



Introduction to NeLS and Galaxy

The national bioinformatics infrastructure

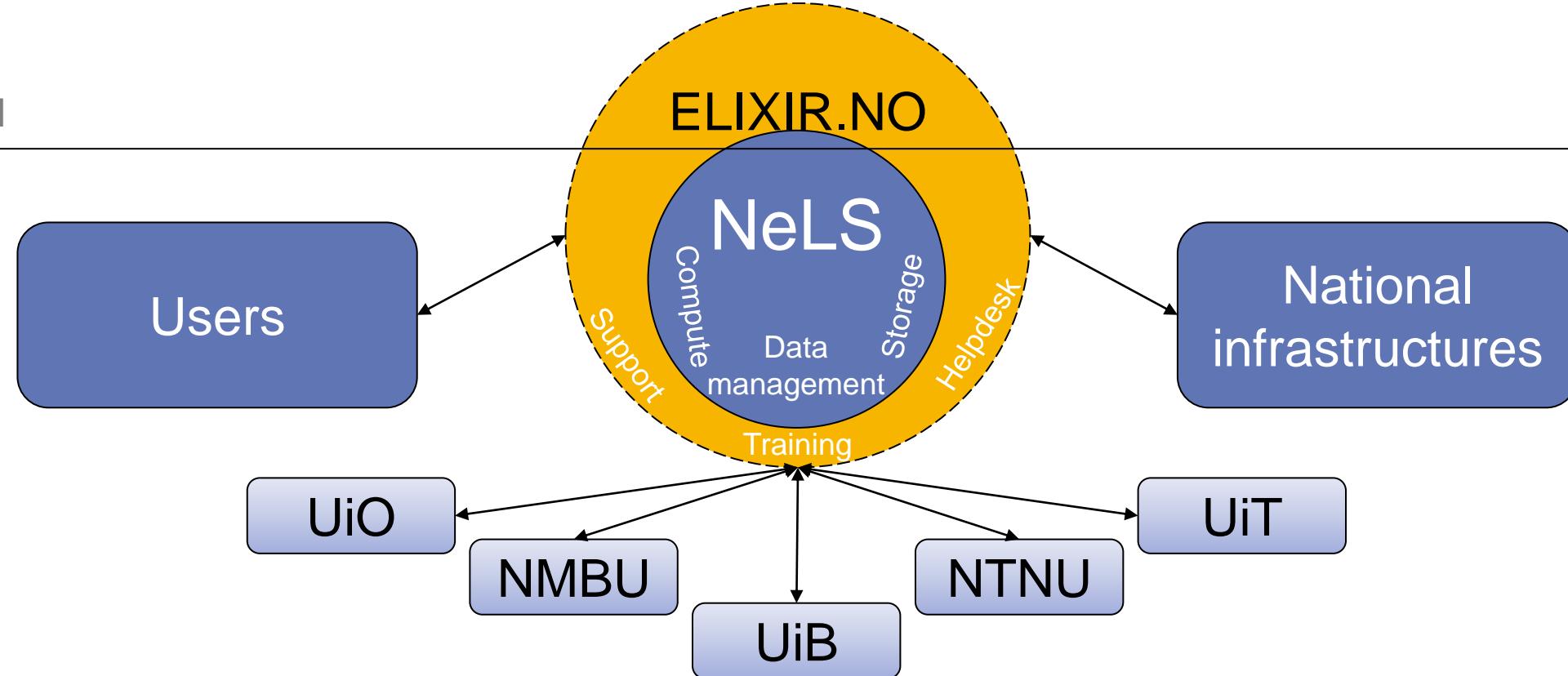


Erik Hjerde

NeLS - the Norwegian e-Infrastructure for Life Science

International

National



Data
management

Data storage

Data sharing

Data analysis

Overview of data management

FAIR data management from sequencing provider to public repository

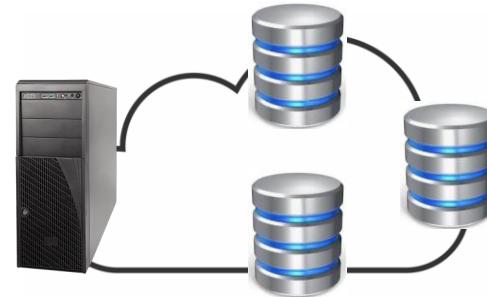
Data production



Data storage/sharing



Data compute



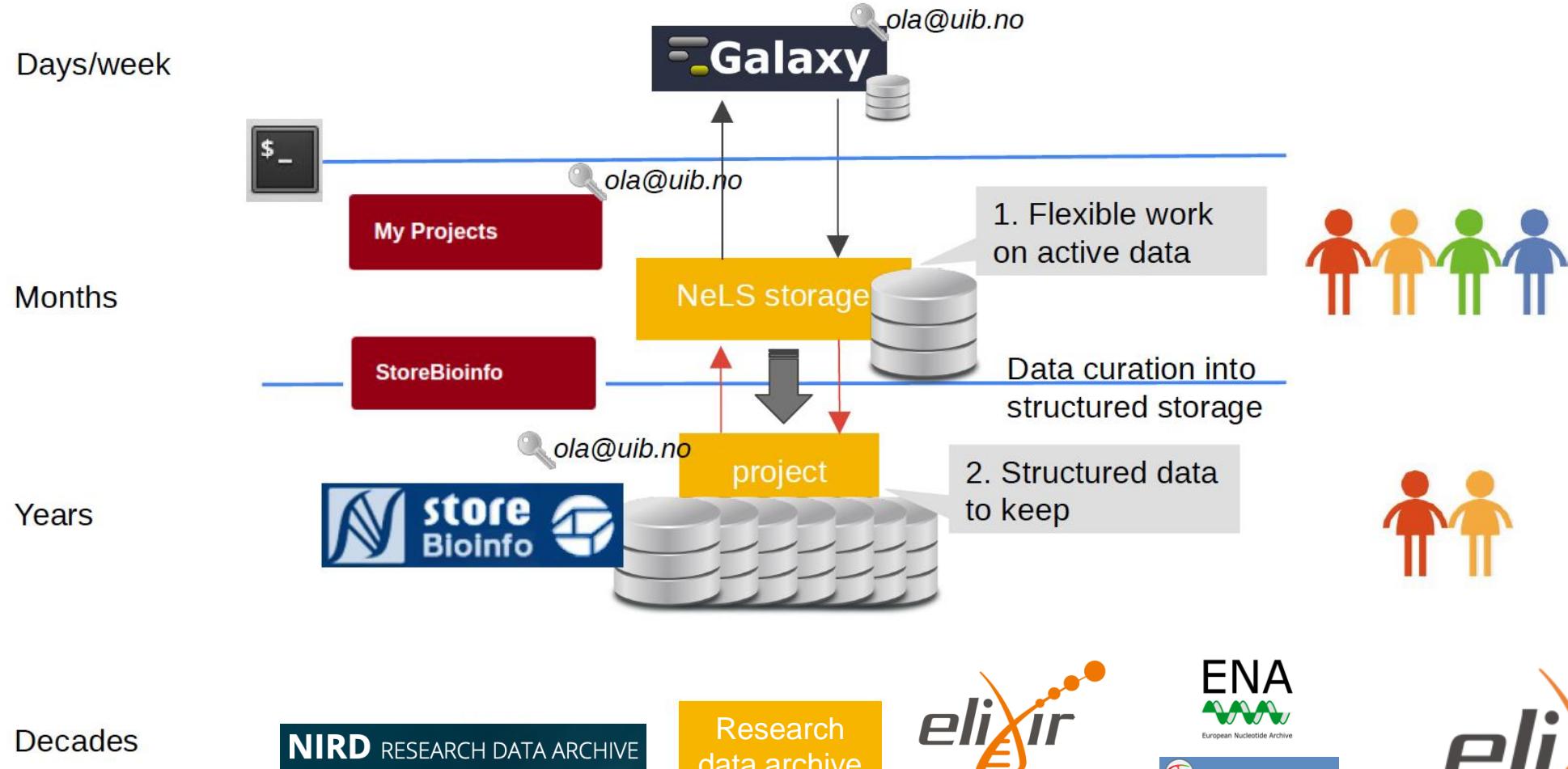
Data archiving



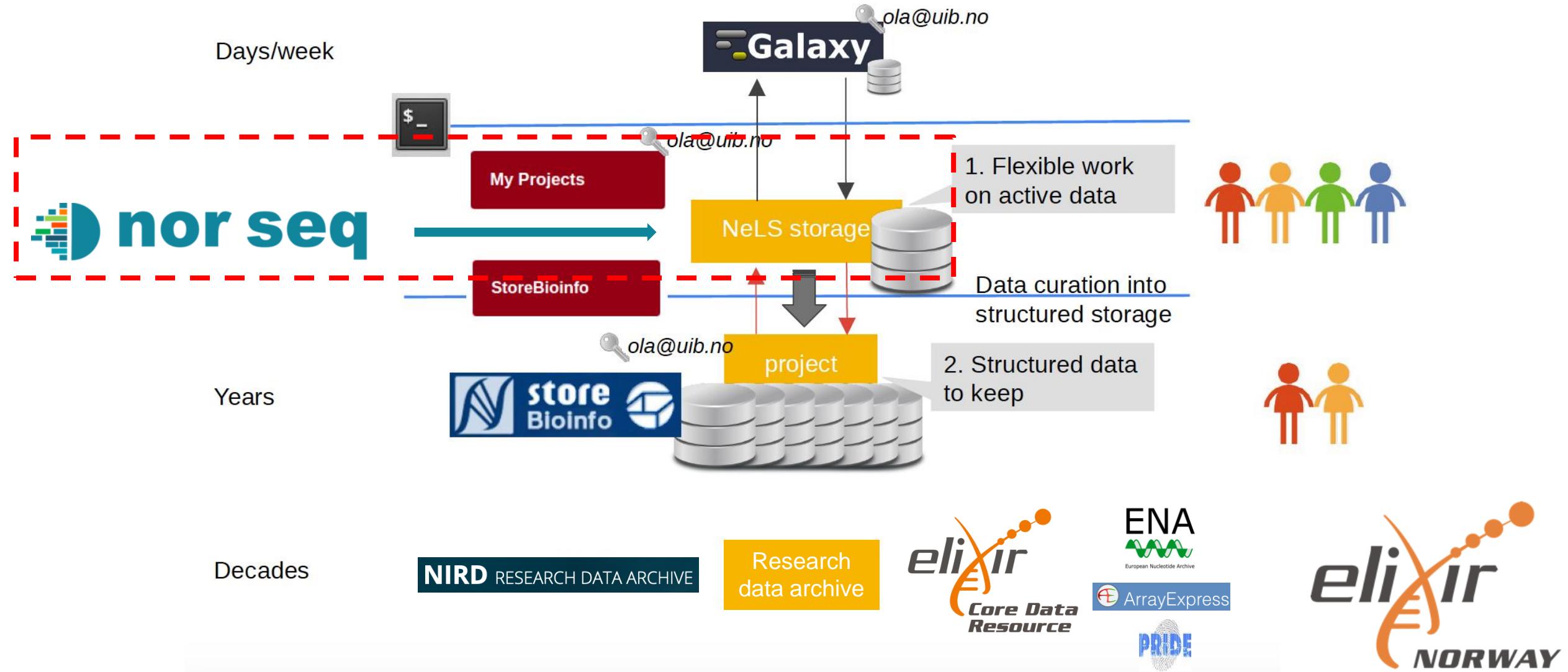
Data management



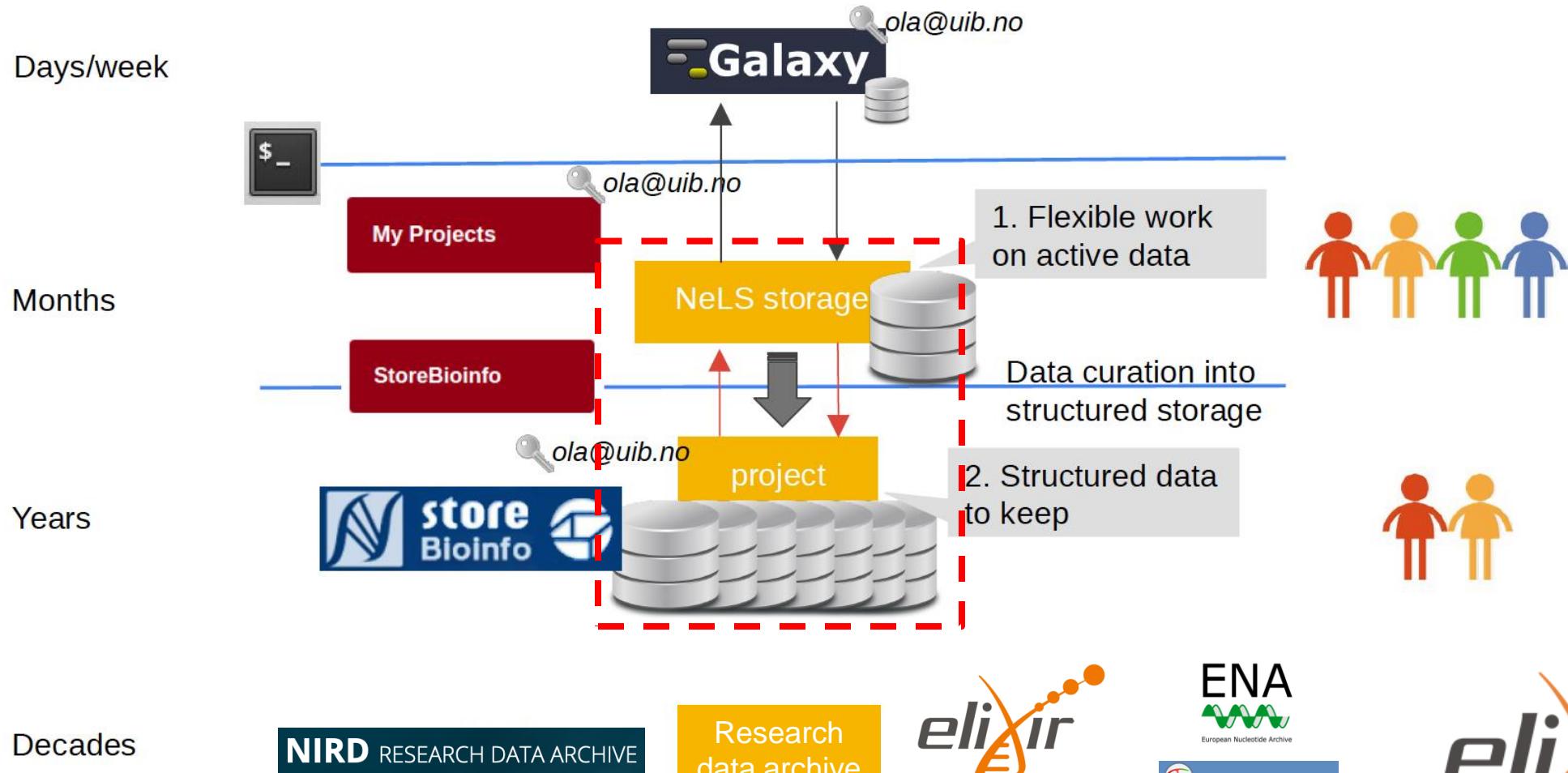
NeLS architecture



Data from data providers



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)

The screenshot shows the NeLS portal interface. At the top, there's a blue header bar with the NeLS logo. Below it, the main content area is divided into three sections:

- Welcome !**: A brief introduction stating "This is NeLS portal for the administration of Norwegian e-Infrastructure for Life Sciences. NeLS is one of the packages of ELIXIR Norway."
- Features**: A list of features:
 - Federated login using FEIDE
 - Sharing of data
 - Bioinformatics pipelines
 - API for integration
- Help (Support)**: Instructions to contact contact@bioinfo.no for support.

In the center, there's a box for **ELIXIR Norway**, which includes logos for the University of Bergen, Oslo University, Trondheim University, Tromsø University, and the Research Council of Norway.

