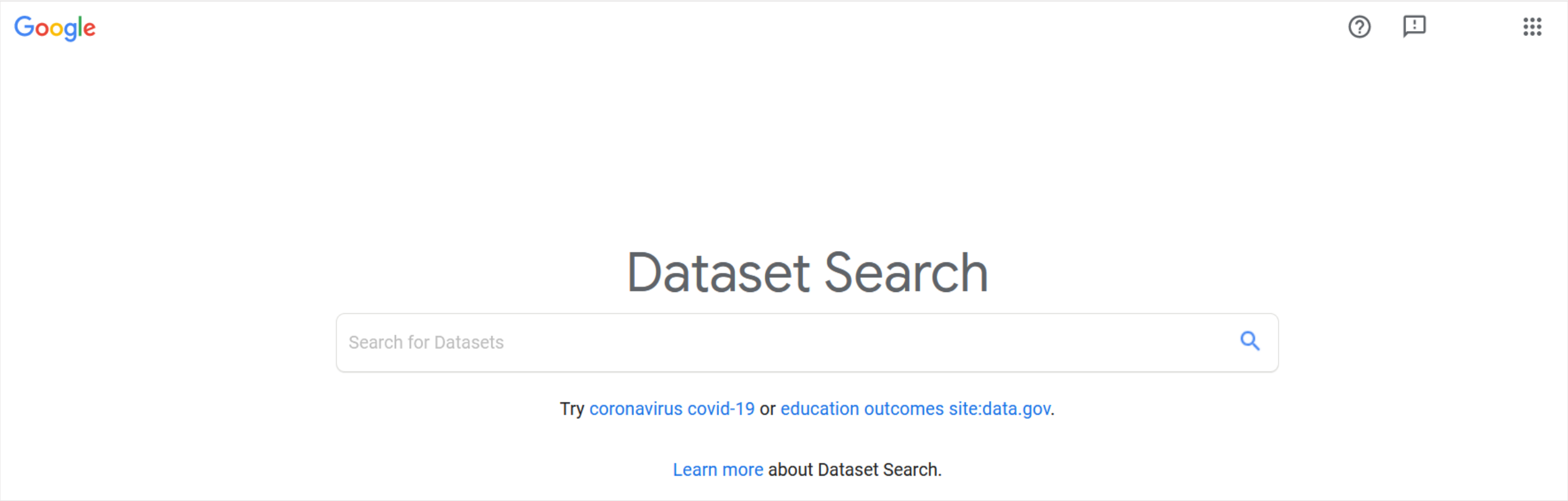
Where do I look for the data?

- Discipline-specific repositories
- General-purpose repositories
- A search engine or (meta)data aggregator
- A data journal

Where do I look for the data?

