Repository Submissions

Introduction to Data Management Practices course

NBIS DM Team
data@nbis.se

https://nbisweden.github.io/module-open-science-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

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

What is data?

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

Types of repositories

- 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.

Domain specific repositories

How find a domain specific repository?

- EBI 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

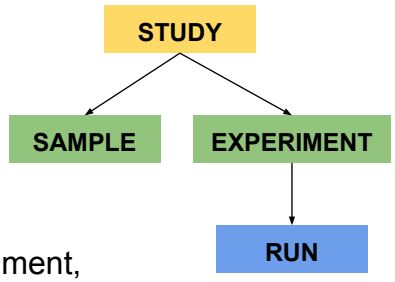
How do you find a domain specific repository?

- 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.

European Nucleotide Archive

ENA repository for (non-human) DNA & RNA sequences

- **Study**: groups together the submitted data
- **Sample**: information about the sequenced source material, provided via a metadata standard (checklist)
- **Experiment**: information about a sequencing experiment, including library and instrument details
- **Run**: data files containing sequence reads



The ENA is a repository providing submission of, and access to, annotated DNA and RNA sequences.

It also stores complementary information such as experimental procedures, details of sequence assembly and other metadata related to sequencing projects.

Submissions are represented using a number of different metadata objects:

- Study: A study (project) groups together data submitted to the archive and controls its release date. A study accession is typically used when citing data submitted to ENA. Note that all associated data and other objects are made public when the study is released.
- Sample: A sample contains information about the sequenced source material. Samples are associated with checklists, which define the fields used to annotate the samples. Samples are always associated with a taxon.
- Experiment: An experiment contains information about a sequencing experiment including library and instrument details.
- Run: A run is part of an experiment and refers to data files containing sequence reads.

Ways to submit to ENA

- Interactive - using browser
- Webin-CLI - command-line submission interface using manifest file
- Programmatic submission - XML document submitted using cURL

Test site: https://wwwdev.ebi.ac.uk/ena/submit/webin/
Production site: https://www.ebi.ac.uk/ena/submit/webin/

Note: Test first when doing new submission, but it is restarted nightly ⇒ submissions will be gone next day

There are three ways to submit:

- Interactive using a web browser. This is the way we will do it in the following excercise
- Via command-line using Webin-CLI. Used for the read data only, i.e. Experiment and Run objects. Requires a manifest file to be created. Convenient when there are many sequence files to be submitted.
- Programmatical submission, via creation of XML file which is then submitted to the repository using cURL command. Very convenient when you do submissions on a regular basis.

Use the test submission site when you want to test, and the production site for real submissions:

- Test site: https://wwwdev.ebi.ac.uk/ena/submit/webin
- Production site: https://www.ebi.ac.uk/ena/submit/webin

Note: It is always good to use the test site first when doing a new submission. However, the test service is restarted every night, any submissions made to the test service will be removed by the following day. Hence, do not start a test submission one day, and expect to continue the next day.

Key Points

- There are different types of data e.g. raw, processed and metadata.
- 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.
- In ENA, submissions are represented using a number of different metadata objects: Study, sample and raw reads.
- Submissions can be done via browser, command-line interface or programmatically.