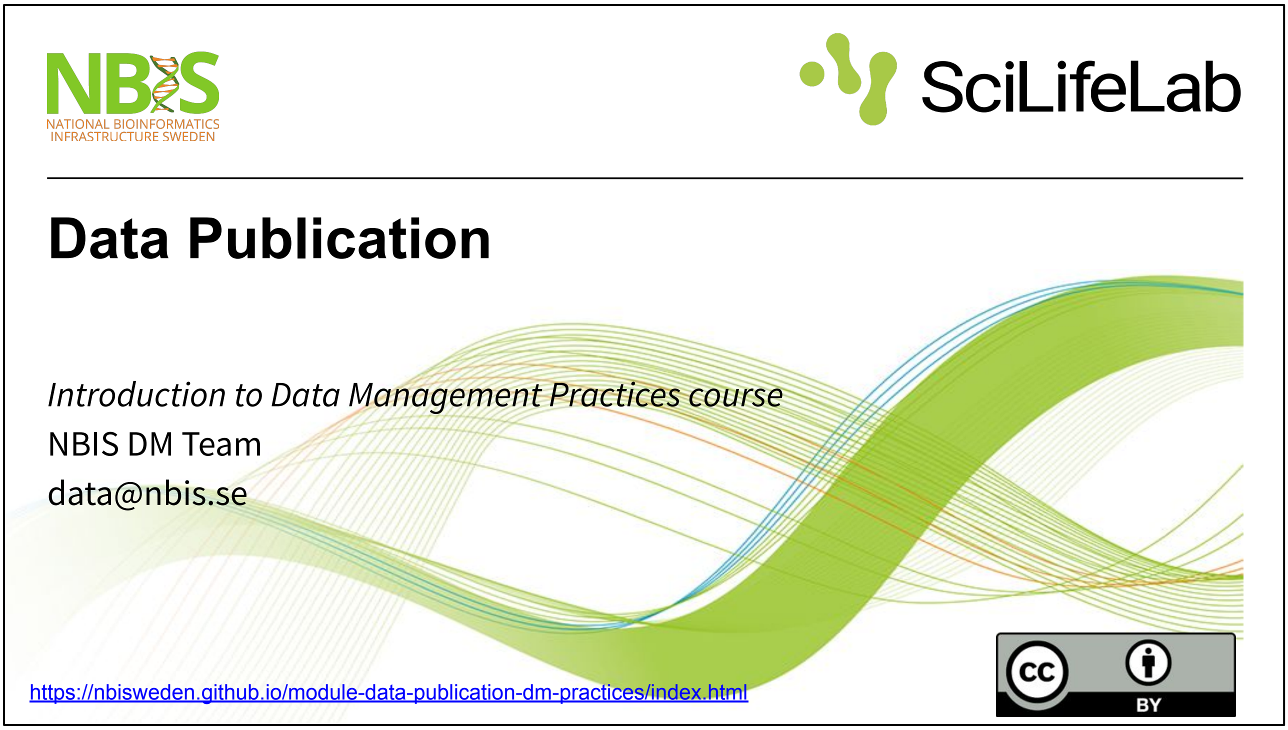


NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

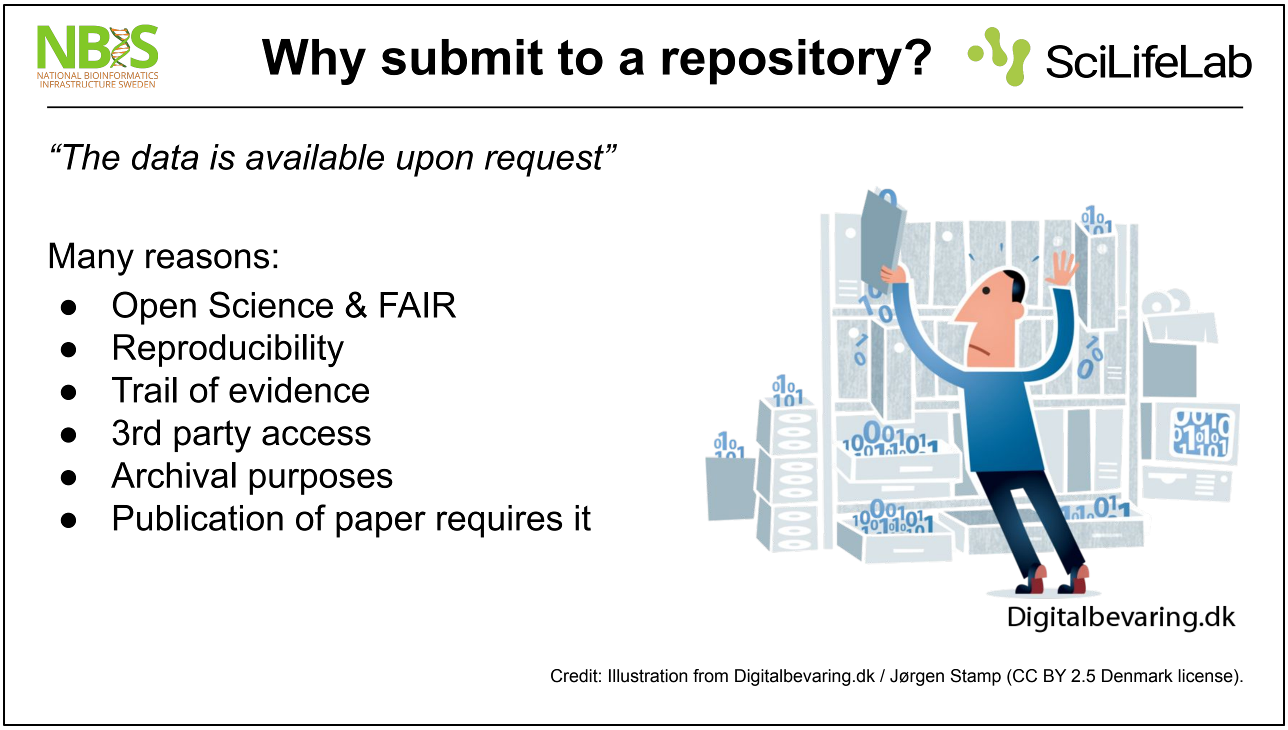
Data Publication

