



# Data Reuse

**Speakers:** Jenny Ostrop (UiB)

**Moderator:** Xian Hu (Edna) (NCMM, UiO)

# Discovering & reusing research data

[Jenny Ostrop](#)

University of Bergen Library



## Learning Objectives

At the end of this session, you will be able to:

- outline benefits of reusing research data
- explain prerequisites for data reuse
- cite datasets
- apply different strategies to find scientific datasets



«Suggests measures to promote more use and reuse of data»

The Research Council of Norway  
Press release, 31.05.2022

# Foreslår tiltak for mer bruk og gjenbruk av data

For at data fra forskning og forvaltning skal komme hele samfunnet til nytte, må det satses mer koordinert og på tvers av sektorer, mener utvalg.

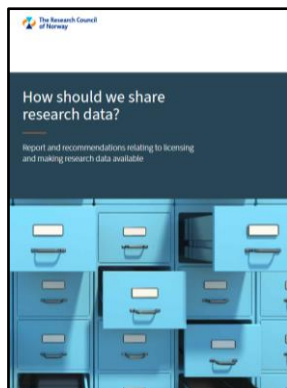
Pressemelding | Publisert 31. mai 2022

Del  Last ned 



# Research data are at the core of academic value creation

*Research and innovation are increasingly driven by access to new and large quantities of data.*



The Research Council of Norway (2021):  
How should we share research data? Report and recommendations  
related to licensing and making research data available. CC BY.



# Research data are at the core of academic value creation

*Research data should be managed and curated to take full advantage of their potential.*



Norwegian Ministry of Education and Research (2017):  
National strategy on access to and sharing of research data



# Research funders require data sharing



*The data used as the basis for scientific articles should be made accessible as soon as possible, and never later than at the time of publication.*

The Research Council of Norway's Policy for Open Access to Research Data (2017)



*Data should be deposited in a trusted repository as soon as possible after data production and at the latest by the end of the project.*

Horizon Europe Programme Guide v2.0 (2022)

# The research data life cycle



Research projects can:

1. Generate novel data
2. Reuse existing datasets (secondary data)



# Agenda

- Research data as resource
- Benefits of reusing data
- Data citation
- Prerequisites for data reuse
- Discovering datasets





# Benefits of reusing data



***Our intention is to make all raw data from all published studies available. The data contain a lot more interesting information than what has been published and we encourage users to dig further.***

The Moser group, Kavli Institute for Systems Neuroscience

# Benefits of reusing data

- many published datasets contain information that was not followed up in the connected research articles



# Benefits of reusing data

- many published datasets contain information that was not followed up in the connected research articles
- allows to apply new questions/angles to a published dataset



# Benefits of reusing data

- many published datasets contain information that was not followed up in the connected research articles
- allows to apply new questions/angles to a published dataset
- allows researchers to work with data they would not have the expertise/infrastructure/resources to produce themselves



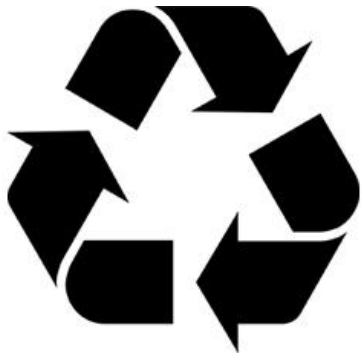
# Benefits of reusing data

- many published datasets contain information that was not followed up in the connected research articles
- allows to apply new questions/angles to a published dataset
- allows researchers to work with data they would not have the expertise/infrastructure/resources to produce themselves
- allows to integrate data from different studies, labs, disciplines,...



# Benefits of reusing data

- can inspire new avenues of research
- avoids unnecessary duplication of efforts, cost-effective



# Agenda

- Research data as resource
- Benefits of reusing data
- Data citation
- Prerequisites for data reuse
- Discovering datasets



# Benefits of sharing data: Data credit

- Joint Declaration of Data Citation Principles (JDDCP)

**Example:** The plots shown in Figure X show the distribution of selected measures from the main data [Author(s), Year, portion or subset used].

*Example:*

## References Section

Author(s), Year, Article Title, Journal, Publisher, DOI.

Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier.

Author(s), Year, Book Title, Publisher, ISBN.





# Benefits of sharing data: Data credit

- Best practices for attribution



## Title, Author, Source, License

A good rule of thumb is to use the acronym **TASL**, which stands for **T**itle, **A**uthor, **S**ource, **L**icense.




