Data (including metadata) Collection

Speakers: Korbinian Bösl (UiB),
Siri Kallhovd (UiB),
Erik Hjerde (UiT)
Federico Bianchini (UiO)
Moderator: Michal Torma / Rukaya Johaadein (UiO)

14 June 2022

FAIR Data Management in Life Sciences

Learning Objectives

In this talk, we will learn:

- differences in data capture methods
- necessary considerations before you start collecting data
- to navigate different resources for more information

Session Take-Away

After completing this session, you will be:

- able to understand considerations related to data capture
- able to identify storage infrastructures for your project
- understand the concepts underlying metadata schemas

General considerations

Data transfer
Data security
Data safety and consistency
Standardization

Pre-existing data?

Will you use already existing data?

- This might include reference data
- Can you already access the data? - Will it stay available?
- What are the conditions & limitations ?

-> licensing & reuse session

https://rdmkit.elixir-europe.org/existing_data

Data transfer

How to get your data to your storage/workspace?

Is the transfer protocol fast and secure enough for our needs?

Will data come in bulk or 1 by 1?

https://rdmkit.elixir-europe.org/data_transfer.html#solutions
https://en.wikipedia.org/wiki/MD5
https://en.wikipedia.org/wiki/SHA-2

Data security

Is the environment secure enough for our needs?

Does your institution have explicit requirements for this kind of data?

(-> specialized platforms)

Who needs access to the data (and how to prevent access by others)?

Is remote access necessary/not allowed?

Do we have to take active steps to increase security beyond the institute standard?

Secure platforms

Special category data has to be stored in special environments:

- SAFE (UiB)
- HUNT (NTNU)
- TSD (UiO/national)

..

https://rdmkit.elixir-europe.org/no_resources#institutional-policies-on-research-data

Data Classification Icons from UiO

Data safety & consistency

Can measurements/capture be repeated?

How will you backup|snapshot the data?

What will be the consequence of data loss?

How to ensure integrity of the data across transfer and storage? (checksums!)

Can the data be ‘read-only’ in the workspace?

Data standardization

What is the format of the data?

Is there and open format?

Do you have convert the data format? -> How to make this consistent?

(Standard metadata -> later in this session)

Folder structure

Organize your data files & folders in a structured manner
Be consistent!

```
project/
  code/                code needed to go from input files to final results
  data/                raw and primary data (never edit!)
     raw_external/
     raw_internal/
     meta/
  doc/                 documentation of the study
  intermediate/        output files from intermediate analysis steps
  logs/                logs from the different analysis steps
  notebooks/           notebooks that document your day-to-day work
  results/             output from workflows and analyses
     figures/
     reports/
     tables/
  scratch/             temporary files that can safely be deleted or lost
  README.txt           file and folder description
```

https://rdmkit.elixir-europe.org/data_organisation.html#what-is-the-best-way-to-name-a-file

File names

Order the elements from general to specific.

Use meaningful abbreviations.

Use underscore (_), hypen (-) or capitalized letters - Don’t use spaces or special characters

Use date format ISO8601: YYYYMMDD, and time if needed HHMMSS.

Include a unique identifier

Include a version number if appropriate: minimum two digits (V02) and extend it, if needed for minor corrections (V02-03). The leading zeros, will ensure the files are sorted correctly.

Write your file naming convention down and explain abbreviations in your data documentation.

https://rdmkit.elixir-europe.org/data_organisation.html

File names

Honeybee project, experiment 2 done in Helsinki, data file created on the second of December 2020

- File name: `20201202_HB_EXP2_HEL_DATA_V03.xls`
- Explanation: `Time_ProjectAbbreviation_ExperimentNumber_Location_TypeOfData_VersionNumber`

https://rdmkit.elixir-europe.org/data_organisation.html

File names

Honeybee project, experiment 2 done in Helsinki, data file created on the second of December 2020

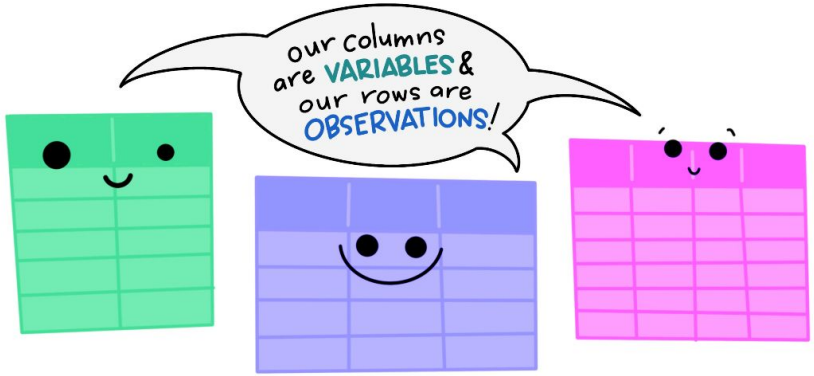
- File name: `20201202_HB_EXP2_HEL_DATA_V03.xls`
- Explanation: `Time_ProjectAbbreviation_ExperimentNumber_Location_TypeOfData_VersionNumber`

