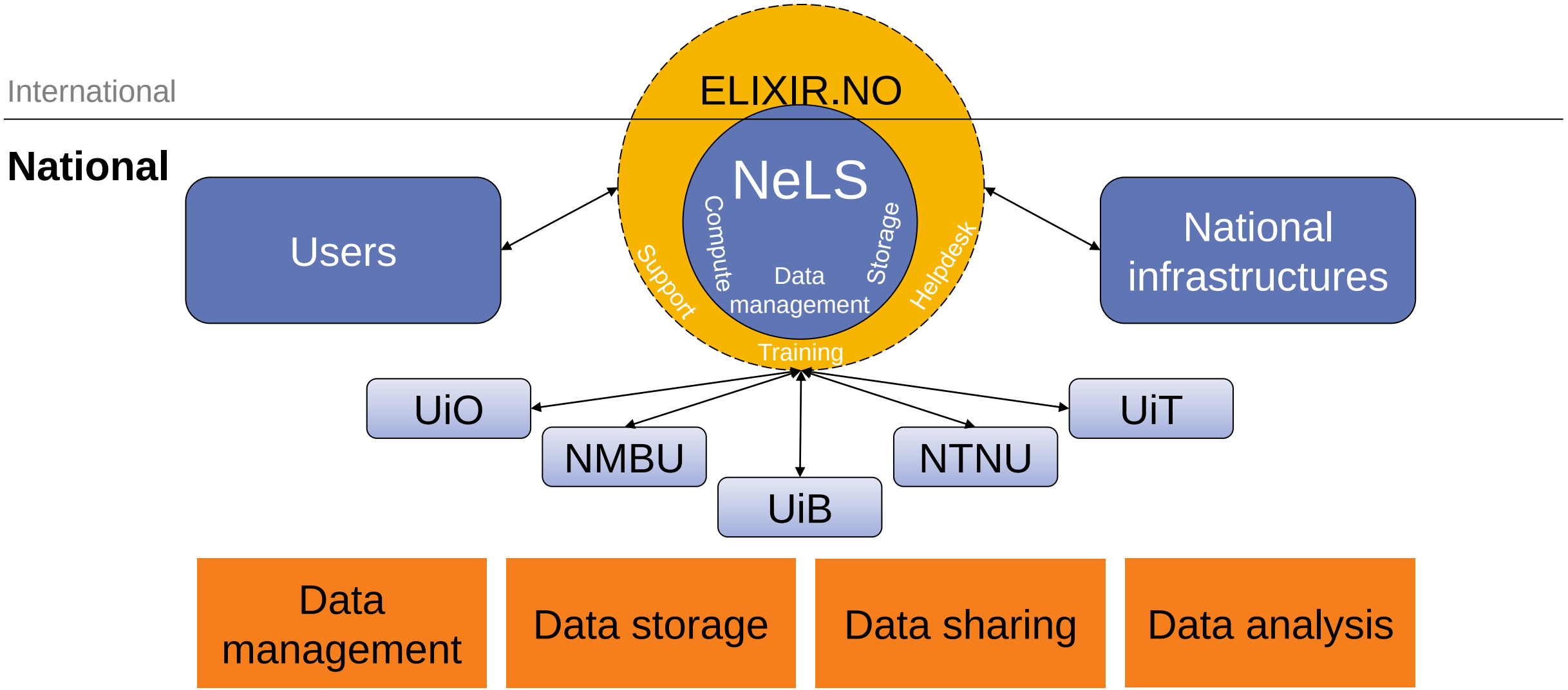
Introduction to NeLS and Galaxy

The national bioinformatics infrastructure

Erik Hjerde

NeLS - the Norwegian e-Infrastructure for Life Science



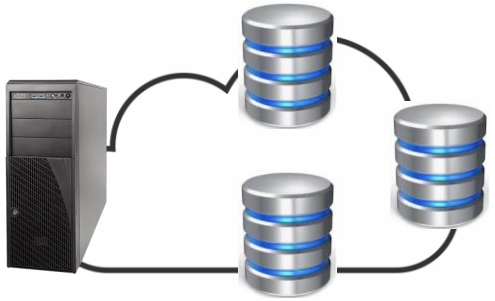
Overview of data management

FAIR data management from sequencing provider to public repository

Data production

Data storage/sharing

Data compute



Data archiving

nor seq

NeLS
Norwegian e-Infrastructure for Life Sciences

TSD

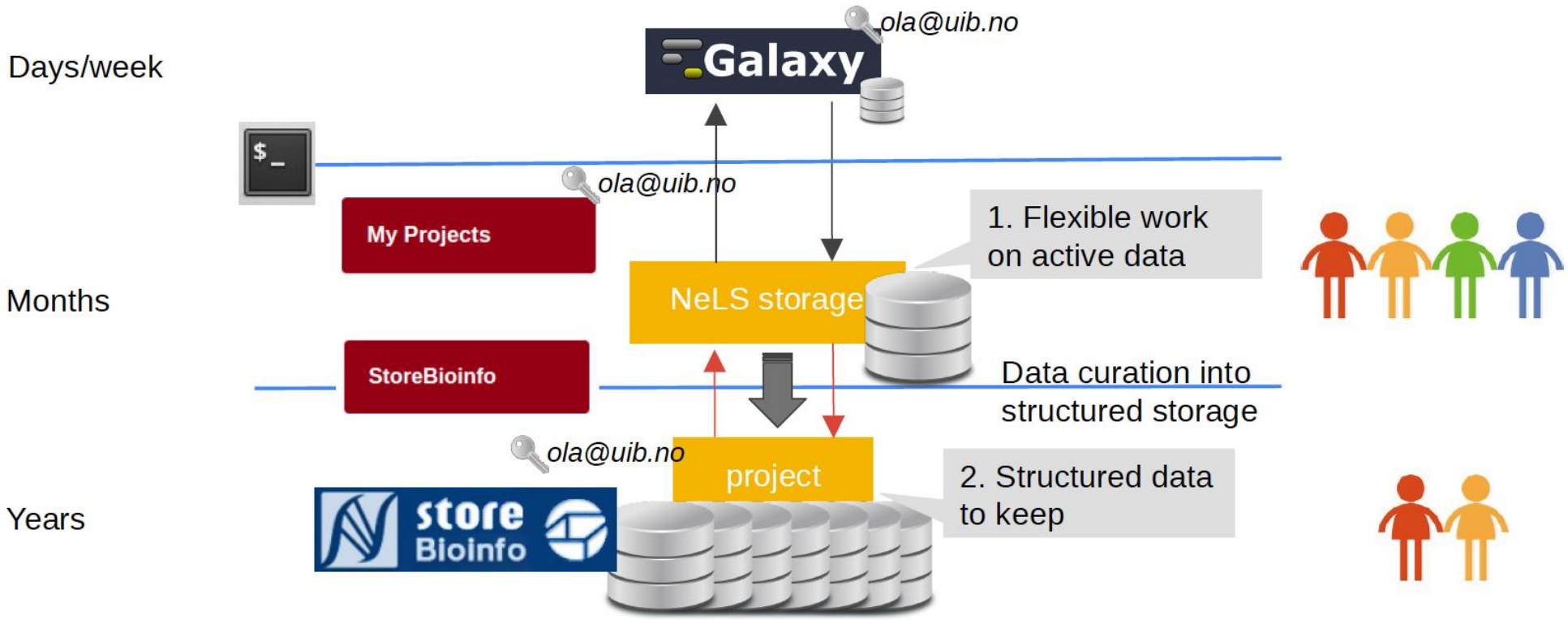
Galaxy

elixir
NIRD RESEARCH DATA ARCHIVE
Core Data Resource

Data management

elixir
NORWAY

NeLS architecture



NIRD RESEARCH DATA ARCHIVE

Research data archive

elixir Core Data Resource

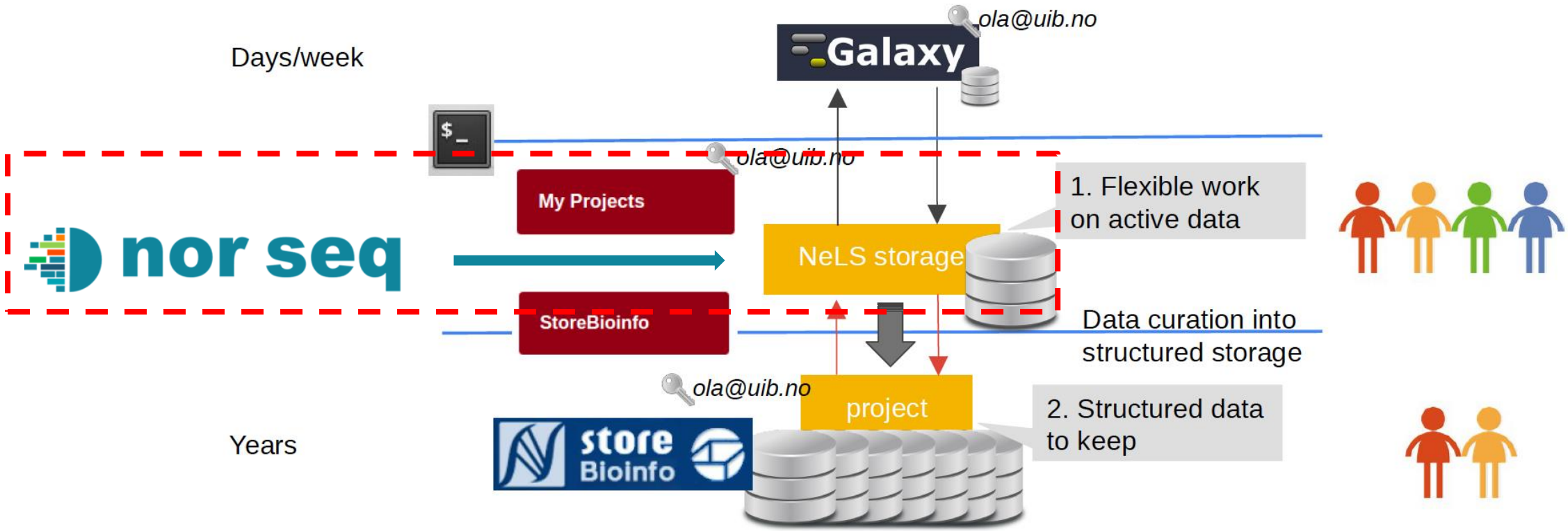
ENA European Nucleotide Archive

ArrayExpress

PRIDE

elixir NORWAY

Data from data providers



