NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

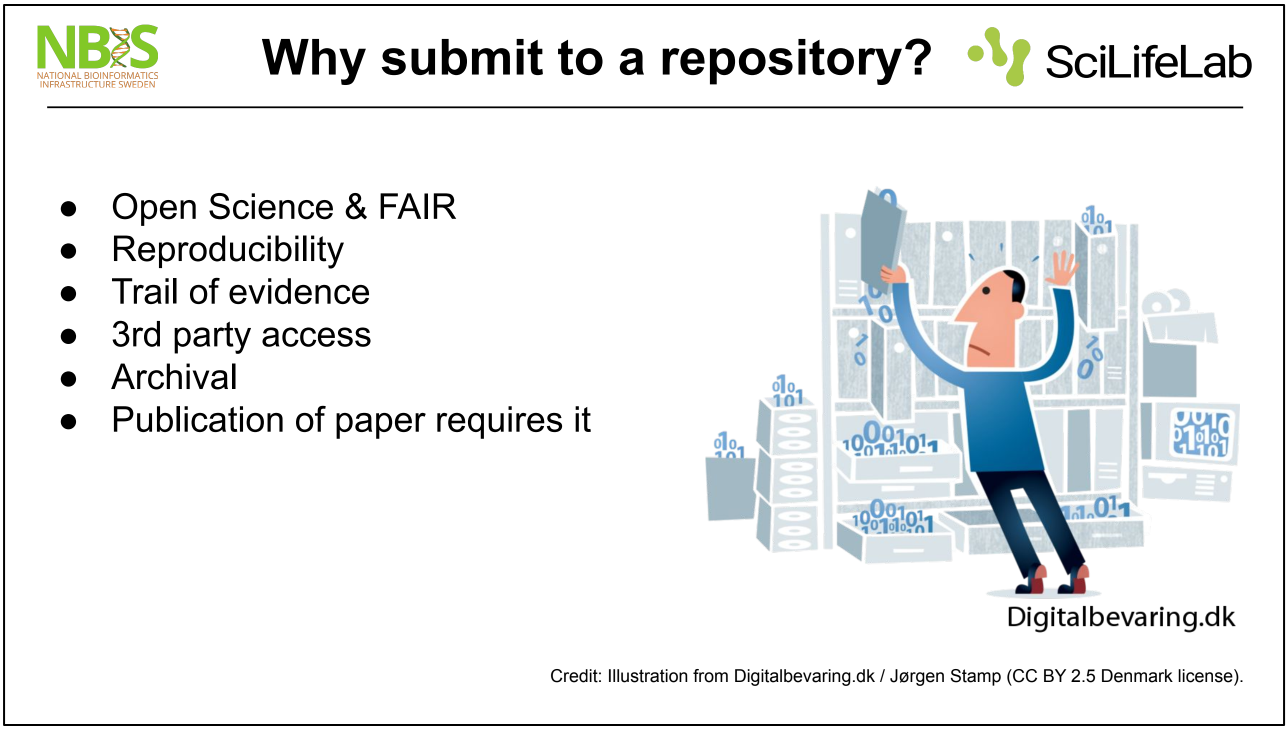
SciLifeLab

Data Publication

Introduction to Data Management Practices course

NBIS DM Team
data@nbis.se

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



Why submit to a repository?

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

Digitalbevaring.dk

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

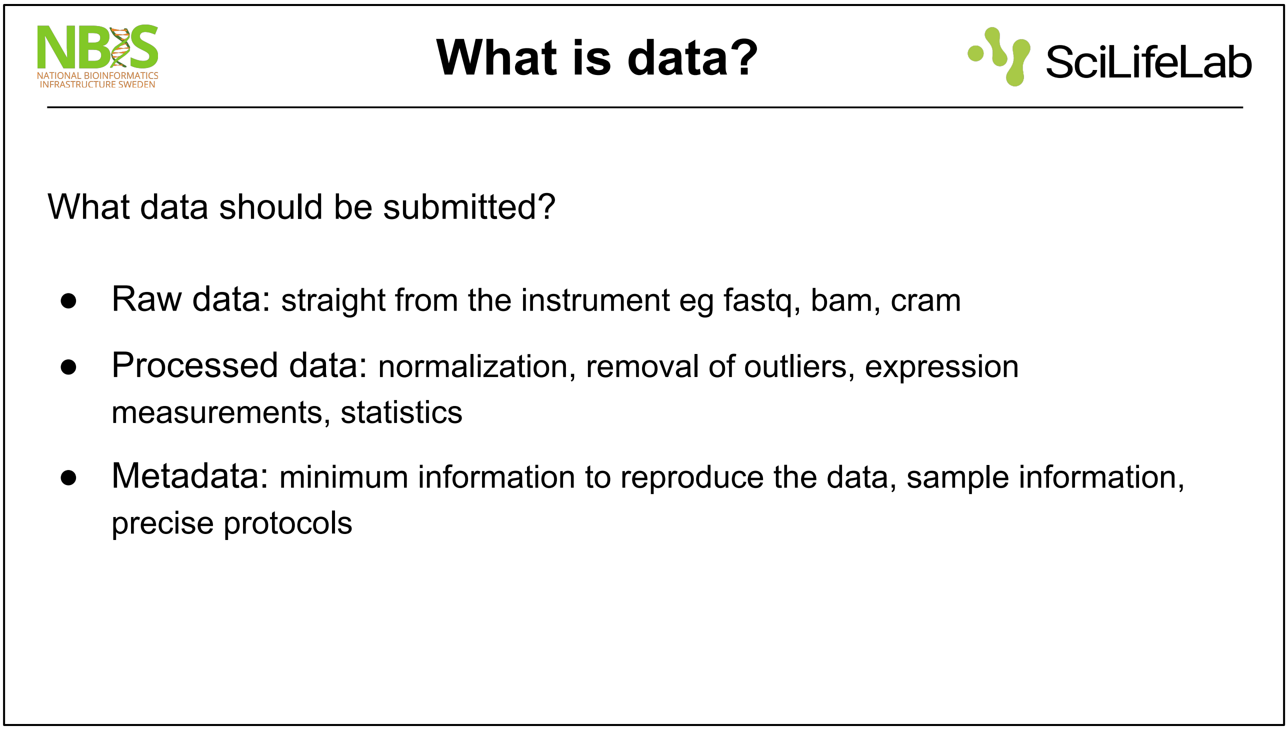
Why submit your datasets to a repository?

