NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Data Publication

Introduction to Data Management Practices course

NBIS DM Team
data-management@scilifelab.se

https://nbisweden.github.io/module-data-publication-dm-practices/index.html

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

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

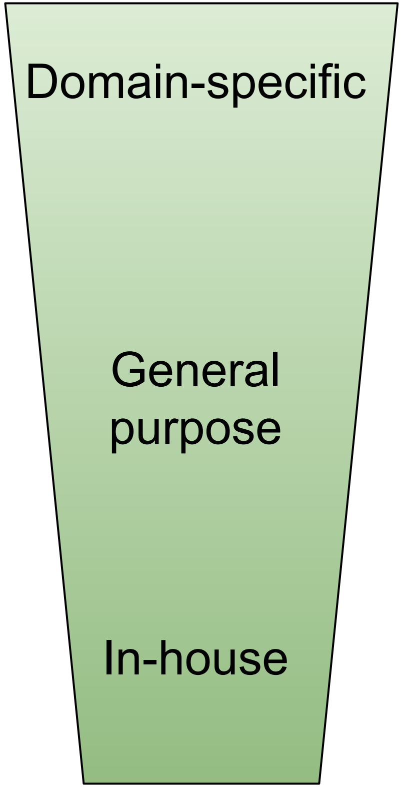
What is data?

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



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 (DNA/RNA sequence)
  - need for controlled access (No)
  - if experimentally produced by you (Yes)
  - type of study (Other)
- Solution: European Nucleotide Archive (ENA)

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.