NIRD RESEARCH DATA ARCHIVE

Research data archive

elixir Core Data Resource

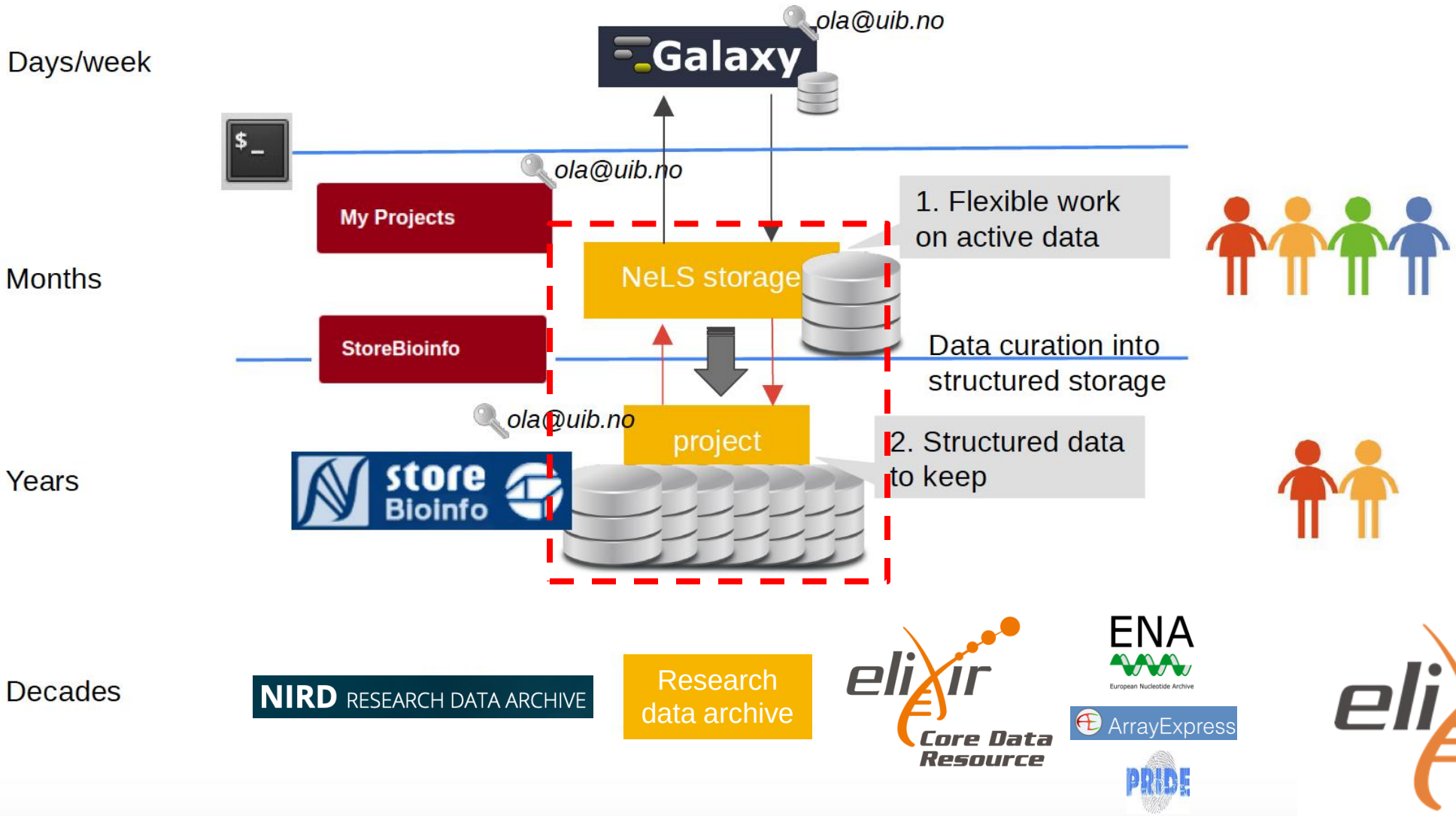
ENA European Nucleotide Archive

ArrayExpress

PRIDE

elixir NORWAY

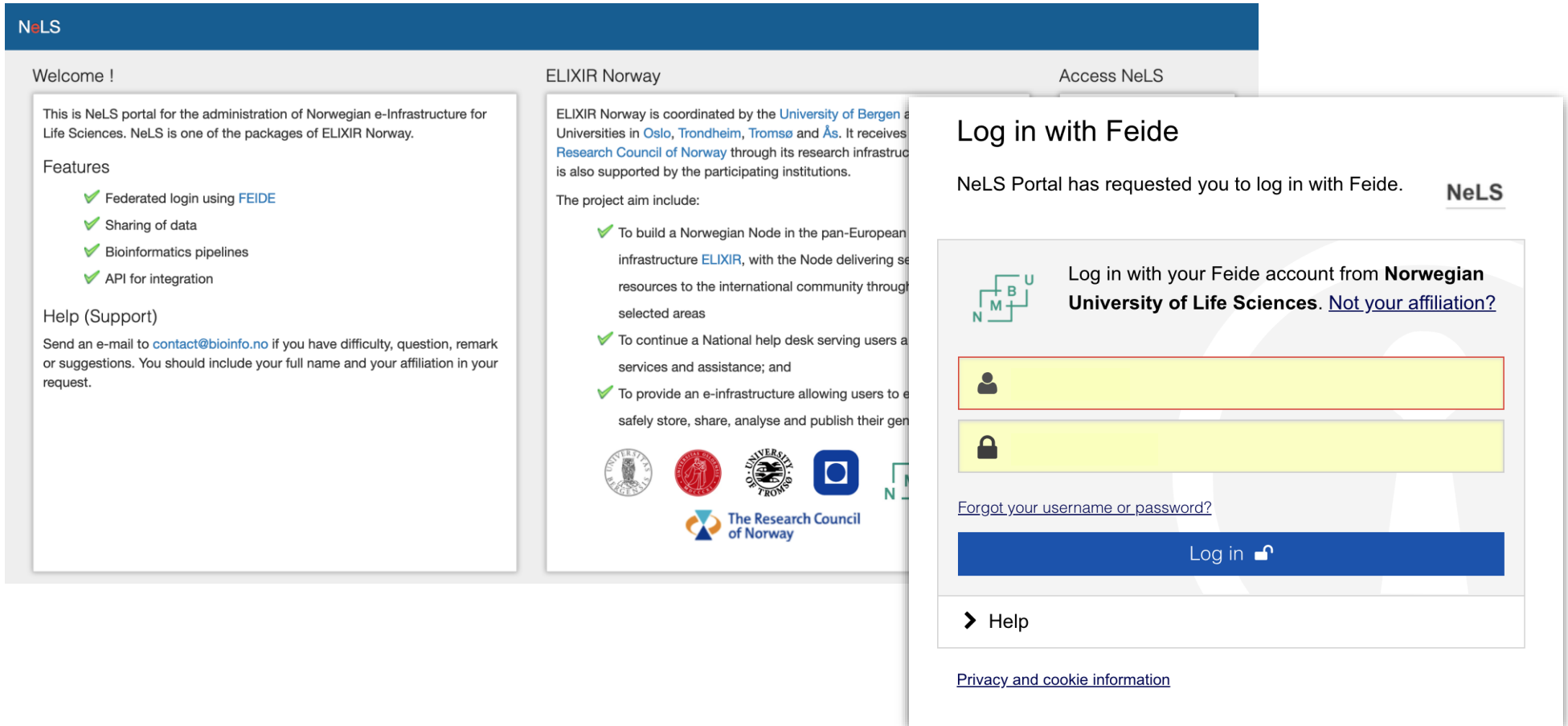
Data storage in NeLS and StoreBioInfo



Access to storage and sharing of data in NeLS

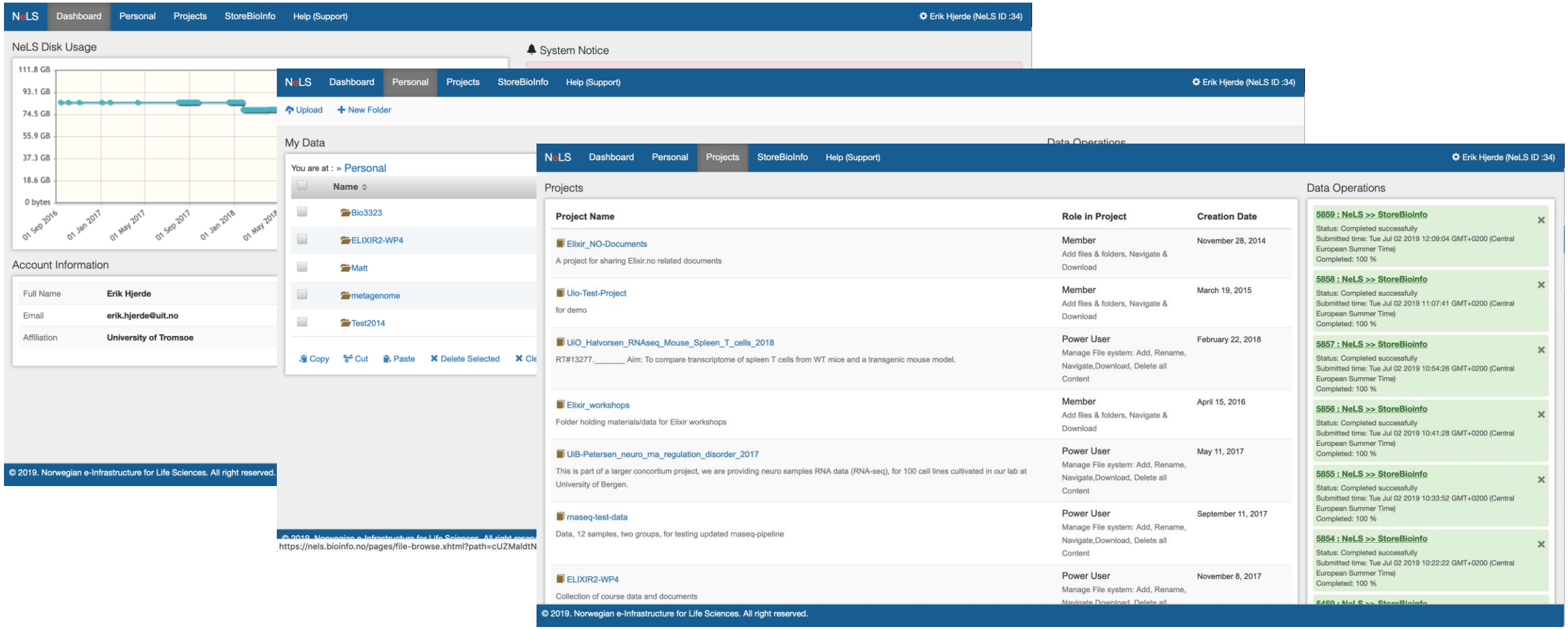
Temporary storage – typically during the data analysis and manuscript preparation

Users with permissions can access data in a NeLS project (FEIDE or NeLS ID)



Access to storage and sharing of data in NeLS

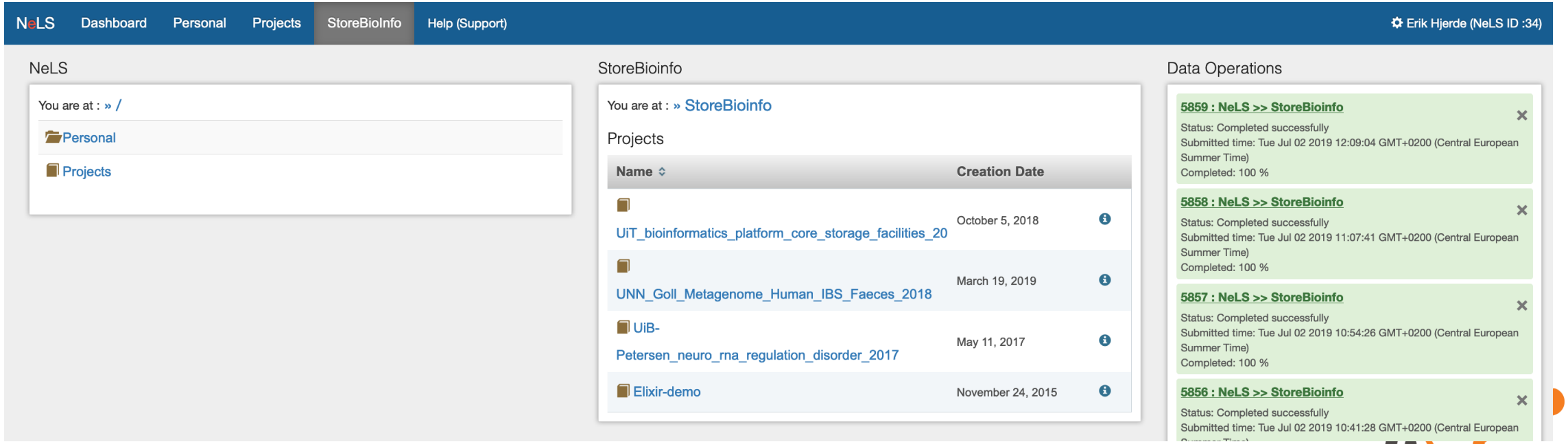
Access to personal storage and projects



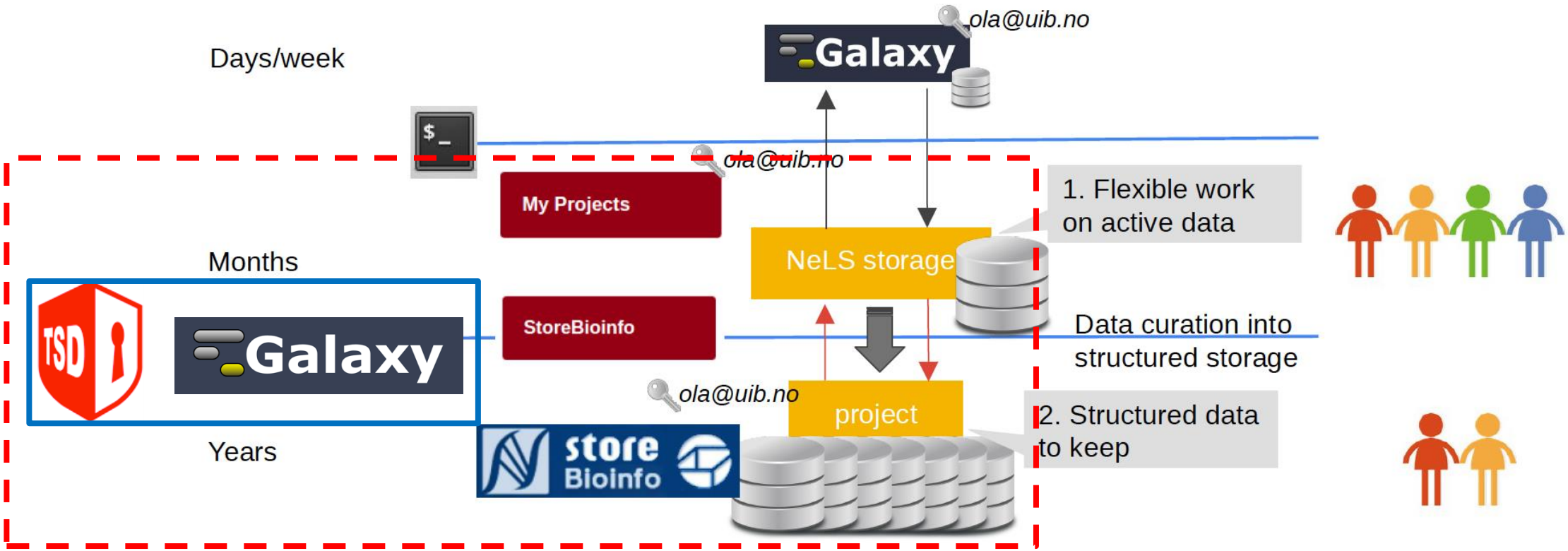
Data storage in StoreBioInfo

Long term storage – typically for large projects where data is used in several publications