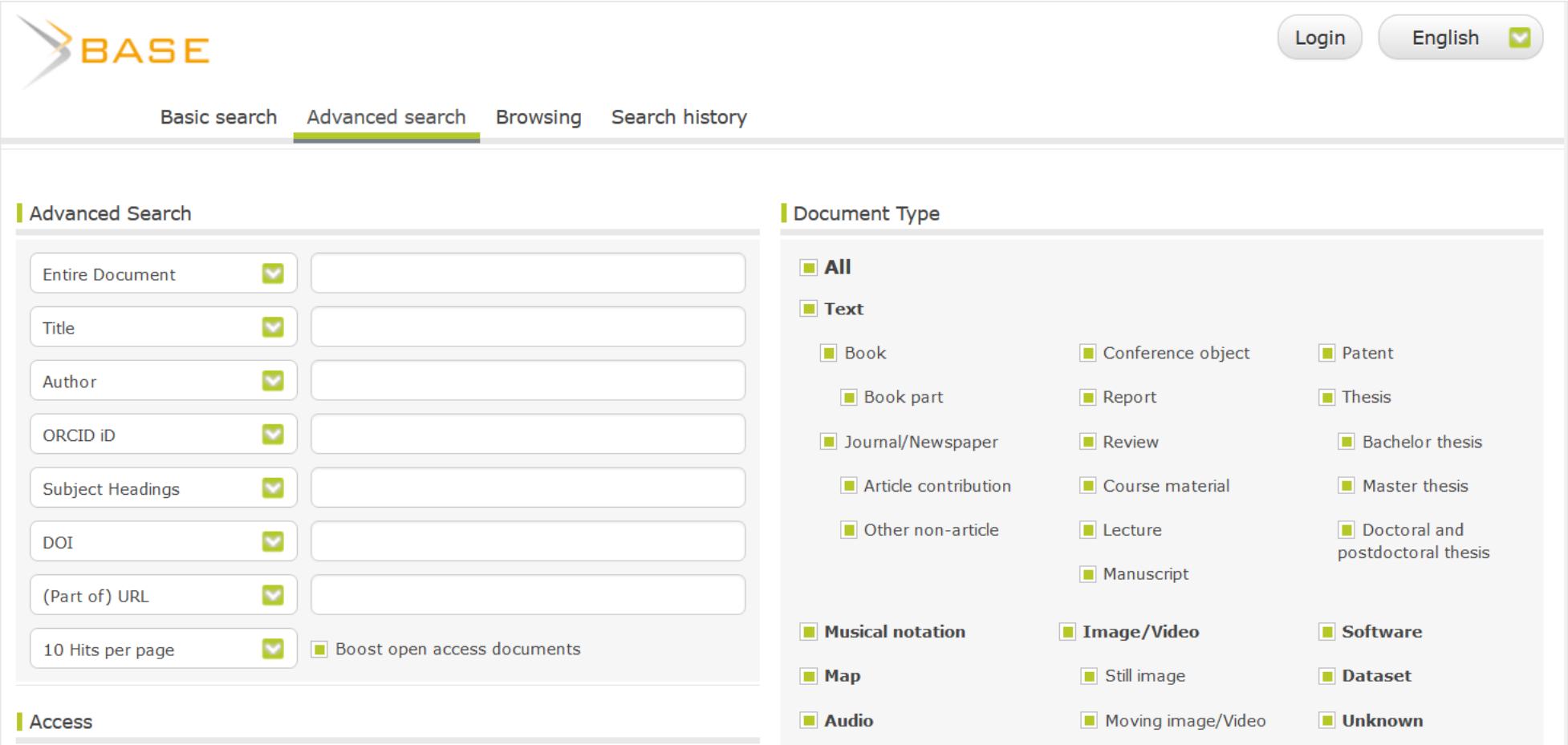
- Discipline-specific repositories
- General-purpose repositories
- **A search engine or (meta)data aggregator**
- A data journal

Search engines



https://datasetsearch.research.google.com/

Search engines



UNIVERSITY
OF OSLO

https://www.base-search.net/Search/Advanced

Where do I look for the data?

- Discipline-specific repositories
- General-purpose repositories
- A search engine or (meta)data aggregator
- A data journal

Where do I look for the data?

- Discipline-specific repositories
- General-purpose repositories
- A search engine or (meta)data aggregator
- **A data journal**

Data journals

scientific **data**

View all journals | Search | Login

Explore content | **Journal information** | **Publish with us** | **Sign up for alerts** | **RSS feed**

nature > scientific data > about

About

Principles

Open Access

FAQ

About

Scientific Data is a peer-reviewed, open-access journal for descriptions of scientifically valuable datasets, and research that advances the sharing and reuse of scientific data.

Read our key principles ▸

Geoscience Data Journal

Open Access

RMetS

Co-Editors-in-Chief: Katherine Royse & Jian Peng

Impact factor: 2.714

2019 Journal Citation Reports (Clarivate Analytics): 74/200 (Geosciences, Multidisciplinary) 42/93 (Meteorology & Atmospheric Sciences)

Online ISSN: 2049-6060

LATEST ISSUE >

Volume 7, Issue 2
November 2020

Where do I look for the data?