Cropped image of an ant head taken on the third of December 2020 by Meg Megson

- File name: `20201203_MM_HEAD_CROPPED_V1.psd`
- Explanation: `Time_CreatorData_TypeModification_Version`

https://rdmkit.elixir-europe.org/data_organisation.html

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM

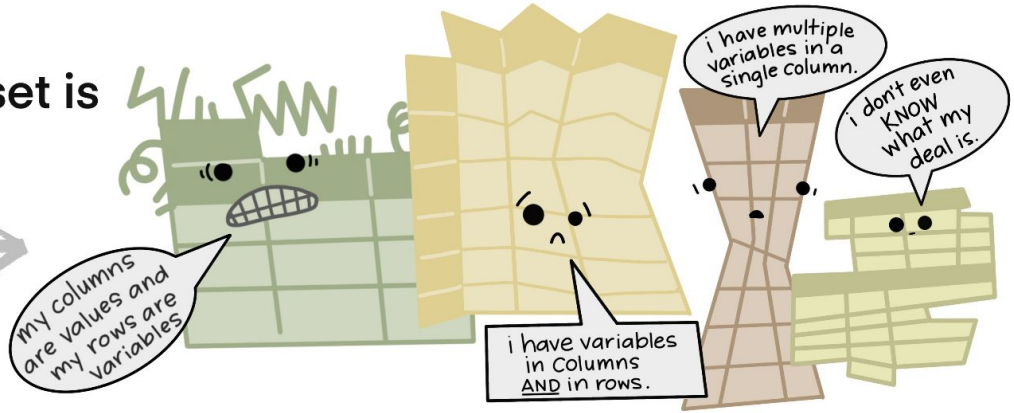


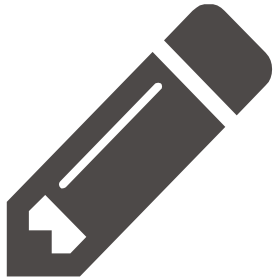
Illustration by allison_horst (CC-BY). Wickham 2014 – Tidy Data: https://doi.org/10.18637/jss.v059.i10 . Library Carpentry – Formatting data tables in spreadsheets

Types of data capture

Instrument driven

Non-instrument

- (Lab-) Notebooks
- Field observations
- Case report forms (CRFs)
- Questionnaires/Surveys



Instrument capture

Is there specialized software necessary?

Can you export (directly) to an open format?

What information on instrument settings is part of the metadata and what is not?

Do you have to set instrument parameters in a certain way to comply with a specific standard?

Non-instrument data capture

How will you digitalize the information?

Compliance with standards, is the information unambiguous?

Special considerations on:

- Surveys -> coding of results, thesauri etc. - support
- CRFs - special tools (<- semantical backing!)
- …