# Benefits of sharing data: Data credit

- Many archives contain information how a dataset should be cited



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



# Agenda

- Research data as resource
- Benefits of reusing data
- Data citation
- Prerequisites for data reuse
- Discovering datasets



# Data reuse requirements

1. Discovering suitable datasets



# Data reuse requirements

1. Discovering suitable datasets
2. Retrieving the data
  - Scale? Manual, automated, or API-retrieval?



# Data reuse requirements

1. Discovering suitable datasets
2. Retrieving the data
  - Scale? Manual, automated, or API-retrieval?
3. Understanding the data
  - Human-readable vs. machine-readable (metadata, data files)



# Data reuse requirements

1. Discovering suitable datasets
2. Retrieving the data
  - Scale? Manual, automated, or API-retrieval?
3. Understanding the data
  - Human-readable vs. machine-readable (metadata, data files)
4. Permission to build upon the data



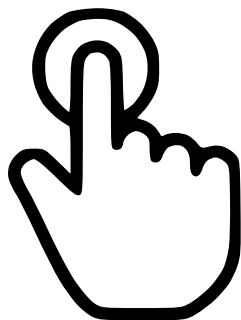
# FAIR: prerequisites for data reuse

F<sub>indeable</sub> A<sub>ccessible</sub> I<sub>nteroperable</sub> R<sub>euseable</sub>

