Data Reuse

Speakers: Jenny Ostrop (UiB)
Moderator: Xian Hu (Edna) (NCMM, UiO)

15 June 2022

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

Forskningsrådet

«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

https://www.forskningsradet.no/om-forskningsradet/pressekontakt/pressemeldinger/2022/foreslar-tiltak-for-mer-bruk-og-gjenbruk-av-data/

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.

https://www.forskningsradet.no/siteassets/publikasjoner/2021/how-should-we-share-research-data.v2.pdf

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

https://www.regjeringen.no/en/dokumenter/national-strategy-on-access-to-and-sharing-of-research-data/id2582412

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

Modified from Research data management training @ UiT (2019). Kavli Institute for Systems Neuroscience: Grid cell data.

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

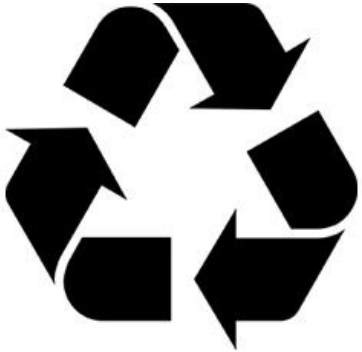
- many published datasets contain information that was not followed up in the connected research articles
- allows to apply new questions/angles to a published dataset
- allows researchers to work with data they would not have the expertise/infrastructure/resources to produce themselves

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.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

Cousijn, Clark et al., 2018: https://doi.org/10.1038/sdata.2018.259

Benefits of sharing data: Data credit

- Best practices for attribution

creative commons

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.

https://wiki.creativecommons.org/wiki/best_practices_for_attribution

Benefits of sharing data: Data credit

- Many archives contain information how a dataset should be cited

DataverseNO

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

https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/NMKI2B

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)

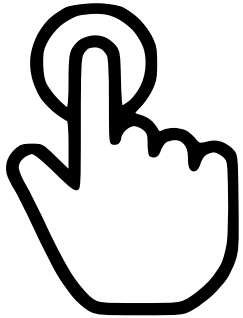
4. Permission to build upon the data

FAIR: prerequisites for data reuse

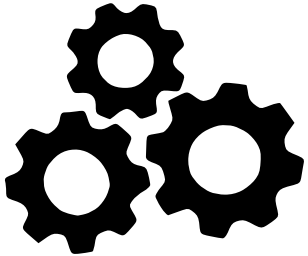
Findeable **A**ccessible **I**nteroperable **R**euseable