To the right, there's a box for **Access NeLS** titled "Log in with Feide". It displays a message from the NeLS Portal requesting log in with Feide, followed by a Feide login form with fields for username and password, and links for forgot password and help.



Access to storage and sharing of data in NeLS

Access to personal storage and projects

The screenshot displays the NeLS (Norwegian e-Infrastructure for Life Sciences) web application interface. The top navigation bar includes links for NeLS, Dashboard, Personal, Projects, StoreBioInfo, and Help (Support). A user profile for "Erik Hjerde (NeLS ID :34)" is shown in the top right.

NeLS Disk Usage: A chart showing storage usage over time from September 2016 to May 2018. The total capacity is 111.8 GB, with usage starting at 93.1 GB and decreasing to 74.5 GB by May 2018. Buttons for "Upload" and "New Folder" are available.

My Data: A sidebar listing personal data items such as "Bio3323", "ELIXIR2-WP4", "Matt", "metagenome", and "Test2014". Actions like Copy, Cut, Paste, Delete Selected, and Clear Selection are provided.

Account Information: Displays full name (Erik Hjerde), email (erik.hjerde@uit.no), and affiliation (University of Tromsø).

System Notice: A message indicating "Completed successfully" for a task.

Projects: A table listing projects with columns for Project Name, Role in Project, and Creation Date.

Project Name	Role in Project	Creation Date
Elixir_NO-Documents	Member Add files & folders, Navigate & Download	November 28, 2014
Uio-Test-Project	Member Add files & folders, Navigate & Download	March 19, 2015
UiO_Halvorsen_RNAseq_Mouse_Spleen_T_cells_2018	Power User Manage File system: Add, Rename, Navigate, Download, Delete all Content	February 22, 2018