- Discipline-specific repositories
- General-purpose repositories
- A search engine or (meta)data aggregator
- A data journal

DATA DISCOVERY PROCESS

1. Develop a clear **picture** of the **data** you need
2. Locate appropriate data **resources**
3. Set up a **search query** and search the resource
4. Select data **candidates**
5. Evaluate data **quality**

Based on: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/The-process-of-data-discovery

Set up a search query and search the data resource

How to search the data resource?

- Familiarize yourself with the **structure** of the data resource
- **Register** yourself as a user
- Learn how the data repository **advanced search functions** work
- Ask for help!
  - Ask your subject librarian: https://www.ub.uio.no/english/using/guidance/index.html
  - Consult information pages: https://sokogskriv.no/en/searching/

How to set up search queries?

Choose **keywords**

- Use the terms from your discipline
- Focus on main concepts
- Think of possible synonyms

Use **boolean operators** (if allowed)

- Terms such as AND, OR

In general search engines (e.g. Web of Science) add «data» or «dataset» to the search query or choose the type of document in the filters.

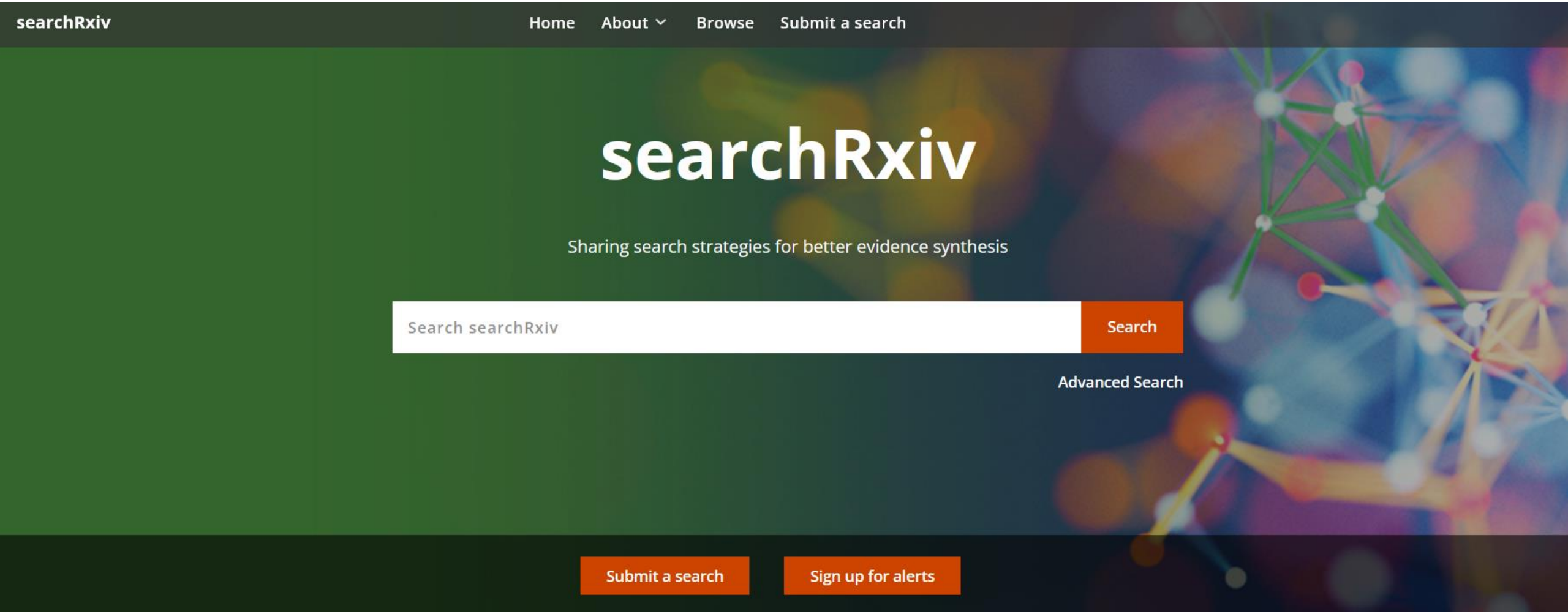
Adjusting your search: you might have to broaden or narrow down your scope

