

Repository Submissions

NBIS DM Team
data-management@scilifelab.se

<https://doi.org/10.17044/scilifelab.c.6820587>



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



Digitalbevaring.dk

What data should be submitted?

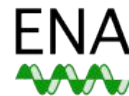
- “Raw” data: straight from the instrument
 - e.g. fastq, bam, cram
- Processed data: derived from raw data
 - e.g. genome assemblies, genome annotations, sequence variation data, expression measurements, etc
- Metadata: minimum information to reproduce the data
 - e.g. sample information, precise protocols, instruments, etc

- Domain specific - Best choice
 - domain-specific metadata standards
 - maximum reach
 - long-term sustainability plan
 - typically free
 - E.g. ENA/EGA, ArrayExpress, PRIDE
- General purpose - Second best
 - (Often) long-term sustainability 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](#) - Nucleotide sequence data
 - INSDC partner (with NCBI, DDBJ)
- [EGA](#) - *Controlled access* nucleotide sequence data
 - [FECA-SE](#) - Swedish node
- [ArrayExpress](#) - RNAseq and microarray data
- [EVA](#) - genetic variation data
- [DGVa](#): structural genetic variation data
- [BioStudies](#) - any type of biological studies / collate different data types
- [BioSamples](#) - links to associated experimental data (ENA, ArrayExpress,...)