Elixir_workshops	Member Add files & folders, Navigate & Download	April 15, 2016
UiB-Petersen_neuro_ma_regulation_disorder_2017	Power User Manage File system: Add, Rename, Navigate, Download, Delete all Content	May 11, 2017
maseq-test-data	Power User Manage File system: Add, Rename, Navigate, Download, Delete all Content	September 11, 2017
ELIXIR2-WP4	Power User Manage File system: Add, Rename, Navigate, Download, Delete all Content	November 8, 2017

Data Operations: A list of completed tasks with details like status, submission time, and completion percentage.

Task ID	Description	Status	Submitted Time	Completion %
5859	NeLS >> StoreBioInfo	Completed successfully	Tue Jul 02 2019 12:09:04 GMT+0200 (Central European Summer Time)	100 %
5858	NeLS >> StoreBioInfo	Completed successfully	Tue Jul 02 2019 11:07:41 GMT+0200 (Central European Summer Time)	100 %
5857	NeLS >> StoreBioInfo	Completed successfully	Tue Jul 02 2019 10:54:26 GMT+0200 (Central European Summer Time)	100 %
5856	NeLS >> StoreBioInfo	Completed successfully	Tue Jul 02 2019 10:41:28 GMT+0200 (Central European Summer Time)	100 %
5855	NeLS >> StoreBioInfo	Completed successfully	Tue Jul 02 2019 10:33:52 GMT+0200 (Central European Summer Time)	100 %
5854	NeLS >> StoreBioInfo	Completed successfully	Tue Jul 02 2019 10:22:22 GMT+0200 (Central European Summer Time)	100 %

Footer: © 2019, Norwegian e-Infrastructure for Life Sciences. All rights reserved. Links to https://nels.bioinfo.no/pages/file-browse.xhtml?path=cUZMaldtN and https://nels.bioinfo.no/pages/file-browse.xhtml?path=cUZMaldtN.

