

Data Publication

Introduction to Data Management Practices course

NBIS DM Team

data-management@scilifelab.se

Learning objectives

1. Know the benefits of data sharing
2. Know how to find a suitable repository for different types of data
3. Know how to make a publication plan for your dataset(s)
4. Know where to get help in future submission adventures

What is a data repository?

A centralized storage for datasets, allowing for easy access, sharing, and reuse of data.

Key features:

- Metadata: Descriptive information about data that makes it easier to find and understand.
- Persistent Identifiers (PIDs): Unique, permanent identifiers (e.g., DOIs) that ensure data can always be found and cited.
- Data repositories support long-term data preservation and research transparency.

Why submit to a repository?

“The data is available upon request”

Many reasons:

- Open Science & FAIR
- Reproducibility
- Trail of evidence
- 3rd party access
- Archival purposes
- Publication of paper requires it



Digitalbevaring.dk

Credit: Illustration from Digitalbevaring.dk / Jørgen Stamp (CC BY 2.5 Denmark license).

Why submit to a repository?

Data publication is the best way to make your research projects FAIR since your data becomes:

- **Findable** by being assigned a persistent identifier, and by being described with rich metadata
- **Accessible** by being put in a resource that is searchable, and enables easy access via internet
- **Interoperable** by using standard format and language to represent both the data and its metadata
- **Reusable** by fulfilling the F, A, and I, and by having a clear and accessible data usage license

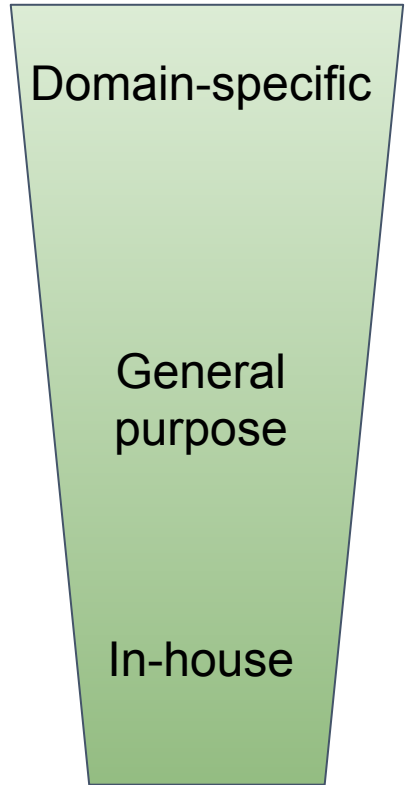
What is data?

What research outputs should be submitted?

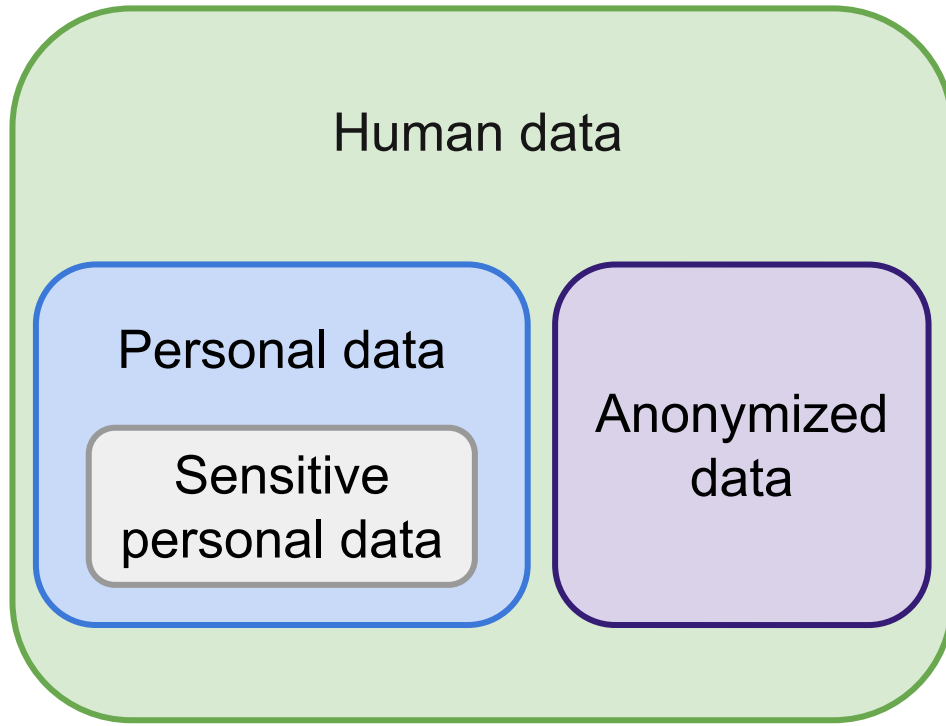
- **Raw data:** straight from the instrument eg fastq, bam, cram
- **Processed data:** normalization, removal of outliers, expression measurements, statistics
- **Metadata:** minimum information to reproduce the data, sample information, precise protocols

Types of repositories

- **Domain-specific:**
 - Best choice - long-term plan, typically free, maximum reach
 - E.g. [European Nucleotide Archive](#), [European Genome Phenome Archive](#), [ArrayExpress](#), [PRIDE](#)
- **General purpose:**
 - Second best - long-term plan, might cost (now or in future), good reach but less specific in metadata → more difficult for future users to judge if a dataset will be useful
 - E.g. [Zenodo](#), [\(SciLifeLab\) Figshare](#), [Dryad](#)
- **In-house/institutional:**
 - For archive/backup purpose mainly, might cost, limited reach unless also published in a data catalogue



Sensitive personal data



Human data is any data associated with or derived from a human being.

