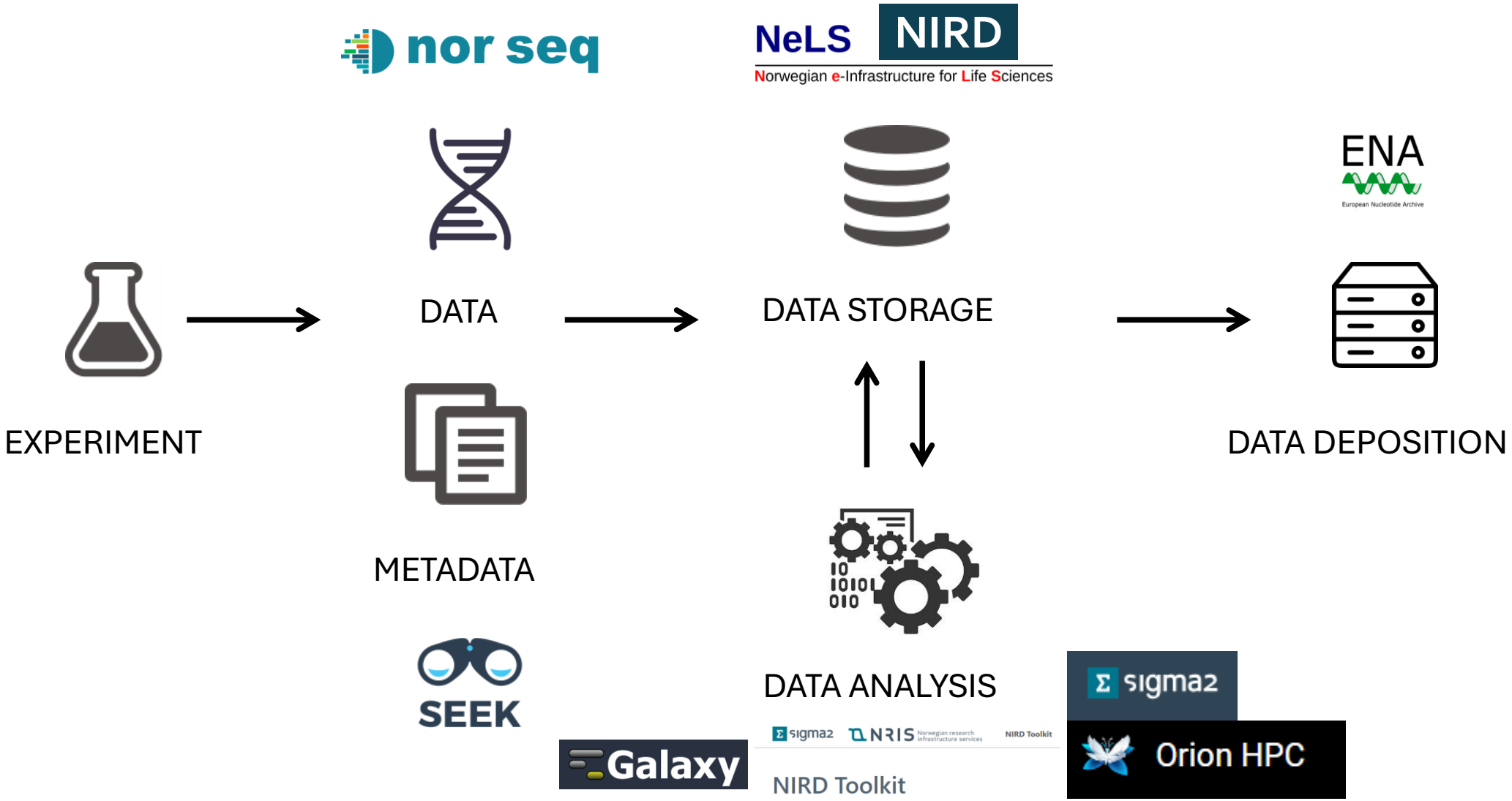
Data deposition to public archives
The ENA example