https://helseforskningsportalen.no/#/

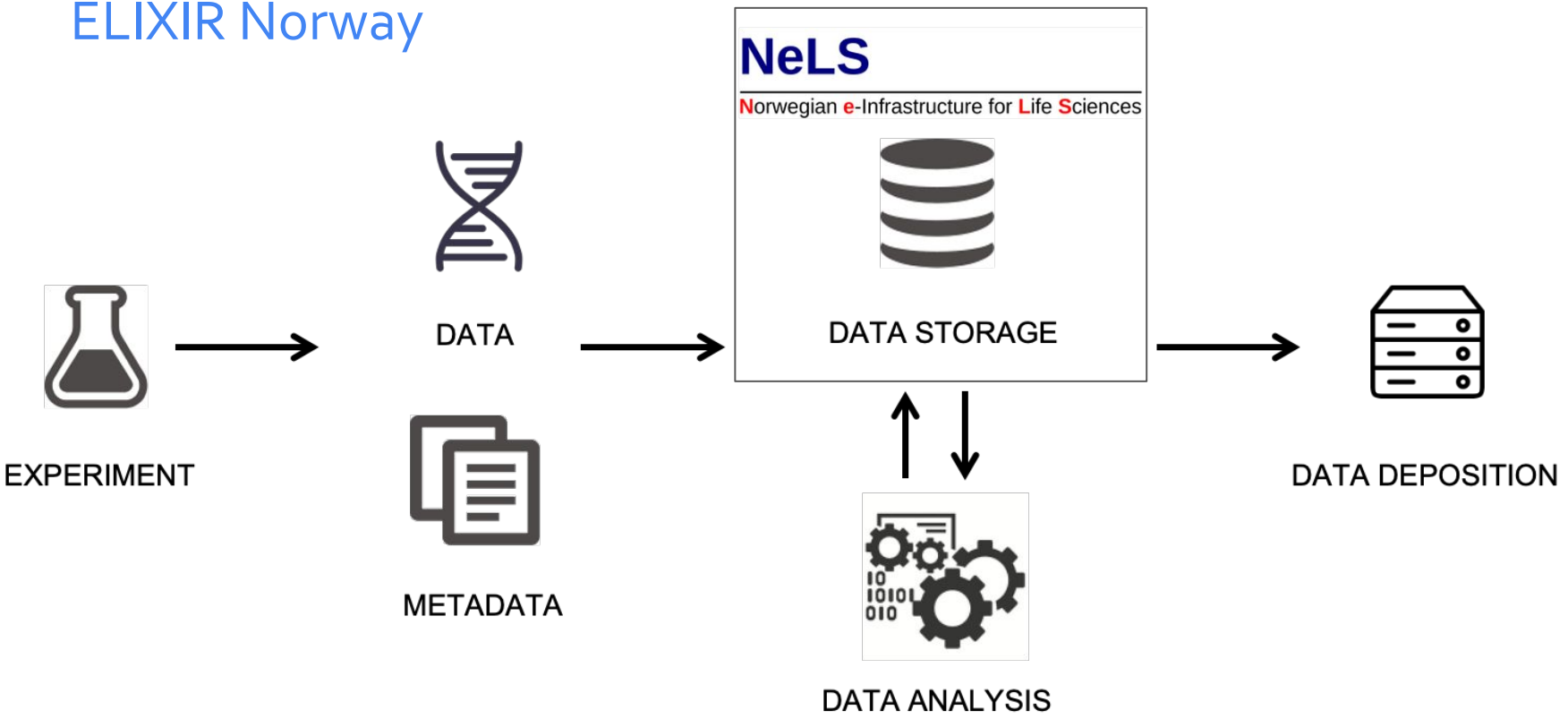
Life science storage infrastructures

At the end of this session, you will be able to:

- Disseminate information that there is a national infrastructure for storing non-sensitive life science data
- Contact the ELIXIR support desk for questions and assistance to create a storage project

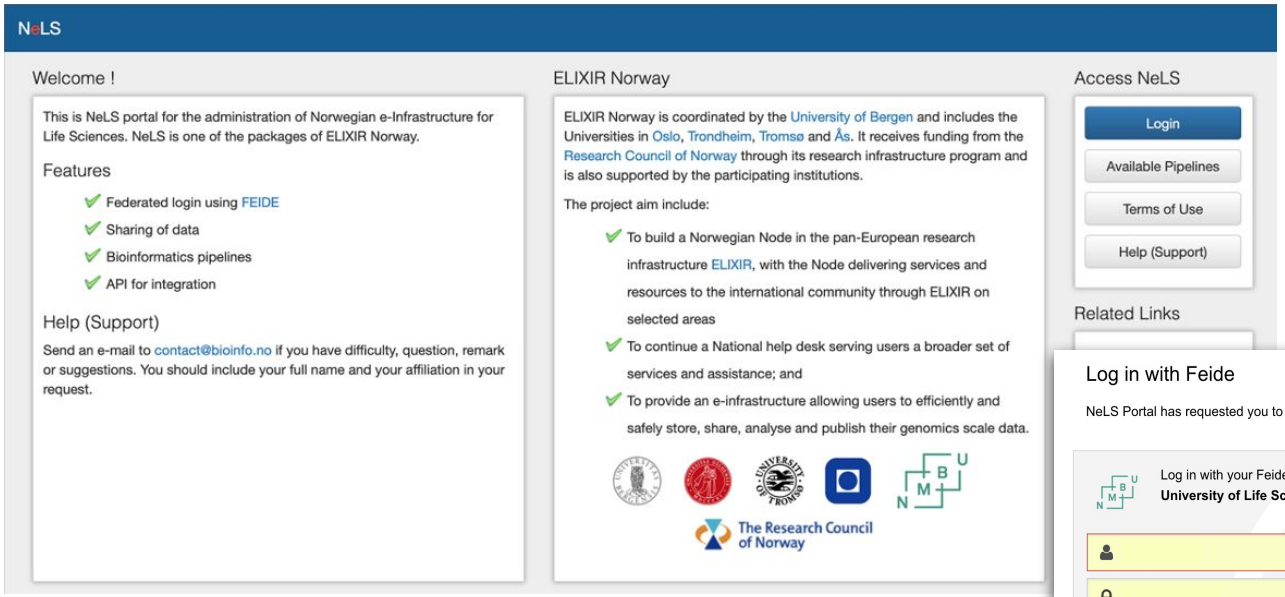
Life science storage infrastructures

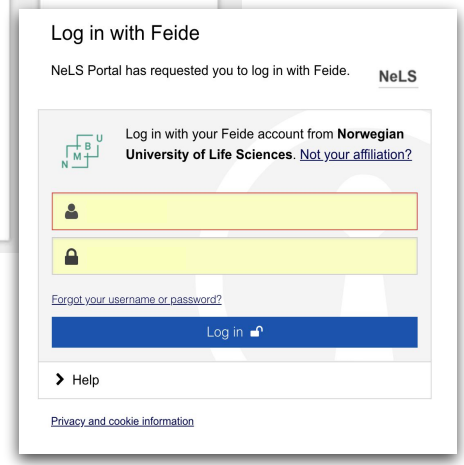
ELIXIR Norway



Storage of non-sensitive data in NeLS and StoreBioInfo (SBI)

Norwegian e-infrastructure for Life Sciences - developed and operated by ELIXIR Norway