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

NeLS Dashboard Personal Projects StoreBioInfo Help (Support) Erik Hjerde (Nels ID :34)

NeLS You are at : » / Personal Projects

StoreBioinfo You are at : » StoreBioinfo

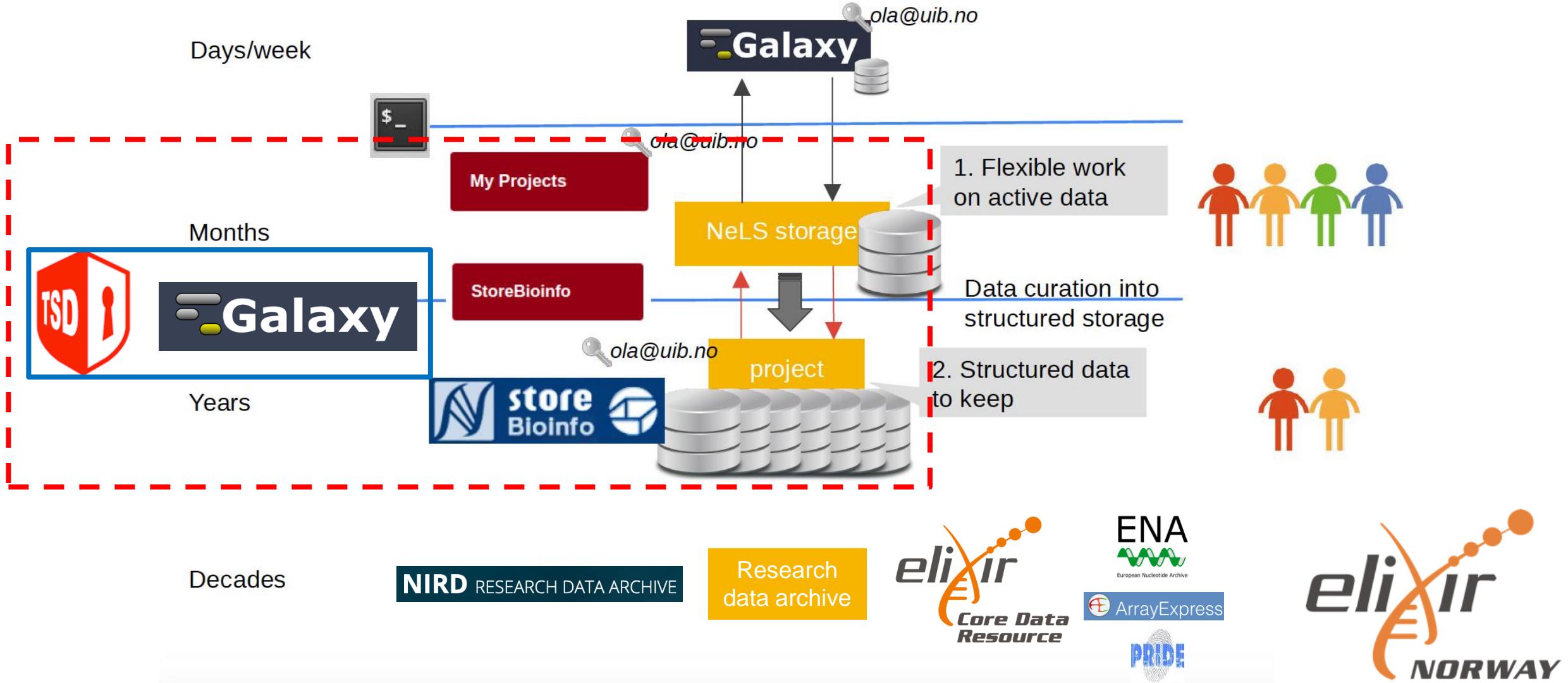
Projects

Name	Creation Date
UiT_bioinformatics_platform_core_storage_facilities_20	October 5, 2018
UNN_Goll_Metagenome_Human_IBS_Faeces_2018	March 19, 2019
UiB-Petersen_neuro_rna_regulation_disorder_2017	May 11, 2017
Elixir-demo	November 24, 2015

Data Operations

- 5859 : NeLS >> StoreBioinfo Status: Completed successfully Submitted time: Tue Jul 02 2019 12:09:04 GMT+0200 (Central European Summer Time) Completed: 100 %
- 5858 : NeLS >> StoreBioinfo Status: Completed successfully Submitted time: Tue Jul 02 2019 11:07:41 GMT+0200 (Central European Summer Time) Completed: 100 %
- 5857 : NeLS >> StoreBioinfo Status: Completed successfully Submitted time: Tue Jul 02 2019 10:54:26 GMT+0200 (Central European Summer Time) Completed: 100 %
- 5856 : NeLS >> StoreBioinfo Status: Completed successfully Submitted time: Tue Jul 02 2019 10:41:28 GMT+0200 (Central European Summer Time)

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

The screenshot shows the University of Oslo website with a dark header. The header includes the university logo, the text "UiO Universitetet i Oslo", and links for "For ansatte", "English website", and a search bar. Below the header, a navigation bar has links for "Forsiden", "Forskning", "Studier", "Livet rundt studiene", "Tjenester og verktøy" (which is highlighted in blue), "Om UiO", and "Personer". The main content area has a sidebar with links for "Tjenester og verktøy", "IT-tjenester", "IT-støtte i forskning", "Tjenester for sensitive forskningsdata" (which is highlighted in dark grey), and "Om TSD". The main content area is titled "Tjenester for Sensitive Data (TSD)" and contains text about the service, a logo for "TSD" (a red square with a white keyhole icon), and a list of links related to the service.

Tjenester for Sensitive Data (TSD)

TSD gir forskere ved UiO og ved andre offentlige forskningsinstitusjoner en forskningsplattform som oppfyller lovens strenge krav til behandling og lagring av sensitive forskningsdata. TSD utvikles og driftes av USIT ved UiO, og inngår i NorStore, den nasjonale infrastrukturen for håndtering og lagring av vitenskapelige data.