Arturo Vera-Ponce de Leon
ELIXIR Norway

Example - Data flow/handle using ELIXIR Norway



European Nucleotide Archive

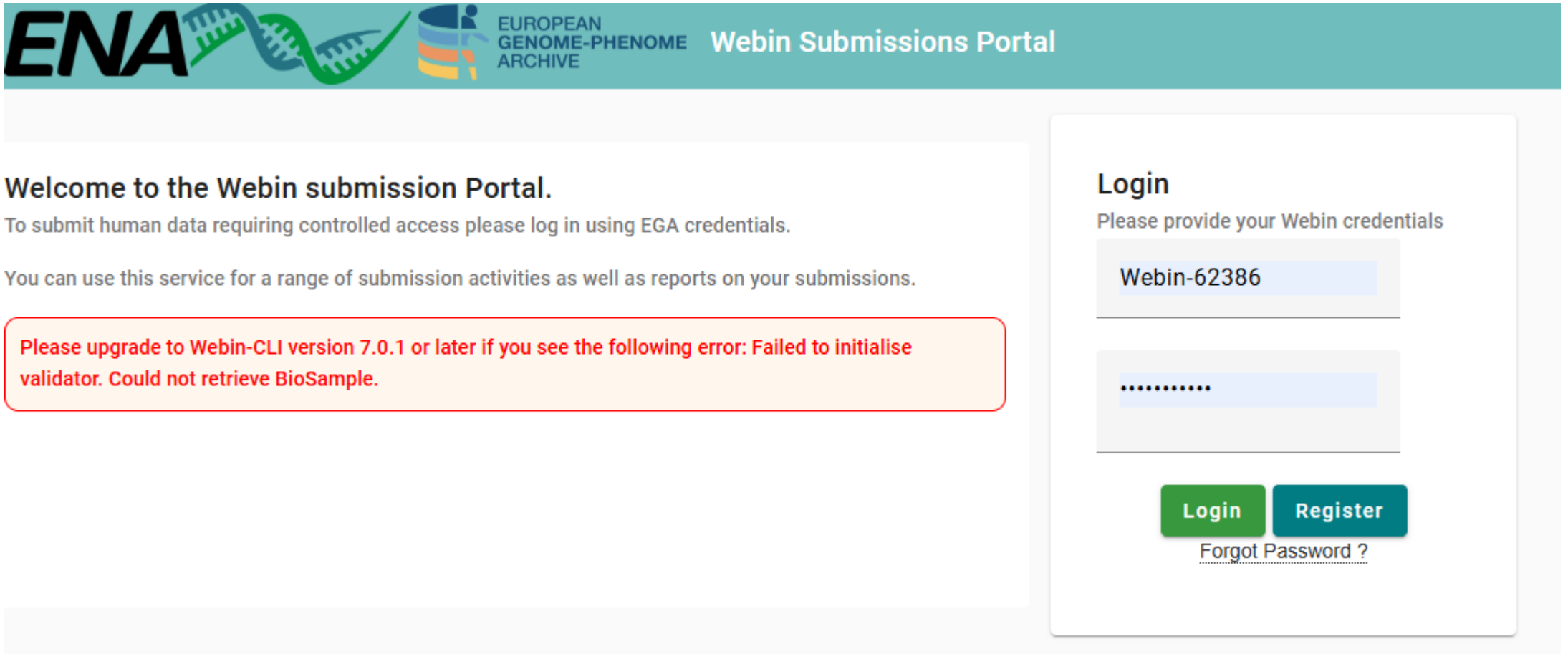
The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