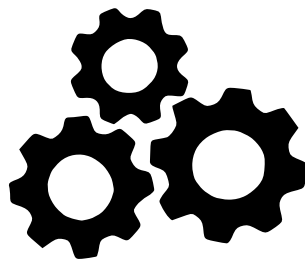


Rich metadata

Persistent identifier

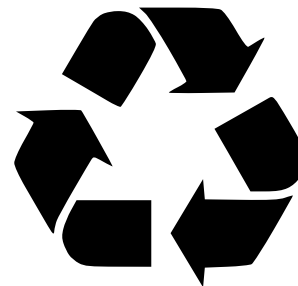


Data Repository  
& Access criteria



Standards

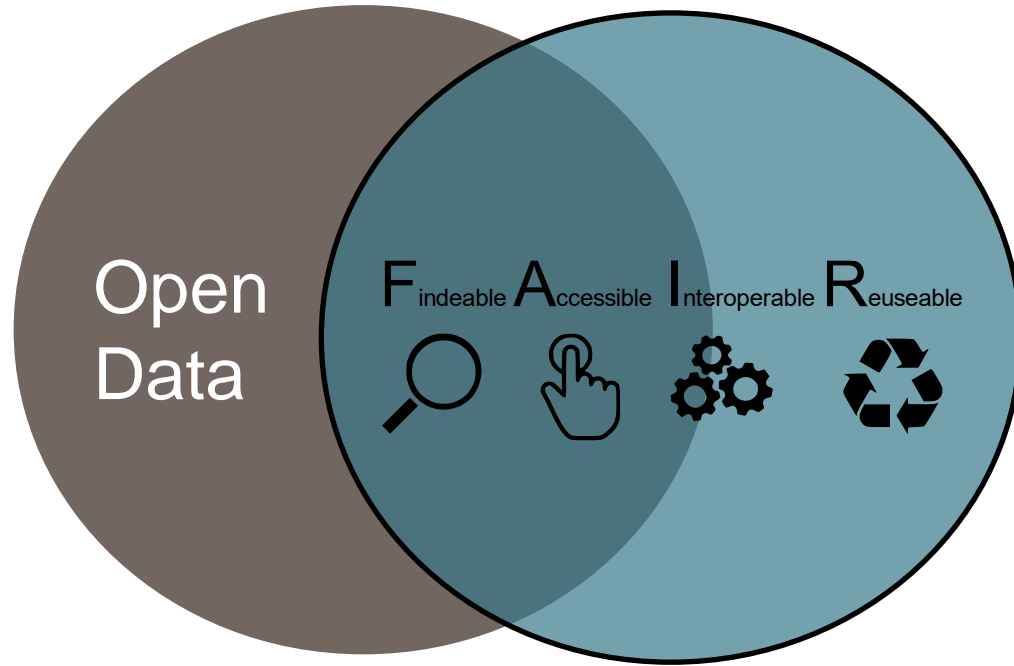
Avoiding ambiguity



License



# Open data and FAIR data

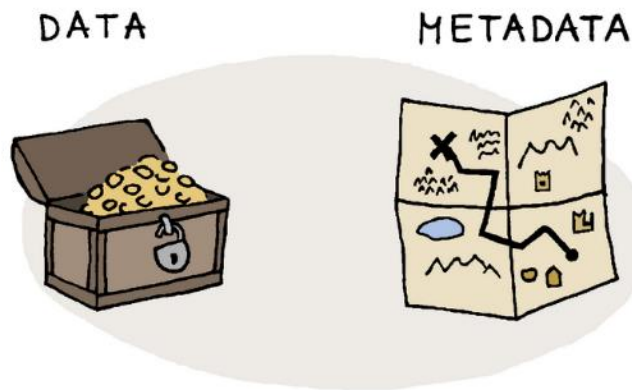


**“As open as possible – as closed as necessary”**



# Agenda

- Research data as resource
- Benefits of reusing data
- Data citation
- Prerequisites for data reuse
- Discovering datasets



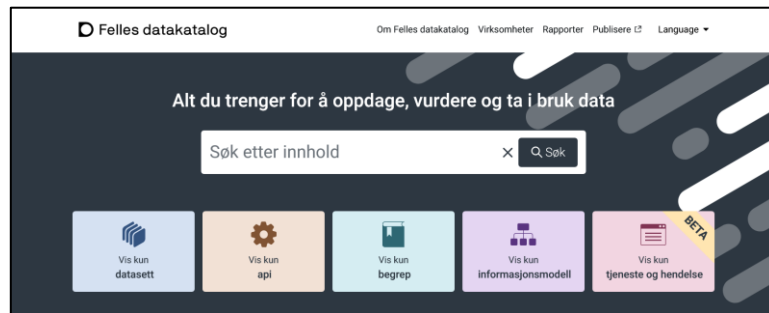
# Research data: an inclusive term

- All [digital] data collected or generated for use for or as a result of research activities (unprocessed and processed).
- Might include methodologies and workflows to reproduce the data.
- Describes range of types of information. Digital data can be structured and stored in variety of file formats.

# Discovering datasets

- Data from the public sector

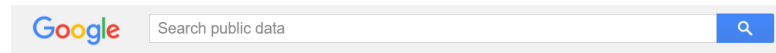
e.g.  Statistisk sentralbyrå  
Statistics Norway



<https://data.norge.no/>



<https://data.europa.eu/en>



Public Data

<https://www.google.com/publicdata/directory>

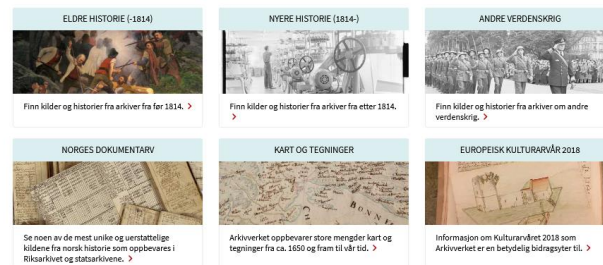
# Discovering datasets

- Data from the public sector
- Data in digital archives & collections



## Historier fra arkivene

Her finner du historiske smakebiter fra arkivene



<https://www.arkivverket.no/utforsk-arkivene>

## Utvalgte samlinger



### Manuskriptsamlingen

Manuskriptsamlingen inneholder håndskrevet og upublisert arkivmateriale. Store deler av den har bergensk og vestnorsk interesse, men samlingens lokalhistoriske stoff har rikspolitisk interesse på grunn av den dominerende stilling Bergen har spilt i norsk økonomisk, sosial og kulturell historie.



### Diplomsamlingen

Diplom- og dokumentsamlingen omfatter 1250 nummer, og er antagelig, etter Riksarkivets, den nest største i landet. Av dette er ca. 300 nummer diplomer fra tiden før 1600. Resten av samlingen består av skjøter, skillepapirer og kontrakter, de fleste fra 1600- og 1700-tallet.



### Billedsamlingen

Billedsamlingen ved Universitetsbiblioteket er en av landets største og mest anerkjente arkiv av historisk fotografi. Samlingen består av enkeltbilder og arkiver av varierende størrelse fra fotografer, samlere og private givere, til sammen omkring en halv million fotografiske bilder.



### Harmonien-samlingen

I forbindelse med Bergen Filharmoniske Orkester (Harmonien) sitt 250 års jubileum har Universitetsbiblioteket digitalisert en stor samling med dokumenter vedrørende dets første 150 år. Blant de interessante tingene man kan finne henvisninger til er Ole Bulls opptakspåre her som åtteåring i 1818.