Storage quotas has to be applied for – typically > 1TB data



Sensitive data in TSD



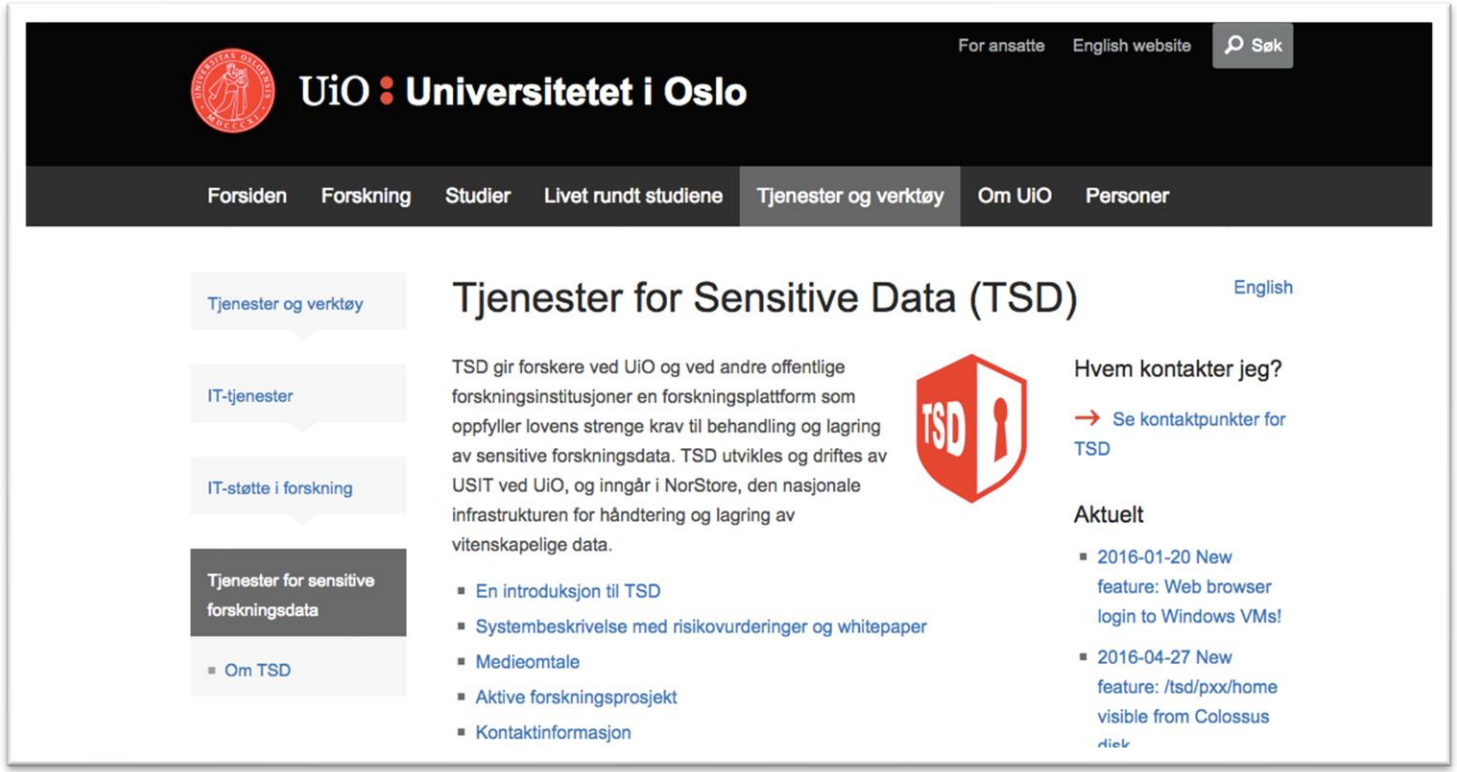
TSD – Tjenester for Sensitive Data

National infrastructure for computing and data management of sensitive data

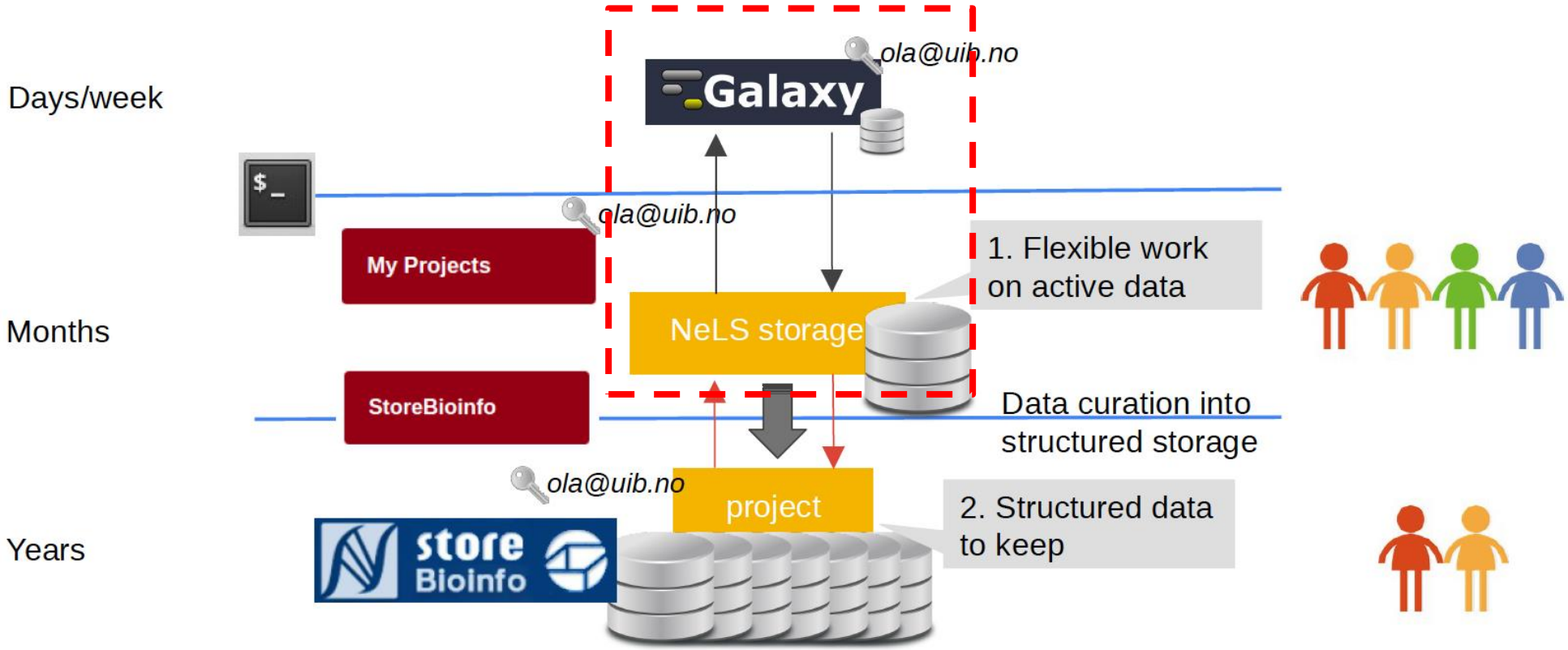
Fulfils the requirements by Norwegian law for treatment and storage of sensitive data

TSD is developed and hosted by USIT at the UiO

The NeLS-portal is the only web entry point to get data in and out of TSD

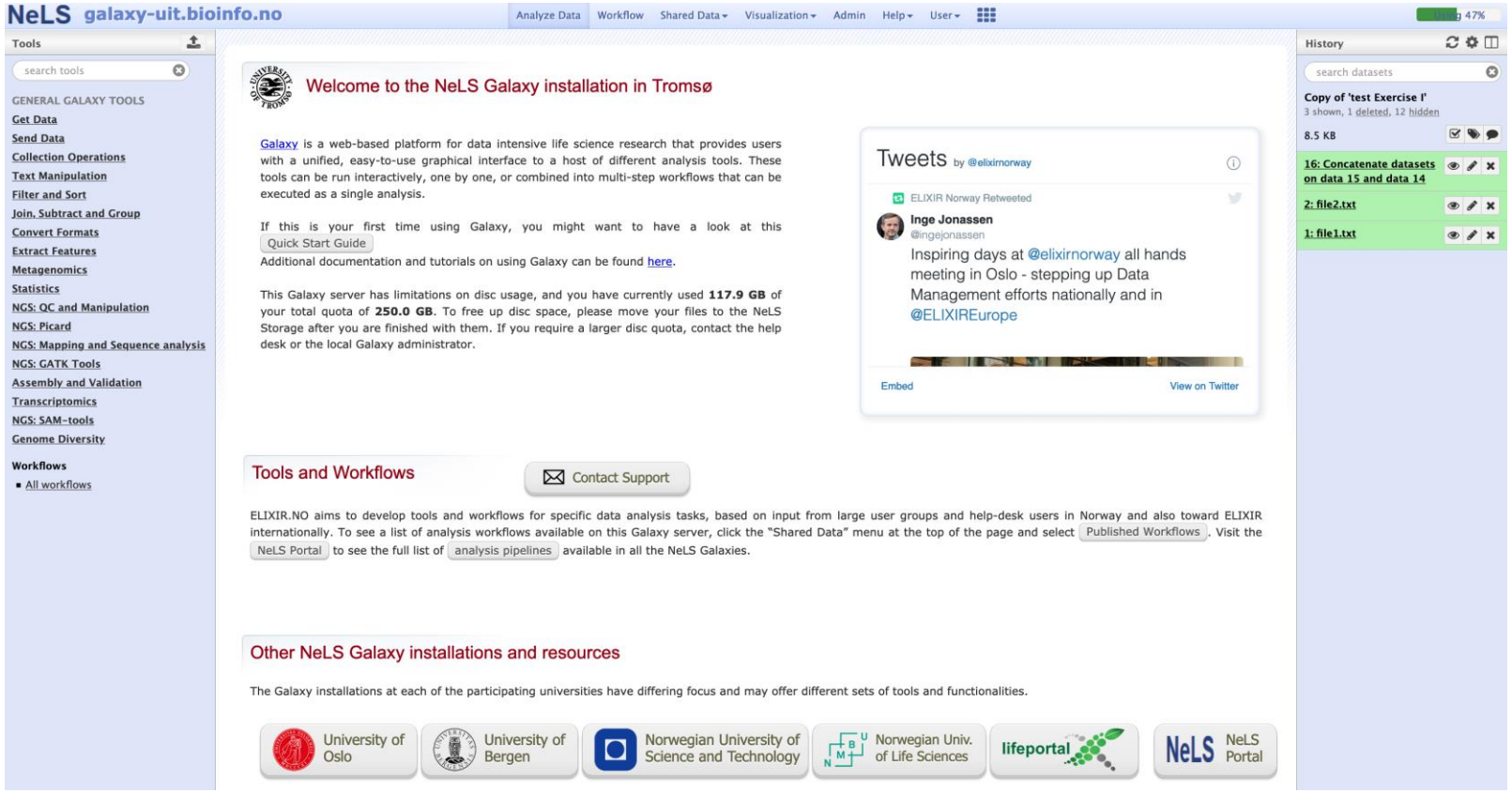


Data analysis



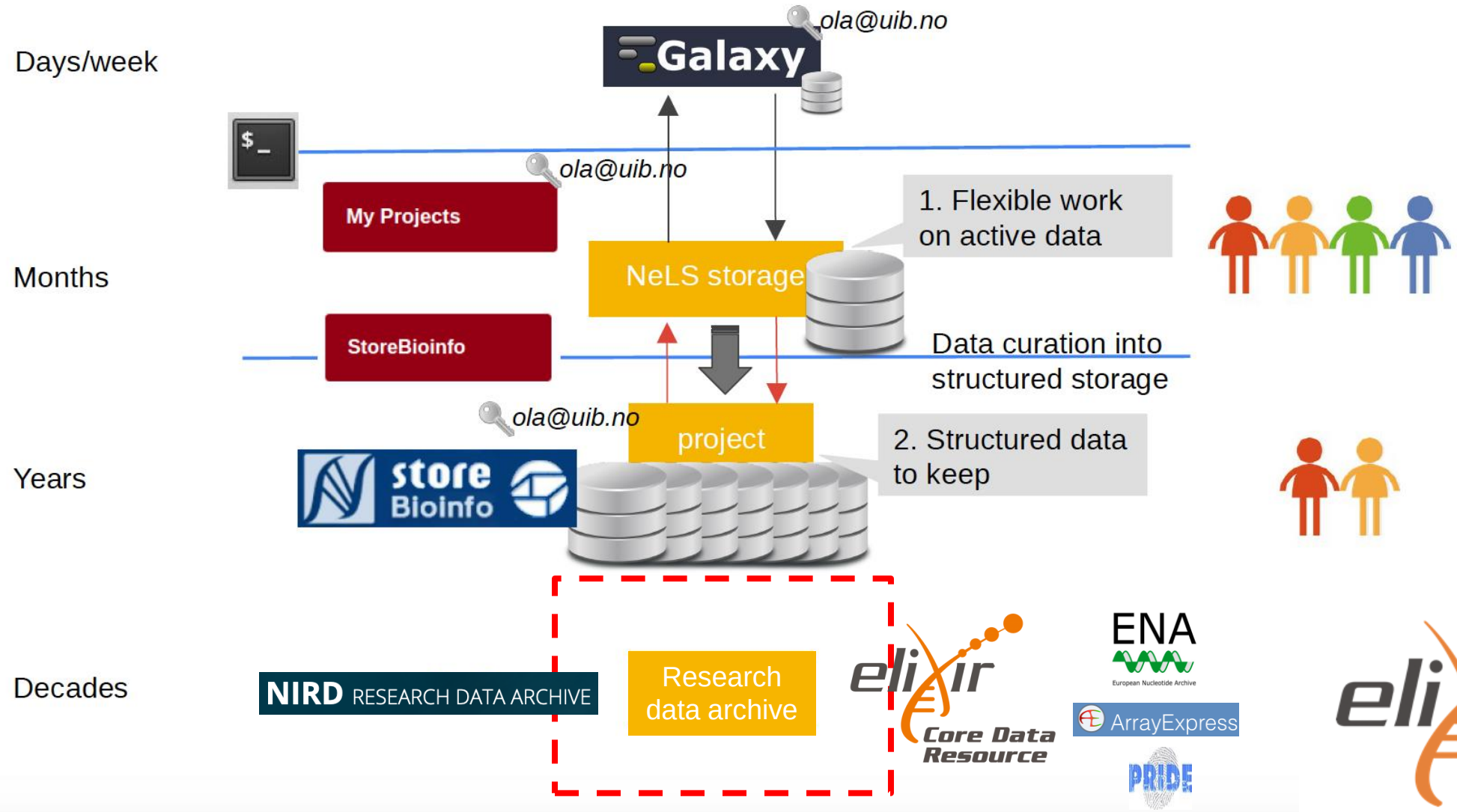
Data analysis in NeLS Galaxy

Each node host a Galaxy server with unique “flavour”