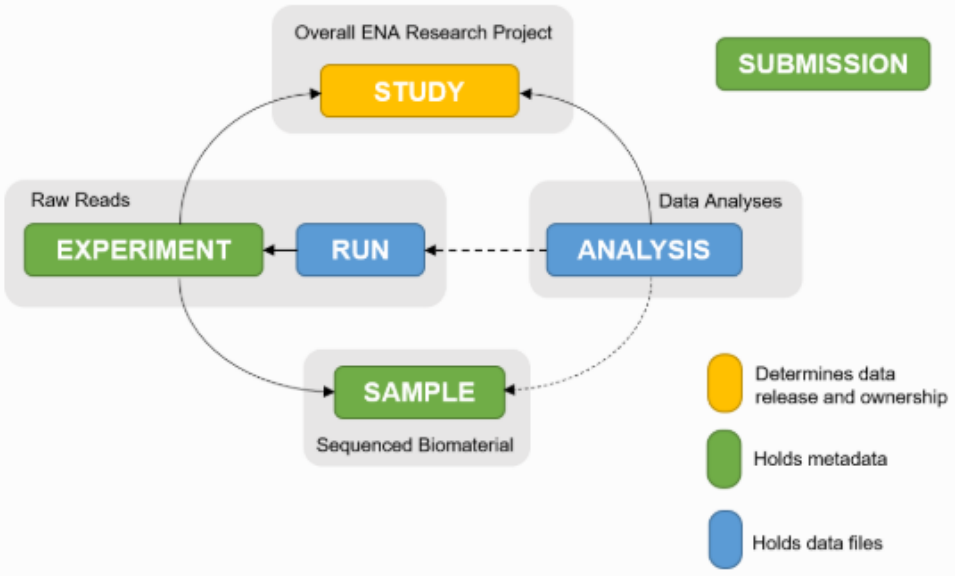
Submit | Search | Rulespace | Support

How to transfer data from NeLS, NIRD or Orion filesystem to ENA?