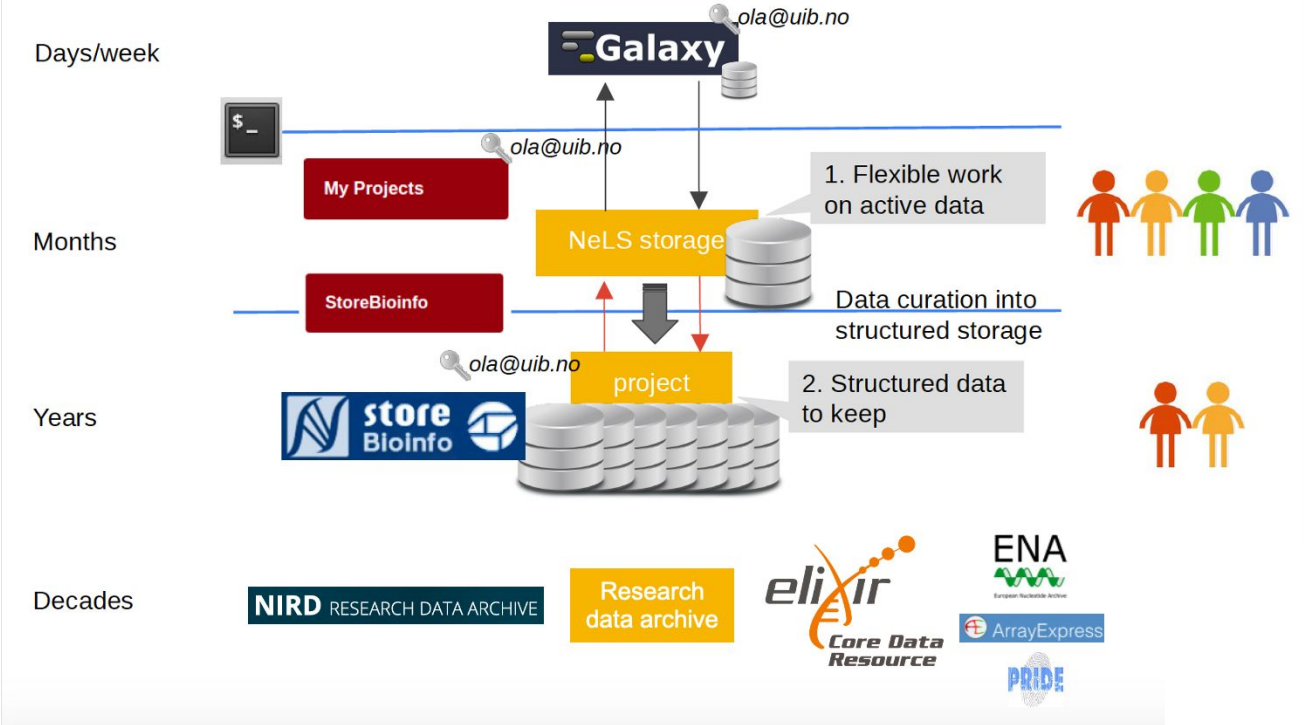


NeLS portal https://nels.bioinfo.no/

NeLS architecture - two layer storage structure

NeLS – typically during the data analysis and manuscript preparation

SBI – typically for large projects where data is used in several publications

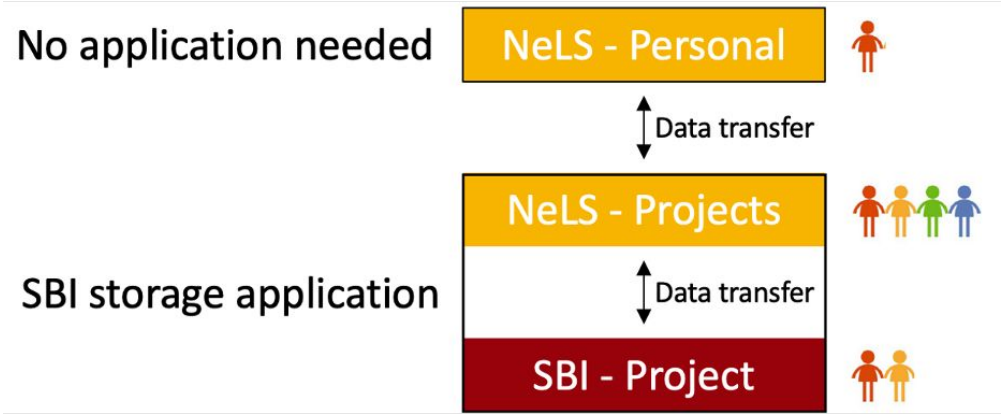


Apply for storage project and access

Storage application needed for projects where data is shared by many users

Access via FEIDE user or NeLS idp can be made for non-FEIDE users