Rich metadata

Persistent identifier

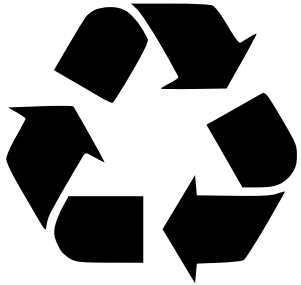


Data Repository
& Access criteria



Standards

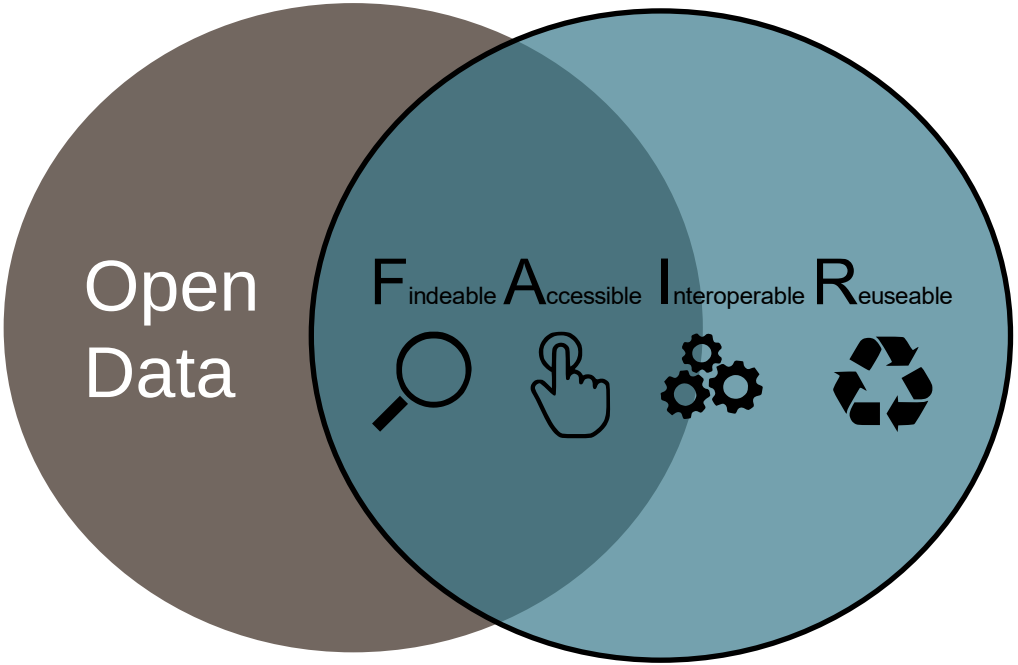
Avoiding ambiguity



License

Wilkinson et al., 2016 Sci Data; https://www.go-fair.org/fair-principles/

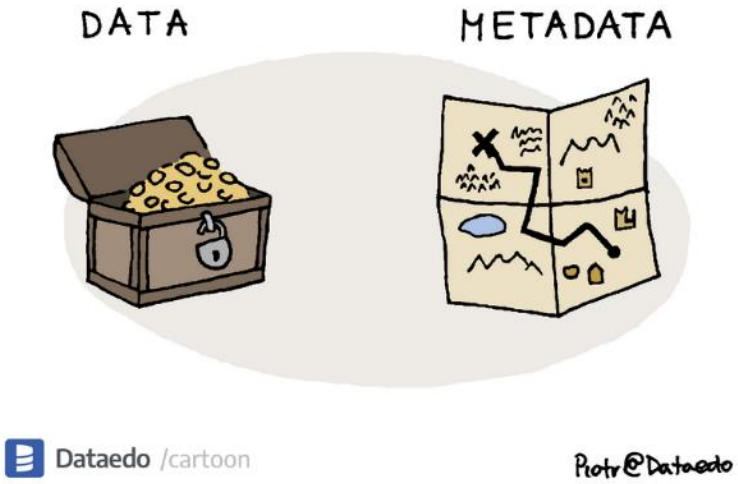
Open data and FAIR data



“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



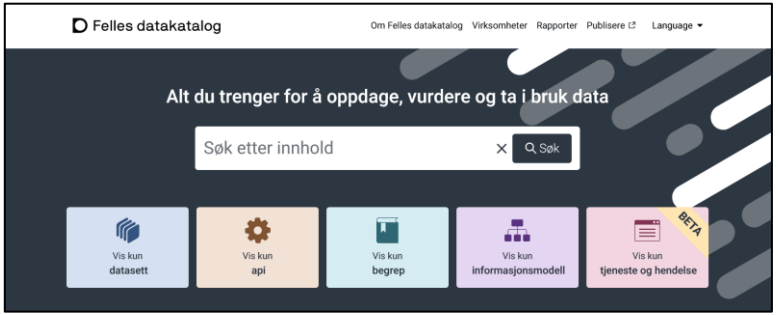
Research data: an inclusive term

- 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

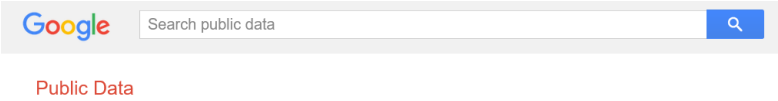
e.g. Statistisk sentralbyrå
Statistics Norway



https://data.norge.no/

data.europa.eu

https://data.europa.eu/en

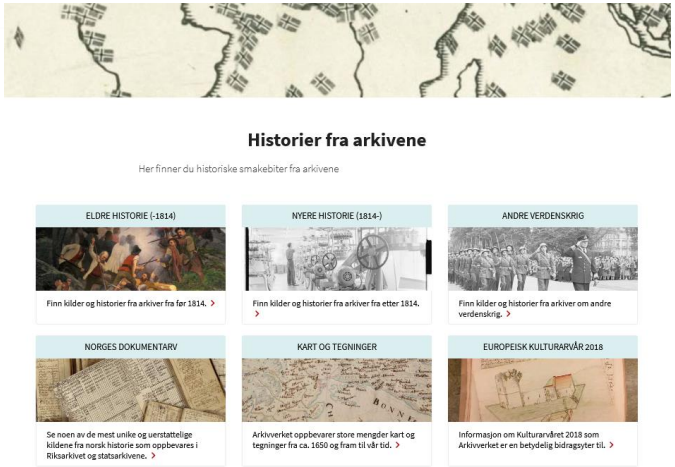


Public Data

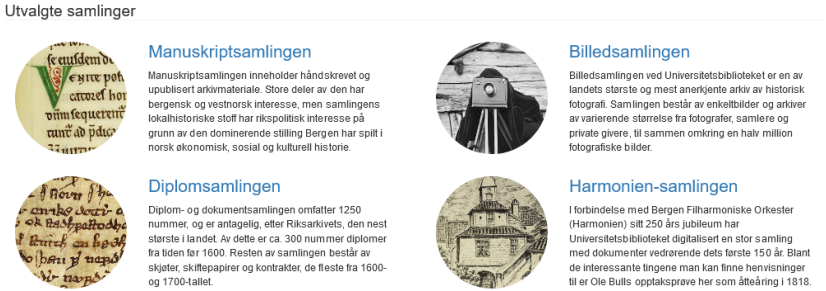
https://www.google.com/publicdata/directory