- En introduksjon til TSD
- Systembeskrivelse med risikovurderinger og whitepaper
- Medieomtale
- Aktive forskningsprosjekt
- Kontaktinformasjon

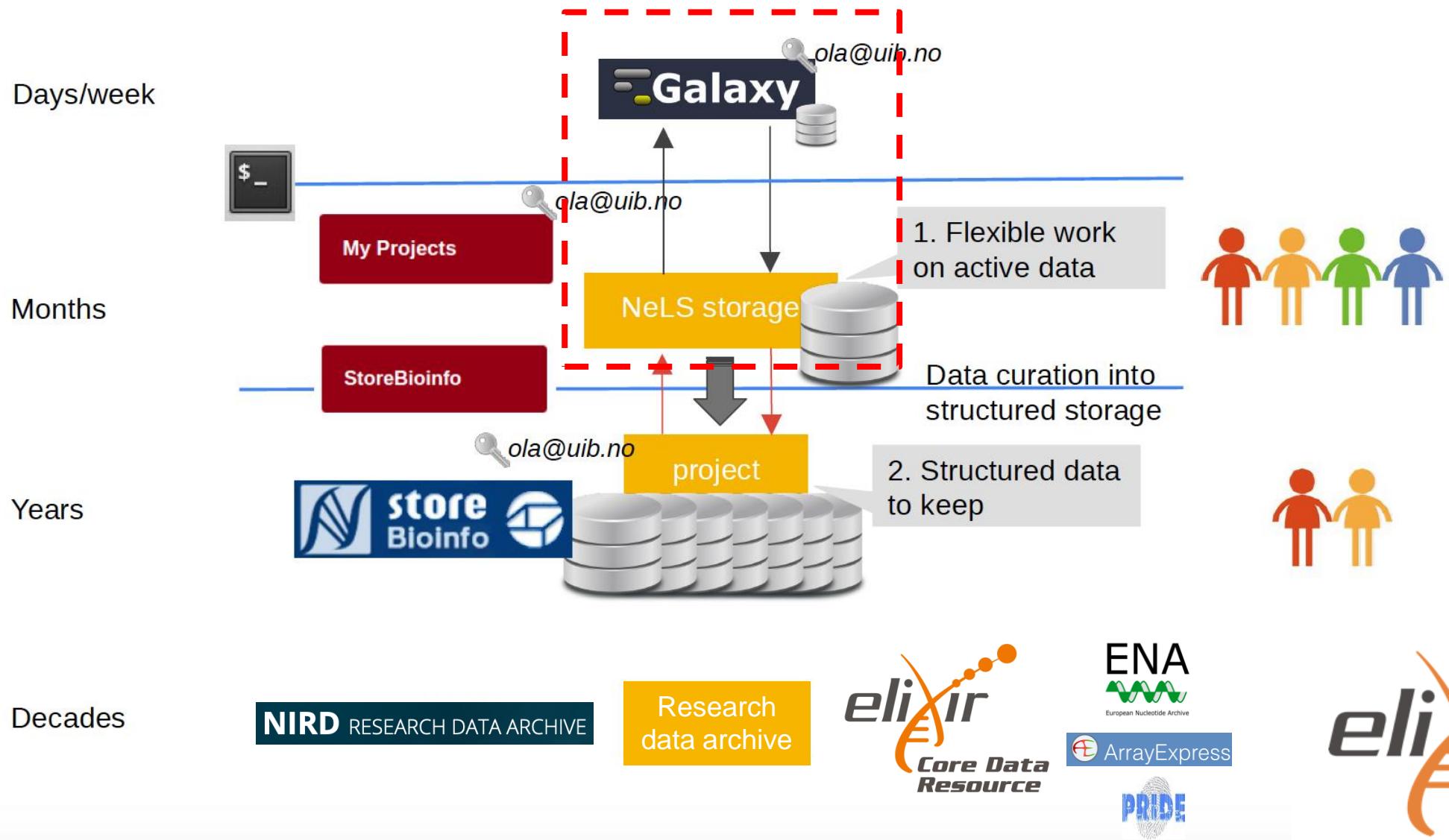
Hvem kontakter jeg?
→ Se kontaktpunkter for TSD

Aktuelt

- 2016-01-20 New feature: Web browser login to Windows VMs!
- 2016-04-27 New feature: /tsd/pxx/home visible from Colossus