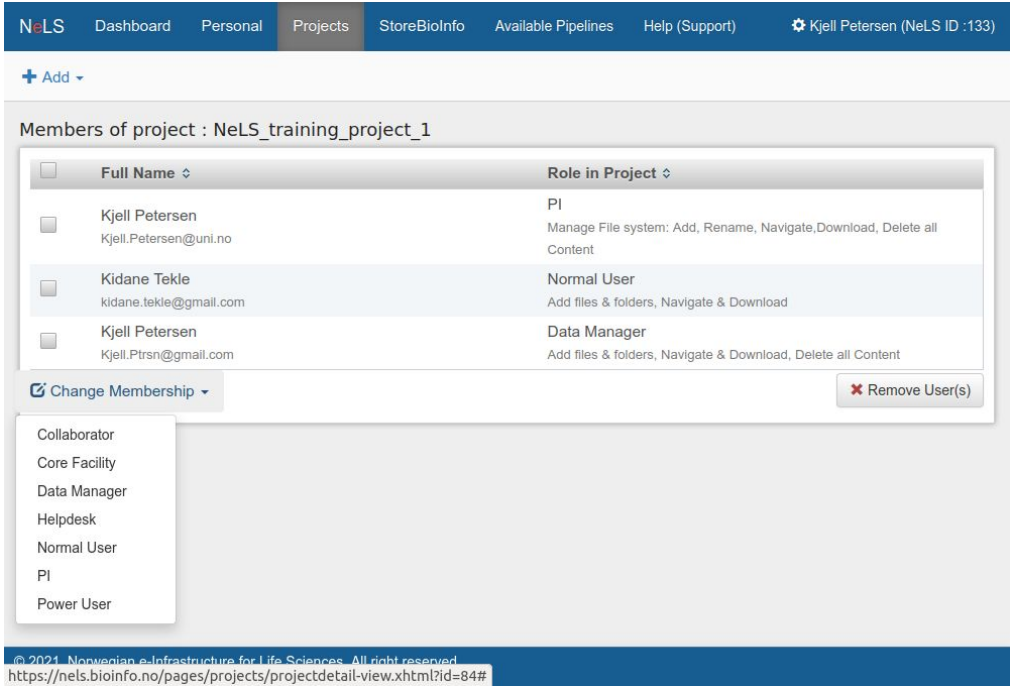
Access to personal storage and collaborative projects - and SBI



Apply for storage: contact@bioinfo.no
Wiki for usage: https://nels-docs.readthedocs.io/en/latest/

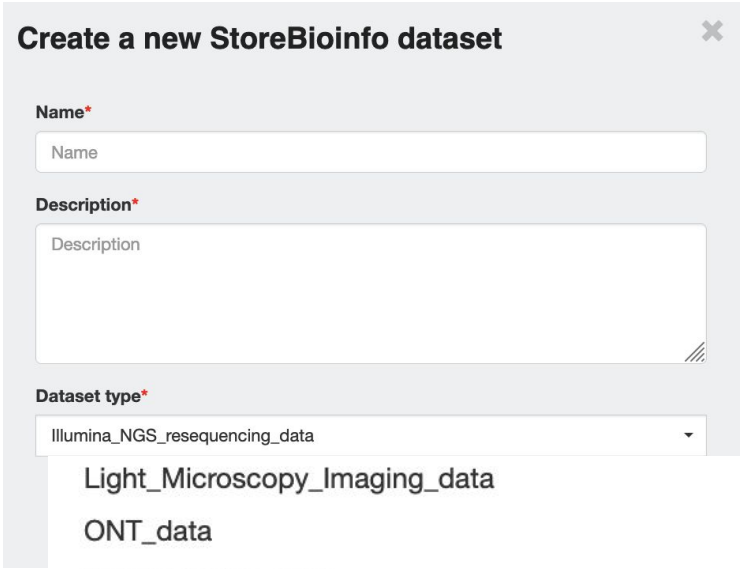
Users can have different roles and permissions in a project

In both NeLS and StoreBioinfo projects a user profile gets permissions to work with data (upload, rename, move, delete, update) based on a role in a project