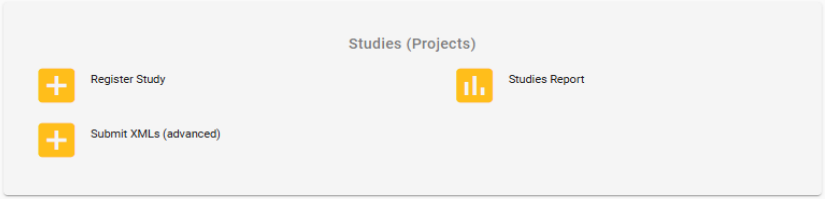
Use the Webin GUI Portal

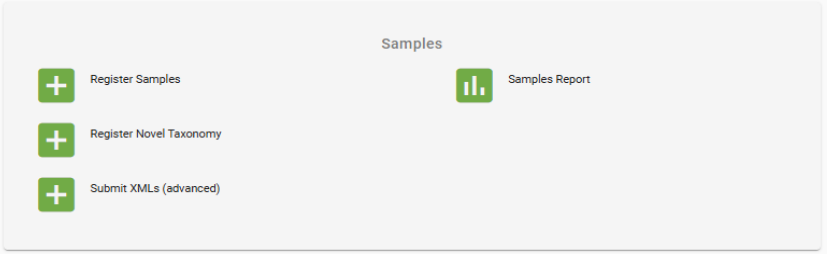


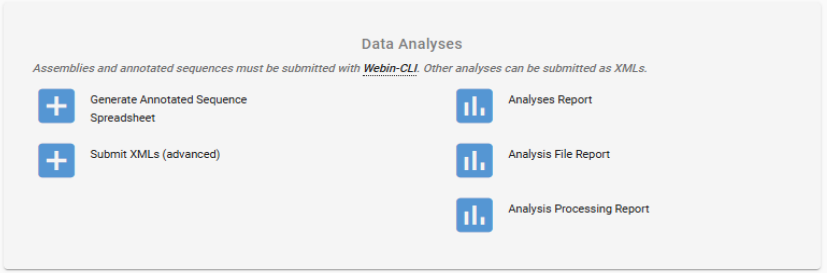
The ENA Metadata Model



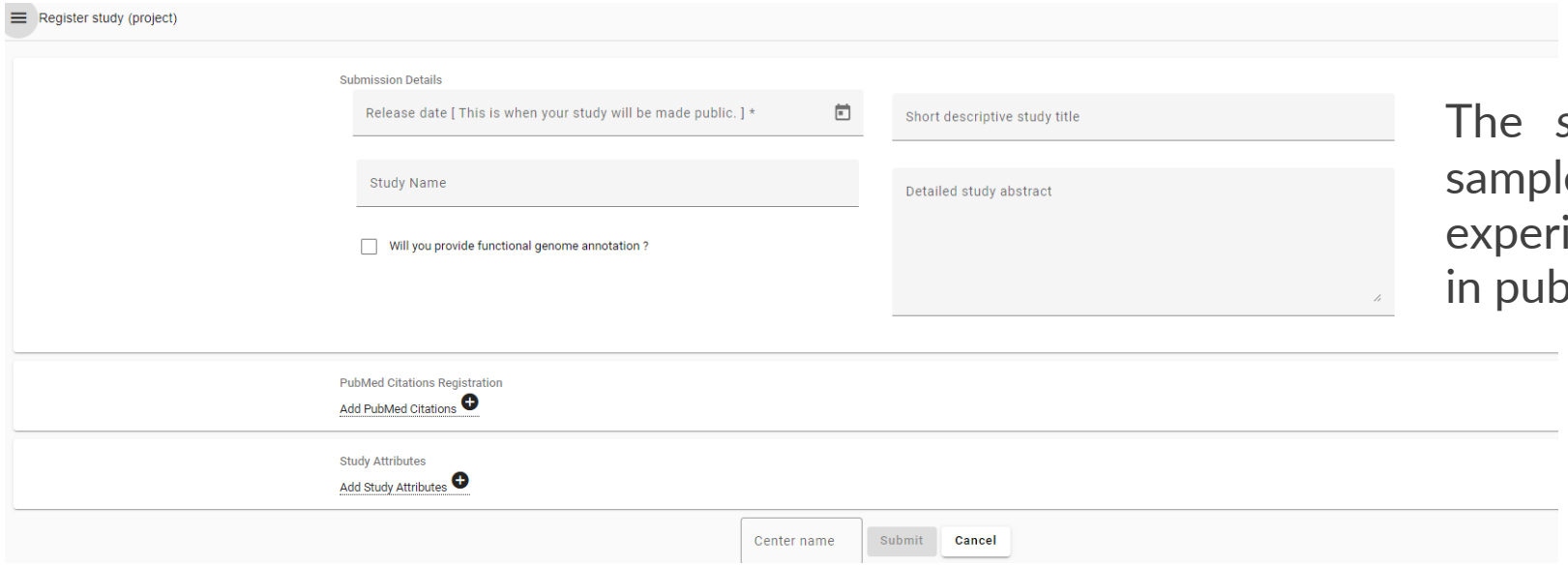
- **Study**: A study (project) groups together data submitted to the archive and controls its release date. A study accession is typically used when citing data submitted to ENA. Note that all associated data and other objects are made public when the study is released.
- **Sample**: A sample contains 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.
- **Analysis**: An analysis contains secondary analysis results derived from sequence reads (e.g. a genome assembly).
- **Submission**: A submission contains submission actions to be performed by the archive. A submission can add more objects to the archive, update already submitted objects or make objects publicly available.



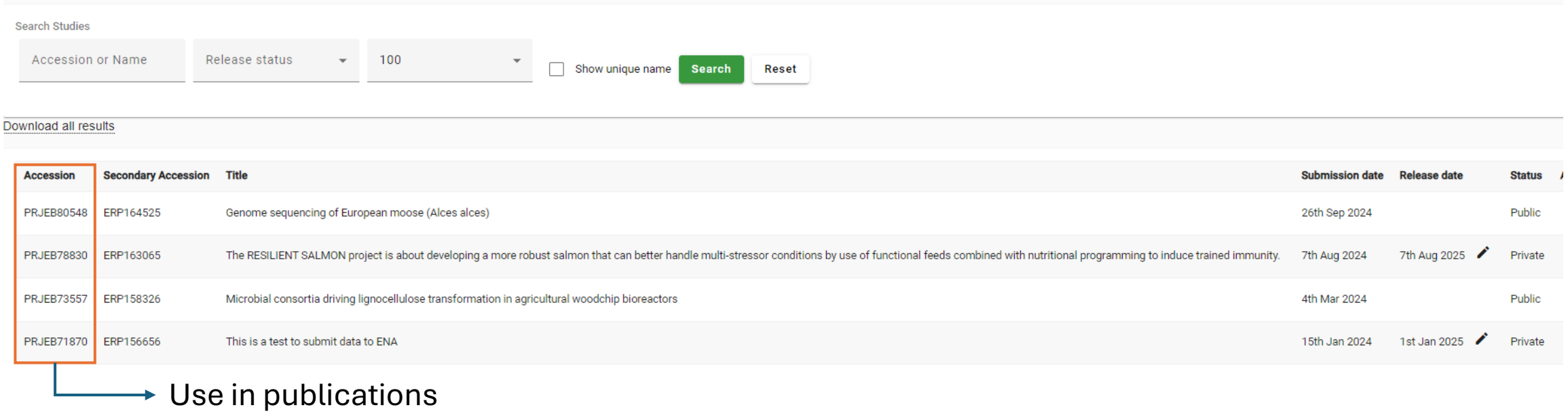




Registering study



The *study (project)* object is linked to samples and sequence reads via experiments, and is typically what you cite in publications.