- To meet the requirements from funders and society on Open Science & FAIR
- So that your published research results can be reproduced
- To provide a trail of evidence, a provenance of the data
- To give others access to your data (3rd party access)
- For archival purposes, research data should be available for as long as it is useful to someone
- Nowadays most publishers require you to submit the data to a repository when publishing a paper

FAIR Data

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



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

What is data?

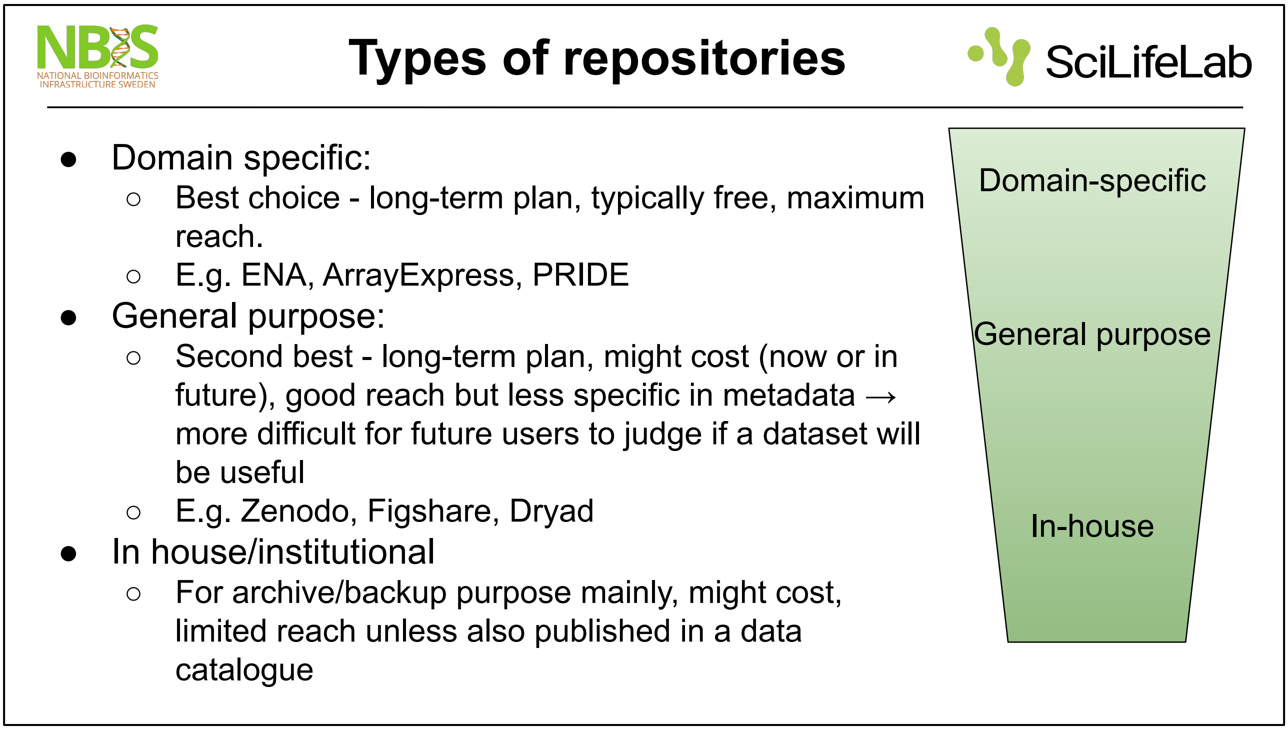
SciLifeLab

What data 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

What data should be submitted?