Data types in SBI

A project can store multiple different data types



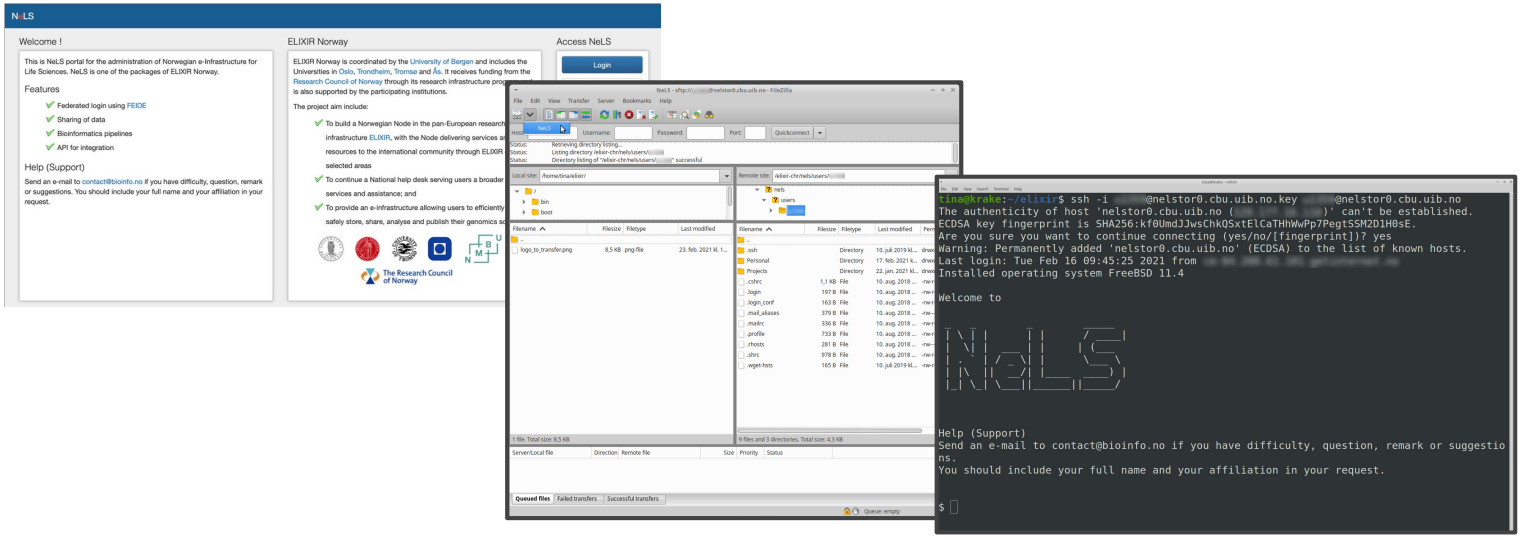
- Light_Microscopy_Imaging_data
- ONT_data
- Metabolomics_data
- Illumina_seq_data
- Microarray_Methylation_data
- Sanger_seq_data
- 454_seq_data
- Molecular_Dynamics_data
- AB_SOLiD_NGS_data
- Proteomics_data
- Track
- Microarray_Gene_Expression_data
- Pacific Biosciences seq Data
- Illumina_NGS_resequencing_data