If your search is too narrow:

- Check your spelling
- Use more general search terms
- Turn off some of the filters you applied
- Use more synonyms

If your search is too broad:

- Use more specific search terms
- Use more search terms
- Use more filters
- Check the use of boolean logic (is it applied correctly?)



https://www.cabidigitallibrary.org/journal/searchrxiv

DATA DISCOVERY PROCESS

1. Develop a clear **picture** of the **data** you need
2. Locate appropriate data **resources**
3. Set up a **search query** and search the resource
4. Select data **candidates**
5. Evaluate data **quality**

Based on: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/The-process-of-data-discovery

Select data candidates

Can I use these data?

- Are the data **relevant** to your research questions?
- Are the **concepts** appropriate?
- Are the **variables** and the indicators appropriate?

*Check dataset **documentation** (e.g. README files, data dictionaries or codebooks) very carefully!

DATA DISCOVERY PROCESS

1. Develop a clear **picture** of the **data** you need
2. Locate appropriate data **resources**
3. Set up a **search query** and search the resource
4. Select data **candidates**
5. Evaluate data **quality**

Based on: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/The-process-of-data-discovery

Evaluate data quality

What is the quality of the data?

- **What** information was collected, from whom, when and where?
- **Who** collected the data and when?
- **Why** was the data created? (research, social policy, marketing?)
- **How** was the data **collected**? (methodology)
- **How** was the data **processed**? Were there any changes in data?
- Is the data "**clean**" (were nonlogical and erroneous values deleted?)
- What **quality assurance procedures** were used? Did researchers use verified measurement tools?

DATA DISCOVERY PROCESS

1. Develop a clear **picture** of the **data** you need
2. Locate appropriate data **resources**
3. Set up a **search query** and search the resource
4. Select data **candidates**
5. Evaluate data **quality**

Based on: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/The-process-of-data-discovery

Other considerations

Access the data: is it free? Do I need to register? Is the access restricted? Do I need to apply to get access?

Data format: is the format of the files correct for your analyses? Do you need to transform the files or the dataset?

Missing data: are there any missing data in the dataset? How are you going to handle missing data?

Kaitlyn M. Werner, PhD
@kaitlynmwerner

Open science truly is beautiful. Someone recommended a paper w. open data relevant to this question. Within minutes I was able to analyze my question because the data/code was so beautifully and efficiently organized -- the best I've seen! Major props to @russpoldrack and team.

> **Kaitlyn M. Werner, PhD** @kaitlynmwerner · May 1
> I have been thinking a lot about socioeconomic status and self-control/self-regulation. I'm starting to plan an esm+diary study where I can start digging into this topic in more detail, but in the meantime I'm curious: what are the interesting papers you've read in this space?
> Show this thread

11:52 PM · May 9, 2022 · Twitter Web App

Cite the data

Harvard citation style:
Author names. Year. Title of resource. [medium type]. Host institution name, Physical location. Date of access. Identifier

Vancouver citation style
Author names. Title of resource [medium type]. Host institution name: Physical location; Year of publication. [Date accessed]. Available from: Identifier

Cite the data

Harvard citation style example:
Scarrow, S., Webb, P., Poguntke, T., 2017, Political Party Database, 2011-2014, [data collection], UK Data Service, Accessed 17 October 2018. SN: 8265, http://doi.org/10.5255/UKDA-SN-8265-1

DataverseNO

Dataverse Network Norway

Solvang, Øystein; Stein, Jonas; Brattland, Camilla, 2020, "Covid-19 Municipal Level (Norway) Social Science Dataset", https://doi.org/10.18710/NMKI2B, DataverseNO, V2

Cite Dataset

Learn about Data Citation Standards.