<https://marcus.uib.no/home>



<https://www.europeana.eu/en>

# Discovering datasets

- Data from the public sector
- Data in digital archives & collections
- Scientific datasets
  - Data underlying a scientific article
  - Data not connected to a publication (e.g. negative data)

➤ **Research data repositories**



# Research data repositories

- Community repository



- Institutional repository



UiB Open Research Data

- Multidisciplinary repository



# Strategies to find scientific datasets

1. Data underlying a scientific article
2. Data in a relevant community archive
3. Dataset metasearch across archives

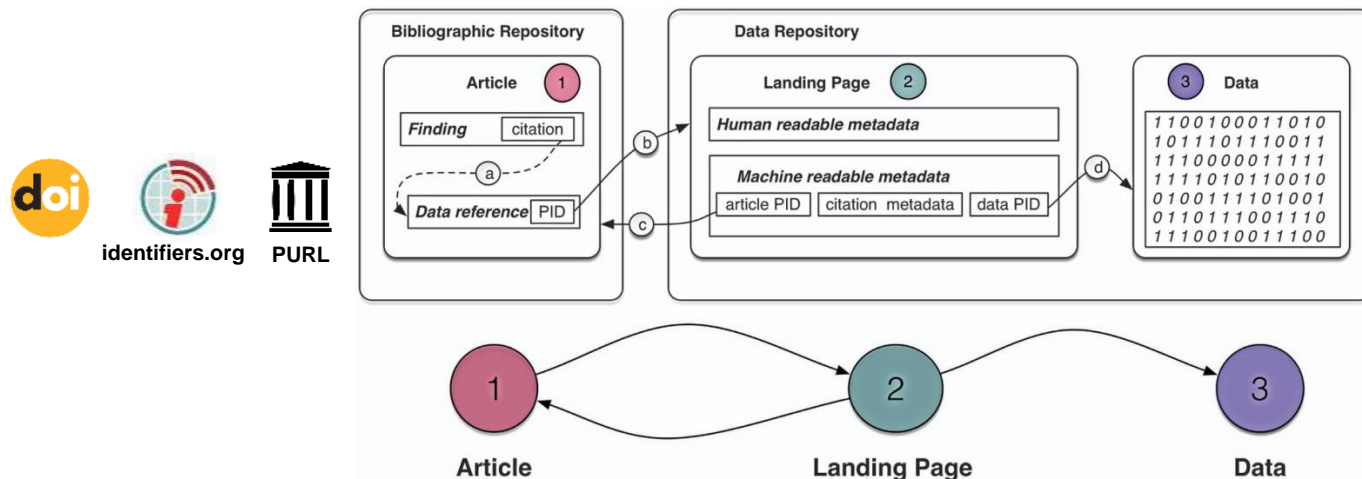




# Strategies to find scientific datasets

## 1. Data underlying a scientific article

- ~~Supplemental material~~
- Data repository



# Strategies to find scientific datasets

1. Data underlying a scientific article
  - Data availability statement?

Example:

**Cell**

Article

**Genetic Screens Identify Host Factors for SARS-CoV-2 and Common Cold Coronaviruses**

STAR★Methods

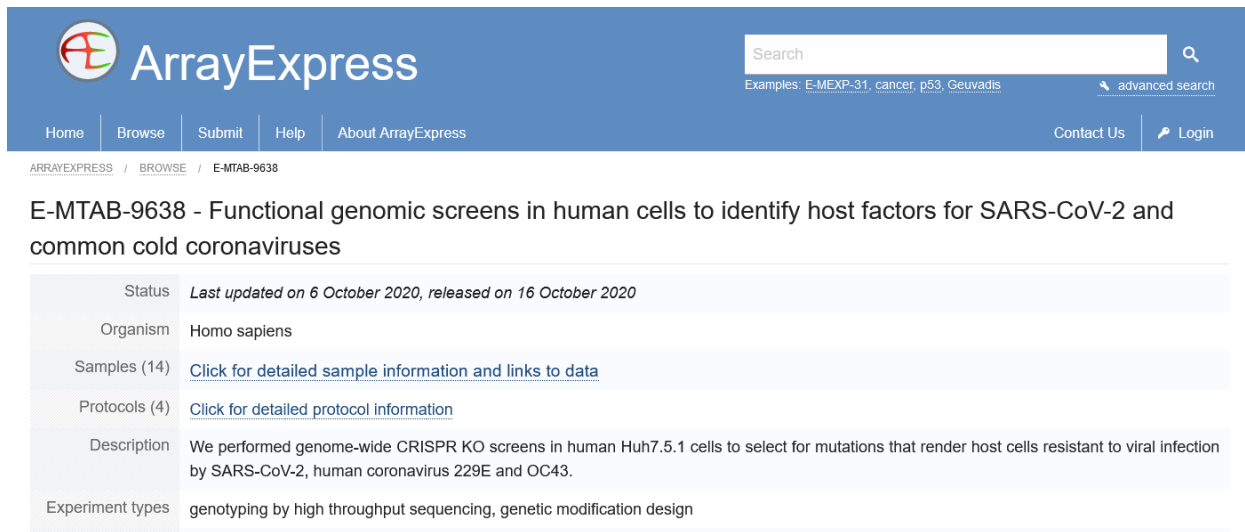
Data and Code Availability

The accession number for the raw sequencing data of the CRISPR KO screens reported in this paper is EMBL-EBI ArrayExpress: E-MTAB-9638.

# Strategies to find scientific datasets

1. Data underlying a scientific article
  - Data availability statement?

Example:



The screenshot shows the ArrayExpress website interface. The top navigation bar is blue with the ArrayExpress logo and name. A search bar is on the right. Below the navigation bar, there are links for Home, Browse, Submit, Help, and About ArrayExpress. The main content area displays the experiment title "E-MTAB-9638 - Functional genomic screens in human cells to identify host factors for SARS-CoV-2 and common cold coronaviruses". Below the title, there is a table with details about the experiment.

Status	<i>Last updated on 6 October 2020, released on 16 October 2020</i>
Organism	Homo sapiens
Samples (14)	<a href="#">Click for detailed sample information and links to data</a>
Protocols (4)	<a href="#">Click for detailed protocol information</a>
Description	We performed genome-wide CRISPR KO screens in human Huh7.5.1 cells to select for mutations that render host cells resistant to viral infection by SARS-CoV-2, human coronavirus 229E and OC43.
Experiment types	genotyping by high throughput sequencing, genetic modification design

# Strategies to find scientific datasets

## 1. Data underlying a scientific article

- Some literature search engines link to the data

The screenshot shows the Europe PMC search results page. The search bar contains 'j.cell.2020.12.004'. The search results show 1 result found. The article title is 'Genetic Screens Identify Host Factors for SARS-CoV-2 and Common Cold Coronaviruses'. The authors listed are Wang R<sup>1</sup>, Simoneau CR<sup>2</sup>, Kulsuptrakul J<sup>1</sup>, Bouhaddou M<sup>3</sup>, Travisano KA<sup>1</sup>, Hayashi JM<sup>4</sup>, Carlson-Stevermer J<sup>5</sup>, Zengel JR<sup>6</sup>, Richards CM<sup>6</sup>, Fozouni P<sup>7</sup>, Oki J<sup>5</sup>, Rodriguez L<sup>8</sup>, Joehnk B<sup>9</sup>, Walcott K<sup>9</sup>, Holden K<sup>5</sup>, Sil A<sup>9</sup>, Carrette JE<sup>6</sup>, Krogan NJ<sup>3</sup>, Ott M<sup>4</sup>, and Puschnik AS<sup>1</sup>. The article is from Cell, 09 Dec 2020, 184(1):106-119.e14. The DOI is 10.1016/j.cell.2020.12.004, PMID is 33333024, and PMCID is PMC7723770. The article is free to read & use.

The screenshot shows the 'Data' section on the Europe PMC article page. It lists various data resources linked to the article:

- Data behind the article**  
This data has been text mined from the article, or deposited into data resources.
- BioStudies: supplemental material and supporting data**  
<http://www.ebi.ac.uk/biostudies/studies/S-EPMC7723770?xr=true>
- Data Citations**  
DOI - 10.17632/r49yg49ddc (1 citation)
- Functional Genomics Experiments**  
ArrayExpress - E-MTAB-9638 (2 citations)
- HPA**  
HPA - HPA058342 (2 citations)
- Nucleotide Sequences (2)**  
ENA - R70007 (1 citation)  
ENA - MN908947 (1 citation)

# Strategies to find scientific datasets

1. Data underlying a scientific article
2. Data in a relevant community archive
3. Dataset metasearch across archives



# Strategies to find scientific datasets

## 2. Data in a relevant community archive

- Advantage: uniform format & metadata schemes, discipline-specific standards, sometimes curated