Access and transfer data

Uploading and saving files via web browser in NeLS

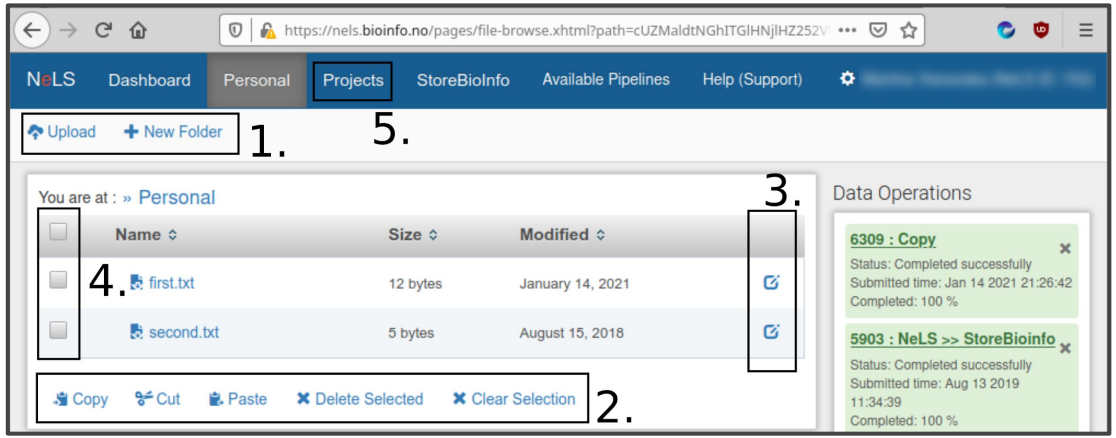
Transferring data with FileZilla

Transferring data with command line tools: ssh,scp



NeLS portal - quick usage intro

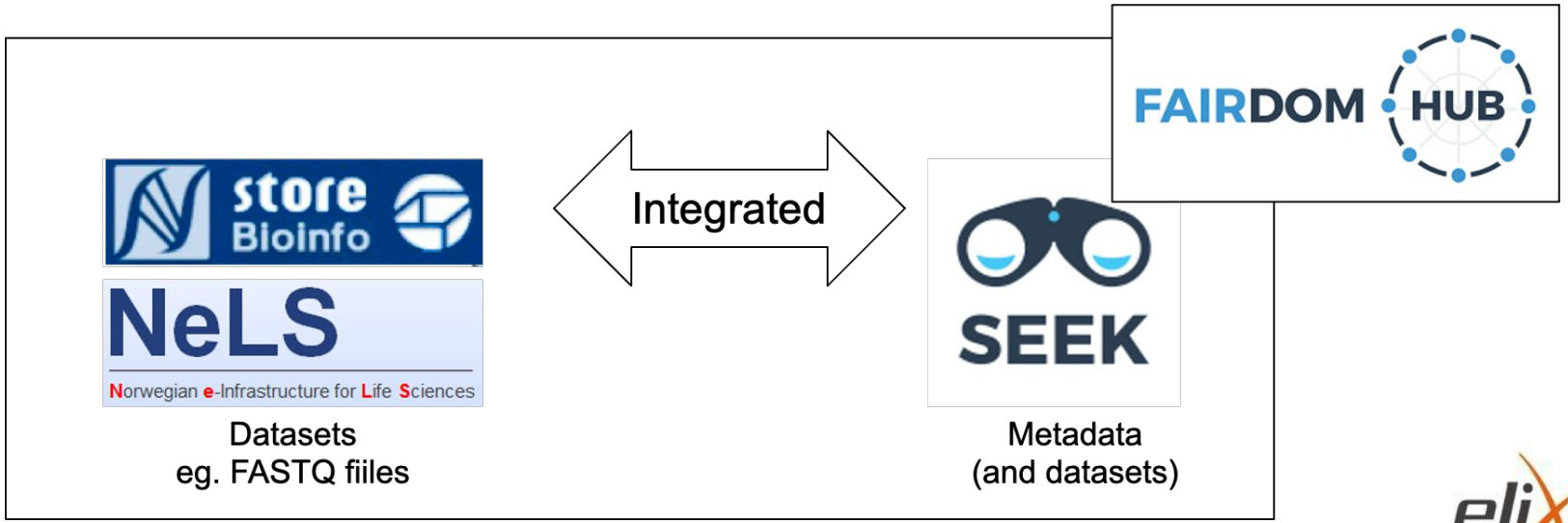
1. Upload - only one file at a time, add new folder
2. File/folder manipulation
3. Rename file/folder
4. (De)select all/some items
5. Projects - containing projects you have access to



SEEK integration for metadata

The SEEK platform is a web-based tool for organising and storing data, and for exploring and annotating data

Norwegian users can link datasets stored in NeLS to a SEEK project using FAIRdom hub