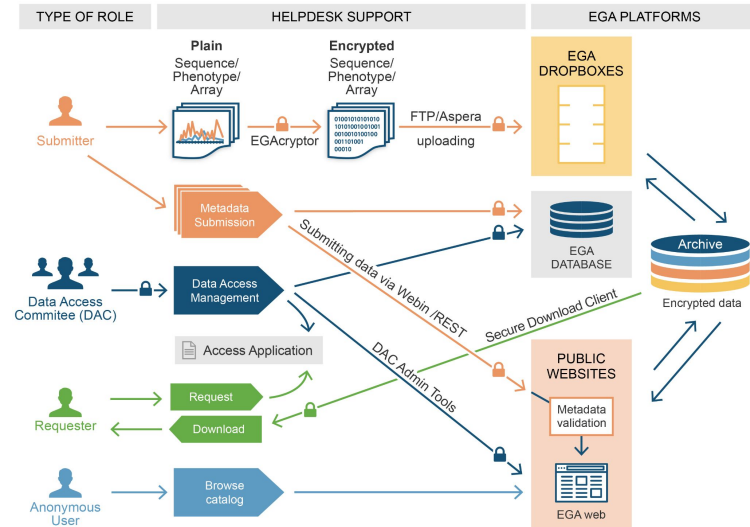
EMBL-EBI



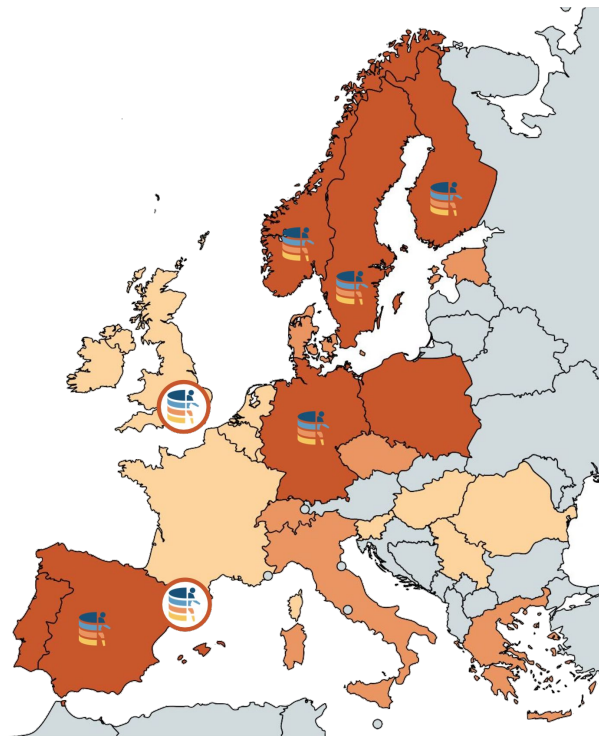
<https://www.ebi.ac.uk/>

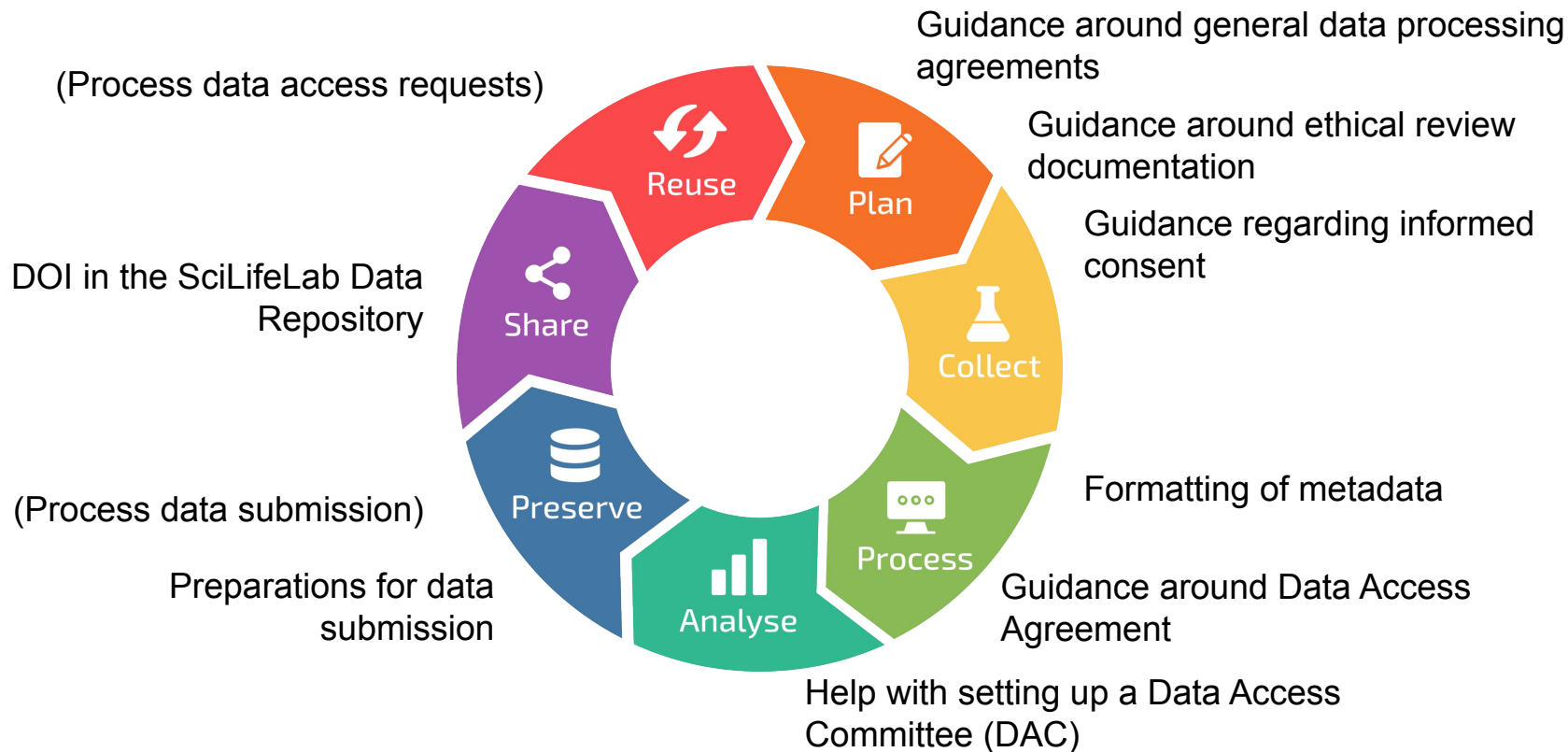
 **BioStudies.**

- The [European Genome-Phenome Archive](#) (EGA) is a repository for archiving and sharing sensitive personal data from biomedical research projects with a standardized application process involving a Data Access Committee.



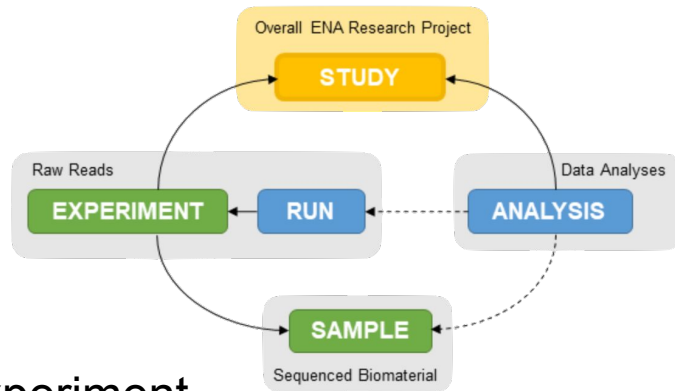
- FEAGA Sweden is the Swedish node of the Federated EGA where the public metadata is accessible via EGA but the sensitive data is stored in Sweden.
- FEAGA Sweden is expected to become operational in September 2023
- We are currently working with two pilots: Swedish Childhood Tumor Biobank and Human Developmental Cell Atlas
- Any data submitted to the archive is subject to controlled access
- FEAGA Sweden is hosted by the National Bioinformatics Infrastructure Sweden (NBIS)