Discovering datasets

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



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



https://marcus.uib.no/home

europeana
think culture

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
- Multidisciplinary repository

Strategies to find scientific datasets

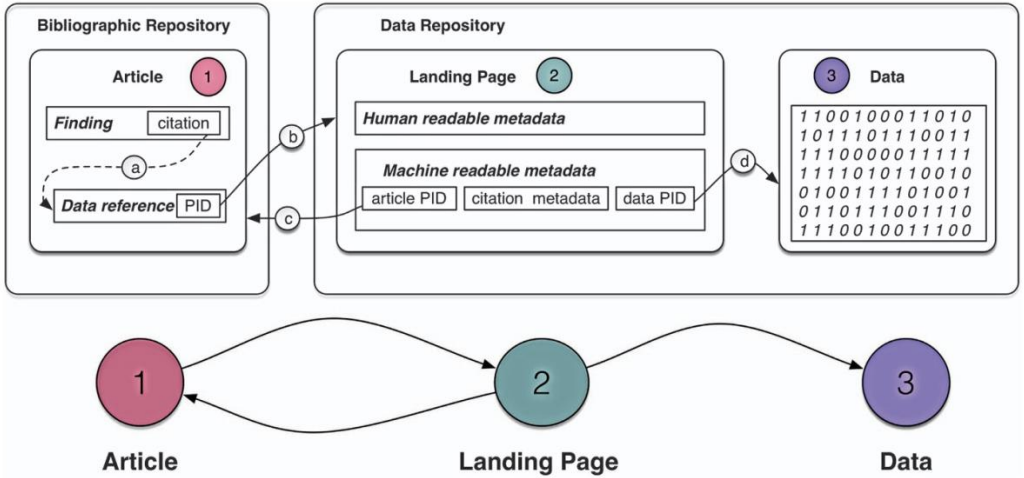
1. Data underlying a scientific article
2. Data in a relevant community archive
3. Dataset metasearch across archives

Strategies to find scientific datasets

1. Data underlying a scientific article
   - ~~Supplemental material~~
   - Data repository

identifiers.org

PURL



Cousijn, Clark et al., 2018: https://doi.org/10.1038/sdata.2018.259

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

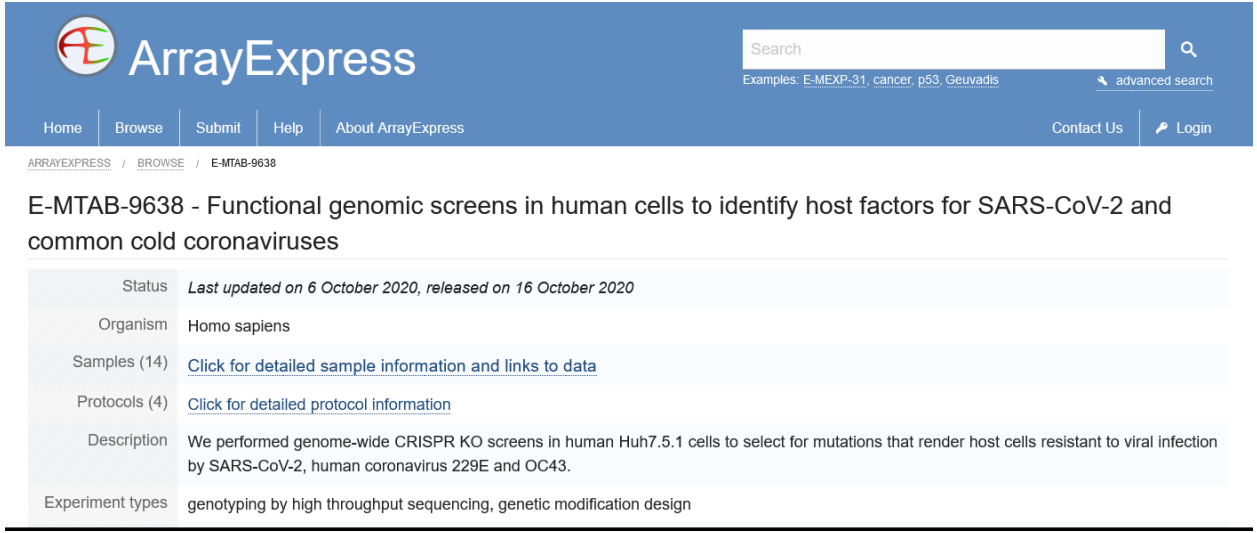
The accession number for the raw sequencing data of the CRISPR KO screens reported in this paper is EMBL-EBI ArrayExpress: E-MTAB-9638.