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?

“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).

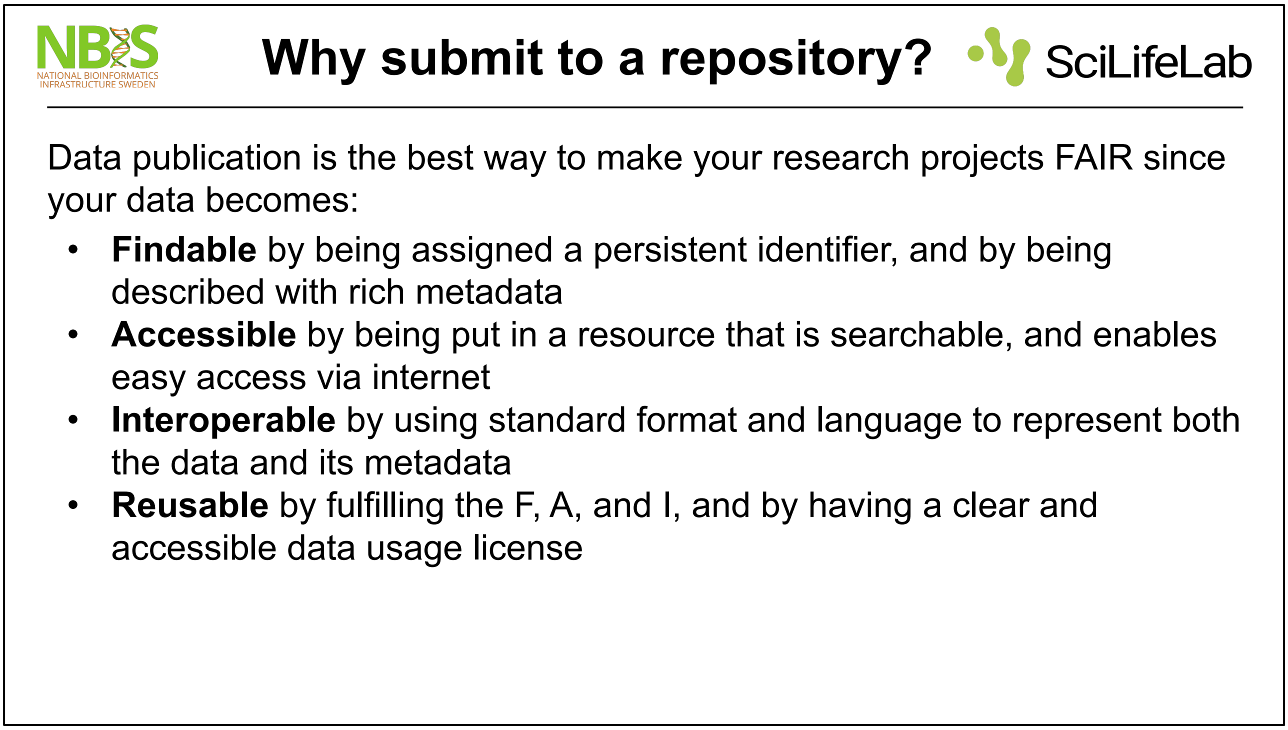
Why submit your datasets to a repository?

