

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>



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



Digitalbevaring.dk

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

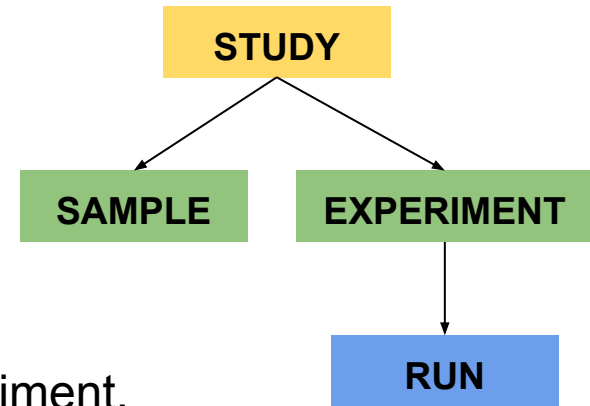
- Domain specific:
 - Best choice - long-term plan, typically free, maximum reach.
 - E.g. ENA, ArrayExpress, PRIDE
- General purpose:
 - Second best - long-term plan, might cost (now or in future), good reach but less specific in metadata → more difficult for future users to judge if a dataset will 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

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

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



- [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 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.