Wang, Puschnik et al., 2021: https://doi.org/10.1016/j.cell.2020.12.004

Strategies to find scientific datasets

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

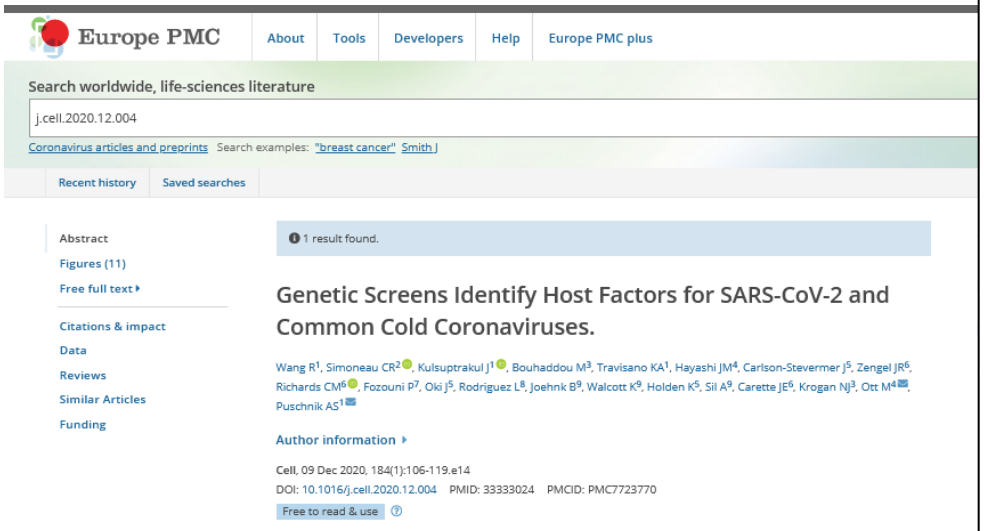
Example:

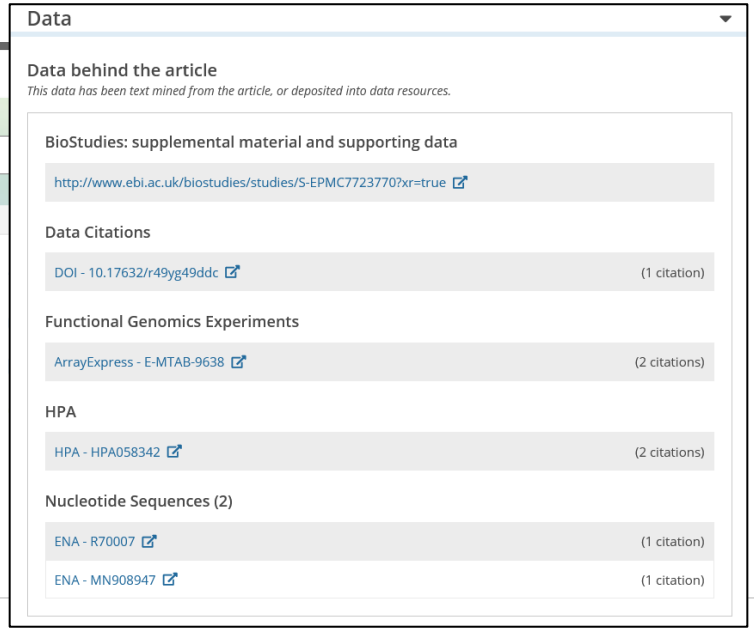


https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9638/

Strategies to find scientific datasets

1. Data underlying a scientific article
   - Some literature search engines link to the data





https://europepmc.org/article/MED/33333024

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

fairsharing.org

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

DataCite
FIND, ACCESS, AND REUSE DATA

BASE

OpenAIRE | EXPLORE

Google
Dataset Search Beta

- 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

https://eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf

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)

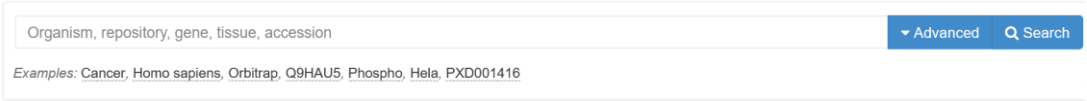
Life Science-specific metasearch

- EBI Search: uniform access to the biological data resources hosted ad EMBL-EBI
  - Web search
  - RESTful API

Life Science-specific metasearch

- EBI Search: uniform access to the biological data resources hosted ad EMBL-EBI
  - Web search
  - RESTful API

- OmicsDI: knowledge discovery framework across heterogeneous omics data (genomics, proteomics, transcriptomics and metabolomics).
  - Web search
  - RESTful API



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!

elixir-norway.org

@jennyostrop

@elixirnorway

jenny.ostrop@uib.no

contact@bioinfo.no