# Strategies to find scientific datasets

## 2. Data in a relevant community archive

- Advantage: uniform format & metadata schemes, discipline-specific standards, sometimes curated
- Where do researchers in the field publish their data?
- Curated registries can help to identify data repositories:

[re3data.org](https://re3data.org)

**re3data.org**  
REGISTRY OF RESEARCH DATA REPOSITORIES

[fairsharing.org](https://fairsharing.org)

**FAIRsharing.org**  
standards, databases, policies



# Strategies to find scientific datasets

1. Data underlying a scientific article
2. Data in a relevant community archive
3. Dataset metasearch across archives





# Strategies to find scientific datasets

## 3. Data metasearch engines

- Searching across disciplines
- Data in institutional archives
- Data in multidisciplinary archives

- Metadata quality is critical!
- Repository coverage varies
- Persistent identifiers: datasets with DOI are easiest to find



# The vision: EOSC

- Making data and any other digital research artefact (such as documents, algorithms, tools and workflows) as FAIR as possible across all European research infrastructures.



**EUROPEAN OPEN  
SCIENCE CLOUD**

- Core data infrastructure should be in place 2024-2025



# Data metasearch engines

- Non-commercial

- [DataCite](#)



- [BASE](#)



- [OpenAIRE](#)



- Commercial

- [Google Dataset Search](#)



- [Mendeley Data](#)



- [WOS Data Citation Index](#)



## NB!

- Not every search result will be a “real” dataset
- Some journals deposit articles figures/tables as dataset (e.g. to FigShare)



# Life Science-specific metasearch

- [EBI Search](#): uniform access to the biological data resources hosted ad EMBL-EBI
  - Web search
  - [RESTful API](#)

**EBI Search**

A screenshot of the EBI Search web interface. It features a dark teal header bar. On the left, there is a white search input field with the placeholder text "Search". To the right of the input field is a green square button with a white magnifying glass icon. Below the input field, there is a line of small text providing examples: "Examples: VAV\_HUMAN, tp53, Sulston...". In the bottom right corner of the header bar, there is a link that says "Build Query".

Search

Examples: VAV\_HUMAN, tp53, Sulston...

Build Query



# Life Science-specific metasearch

- [EBI Search](#): uniform access to the biological data resources hosted at EMBL-EBI

**EBI Search**

- Web search

- [RESTful API](#)



Search

Examples: VAV\_HUMAN, tp53, Sulston...

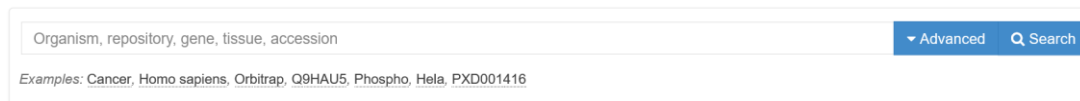
Build Query

- [OmicsDI](#): knowledge discovery framework across heterogeneous omics data (genomics, proteomics, transcriptomics and metabolomics).



- Web search

- [RESTful API](#)



Organism, repository, gene, tissue, accession

Examples: Cancer, Homo sapiens, Orbitrap, Q9HAU5, Phospho, HeLa, PXD001416

Advanced Search

# Take-home

- Reusing existing data can inspire new avenues of research & avoids unnecessary duplication of efforts.
- FAIR principles are prerequisites for data reuse.
- In addition to scientific datasets, data from the public sector and data in digital archives can be interesting sources.
- Scientific datasets are shared in community archives, institutional archives, and general-purpose archives.





## Useful Resources

- **RDMkit - Reusing:** <https://rdmkit.elixir-europe.org/reusing>

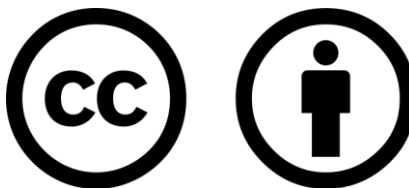


- [OpenAire: Can I reuse someone else research data?](#)
- [CESSDA Data Management Expert Guide: Access, use and cite data](#)
- [Digital Curation Centre: How to Cite Datasets and Link to Publications](#)
- [PhD-on-track: Citing research data](#)



# Thank you!

 @jennyostrop	 elixir-norway.org
 jenny.ostrop@uib.no	 @elixirnorway
	 contact@bioinfo.no



**Except where otherwise noted, this work is licensed under a  
Creative Commons Attribution 4.0 International License**

**<https://creativecommons.org/licenses/by/4.0/>**