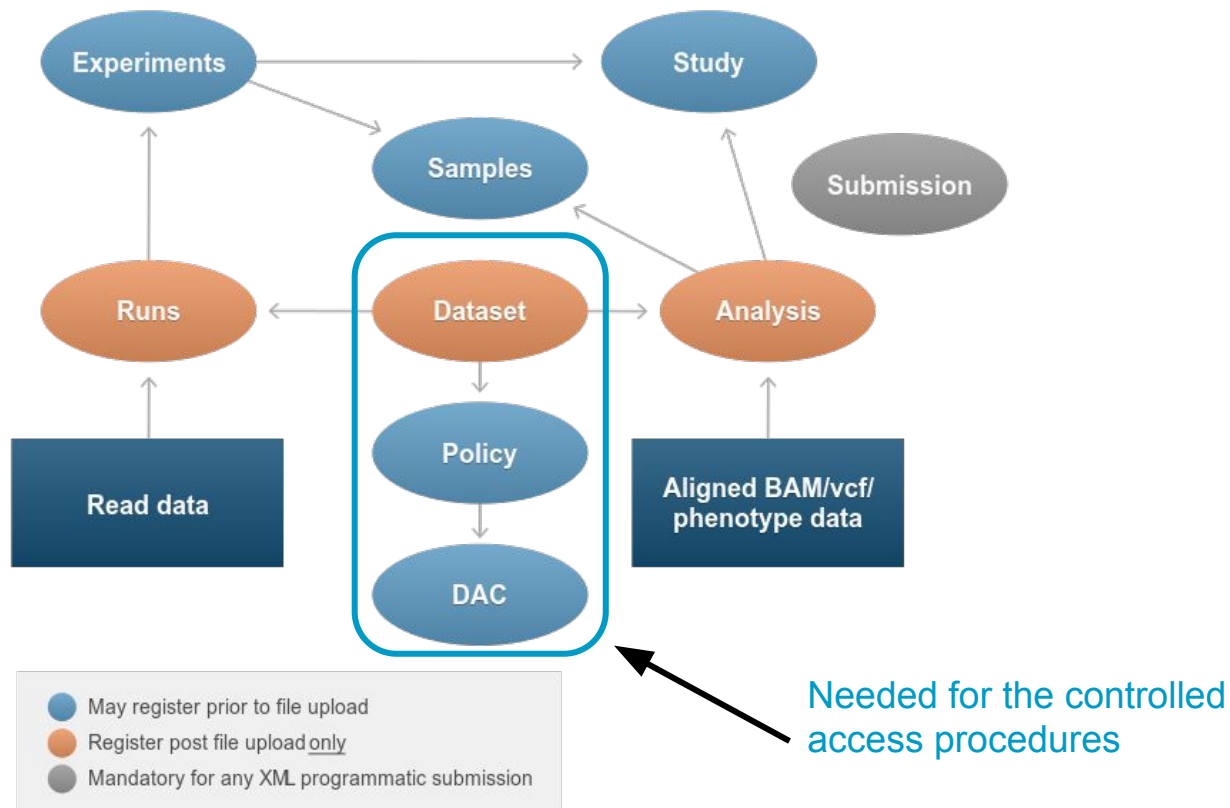




Metadata model

- **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
- **Analysis**: seq assemblies, seq annotations, targeted sequences, reference alignments, PACBio methylation, etc





- [Interactive](#) - using browser using tsv files
 - *Data files need to be uploaded separately by ftp or Aspera*
- [Webin-CLI](#) - command-line submission interface using manifest file
 - *Webin-CLI does data file uploads*
- [Programmatic submission](#) - XML document submitted using cURL
 - *Data files need to be uploaded separately by ftp or Aspera*