elixir NORWAY

Data archiving in data repositories



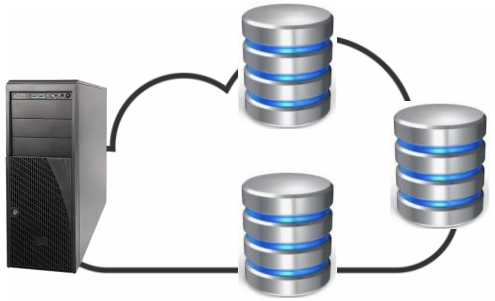
Overview of data management

FAIR data management from sequencing provider to public repository

Data production

Data storage/sharing

Data compute



Data archiving

nor seq

NeLS
Norwegian e-Infrastructure for Life Sciences

TSD

Galaxy

elixir
NIRD RESEARCH DATA ARCHIVE
Core Data Resource

Data management

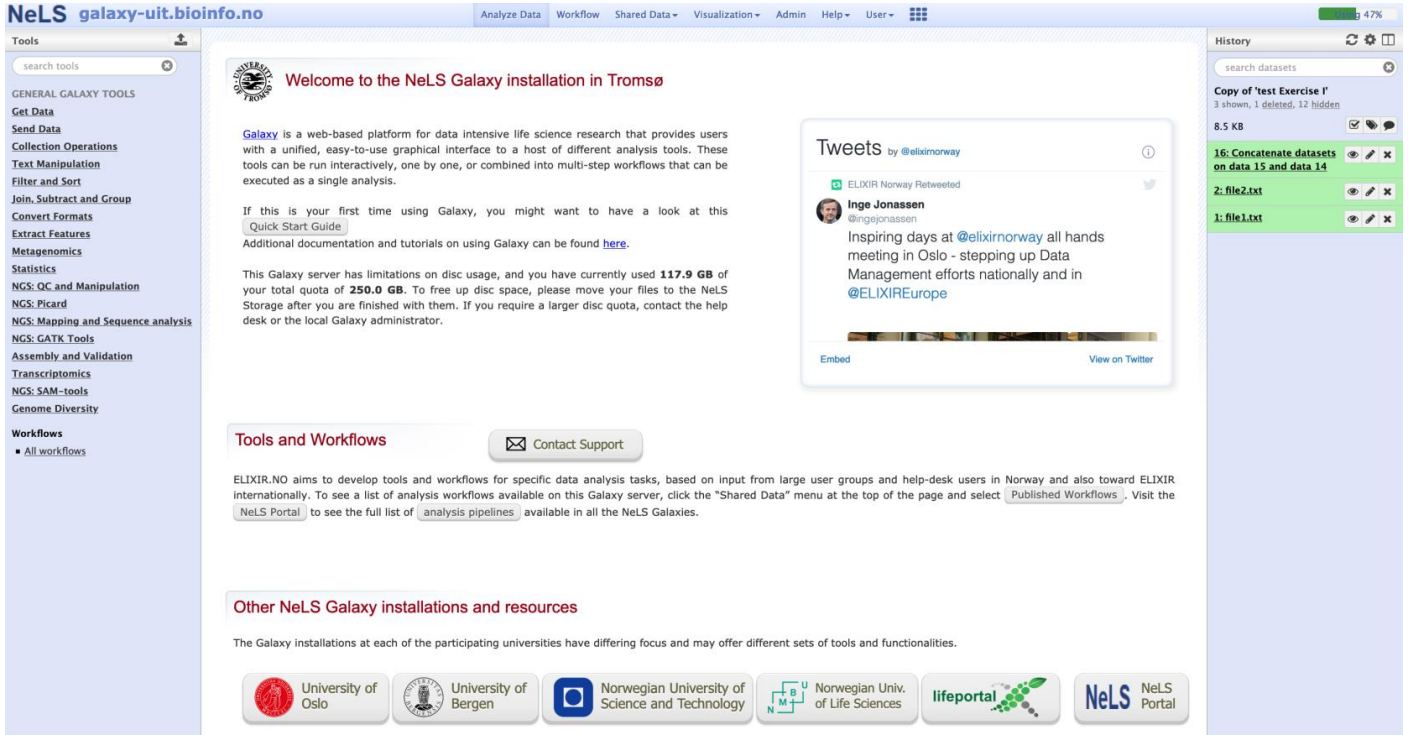
elixir
NORWAY

Now Galaxy…

Data analysis in NeLS Galaxy

Galaxy: open, web-based platform for accessible, reproducible, and transparent computational biomedical research

This give access to multiple tools and data analysis workflows



Data analysis in NeLS Galaxy

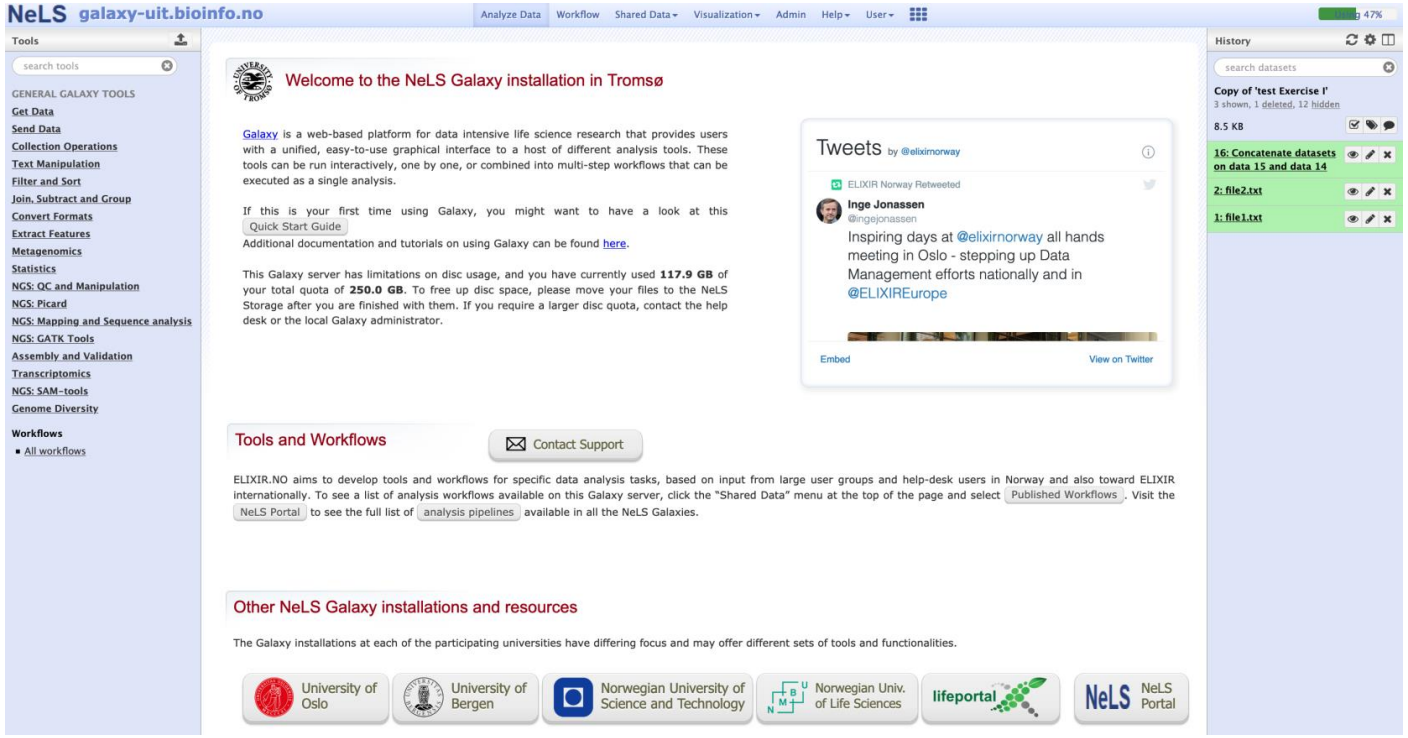
Quick start guide:

- Galaxy basics
- Uploading local data files into Galaxy
- How to import/export data from NeLS storage
- Dataset collections
- Running tools
- Workflows