Why should you care? You could argue that you do share your data if someone asks for it (aka the (in)famous phrase 'available upon request'). Submitting your data to a repository likely takes a lot of time, time you could spend on doing research, and employers as well as funders are only interested in how many articles you have published, not how many datasets you have published. Well, times are changing, there are many institutions with data policies on sharing, and funders want maximum value for their invested money. Besides, if *you* do not share your data, you cannot ask of others to share either, it's as simple as that.

Some of the reasons in short:

- 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



Why submit to a repository?

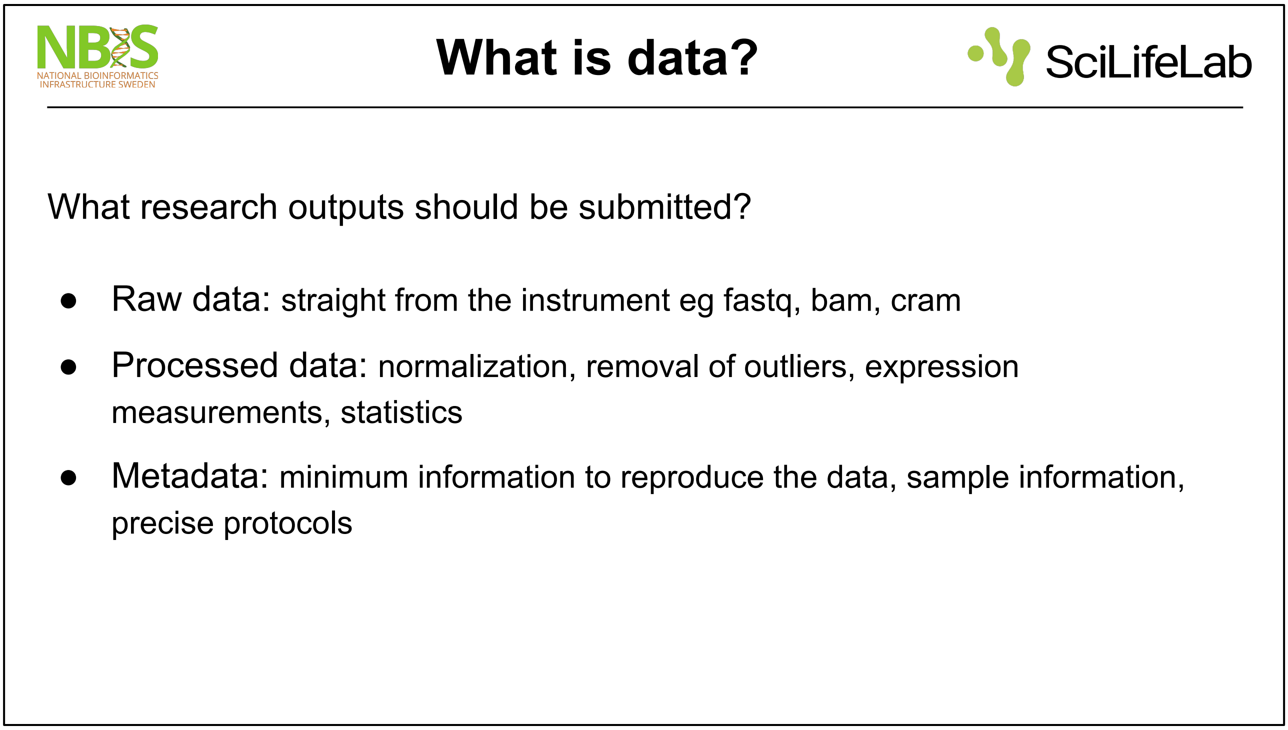
Data publication is the best way to make your research projects FAIR since your data becomes:

- **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

We do repeat FAIR a lot, and here it comes again…Repositories provides the technical solution to FAIR data:

- 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

Hence, by submitting data to a repository, your data becomes FAIR and you do not have to provide a solution on your own.



NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

What is data?

SciLifeLab

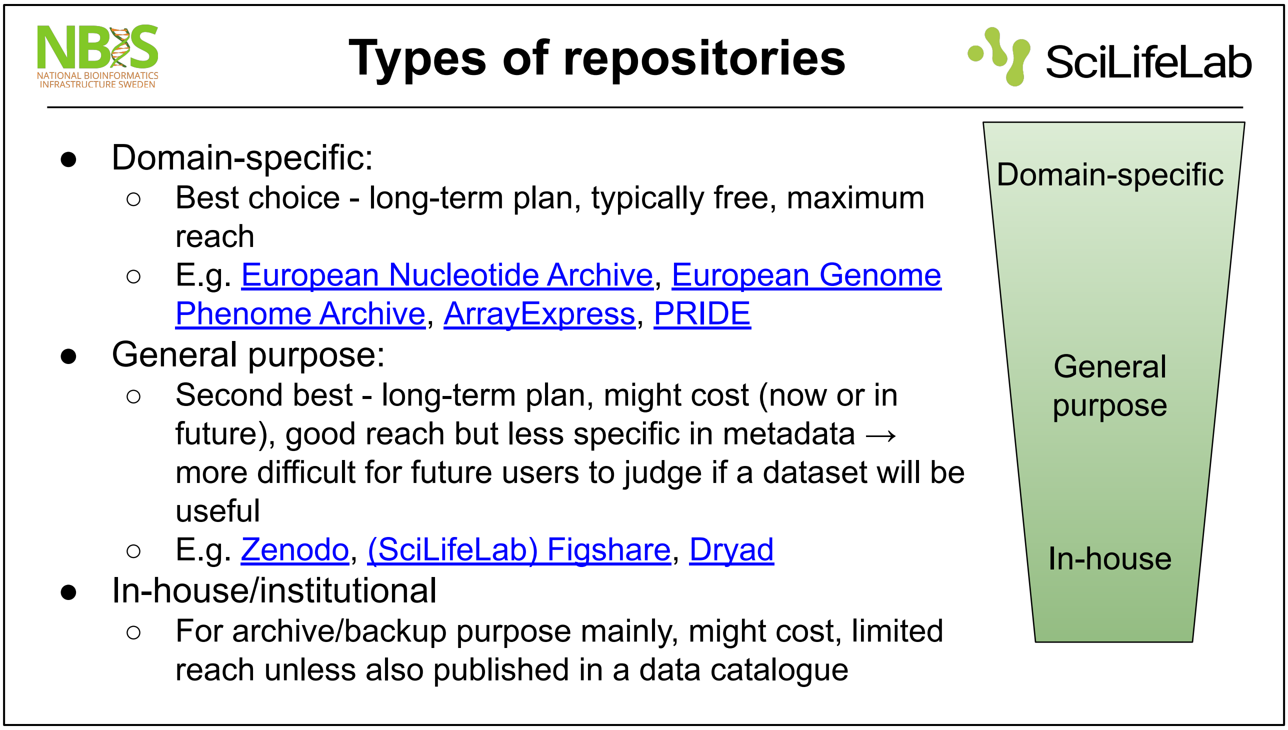
What research outputs should be submitted?

- 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

What research outputs should be submitted?