Accession	Secondary Accession	Title	Submission date	Release date	Status
PRJEB80548	ERP164525	Genome sequencing of European moose (Alces alces)	26th Sep 2024		Public
PRJEB78830	ERP163065	The RESILIENT SALMON project is about developing a more robust salmon that can better handle multi-stressor conditions by use of functional feeds combined with nutritional programming to induce trained immunity.	7th Aug 2024	7th Aug 2025	Private
PRJEB73557	ERP158326	Microbial consortia driving lignocellulose transformation in agricultural woodchip bioreactors	4th Mar 2024		Public
PRJEB71870	ERP156656	This is a test to submit data to ENA	15th Jan 2024	1st Jan 2025	Private

Use in publications

Registering samples

Samples are the source material from which your sequences derive, and the searchability and usability of your submitted data will depend on how well you document these samples

Sample Checklists

There is a minimum amount of information required during ENA sample registration and all samples must conform to a defined checklist of expected metadata values. The most suitable checklist for sample registration depends on the type of the sample.

Filter by accession/name/description/field name

Sampled environment	Recommended checklist
Air or general, above-ground, terrestrial	GSC MixS air
Epi- or endophytic (e.g. leaf, root)	GSC MIxS plant associated
Epi- or endozoic (e.g. spider gut, animal skin)	GSC MIxS host associated
Fresh- or seawater	GSC MixS water
Human gut / oral / skin / vaginal	GSC MIxS human gut / oral / skin / vaginal
Human non- gut / oral / skin / vaginal	GSC MIxS human associated
Sediment	GSC MixS sediment
Soil	GSC MixS soil

Select and fill the Checklist via spreadsheet

(Only one sample even for pair-end sequencing is a single sample)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Checklist	ERC000013	GSC MIxS host associated															
2	tax_id	scientific_name	sample_a	sample_ti	sample_d	project na	target gen	collection	geographi	geographi	geographi	broad-scale environmental context	local environmental context	environm	host common name	host taxid	pcr primers	target gene
3	#units									DD	DD							
4	256318	environmental sequence	1A_Ctr_gai	1A	metabarc	Resilient s	16S rRNA	12/1/2020	Norway	59.66934	10.75799	fecal material ENVO_00002003	digestive tract environment ENVO_01001033	feces	Salmo salar	8030	[515FB] GTGY	abV4-C
5	256318	environmental sequence	2A_Ctr_gai	2A	metabarc	Resilient s	16S rRNA	12/1/2020	Norway	59.66934	10.75799	fecal material ENVO_00002003	digestive tract environment ENVO_01001033	feces	Salmo salar	8030	[515FB] GTGY	abV4-C

Submit and check the result

Accession	BioSample	Title	Organism	Tax id	Submission date	Status	Action
ERS20855763	SAMEA115911915	1A	metagenome	256318	8th Aug 2024	Private	

Prepare your data to be submitted

NeLS NIRD
Norwegian e-Infrastructure for Life Sciences

DATA STORAGE

ENA
European Nucleotide Archive

DATA DEPOSITION

▸ Check you have the compressed fastq files (.fq.gz) and all of them are in the same directory:

```
cd $LABFILES/Orion101-2022/TransferToENA
ls -lrth
total 68M
-rw-r--r-- 1 auve root 36M Mar 15  2022 1A.R2.fq.gz
-rw-r--r-- 1 auve root 30M Mar 15  2022 1A.R1.fq.gz
```

> If your files are not compressed you can compress using pigz command as follow:

```
pigz -p 4 *.fq
```

▸ To enable verification of file integrity after upload, calculate the md5 checksum hash of each (compressed) read file:

```
for f in *.gz; do md5sum $f ;done|awk '{print $2"\t"$1}' > MD5file.tsv
```