Data analysis



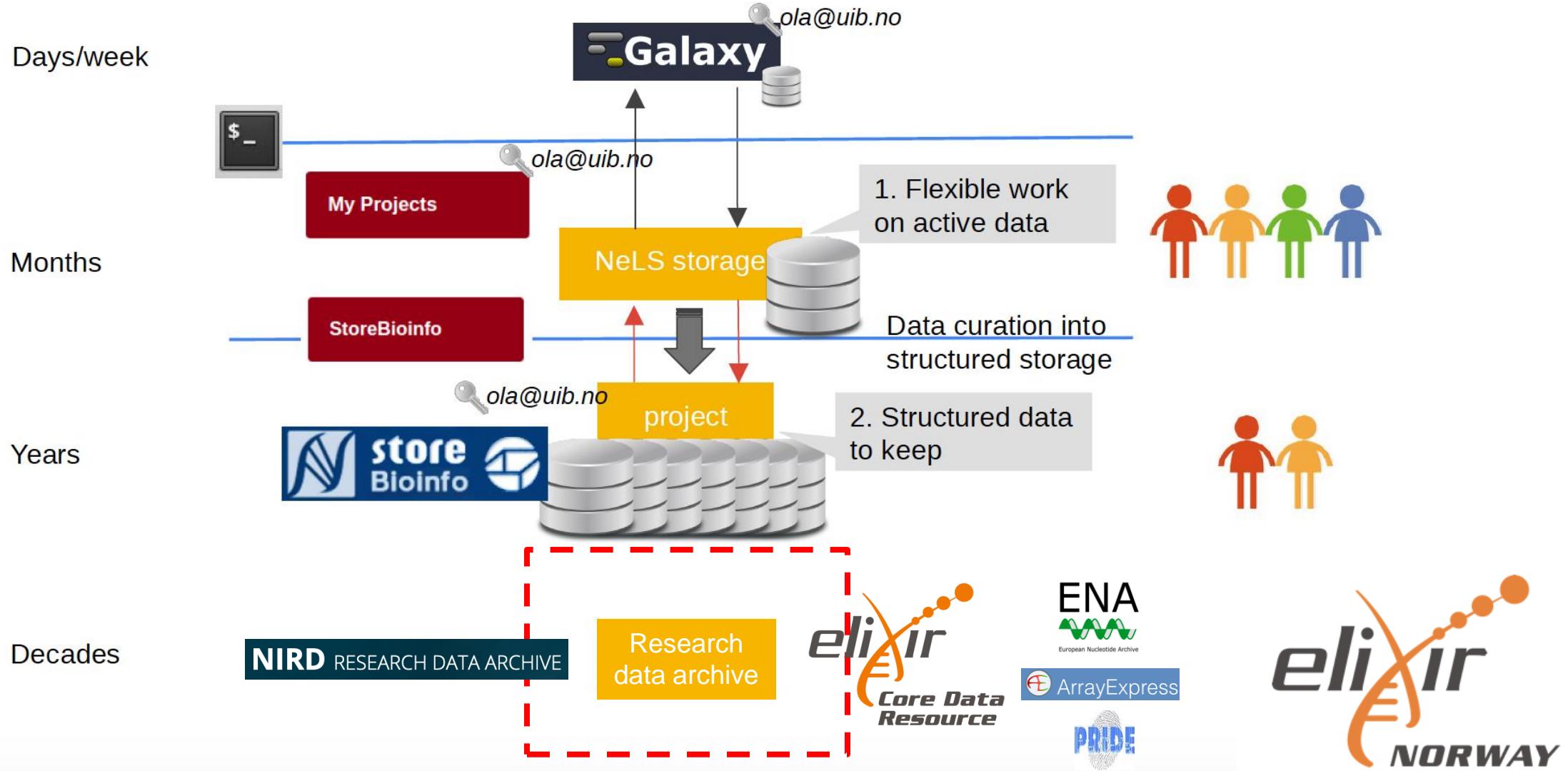
Data analysis in NeLS Galaxy

Each node host a Galaxy server with unique “flavour”

The screenshot shows the NeLS Galaxy interface for the Tromsø node. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The left sidebar lists various tools and workflows, such as 'GENERAL GALAXY TOOLS', 'Get Data', 'Collection Operations', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Metagenomics', 'Statistics', 'NGS: QC and Manipulation', 'NGS: Picard', 'NGS: Mapping and Sequence analysis', 'NGS: GATK Tools', 'Assembly and Validation', 'Transcriptomics', 'NGS: SAM-tools', and 'Genome Diversity'. A 'Workflows' section shows a list of all workflows. The main content area features a 'Welcome to the NeLS Galaxy installation in Tromsø' message, a 'Tweets' feed from @elixirnorway, and a 'History' panel showing a dataset named 'Copy of 'test Exercise I''. At the bottom, there's a section for 'Other NeLS Galaxy installations and resources' featuring logos for University of Oslo, University of Bergen, Norwegian University of Science and Technology, Norwegian Univ. of Life Sciences, lifeportal, NeLS, and NeLS Portal.



Data archiving in data repositories



Overview of data management

FAIR data management from sequencing provider to public repository

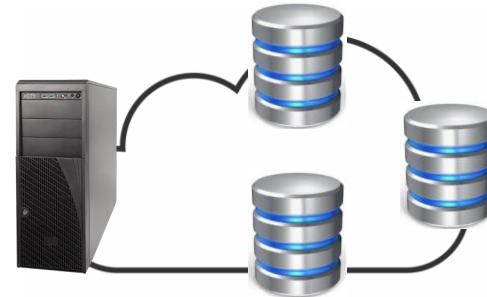
Data production



Data storage/sharing



Data compute



Data archiving



Data management



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

The screenshot shows the NeLS Galaxy installation in Tromsø. The interface includes:

- Tools Catalog:** A sidebar on the left lists various tool categories such as General Galaxy Tools, NGS, Metagenomics, Statistics, and Workflows.
- Welcome Page:** The main content area displays a welcome message, system statistics (117.9 GB used of 250.0 GB), and links for Quick Start Guide and additional documentation.
- History Panel:** On the right, a panel titled "History" shows a list of recent dataset operations, including "Copy of 'test Exercise I'" and "Concatenate datasets on data 15 and data 14".
- Tweets:** A section titled "Tweets" by @eliximorway shows a tweet from Inge Jonassen (@ingejonassen) about inspiring days at ELIXIR Norway.
- Tools and Workflows:** A section titled "Tools and Workflows" provides information on ELIXIR.NO aims and links to the NELS Portal.
- Other NeLS Galaxy installations and resources:** Logos for University of Oslo, University of Bergen, Norwegian University of Science and Technology, Norwegian Univ. of Life Sciences, lifeportal, and NeLS Portal.



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

The screenshot shows the NeLS Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User, and a grid icon. The left sidebar contains a 'Tools' section with a search bar and a list of categories: GENERAL GALAXY TOOLS (Get Data, Send Data, Collection Operations, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Metagenomics, Statistics, NGS: QC and Manipulation, NGS: Picard, NGS: Mapping and Sequence analysis, NGS: GATK Tools, Assembly and Validation, Transcriptomics, NGS: SAM-tools, Genome Diversity), and Workflows (All workflows). The main content area features a 'Welcome to the NeLS Galaxy installation in Tromsø' message. It explains Galaxy as a web-based platform for data intensive life science research, mentioning its graphical interface, interactive execution of tools, and support for workflows. It also notes disc usage limitations (117.9 GB used, 250.0 GB total) and encourages users to move files to NeLS Storage after use. A 'Tweets' section displays a tweet from @eliximorway about ELIXIR Norway Retweeting, followed by a reply from Inge Jonassen (@ingejonassen) about inspiring days at the ELIXIR Norway meeting. The right side of the interface shows a 'History' panel with a list of datasets: 'Copy of "test Exercise I"' (3 shown, 1 deleted, 12 hidden), '8.5 KB' (file2.txt), and '1: file1.txt' (both with edit and delete icons). At the bottom, there are logos for other NeLS Galaxy installations: University of Oslo, University of Bergen, Norwegian University of Science and Technology, Norwegian Univ. of Life Sciences, lifeportal, and NeLS Portal.



Galaxy – The basics

Main menu

Tool menu

Main window

The screenshot shows the main interface of the NeLS Galaxy installation. At the top is the main menu bar with options: Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User, and a grid icon. Below the menu is a red box highlighting the top navigation area. The central part of the screen is the main window, which includes a left sidebar titled 'Tools' containing categories like GENERAL GALAXY TOOLS, Collection Operations, Text Manipulation, and Workflows. The main content area displays a welcome message, a 'Tweets' section from @eliximorway, and a 'Tools and Workflows' section. A red box highlights the 'Tools' sidebar. To the right is a 'History' panel showing a list of datasets: 'Copy of "test Exercise I"', '16: Concatenate datasets on data 15 and data 14', '2: file2.txt', and '1: file1.txt'. A red box highlights the 'History' panel.