Apart from the data itself, everything necessary to understand, reproduce and reuse the data should be submitted to a repository:

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



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

Domain-specific

General purpose

In-house

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.

Domain-specific repositories

This type of repository focuses on specific data types and is typically the best choice if you can find one that is suitable for your research data. It will reach your research community, so that others working in your field can find and reuse your data, and incorporates metadata standards in order to make the data as widely useful as possible. The repositories usually have long-term sustainability plan, i.e. they will be available for a long time, and are typically free of charge.

- European Nucleotide Archive (ENA) - for genomic sequence data (non-human)
- European Genome Phenome Archive (EGA) - for human genomic sequence data
- ArrayExpress - for gene expression data
- PRIDE - for proteomics data

General purpose repositories

From time to time you will most likely come across situations when there is no suitable domain-specific repository for your data type, for example if you have a new data type. Another situation you might find yourself in is that you have done registry-based research with human data, and you are not allowed to share this data but still want to publish the methodology openly. Making a metadata record in a general purpose repository will then allow others to easily find it, without violating the agreements with the registry holders.

General purpose repositories usually accepts anything and everything related to research, i.e. they are also useful for other purposes, besides publishing research data, e.g. posters and presentations can be made publicly available and obtain a persistent identifier (DOI).

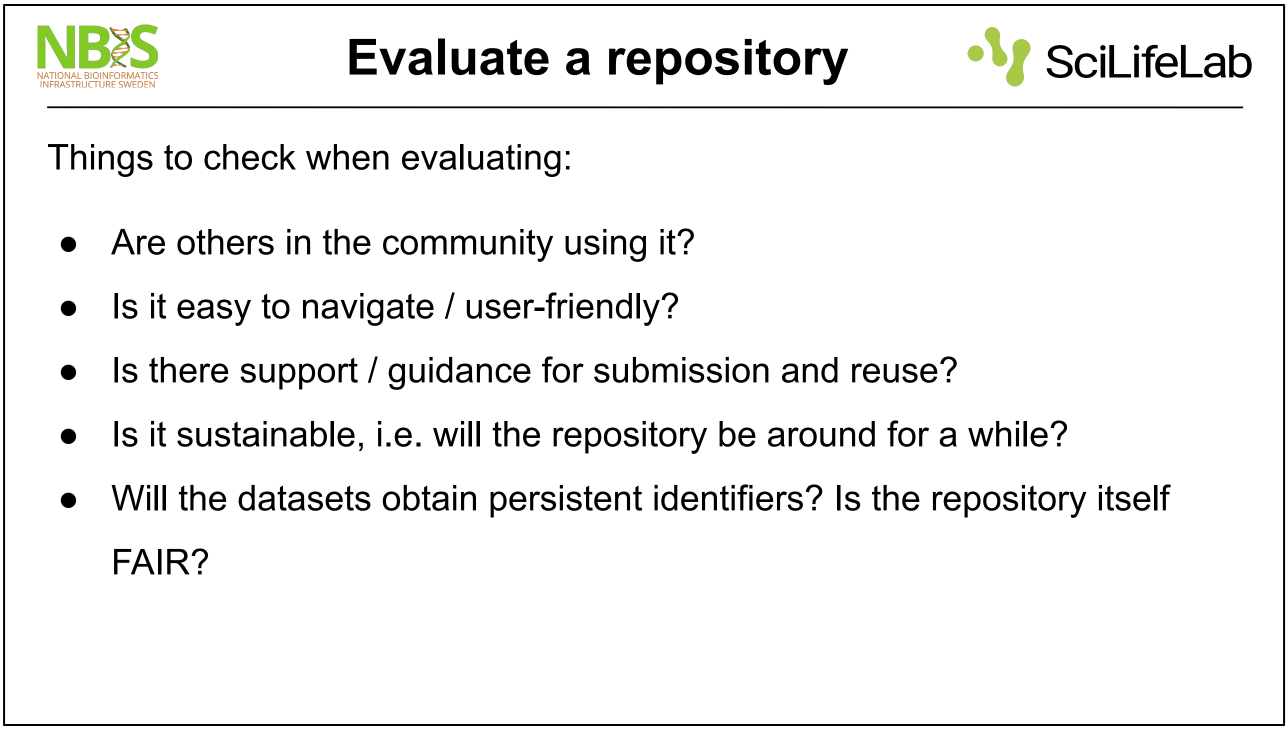
This type of repository is typically indexed, so you can find its content via search engines, and thus it has good reach. However, since the repository accepts many data types, the metadata will be less specific (no or very high level metadata standards), with the result that is more difficult for future users to judge if a dataset will be useful for them. The repositories usually have long-term sustainability plan, i.e. they will be available for a long time, but might cost (now or in future).