Personal data is any information that refers to an identified or identifiable living person.

Anonymized data is personal data that has been stripped of any personally identifiable information.

Sensitive personal data is special categories of personal data that need extra protection.

What data is regarded as personal data?

Not personal data

Name:
'Eva Johansson'

Probably not
personal data

Name:
'Eva Johansson'

City:
'Malmö'

Probably personal data

Name:
'Eva Johansson'

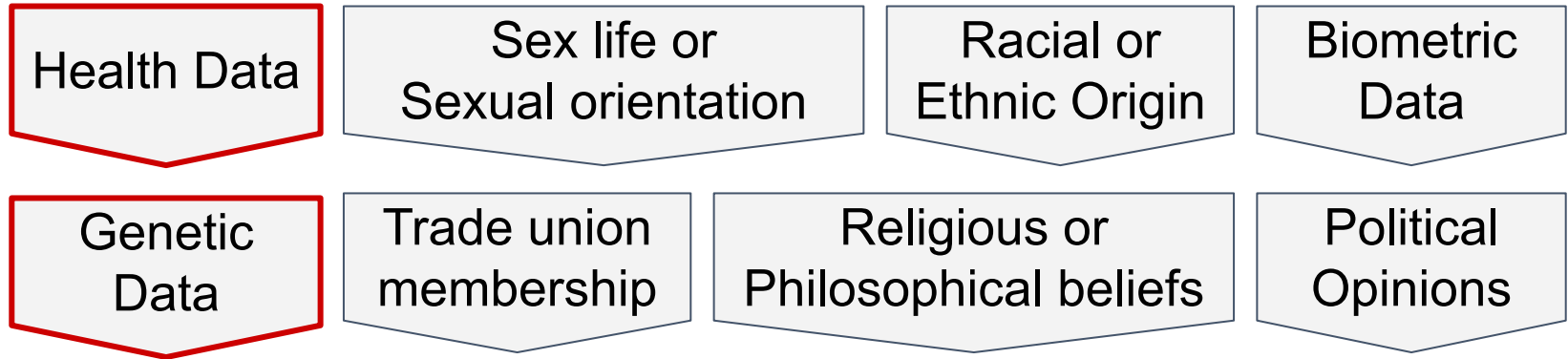
City:
'Malmö'

Date of birth:
'1985-01-01'

Data may or may not be regarded as
personal data depending on the context

What is sensitive personal data?

Special categories of personal data defined by GDPR (Art. 9)



Processing of sensitive personal data is prohibited, except under certain circumstances (e.g. research).

When personal data becomes sensitive

Personal data

Name:
'Eva Johansson'

City: 'Malmö'

Date of birth:
'1985-01-01'



Not personal
data

Diagnosis:
'diabetes, type 1'



Sensitive
personal data

Name:
'Eva Johansson'

City: 'Malmö'

Date of birth:
'1985-01-01'

Diagnosis:
'diabetes, type 1'

Repositories for sensitive personal data

European Genome-Phenome Archive (EGA) is a repository for all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects.

- Hosted by EMBL-EBI and CRG
- Launched 2008
- Legal challenges for Swedish universities



Repositories for sensitive personal data

Federated EGA (FEGA) is a network of national human data repositories (FEGA nodes) in Europe. Sensitive data stored locally in each node.

- Metadata of dataset is available upon search on EGA
- Sensitive data available on request (restricted access)
- [FEGA Sweden](#), the Swedish FEGA node, is hosted by NBIS/Uppsala University
- Provide support and services adapted to the Swedish life science research community



Evaluate a repository

Things to check when evaluating:

- Are others in the community using it?
- Is it easy to navigate / user-friendly?
- Is there support / guidance for submission and reuse?
- Is it sustainable, i.e. will the repository be around for a while?
- Will the datasets obtain persistent identifiers? Is the repository itself FAIR?

Identify repositories

How to find a suitable repository for life science data?

- [EBI repository wizard](#) - guide depending on data type
- [ELIXIR deposition databases](#) - core resources with long-term data preservation and accessibility plans
- [FAIRsharing.org/databases](#) - catalogue of many repositories, with possibility to filter on e.g. domain
- [Scientific Data Repository Guidance](#) - publisher's recommendation
- [re3data.org](#) - registry of research data repositories (not only life science)

Demo: EBI Repository Wizard

Which repository would be suitable if you have a genomics project with mice RNA sequences?

- Go to <https://www.ebi.ac.uk/submission/>
- Answer the questions regarding
 - data type
 - need for controlled access
 - if experimentally produced by you
 - type of study

Key Points

- Publishing data greatly increases the FAIRness of your research.
- Benefits of sharing data are several e.g. reproducibility purposes, follow the Open Science directive, meet requirement from publishers.
- If possible, use a domain-specific repository since it has maximum reach in the research community.
- The research output data types determines which domain-specific repository is suitable.
- **Special conditions apply for sharing sensitive personal data.**

Exercise: Planning for data publication

- **Scenarios listed in Canvas**
 - Project A
 - Project B
 - Project C

- **Answer the questions in the Rolling course notes**
 - What types of outputs will you be creating or collecting?
 - What are suitable repositories for your outputs?
 - What are the documentation guidelines for the repositories? Which formats for data and metadata are required to be able to submit?
 - Under what licenses will your research outputs be shared?