Tool menu

Main menu

Main window

History



Galaxy - Tools

Tools are available from the Tool menu

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

The screenshot shows the Galaxy web interface with the following details:

- Title Bar:** NeLS galaxy-uit.bioinfo.no, SPAdes genome assembler for regular and single-cell projects (Galaxy Version 3.9.0)
- Tool Panel (Left):** A sidebar titled "Tools" with a search bar. It lists various Galaxy tools categorized under "GENERAL GALAXY TOOLS" and "NGS: QC and Manipulation". The "SPAdes genome assembler for regular and single-cell projects" tool is highlighted with a red box.
- Tool Configuration (Main Area):**
 - Single-cell?**: Yes (radio button selected).
 - Run only assembly? (without read error correction)**: Yes (radio button selected).
 - Careful correction?**: Yes (radio button selected).
 - Automatically choose k-mer values**: Yes (radio button selected).
 - K-mers to use, separated by commas**: 21,33,55
 - Coverage Cutoff**: Off
 - Libraries are IonTorrent reads?**: No (radio button selected).
 - Libraries**:
 - 1: Libraries**: Library type: Paired-end / Single reads; Orientation: <-> <- (fr); Files: 1: Sample_R1_pair.fastq, 2: Sample_R2_pair.fastq.
 - 1: Files**: Select file format: Separate input files; Forward reads: 1: Sample_R1_pair.fastq; Reverse reads: 2: Sample_R2_pair.fastq.
- History (Right):** Shows a demo workspace with two datasets: "Sample_R2_pair.fastq" and "Sample_R1_pair.fastq".