- Raw data: this is the data that comes straight from the instrument, eg RNA sequences in fastq format
- Processed data: this is the data where some type of analysis or processing has been done, eg normalization, removal of outliers, expression measurements, statistics
- Metadata: this is the description of the raw and processed data, eg in the form of minimum information to reproduce the data, sample information, precise protocols



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Types of repositories

SciLifeLab

- Domain specific:
  - Best choice - long-term plan, typically free, maximum reach.
  - E.g. ENA, 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, Figshare, Dryad
- In house/institutional
  - For archive/backup purpose mainly, might cost, limited reach unless also published in a data catalogue

There are different types of repositories:
Domain specific repositories:

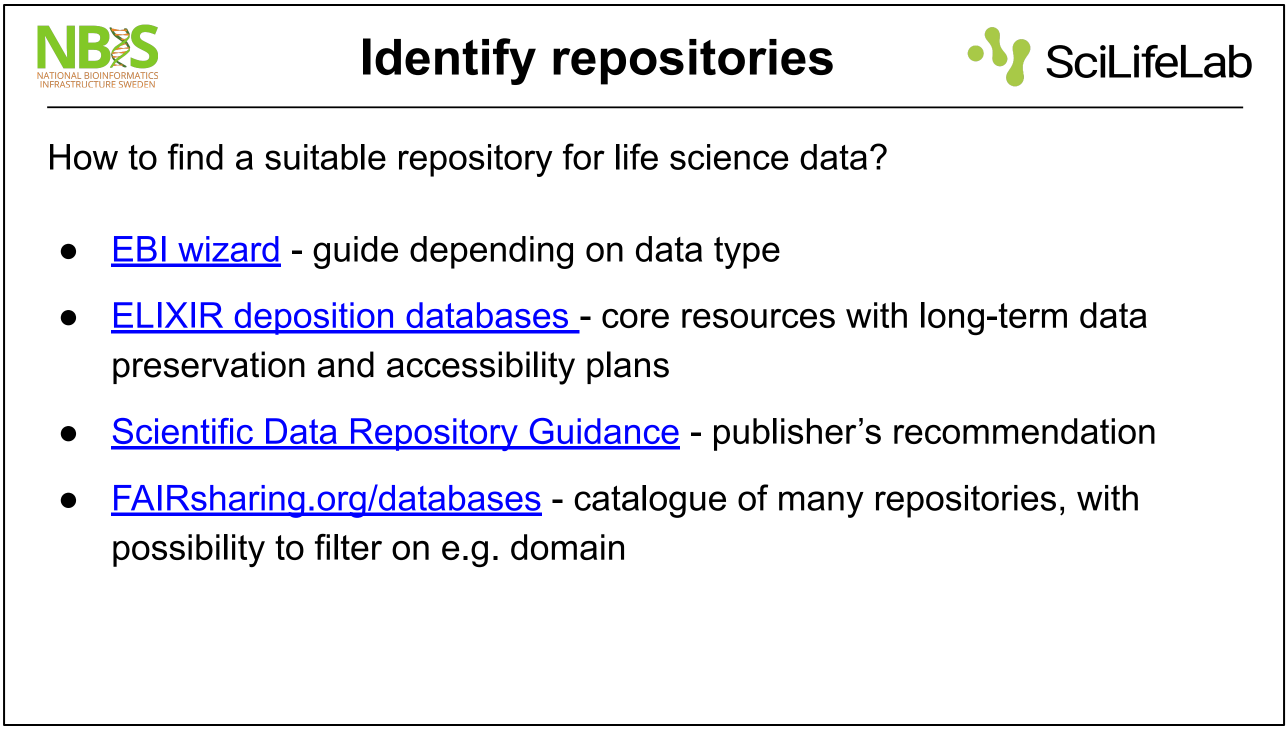
- Best choice if there is a suitable one for your data type. They have long-term plan for sustainability, they are typically free of charge, and has maximum reach in your research community.
- E.g. European Nucleotide Archive, ArrayExpress

General purpose repositories:

- If there is no domain specific repository, a general purpose repository is the best choice. They also have long-term plan for sustainability, but might cost (now or in future), and they do have good reach. However, the metadata is less specific in metadata which means it is more difficult for future users to judge if a dataset will be useful to them or not.
- E.g. Zenodo, Figshare, Dryad

In house/institutional repositories:

- This type of repository is for archive/backup purpose mainly, since it has limited reach outside the institution, they are typically not ‘googable’, unless also published in a public (indexed) data catalogue. Also, it might be associated with a cost.



Identify repositories

How to find a suitable repository for life science data?

- EBI wizard - guide depending on data type
- ELIXIR deposition databases - core resources with long-term data preservation and accessibility plans
- Scientific Data Repository Guidance - publisher’s recommendation
- FAIRsharing.org/databases - catalogue of many repositories, with possibility to filter on e.g. domain

How do you find a domain specific repository?

- EBI provides a wizard, which will guide you to a repository depending on the data type
- ELIXIR deposition databases - core resources with long-term data preservation and accessibility plans, recommended by the European infrastructure for Life Sciences
- FAIRsharing.org/databases - catalogue with many repositories.They provide filtering options on eg domain or recommendations from publishers.

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.