- Zenodo
- Figshare
- SciLifeLab Data Repository (Figshare)
- Dryad

In-house/institutional repositories

indexed (so that search engines can find its content) this type of repository is for archive/backup purpose mainly.

You can also choose to create and host an in-house repository yourself, but that put a lot of responsibility on your shoulders. For how long will you be able to sustain it? It also requires considerable labour in order to make the repository FAIR, and without that the repository will have limited reach unless you also publish in a data catalogue.



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Evaluate a repository

SciLifeLab

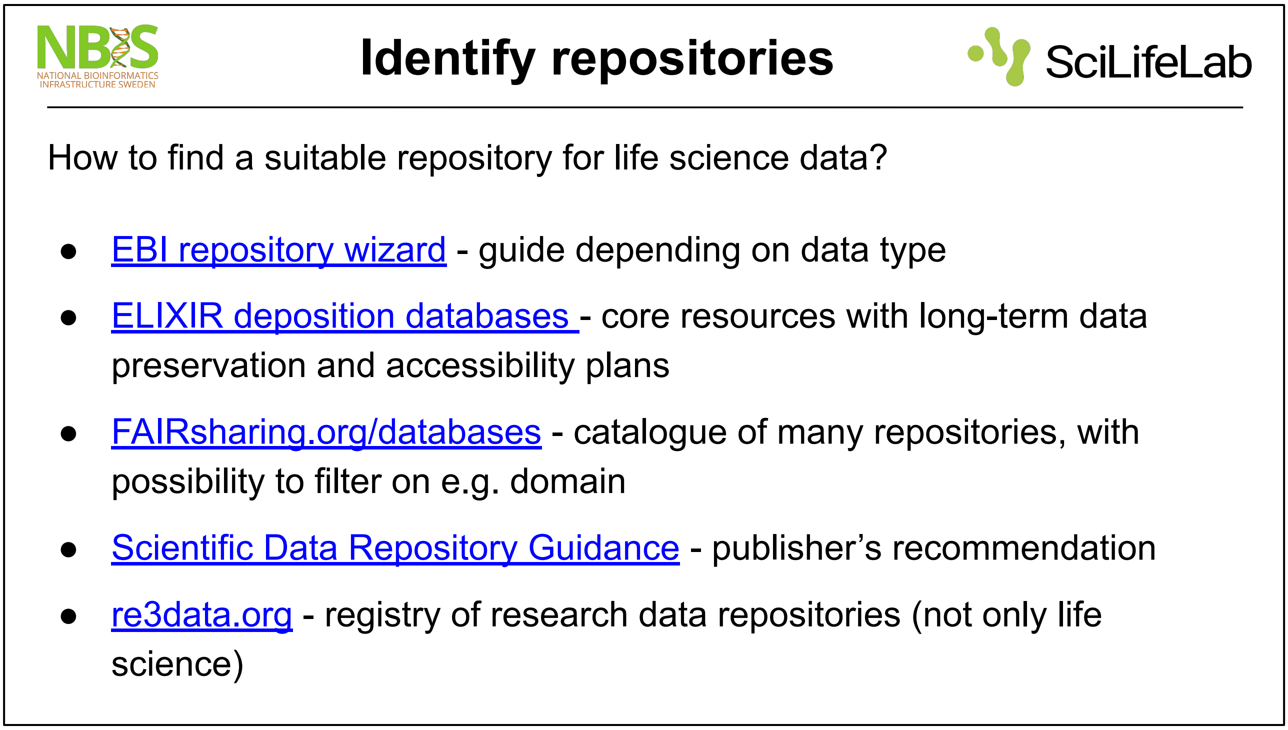
Things to check when evaluating:

- 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?