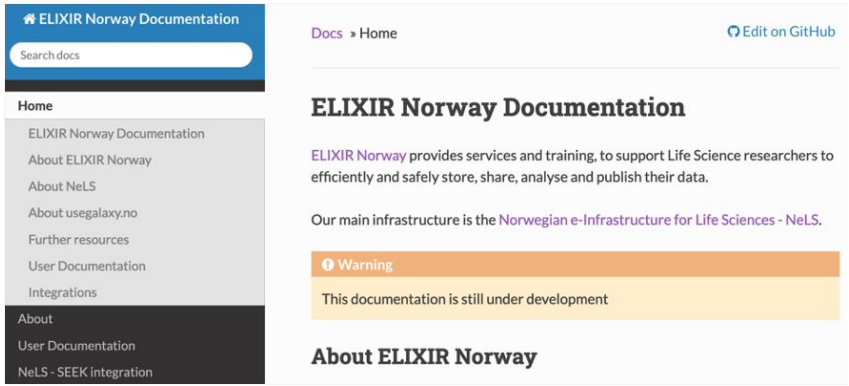


End-user support and training - ELIXIR Norway Support desk

ELIXIR-NO organise courses in the use of the infrastructure regularly

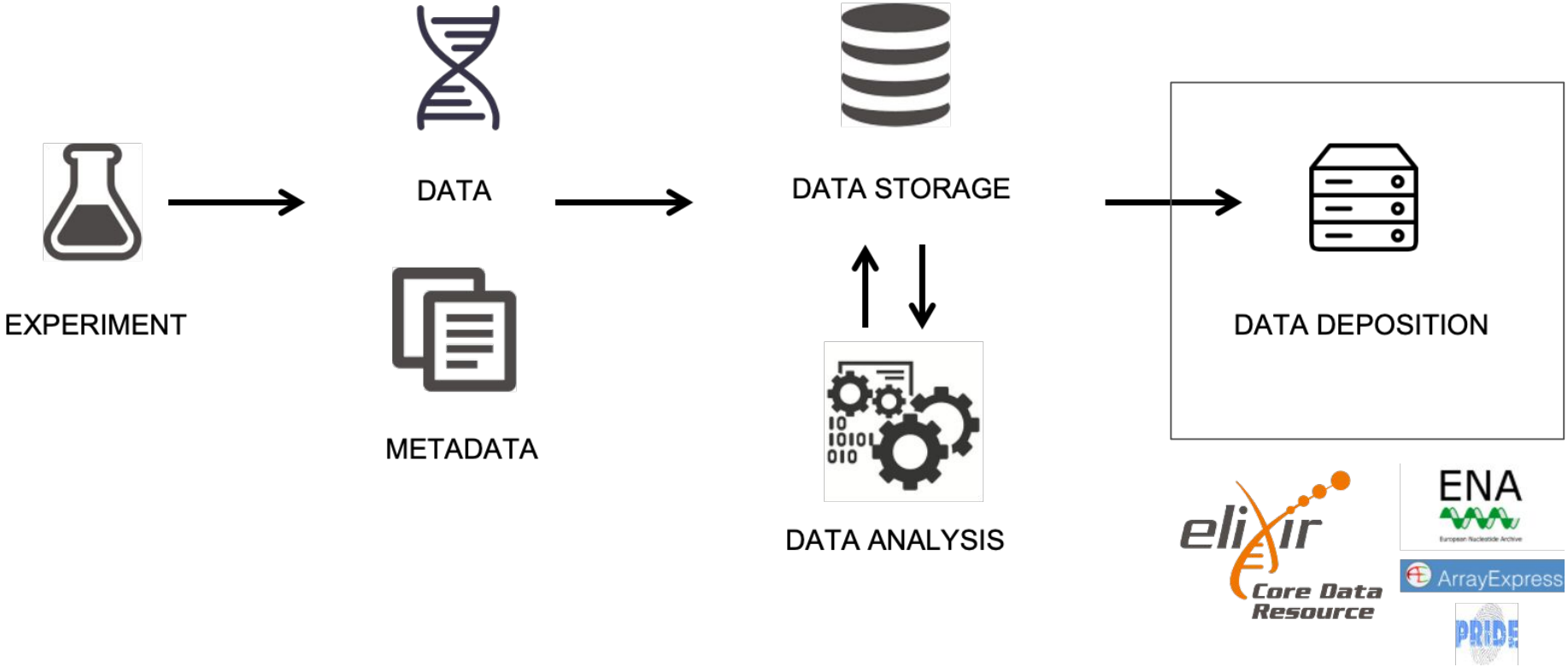
ELIXIR-NO has an active support desk

ELIXIR-NO maintains a wiki for the use of the infrastructure



ELIXIR Norway Support desk: contact@bioinfo.no
Upcoming training: https://elixir.no/events

Support data archiving - ELIXIR core deposition databases



Why use NeLS and SBI?

Life science specific data storage

Easy to use with intuitive navigation

Access via FEIDE user

Connected directly with compute (usegalaxy.no)

Enable collaborative projects

Data are safely backed up

ELIXIR.NO offers training and end-user support

ELIXIR.NO offers support in data deposition

National storage infrastructures

Learning Objectives

At the end of this session, you will be able to:

- Disseminate information that there is a national infrastructure for storing non-sensitive science data
- contact the Sigma2 support for questions and assistance to create a storage project

NRIS

NIRD

NIRD (The National infrastructure for research data) is owned by Sigma 2 and NRIS operates the system

NRIS is a collaboration of the four BOTT universities and Sigma2 to pool competencies, resources and services.

NIRD Storage: Total disk capacity 22PiB, GPFS Parallel filesystem

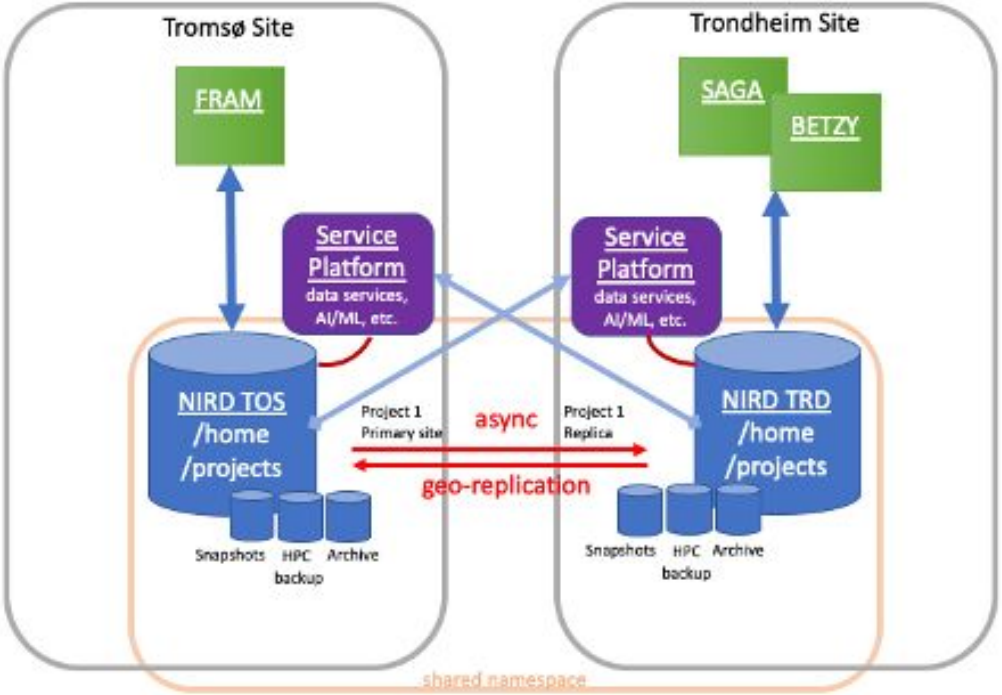
NIRD Service platform: 1152 vcpu, 8192 GiB , 32 v100 GPUs

Apply for resources: https://www.sigma2.no/data-storage#get-access

NRIS

elixir

NIRD Architecture



NIRD Service Platform

Kubernetes based cloud infrastructure

Mounts NIRD project storage so intensive I/O operations can be done on a large pool of data

Portability and reproducibility of tools through containers

Allows pre-/post-processing analysis, data intensive processing, AI/ML

Login nodes of NIRD

Persistent services

NRIS

elixir

Thank you!

elixir-norway.org

@elixirnorway

contact@bioinfo.no