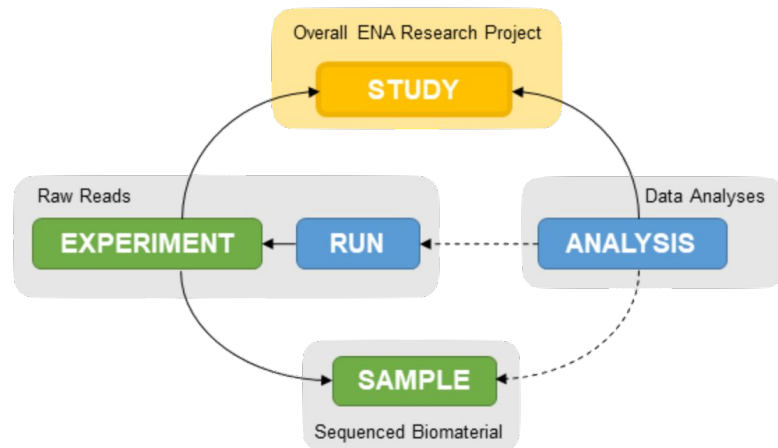
Test site: <https://wwwdev.ebi.ac.uk/ena/submit/webin/>

Production site: <https://www.ebi.ac.uk/ena/submit/webin/>

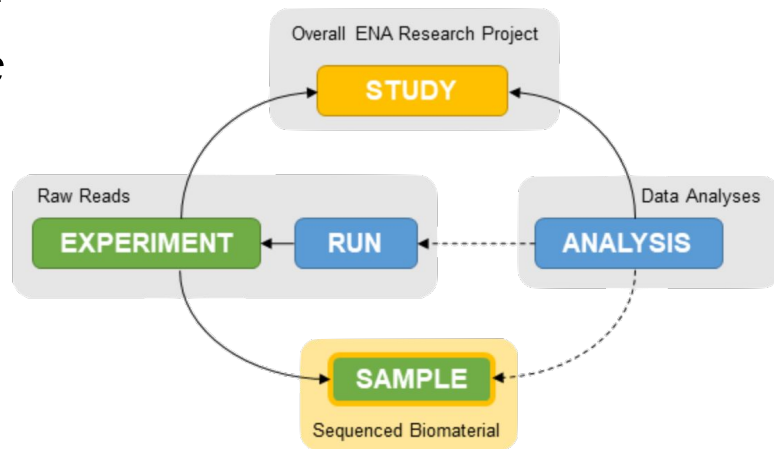
Note: Test first when doing new submission, but...

it is restarted nightly ⇒ submissions will be gone next day

- **Release date**
 - Governs release data for all associated data - Embargo time possible
- Name/Alias
- Title (short descriptive)
- Study abstract
- *Locus Tag Prefixes (only for genome annotations)*
- PubMed Citations - *optional*
- Study Attributes - *optional*
 - Tag - Value pairs
- Interactive & Programmatic submission
- *Researcher needs to provide information*



- Name/Alias
- Title (short descriptive)
- Taxonomic classification
 - Taxon ID (NCBI Taxonomy database)
 - Scientific name - *optional/automatic*
 - Common name - *optional/automatic*
- Sample attributes - *recommended*
 - Tag - Value pairs
 - **Checklists** !
- Interactive & Programmatic submission
- *Researcher needs to provide information*



Checklist: ERC000033



ENA virus pathogen reporting standard checklist

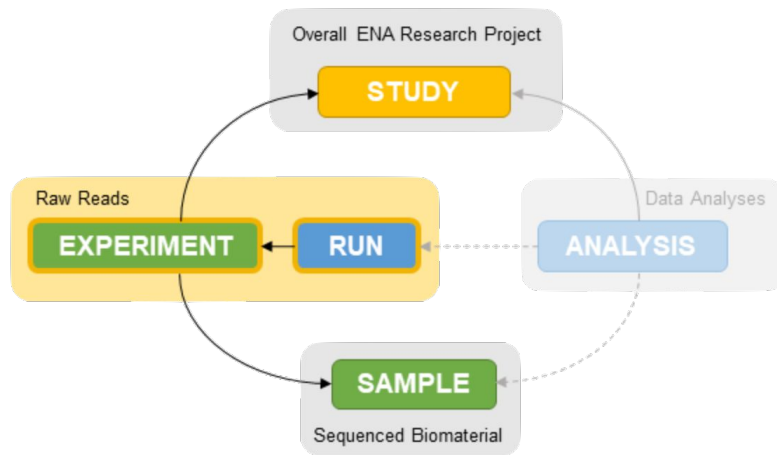
Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

Checklist Fields					
Filter fields...					
Filter by type:					
Human surveillance data					
Collection event information					
sample collection					
host disorder					
host description					
Virus isolate information					
General collection event information					
Serology detection					
Intraspecies					
Field Name	Field Format	(Field Restriction)	Requirement	(Units)	
subject exposure	free text		optional		
subject exposure duration	free text		optional		
type exposure	free text		optional		
personal protective equipment	free text		optional		
hospitalisation	text choice	options	optional		
illness duration	free text		optional		
illness symptoms	free text		optional		
collection date	restricted text	regular expression	recommended		
geographic location (country and/or sea)	text choice	options	mandatory		