Total quota 250 GB

Contact support

Q&A forum

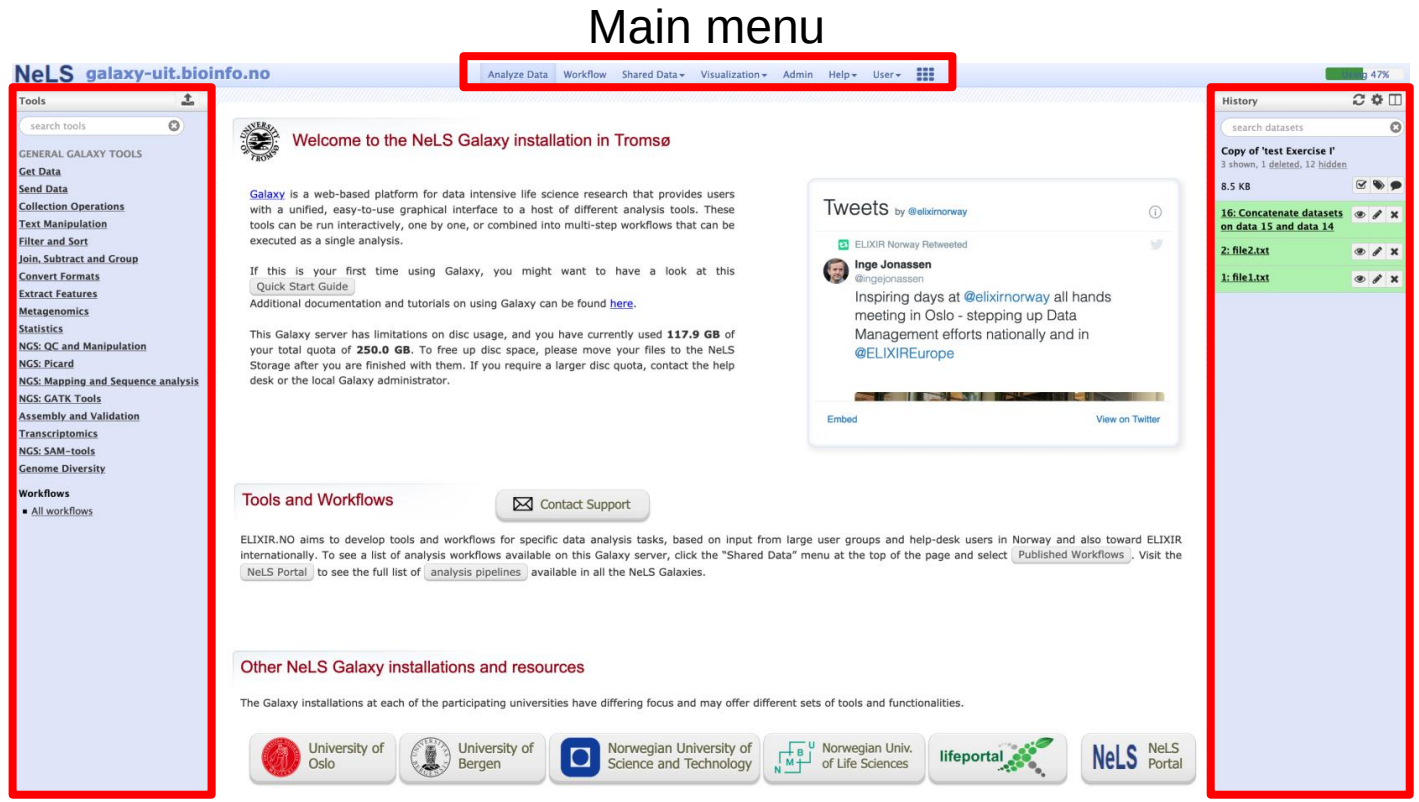


Galaxy – The basics

Main menu

Tool menu

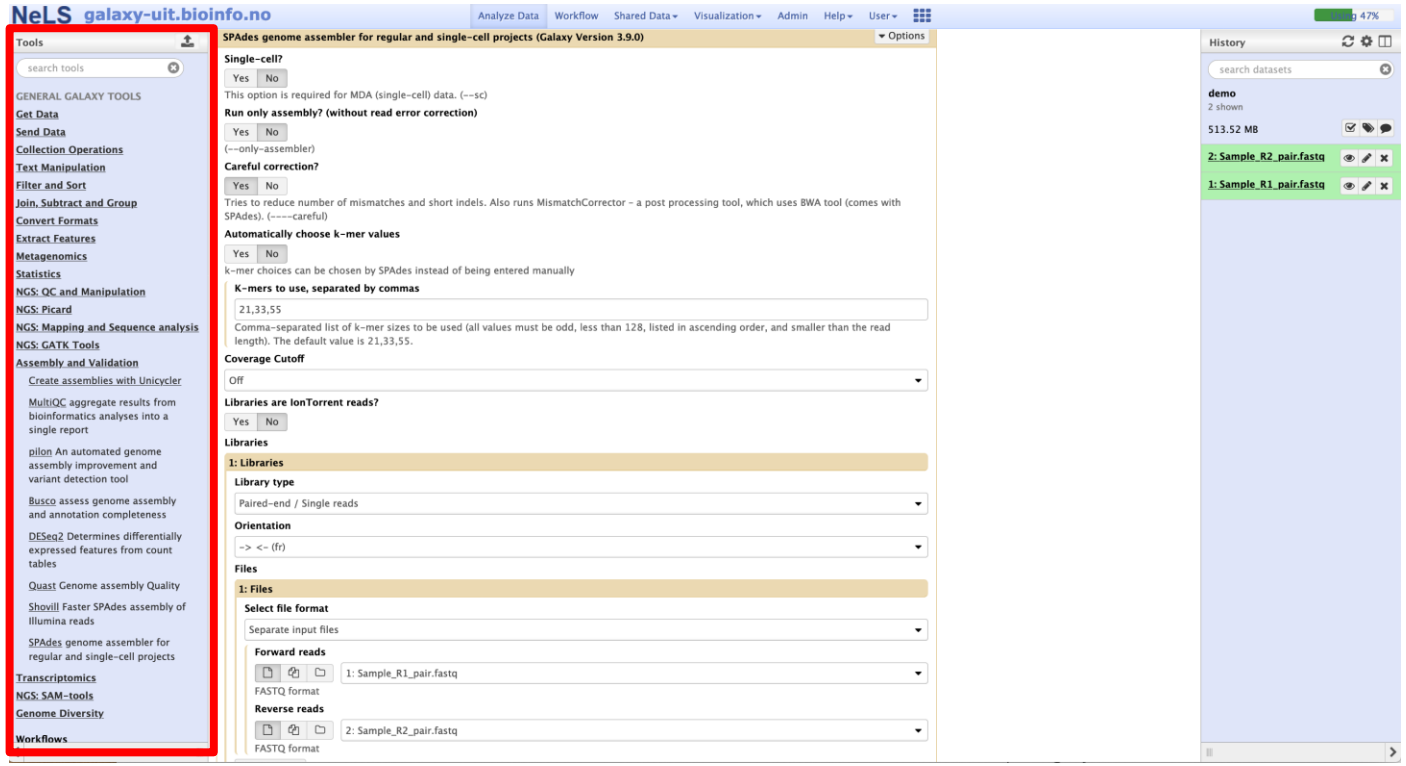
Main window



Galaxy - Tools

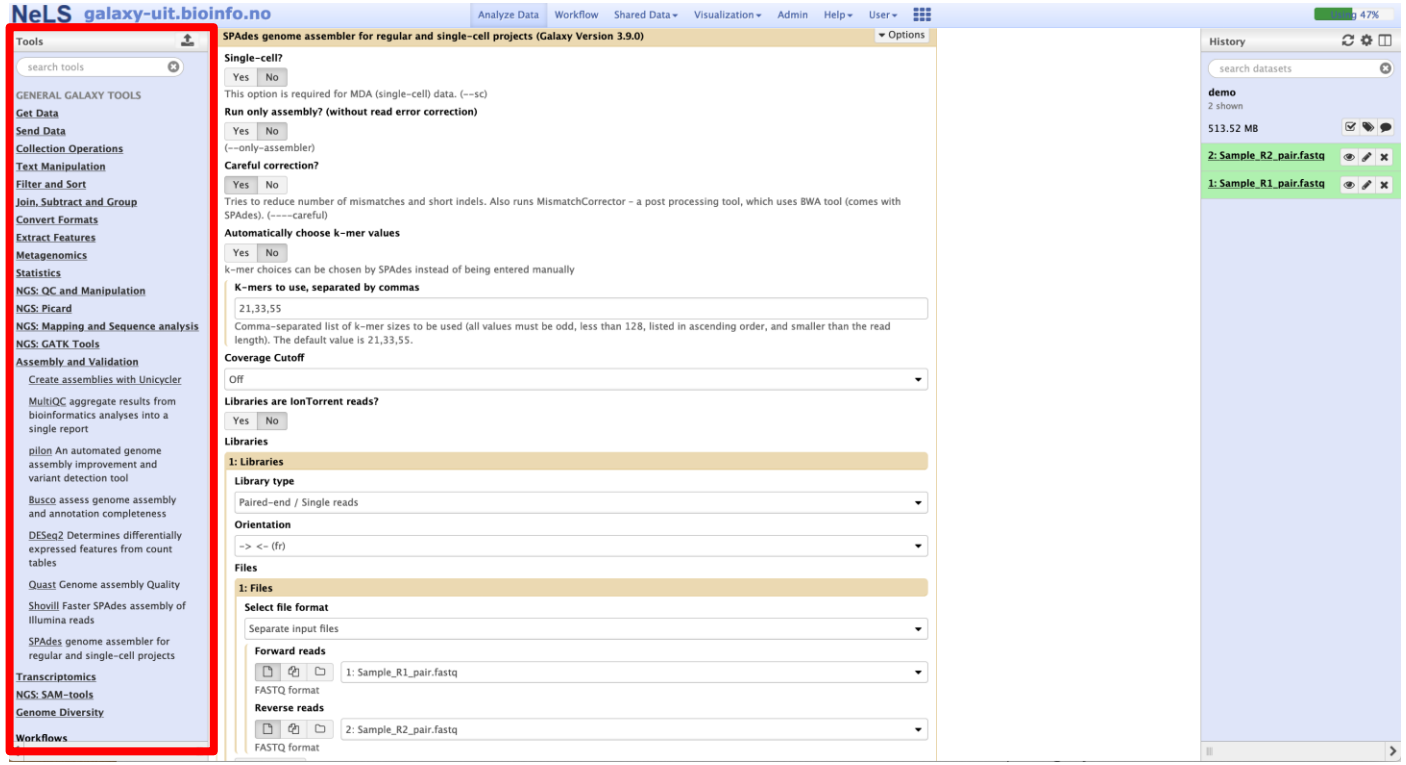
Tools are available from the Tool menu

Command line tools are wrapped into to Galaxy so they become accessible with a GUI



Galaxy - Tools

```
SPAdes genome assembler v3.11.1

Usage: /Users/service/tools/SPAdes-3.11.1-Darwin/bin/spades.py [options] -o <output_dir>

Basic options:
-o      <output_dir>    directory to store all the resulting files (required)
--sc                    this flag is required for MDA (single-cell) data
--meta                  this flag is required for metagenomic sample data
--rna                   this flag is required for RNA-Seq data
--plasmid               runs plasmidSPAdes pipeline for plasmid detection
--iontorrent            this flag is required for IonTorrent data
--test                  runs SPAdes on toy dataset
-h/--help               prints this usage message
-v/--version            prints version

Input data:
--12    <filename>      file with interlaced forward and reverse paired-end reads
-1      <filename>      file with forward paired-end reads
-2      <filename>      file with reverse paired-end reads
-s      <filename>      file with unpaired reads
--pe<#>-12      <filename>      file with interlaced reads for paired-end library number <#> (<#> = 1,2,..,9)
--pe<#>-1       <filename>      file with forward reads for paired-end library number <#> (<#> = 1,2,..,9)
--pe<#>-2       <filename>      file with reverse reads for paired-end library number <#> (<#> = 1,2,..,9)
--pe<#>-s       <filename>      file with unpaired reads for paired-end library number <#> (<#> = 1,2,..,9)
--pe<#>-<or>    orientation of reads for paired-end library number <#> (<#> = 1,2,..,9; <or> = fr, rf, ff)
--s<#>          <filename>      file with unpaired reads for single reads library number <#> (<#> = 1,2,..,9)
--mp<#>-12      <filename>      file with interlaced reads for mate-pair library number <#> (<#> = 1,2,..,9)
--mp<#>-1       <filename>      file with forward reads for mate-pair library number <#> (<#> = 1,2,..,9)
--mp<#>-2       <filename>      file with reverse reads for mate-pair library number <#> (<#> = 1,2,..,9)
--mp<#>-s       <filename>      file with unpaired reads for mate-pair library number <#> (<#> = 1,2,..,9)
--mp<#>-<or>    orientation of reads for mate-pair library number <#> (<#> = 1,2,..,9; <or> = fr, rf, ff)
--hqmp<#>-12    <filename>      file with interlaced reads for high-quality mate-pair library number <#> (<#> = 1,2,..,9)
--hqmp<#>-1     <filename>      file with forward reads for high-quality mate-pair library number <#> (<#> = 1,2,..,9)
--hqmp<#>-2     <filename>      file with reverse reads for high-quality mate-pair library number <#> (<#> = 1,2,..,9)
--hqmp<#>-s     <filename>      file with unpaired reads for high-quality mate-pair library number <#> (<#> = 1,2,..,9)
--hqmp<#>-<or>  orientation of reads for high-quality mate-pair library number <#> (<#> = 1,2,..,9; <or> = fr, rf, ff)
--nxmate<#>-1   <filename>      file with forward reads for Lucigen NxMate library number <#> (<#> = 1,2,..,9)
--nxmate<#>-2   <filename>      file with reverse reads for Lucigen NxMate library number <#> (<#> = 1,2,..,9)
--sanger        <filename>      file with Sanger reads
--pacbio        <filename>      file with PacBio reads
--nanopore      <filename>      file with Nanopore reads
--tslr  <filename>      file with TSLR-contigs
--trusted-contigs       <filename>      file with trusted contigs
--untrusted-contigs     <filename>      file with untrusted contigs

Pipeline options:
--only-error-correction runs only read error correction (without assembling)
--only-assembler        runs only assembling (without read error correction)
--careful               tries to reduce number of mismatches and short indels
--continue              continue run from the last available check-point
--restart-from  <cp>    restart run with updated options and from the specified check-point ('ec', 'as', 'k<int>', 'mc')
--disable-gzip-output   forces error correction not to compress the corrected reads
--disable-rr            disables repeat resolution stage of assembling

Advanced options:
--dataset       <filename>      file with dataset description in YAML format
-t/--threads    <int>           number of threads
                                [default: 16]
-m/--memory     <int>           RAM limit for SPAdes in Gb (terminates if exceeded)
                                [default: 250]
--tmp-dir       <dirname>       directory for temporary files
                                [default: <output_dir>/tmp]
-k              <int,int,...>   comma-separated list of k-mer sizes (must be odd and
```