EndNote XML

RIS

BibTeX

The dataset is a cross-sectional dataset covering social and public health data pertaining
Norwegian municipalities. The dataset was compiled from public register data and media
related fatalities is current as of ultimo July 2020. Data on other variables is from 2018, 2

Document what you find and what you do!

Let’s see an example!

https://dataverse.no/

Example

Dataset: Covid-19 Municipal Level (Norway) Social Science Dataset

DataverseNO
Dataverse Network Norway

Questions?

Contact us at:

research-data@uio.no

UNIVERSITY
OF OSLO

University of Oslo Library

Digital Scholarship Center and Open Research

Open Science Lunch 2023

Each **last Thursday of the month at 12.00** we invite you to join us virtually for an **online open lunch** to hear about **how to make your research more open**.

February 23rd
Open your data with DataverseNO

March 30th
Public health reporting: an open source approach

April 27th
Open knowledge resources: Store norske leksikon and Wikipedia

May 25th
Keep the rights to your work: UiO's Rights Retention Policy

www.ub.uio.no/english/courses-events/events/open-science-lunch/

Time and place: Apr. 27, 2023 12:00 PM–1:00 PM, Hybrid: Georg Sverdrups hus and Zoom

Open knowledge resources: Store norske leksikon and Wikipedia

Join us for a discussion on using, maintaining and contributing to freely available open knowledge resources.

Time and place: May 25, 2023 12:00 PM–1:00 PM, Hybrid: Georg Sverdrups hus and Zoom

Keep the rights to your work: UiO's Rights Retention Policy

Learn about how you can retain the rights to your published work with the new Rights Retention Policy at UiO.

https://www.ub.uio.no/english/courses-events/events/open-science-lunch/

Menu

← Libraries and centres

https://www.ub.uio.no/english/libraries/dsc/

Norwegian version of this page

Digital Scholarship Centre

At the Digital Scholarship Centre (DSC) you get guidance on how you can make the best possible use of digital tools and methods in your research and communication activities.

Open Access →

Information about open access publishing, publisher agreements, self-archiving, requirements, and guidelines.

Open and reproducible research →

Make your research more transparent and reproducible.

Research Data Management →

Managing your data both during and after a research project.

Text-mining →

Information about digital tools for searching, mining, and analysing textual data.

Systematic search →

Information about systematic literature searching, how to get started, and how to get help.

Visualisation →

Use of visual methods to explore, communicate and understand data.

Carpentry@UiO →

Offers workshops in foundational digital skills such as coding and data management.

Reference management →

Styles, tools, and information on reference management.

Menu

← Libraries and centres

https://www.ub.uio.no/english/libraries/dsc/

Norwegian version of this page

Digital Scholarship Centre

At the Digital Scholarship Centre (DSC) you get guidance on how you can make the best possible use of digital tools and methods in your research and communication activities.

https://sympa.uio.no/ub.uio.no/subscribe/dsc-news/subscribe

Materials developed as a part of the *Skills development for research data* project:
https://www.ub.uio.no/english/about/projects/rdm-skills/

Thank you!

Contact us at:

research-data@uio.no

27. April 2023

Sources

CESSDA. The process of data discovery. https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover

UCL Data discovery & re-use: https://www.ucl.ac.uk/library/research-support/research-data-management/best-practices/how-guides/data-discovery-re-use

Gould Library: Data, Datasets and Statistical Resources https://gouldguides.carleton.edu/c.php?g=146834&p=964067

MacInnes, J. (2020). Secondary Analysis of Quantitative Data. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), *SAGE Research Methods Foundations.* https://www.doi.org/10.4135/9781526421036870195

Learn how to use the boolean operators in search queries: https://www.youtube.com/watch?v=lEo96kOKGmA