Evaluate the suitability of a repository

How do you know if a repository is trustworthy? Say that you find a repository that might fit your purposes, how can you evaluate if it is suitable? Apart from accepting your type of data, there are some questions to consider when deciding if a certain repository is suitable or not:

- Are others in the community using it? Explore what datasets are already in it.
- While exploring 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? Is there a long-term plan for financing the repository, is it managed by a trustworthy group?
- Will the datasets obtain persistent identifiers? Is the repository itself FAIR?



Identify repositories

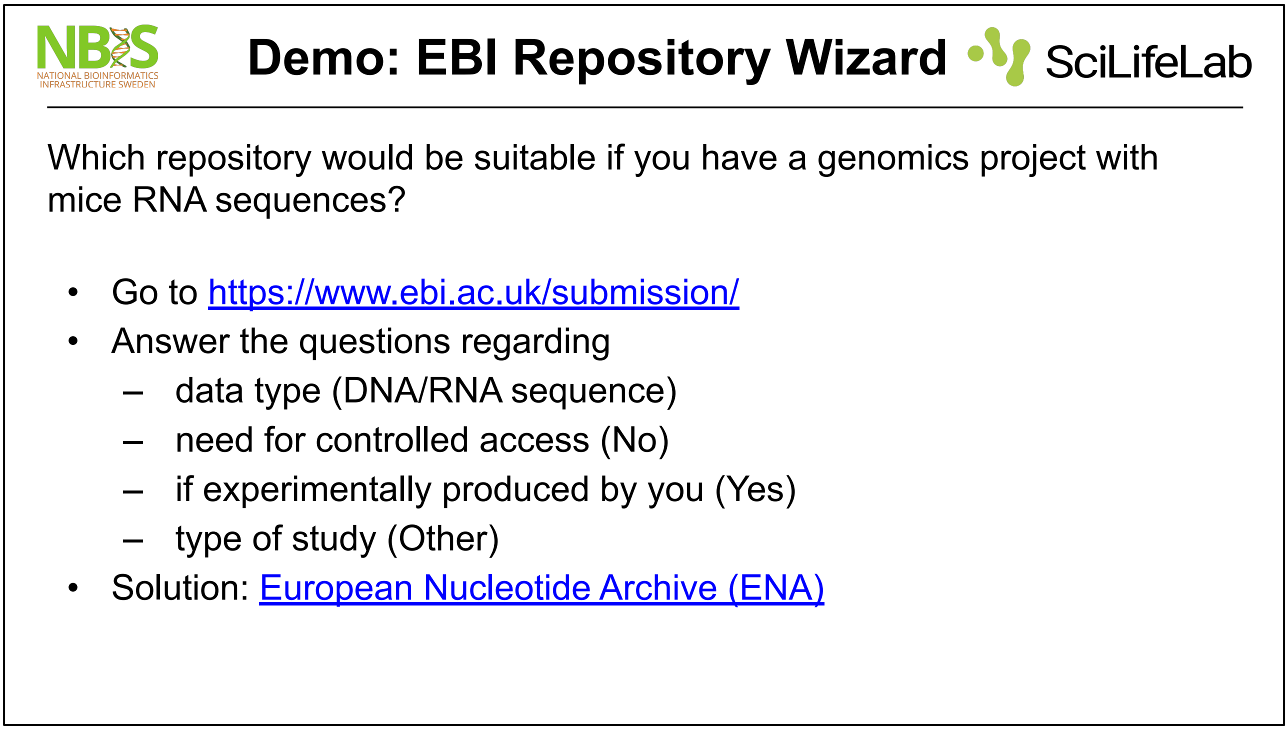
How to find a suitable repository for life science data?

- 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)

How do you find a domain specific repository?

There are several ways to find domain-specific repositories within life science:

- EBI repository wizard - 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 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



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)

EBI hosts several life science repositories, suitable for different types of data. The Repository Wizard helps to identify which one is suitable for your data.

Note: Do this as demonstration

Go to the Wizard and explore the wizard, e.g. which repository would be suitable if you have a genomics project with RNA sequences?

Answer: European Nucleotide Archive (ENA); found via DNA/RNA sequence -> no controlled access -> produced experimentally -> Other

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.