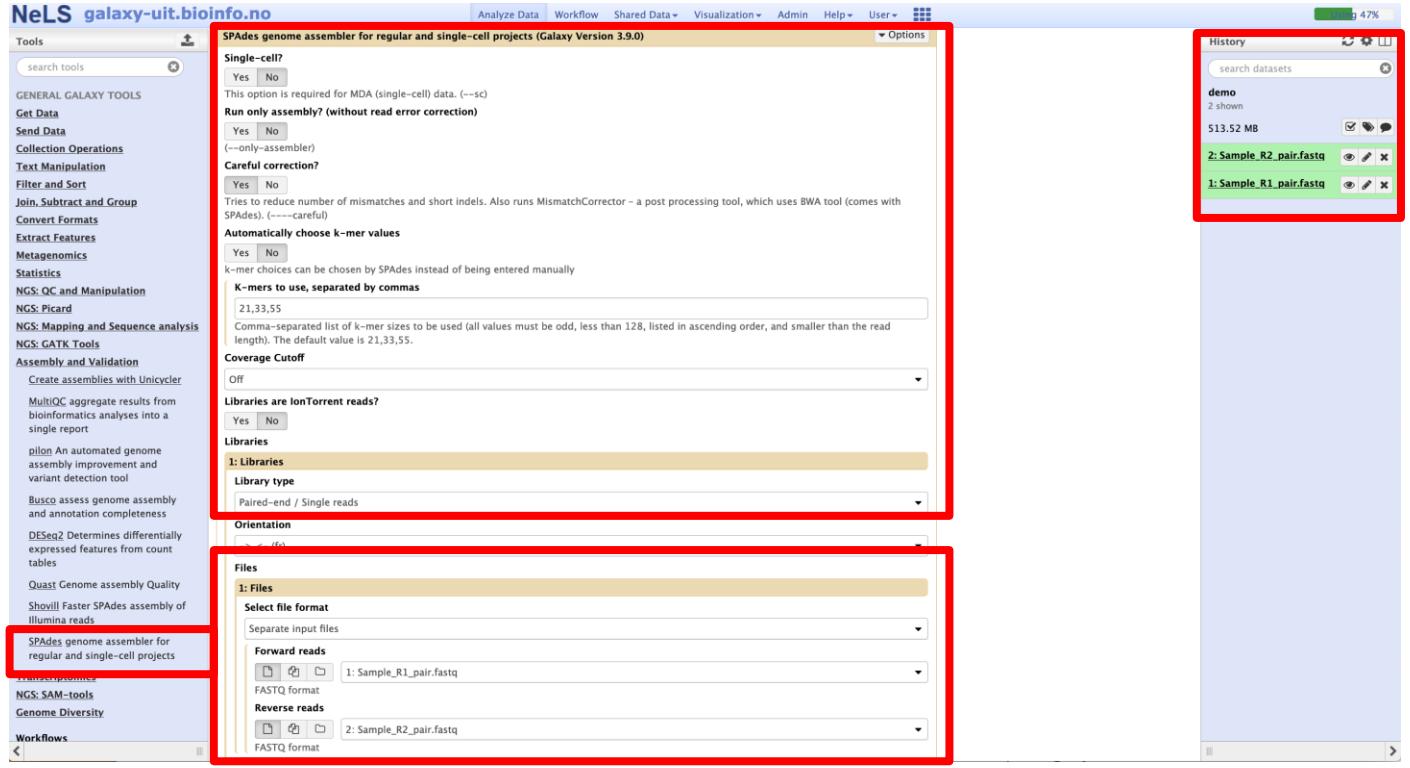
elixir NORWAY

Galaxy - Tools

Example running SPAdes

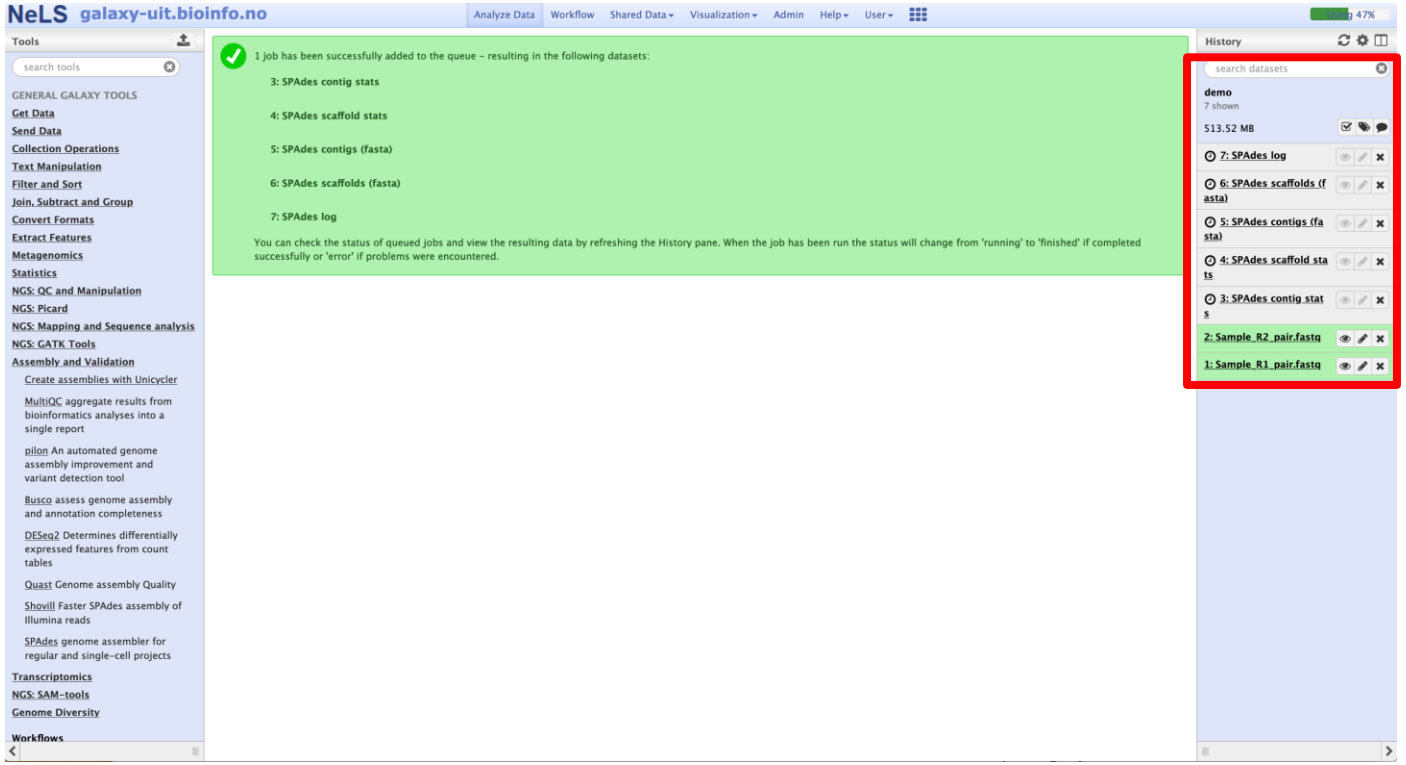
Choose parameter settings

Select input files



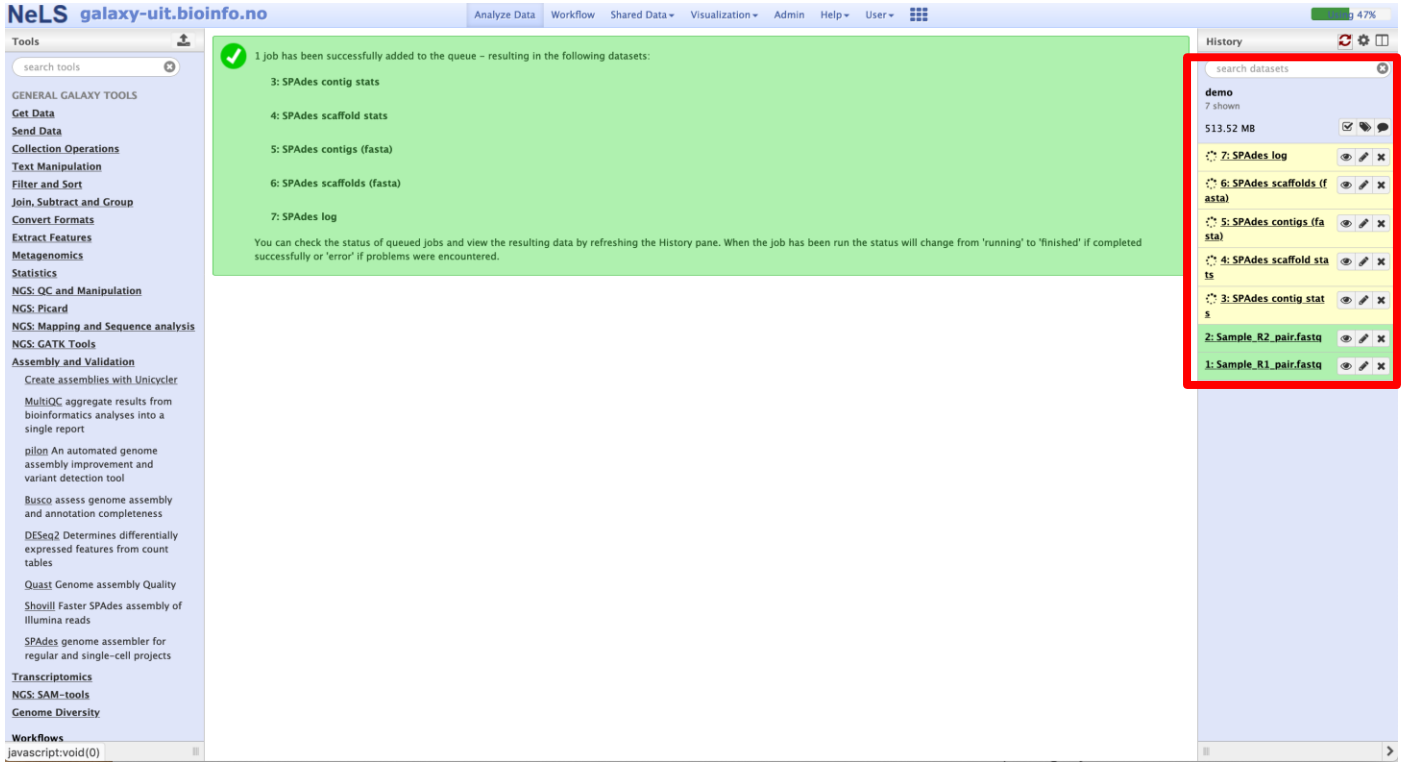
Galaxy – Running tools

The result files are first displayed in grey boxes. This means the job is pending.