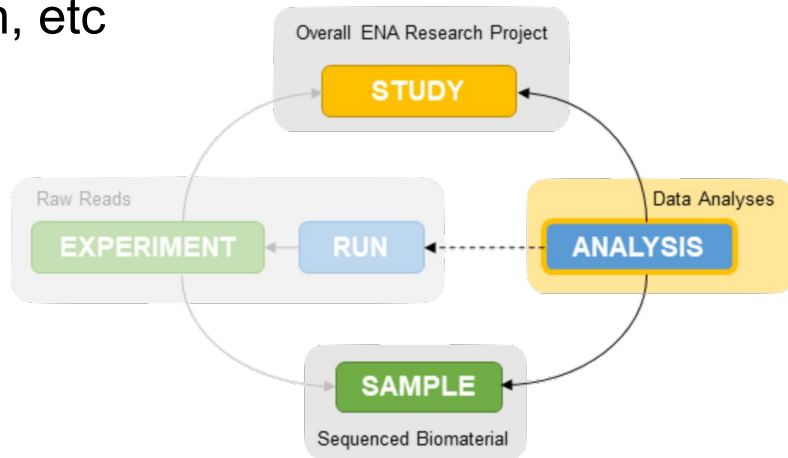
- A headache for the researchers!
- Currently not that FAIR

<https://www.ebi.ac.uk/ena/browser/view/ERC000033>

- STUDY: Study accession or unique name (alias)
- SAMPLE: Sample accession or unique name (alias)
- NAME: Unique experiment name
- PLATFORM: [See permitted values](#). Not needed if INSTRUMENT is provided.
- INSTRUMENT: [See permitted values](#)
- INSERT_SIZE: Insert size for paired reads
- LIBRARY_NAME: Library name (*optional*)
- LIBRARY_SOURCE: [See permitted values](#)
- LIBRARY_SELECTION: [See permitted values](#)
- LIBRARY_STRATEGY: [See permitted values](#)
- DESCRIPTION: free text library description (optional)
- Read data files
 - Formats: single/paired/multi Fastq, BAM, CRAM
- File checksums: md5
- Interactive, Webin-CLI & Programmatic submission
- ***Sequencing facility needs to provide information***



- *Now it's getting messy...*
- Seq assemblies, Seq annotations, targeted sequences, reference alignments, PACBio methylation, etc
- Different for the different types
- Refer to raw reads



- Webin-CLI & Programmatic submission
- *Researcher, Bioinformatics facility(?), and Sequencing facility(?) need to provide information*

- Fill in the **Experiment and Run metadata** for a SNP/SEQ data delivery
- [Data delivery](#)
- Experiment and Run [metadata template](#)

▼ 220922_A00605_0493_AHHTFLDRX2

220922_A00605_0493_AHHTFLDRX2_FU99-2022_multiqc_report_data.zip

220922_A00605_0493_AHHTFLDRX2_FU99-2022_multiqc_report.html

checksums.md5

▼ Sample_FU99-2022-GM12878-PCRfree-1

FU99-2022-GM12878-PCRfree-1_S1_L001_R1_001.fastq.gz

FU99-2022-GM12878-PCRfree-1_S1_L001_R2_001.fastq.gz

FU99-2022-GM12878-PCRfree-1_S1_L002_R1_001.fastq.gz

FU99-2022-GM12878-PCRfree-1_S1_L002_R2_001.fastq.gz

▼ Sample_FU99-2022-GM12878-PCRfree-2

FU99-2022-GM12878-PCRfree-2_S2_L001_R1_001.fastq.gz

FU99-2022-GM12878-PCRfree-2_S2_L001_R2_001.fastq.gz

FU99-2022-GM12878-PCRfree-2_S2_L002_R1_001.fastq.gz

FU99-2022-GM12878-PCRfree-2_S2_L002_R2_001.fastq.gz

SampleSheet.csv

	A	B	C	D	E	F	G	H
1	FileType		Read submission file type					
2	study	sample	design_description	library_name	library_strategy	library_source	library_selection	library_layout
3	Accession number of the study (PR, automatic)	Accession number of the sample (S, automatic)	Goal and setup of the individual library (optional)	The submitter's name for this library. (mandatory)	Sequencing technique intended (mandatory (text choice))	The LIBRARY_SOURCE specifies the source of the library (mandatory (text choice))	Method used to create the library (mandatory (text choice))	Library layout (mandatory (text choice))
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								