This file will be used later and should look something like:

```
1A.R1.fq.gz     7a36fd9ff513b987d254068482905f9e
1A.R2.fq.gz     eb31cf4683b0ef49e0488bc7dea58686
```

https://orion.nmbu.no/en/SubmitReadsToENA

Submit your data using the CLI

Uploading Files Using Command Line FTP Client

```
# Connect to FTP server [replace X:s, and provide password when prompted]
lftp webin2.ebi.ac.uk -u Webin-XXXXX
# Expected response: lftp Webin-XXXXX@webin2.ebi.ac.uk:~>

# Transfer your read files
mput ~/your-read-file-dir/*.fastq.gz
# Expected response: ... Total x files transferred

# Disconnect from server
bye
```

Using Aspera ascp Command Line Program (Orion Users)

▸ Submit the files

```
ascp -QT -l300M -L- *fq.gz Webin-XXXX@webin.ebi.ac.uk:.
```

This will login to ENA server asking for your password:

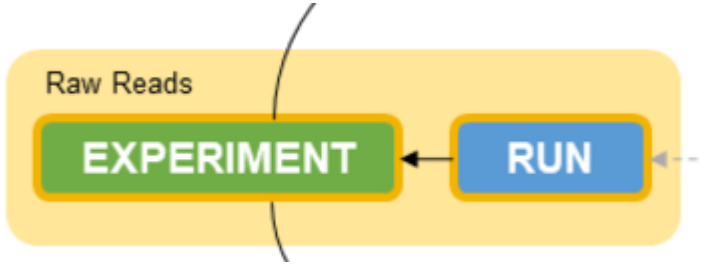
```
LOG Configuration: using v2 configuration file "/net/fs-2/scale/OrionStore/Orion/cluster/software/Aspera-CLI/3.9
LOG Initializing FASP version 3.9.1.168954, license max rate=(unlimited), account no.=1, license no.=2000 produc
LOG Configured symlink actions: create=1, follow=1, follow_wide=0, skip=0
LOG IBM Aspera CLI version 3.9.6.1467.159c5b1

LOG Alternate log directory: "-"
LOG (access key) Not present
LOG [asssh] remote host-key fingerprint f8cb00a109311397d041a88e502f3bb9f3288ec5
Password:
```

The Webin-XXXX is the Webin-Username you created in the Webin Portal. Remember to use the same userID and Password

https://orion.nmbu.no/en/SubmitReadsToENA

Prepare the Read submission spreadsheet template



In the ENA metadata model, an experiment refers to a sequencing event, and contains information on e.g. library construction and instruments, whereas runs represent the read files resulting from an experiment.

	A	B	C	D	E	F	G	H	I	J	K	L
1	FileType	fastq	Read submission file type									
2	sample	study	instrument_model	library_na	library_source	library_se	library_str	library_lay	forward_file_name	forward_file_md5	reverse_file_name	reverse_file_md5
3	ERS20855763	ERP163065	Illumina MiSeq	1A	METAGENOMIC	PCR	AMPLICON	PAIRED	1A.R1.fq.gz	7a36fd9ff513b987d254068482905f9e	1A.R2.fq.gz	eb31cf4683b0ef49e0488bc7dea58686
4	ERS20855764	ERP163065	Illumina MiSeq	2A	METAGENOMIC	PCR	AMPLICON	PAIRED	2A.R1.fq.gz	59dd3ff768b99c69b5d8f1586ed69df4	2A.R2.fq.gz	622e800bb3364a3f5411b75e3468ef61

Search Runs

ERR13509704 | Release status | 100 | Show unique name | Search | Reset

Download all results

Accession	Instrument	Study	Sample	Experiment	Submission date	Status	Action
ERR13509704	Illumina MiSeq	ERP163065	ERS20855763	ERX12879620	15th Aug 2024	Private	

Accession	File name	File format	File size	MD5 checksum	Archive status	Action
ERR13509704	1A.R1.fq.gz	FASTQ	31344276	7a36fd9ff513b987d254068482905f9e	File archived	
ERR13509704	1A.R2.fq.gz	FASTQ	37357088	eb31cf4683b0ef49e0488bc7dea58686	File archived	