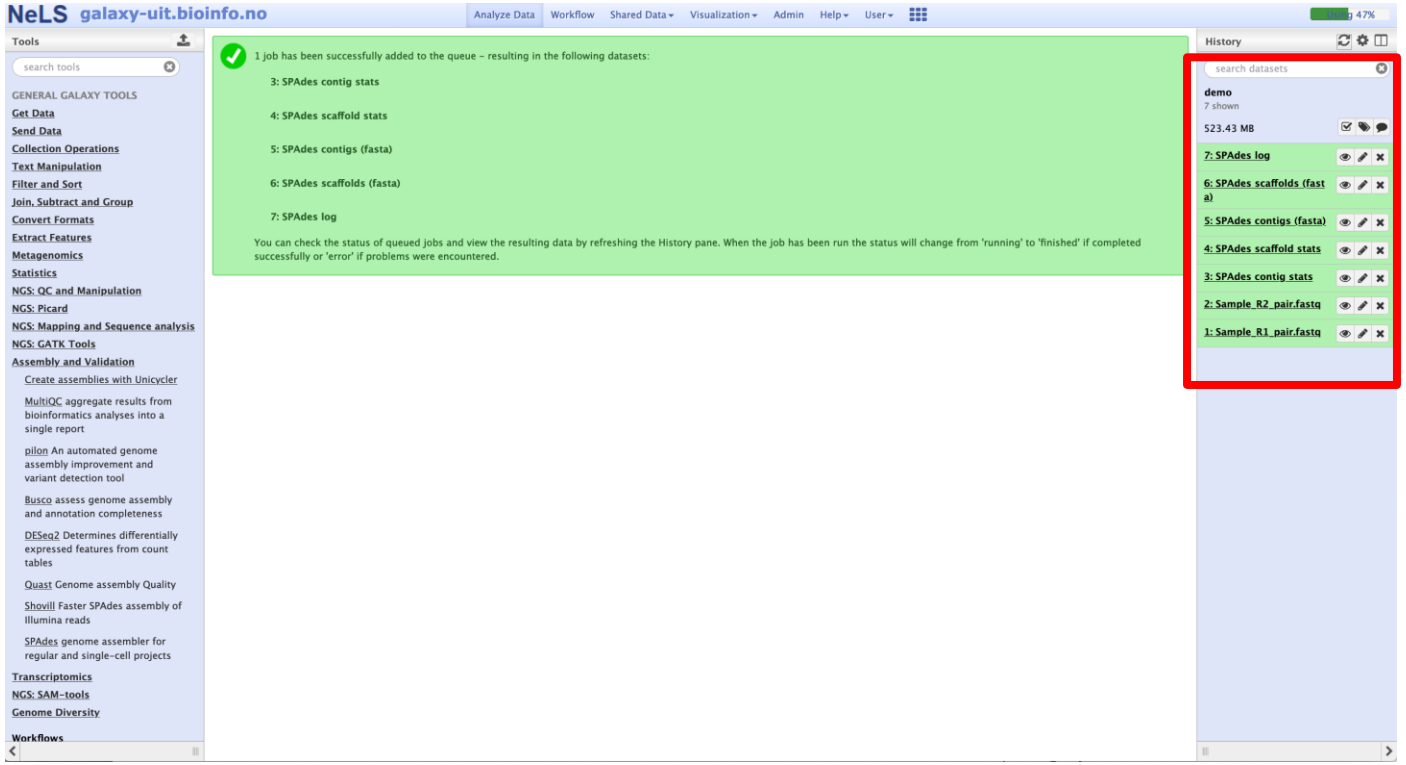
Galaxy – Running tools

When the job is running, the files turns yellow



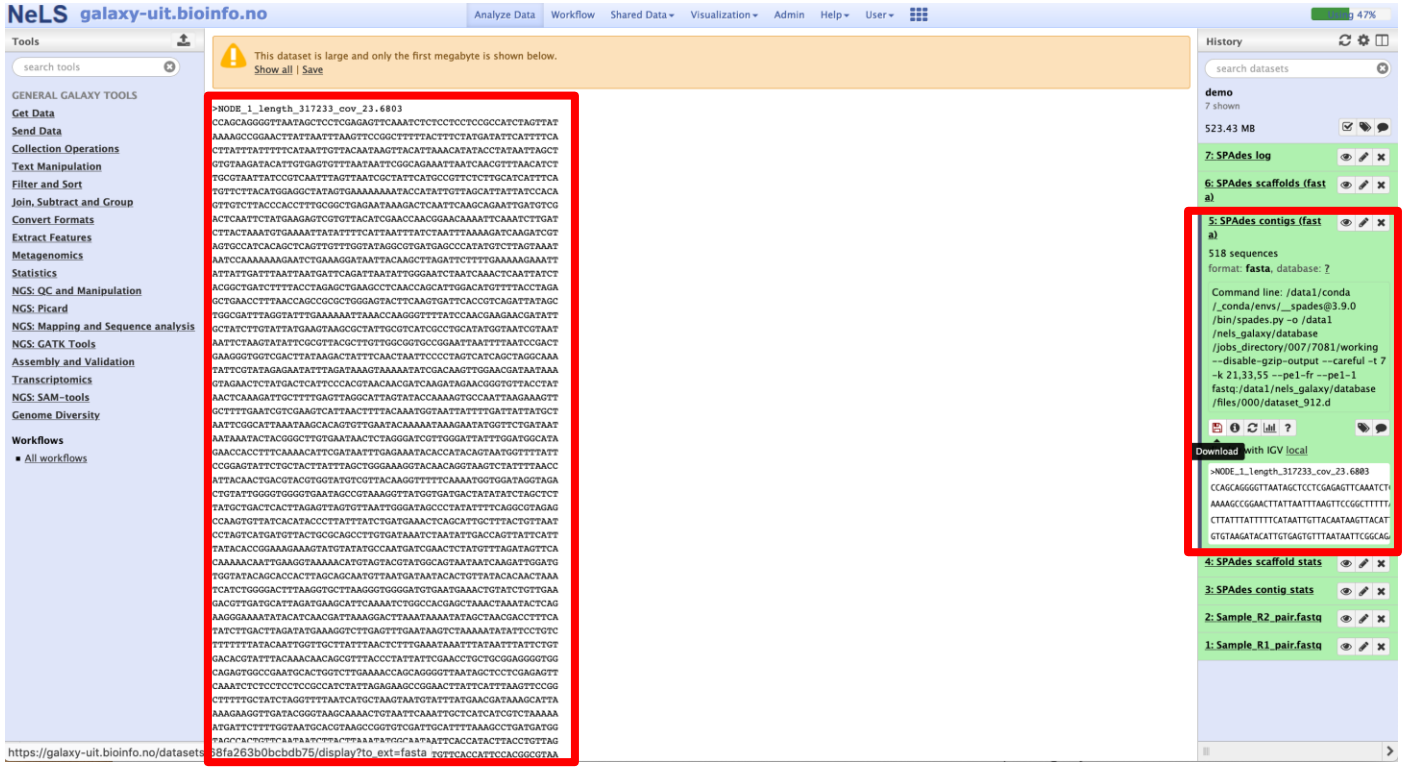
Galaxy – Running tools

When the job finish successfully, the files turn green



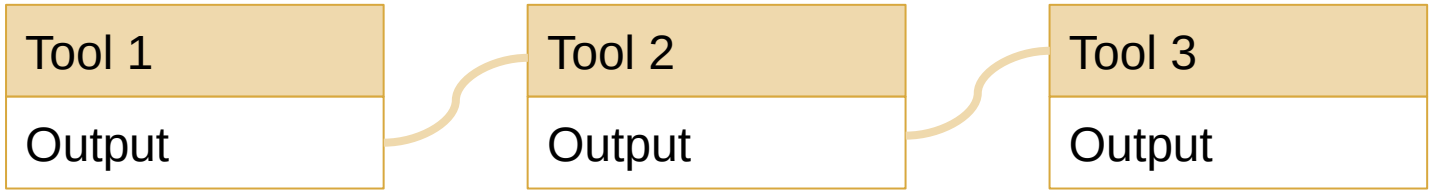
Galaxy – Running tools

You can view and download the result files



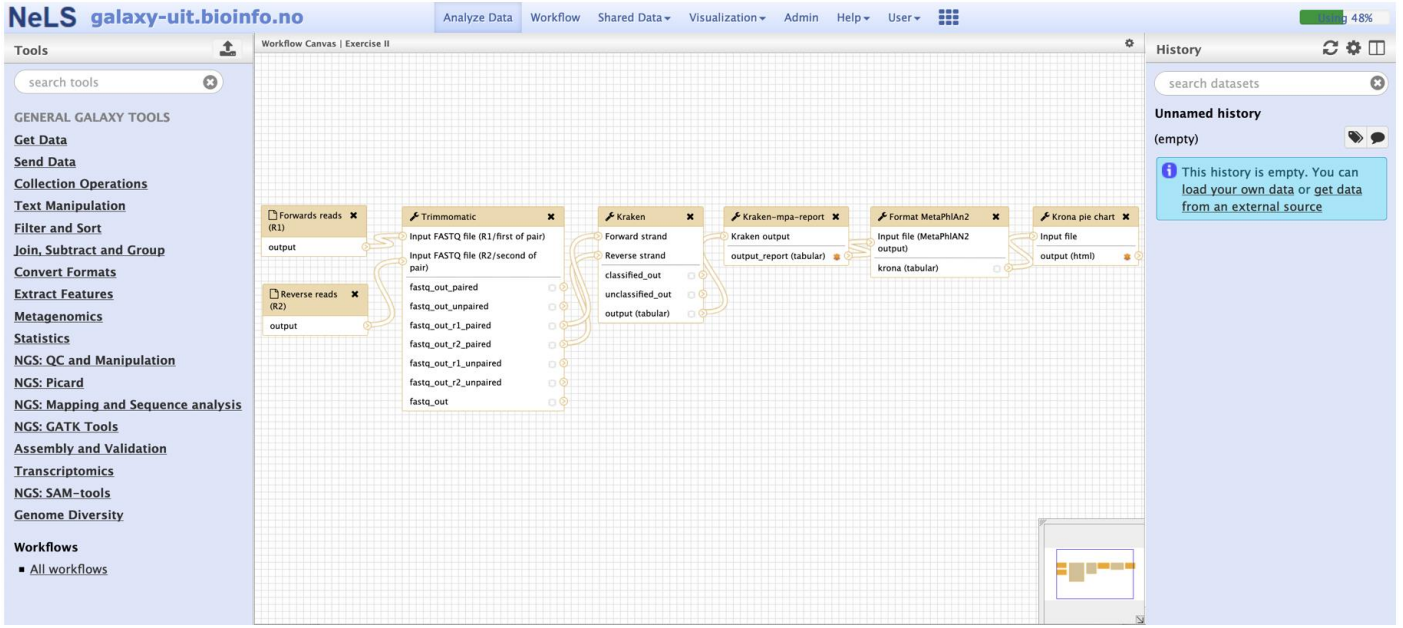
Galaxy workflows

A workflow in Galaxy is basically a string of tools, where the output from one tool becomes the input for the next



Galaxy - Workflows

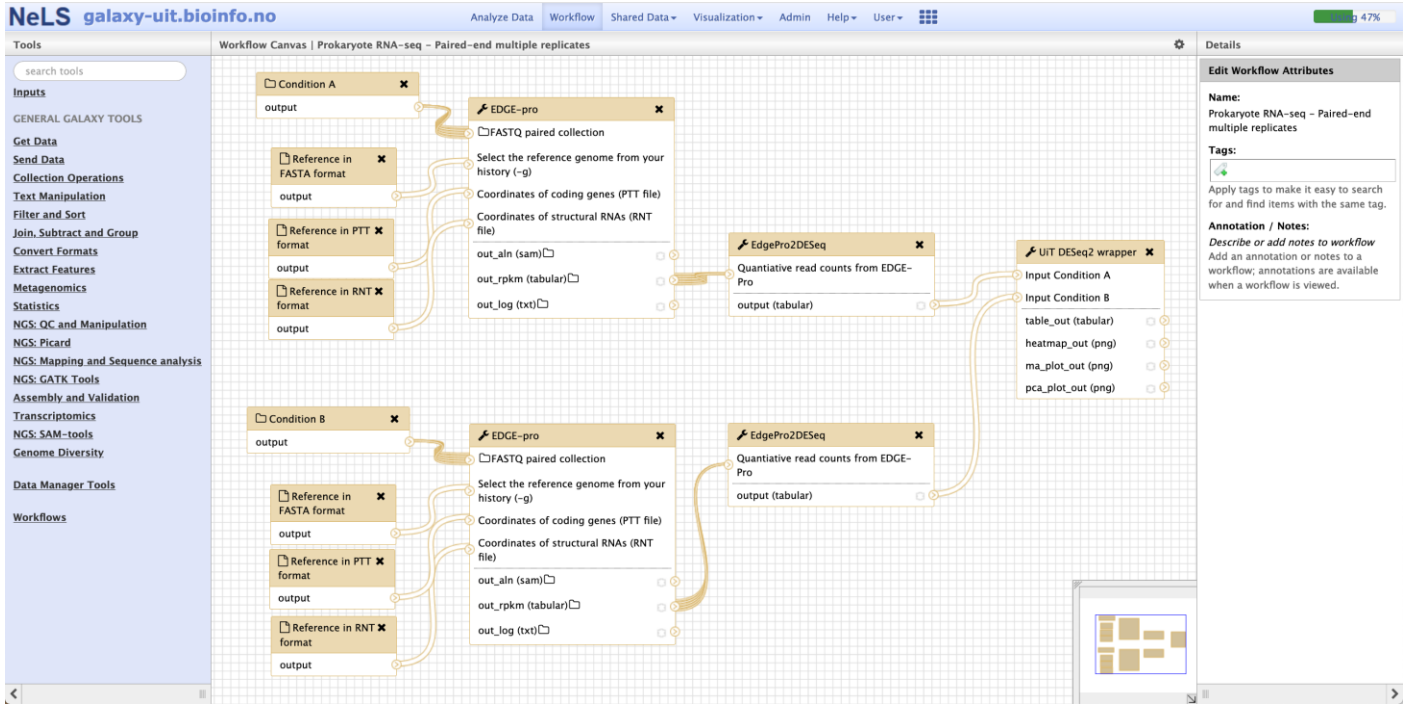
The “nodules” indicate which output file from one acts as input for the next tool



Galaxy - Workflows

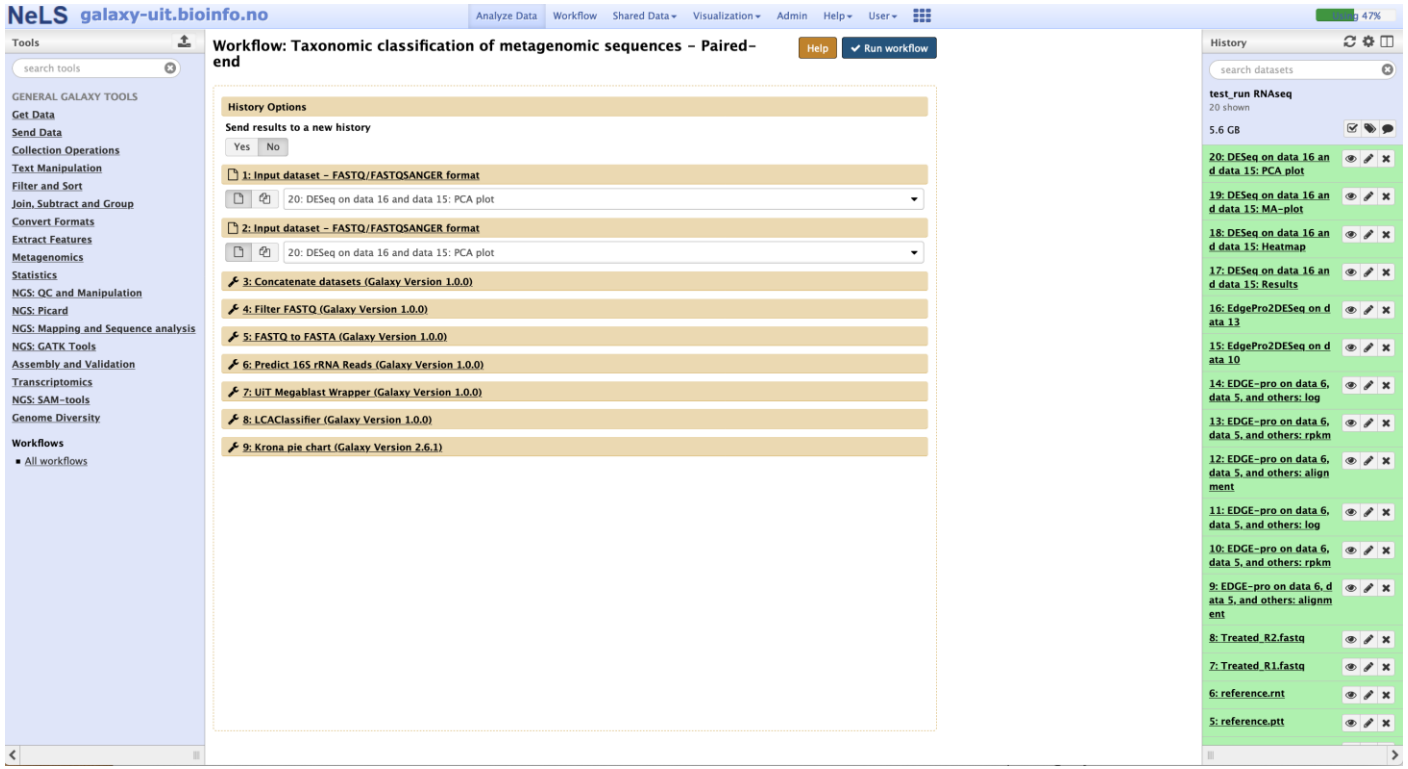
The “nodules” indicate which output file from one acts as input for the next tool

Multiple files can act as input for a tool



Galaxy – Running workflows

You can view and download the result files

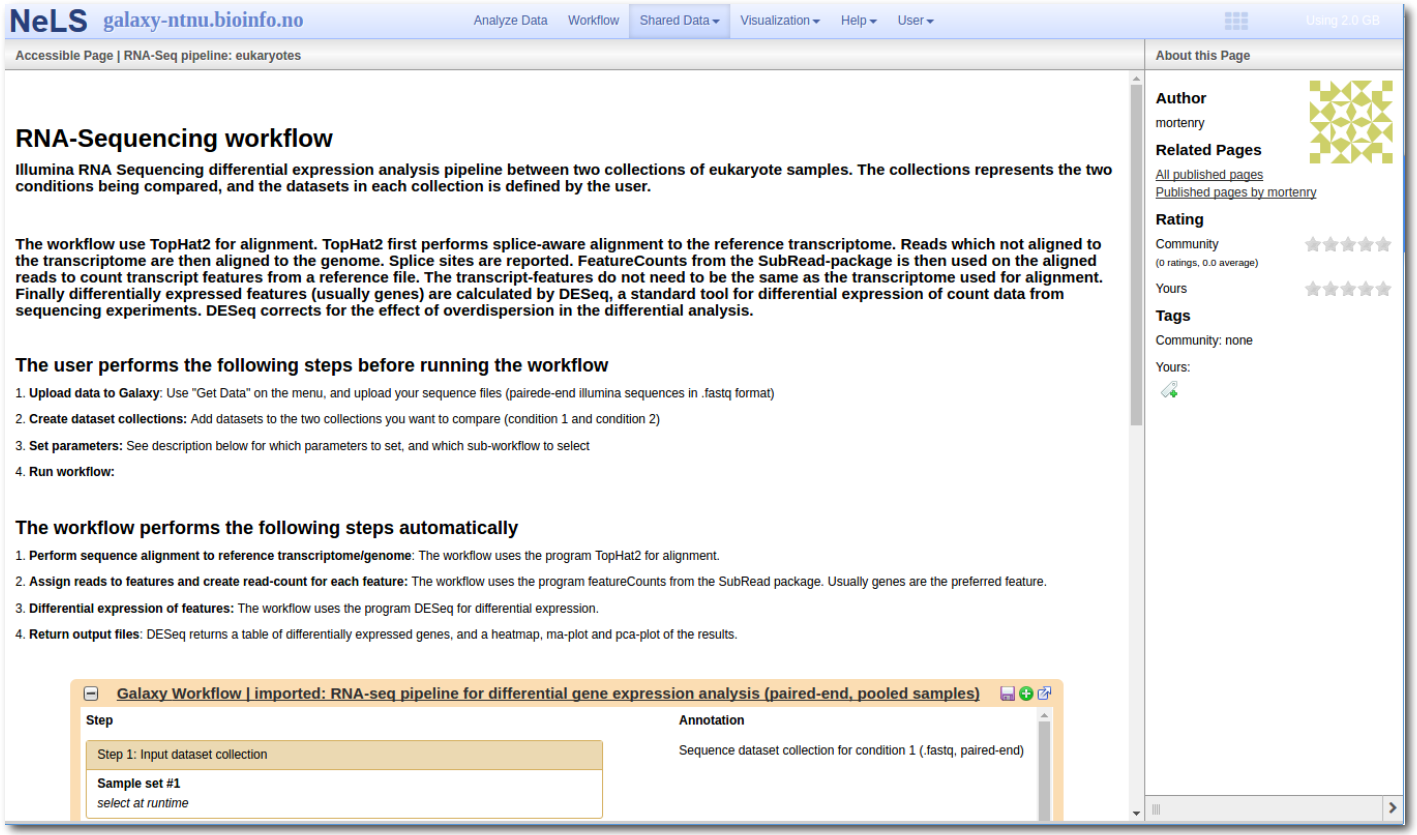


elixir NORWAY

Galaxy – Workflow documentation

Each NeLS supported workflow is documented with instructions how to use it

Also test data sets available for each workflow

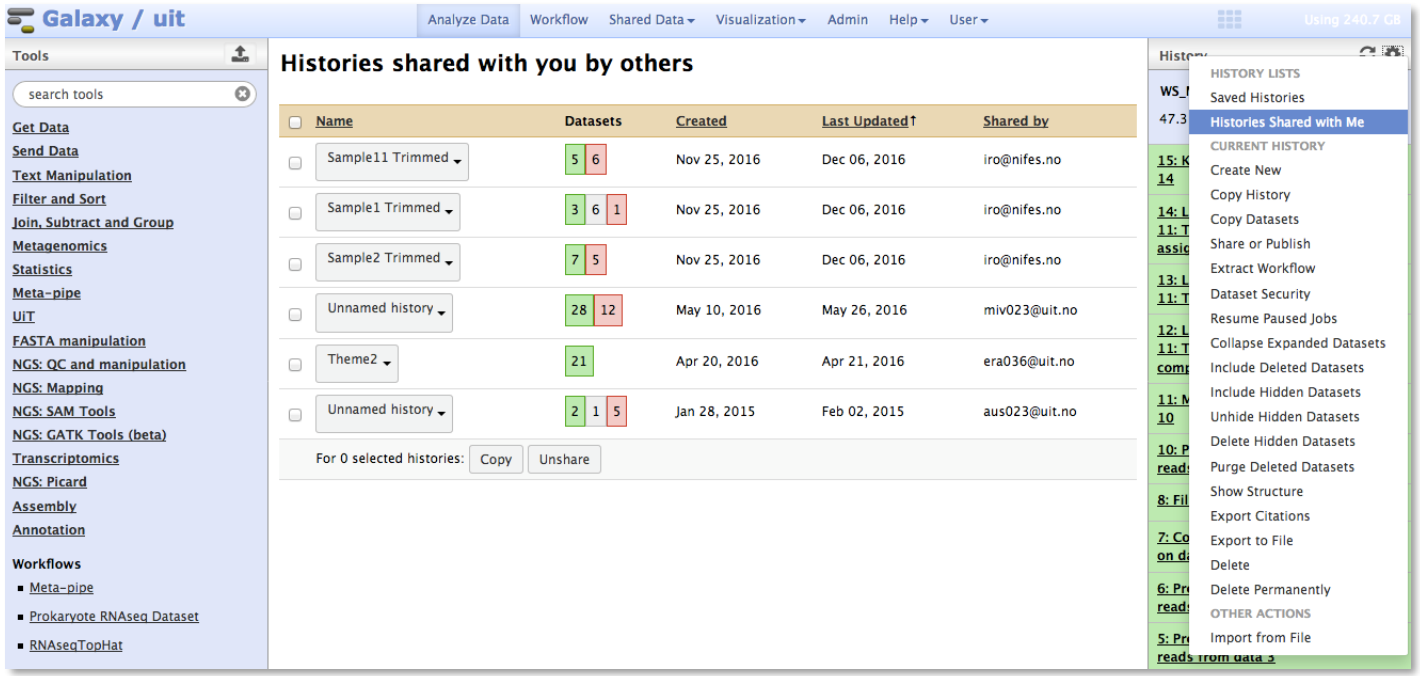


Galaxy - Additional features

Histories can be named and saved

Histories can easily be shared with other users on same galaxy server

Workflows can easily be shared with other users on same galaxy server



Galaxy – New tools

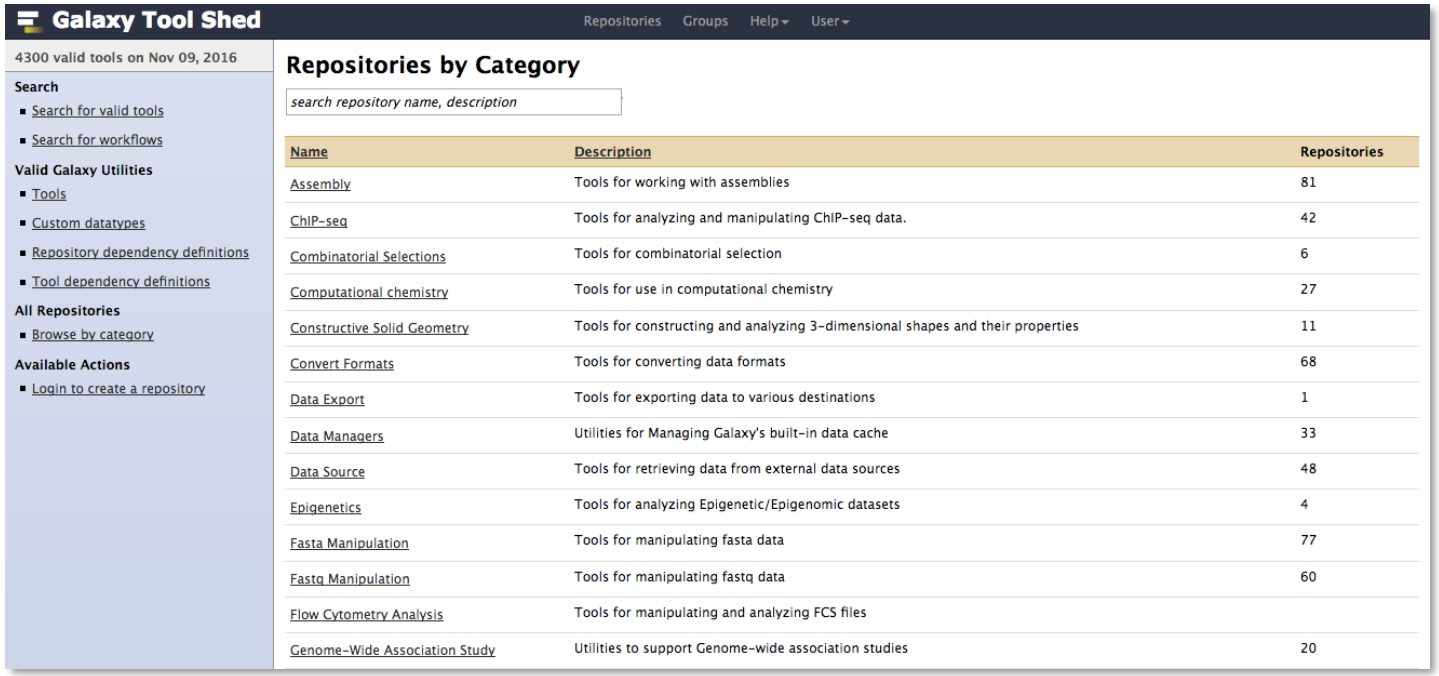
Tool shed = appstore

Central repository for bioinformatic tools

Each tool has a wrapper

Necessary for Galaxy to execute the software

Hosted by the developers of Galaxy



elixir NORWAY

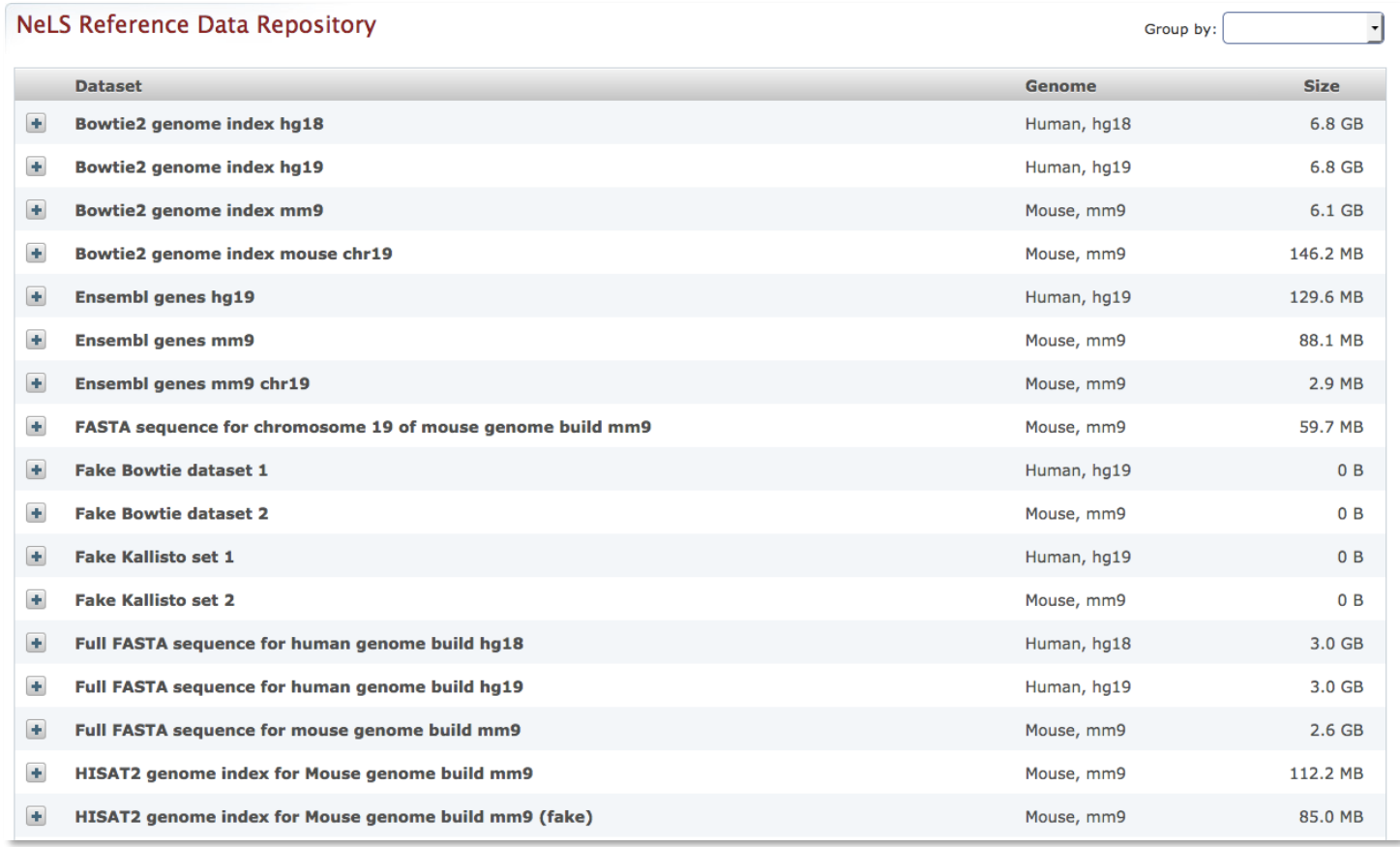
Galaxy – Helpdesk assistance

New reference data

New tools/wrappers

New public workflows if needed by many

Hands-on workshops regularly


NeLS Reference Data Repository

Group by:

Dataset	Genome	Size
Bowtie2 genome index hg18	Human, hg18	6.8 GB
Bowtie2 genome index hg19	Human, hg19	6.8 GB
Bowtie2 genome index mm9	Mouse, mm9	6.1 GB
Bowtie2 genome index mouse chr19	Mouse, mm9	146.2 MB
Ensembl genes hg19	Human, hg19	129.6 MB
Ensembl genes mm9	Mouse, mm9	88.1 MB
Ensembl genes mm9 chr19	Mouse, mm9	2.9 MB
FASTA sequence for chromosome 19 of mouse genome build mm9	Mouse, mm9	59.7 MB
Fake Bowtie dataset 1	Human, hg19	0 B
Fake Bowtie dataset 2	Mouse, mm9	0 B
Fake Kallisto set 1	Human, hg19	0 B
Fake Kallisto set 2	Mouse, mm9	0 B
Full FASTA sequence for human genome build hg18	Human, hg18	3.0 GB
Full FASTA sequence for human genome build hg19	Human, hg19	3.0 GB
Full FASTA sequence for mouse genome build mm9	Mouse, mm9	2.6 GB
HISAT2 genome index for Mouse genome build mm9	Mouse, mm9	112.2 MB
HISAT2 genome index for Mouse genome build mm9 (fake)	Mouse, mm9	85.0 MB

elixir NORWAY

Future developments

Merge all five NeLS Galaxy into one – usegalaxy.no

Direct data upload and archiving in ENA from NeLS

Development of more analysis workflows (on demand)

Integration of Jupyter notebooks in Galaxy