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:
--1 <filename> file with interlaced forward and reverse paired-end reads
--2 <filename> file with forward paired-end reads
--r <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)
--sc<#>
--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)
--mps<#>-s <filename> file with unpaired reads for mate-pair library number <#> (<#> = 1,2,...,9)
--mps<#>-<or> orientation of reads for mate-pair library number <#> (<#> = 1,2,...,9; <or> = fr, rf, ff)
--hqmps<#>-12 <filename> file with interlaced reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmps<#>-1 <filename> file with forward reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmps<#>-2 <filename> file with reverse reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmps<#>-s <filename> file with unpaired reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmps<#>-<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
--memory <int> RAM limit for SPAdes in Gb (terminates if exceeded)
--tmp-dir <dirname> directory for temporary files
--k <int,int,...> comma-separated list of k-mer sizes (must be odd and
```

NeLS galaxy-uit.bioinfo.no

Analyze Data Workflow Shared Data Visualization Admin Help User Options

SPAdes genome assembler for regular and single-cell projects (Galaxy Version 3.9.0)

Single-cell? Yes | No

This option is required for MDA (single-cell) data. (-sc)

Run only assembly? (without read error correction) Yes | No

(--only-assembler)

Careful correction? Yes | No

Tries to reduce number of mismatches and short indels. Also runs MismatchCorrector – a post processing tool, which uses BWA tool (comes with SPAdes). (--careful)

Automatically choose k-mer values Yes | No

k-mer choices can be chosen by SPAdes instead of being entered manually.

K-mers to use, separated by commas 21,33,55

Comma-separated list of k-mer sizes to be used (all values must be odd, less than 128, listed in ascending order, and smaller than the read length). The default value is 21,33,55.

Coverage Cutoff Off

Libraries are IonTorrent reads? Yes | No

Libraries

1: Libraries

Library type Paired-end / Single reads

Orientation <> <- (fr)

Files

1: Files

Select file format Separate input files

Forward reads

FASTQ format 1: Sample_R1_pair.fastq

Reverse reads

FASTQ format 2: Sample_R2_pair.fastq

History search datasets

demo 2 shown 513.52 MB

2: Sample_R2_pair.fastq

1: Sample_R1_pair.fastq



Galaxy - Tools

Example running SPAdes
Choose parameter settings
Select input files

The screenshot shows the Galaxy web interface with the following details:

- Tool Selection:** The "SPAdes genome assembler for regular and single-cell projects (Galaxy Version 3.9.0)" tool is selected.
- Parameter Settings:**
 - "Single-cell?" is set to "No".
 - "Run only assembly? (without read error correction)" is set to "Yes".
 - "Careful correction?" is set to "No".
 - "Automatically choose k-mer values" is set to "Yes".
 - "K-mers to use, separated by commas" is set to "21,33,55".
 - "Coverage Cutoff" is set to "Off".
 - "Libraries are IonTorrent reads?" is set to "No".
 - "Orientation" is set to "Paired-end / Single reads".
- Input Files:**
 - "Select file format" is set to "Separate input files".
 - "Forward reads" is set to "1: Sample_R1_pair.fastq".
 - "Reverse reads" is set to "2: Sample_R2_pair.fastq".
- History:** The history panel shows two datasets:
 - 1: Sample_R1_pair.fastq (green)
 - 2: Sample_R2_pair.fastq (green)

Galaxy – Running tools

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

The screenshot shows the Galaxy web interface with the following details:

- Header:** NeLS galaxy-uit.bioinfo.no, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Left Panel (Tools):** A sidebar listing various bioinformatics tools categorized under GENERAL GALAXY TOOLS, NGS: QC and Manipulation, Assembly and Validation, Transcriptomics, and Workflows.
- Middle Panel:** A message box indicating "1 job has been successfully added to the queue – resulting in the following datasets:" followed by a list of 7 datasets:
 - 3: SPAdes contig stats
 - 4: SPAdes scaffold stats
 - 5: SPAdes contigs (fasta)
 - 6: SPAdes scaffolds (fasta)
 - 7: SPAdes log

A note below states: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."
- Right Panel (History):** A list of datasets in the history pane, with the last two items (2: Sample_R2_pair.fastq and 1: Sample_R1_pair.fastq) highlighted with a red border. The list includes:
 - demo 7 shown 513.52 MB
 - 7: SPAdes log
 - 6: SPAdes scaffolds (asta)
 - 5: SPAdes contigs (fa sta)
 - 4: SPAdes scaffold sta ts
 - 3: SPAdes contig stat s
 - 2: Sample_R2_pair.fastq
 - 1: Sample_R1_pair.fastq

Galaxy – Running tools

When the job is running, the files turns yellow

The screenshot shows the Galaxy web interface with the following details:

- Header:** NeLS galaxy-uit.bioinfo.no, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Left Panel (Tools):** A sidebar listing various bioinformatics tools categorized under GENERAL GALAXY TOOLS, NGS: QC and Manipulation, Assembly and Validation, Transcriptomics, and Workflows.
- Middle Panel:** A message indicating "1 job has been successfully added to the queue – resulting in the following datasets:" followed by a list of 7 datasets:
 - 3: SPAdes contig stats
 - 4: SPAdes scaffold stats
 - 5: SPAdes contigs (fasta)
 - 6: SPAdes scaffolds (fasta)
 - 7: SPAdes log

A note below states: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."
- Right Panel (History):** A list of datasets in the history pane, each with a yellow progress bar indicating completion. The datasets are:
 - demo (513.52 MB)
 - 7: SPAdes log
 - 6: SPAdes scaffolds (fasta)
 - 5: SPAdes contigs (fasta)
 - 4: SPAdes scaffold stats
 - 3: SPAdes contig stats
 - 2: Sample_R2.fastq
 - 1: Sample_R1.fastq



Galaxy – Running tools

When the job finish successfully,
the files turn green

The screenshot shows the Galaxy web interface with the following details:

- Header:** NeLS galaxy-uit.bioinfo.no, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Left Panel (Tools):** A sidebar listing various bioinformatics tools categorized under GENERAL GALAXY TOOLS, NGS: QC and Manipulation, Assembly and Validation, Transcriptomics, and Workflows.
- Middle Panel (Job Queue):** A message indicates "1 job has been successfully added to the queue – resulting in the following datasets:" followed by a list of 7 datasets:
 - 3: SPAdes contig stats
 - 4: SPAdes scaffold stats
 - 5: SPAdes contigs (fasta)
 - 6: SPAdes scaffolds (fasta)
 - 7: SPAdes log

A note below states: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."
- Right Panel (History):** A list of datasets with their status and file sizes. The last four items in the list are highlighted with a red border, indicating they are finished and successful.

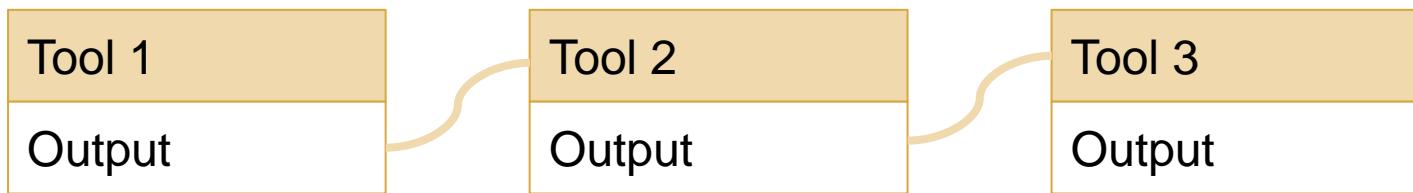
Dataset ID	File Name	Status	Size
demo			523.43 MB
7: SPAdes log		✓	
6: SPAdes scaffolds (fast a)		✓	
5: SPAdes contigs (fasta)		✓	
4: SPAdes scaffold stats		✓	
3: SPAdes contig stats		✓	
2: Sample_R2_pair.fastq		✓	
1: Sample_R1_pair.fastq		✓	

Galaxy – Running tools

You can view and download the result files

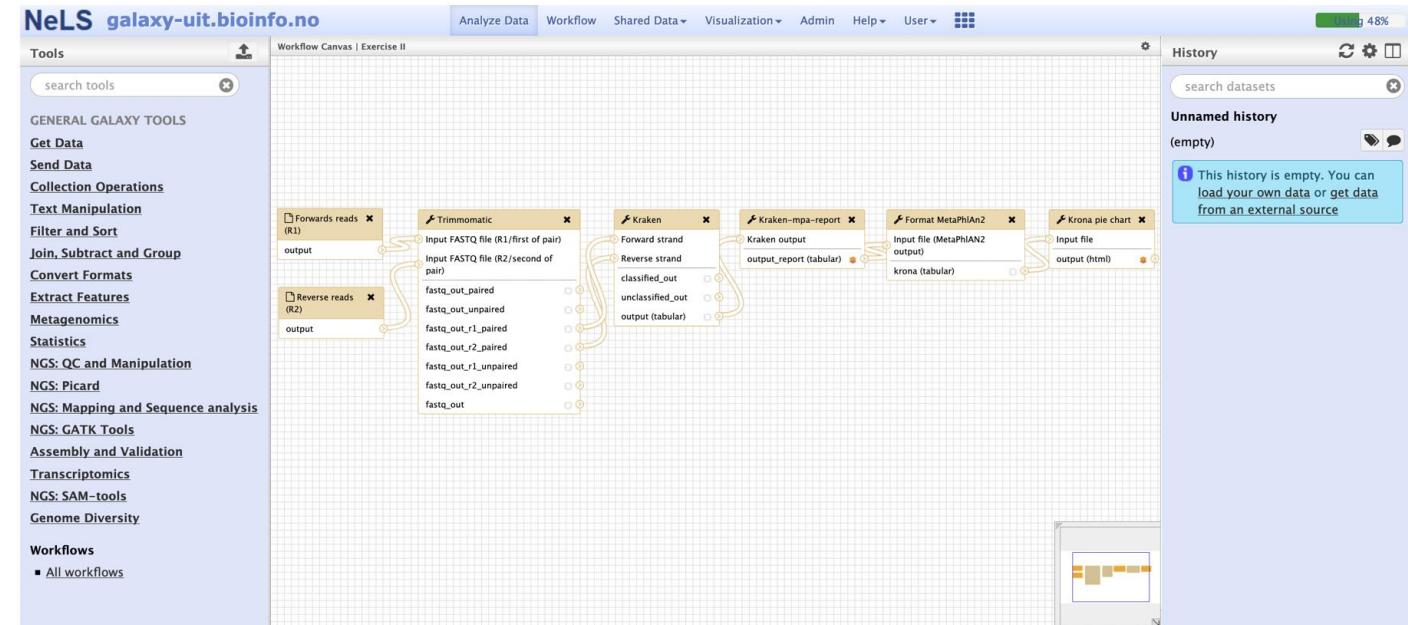
Galaxy workflows

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



Galaxy - Workflows

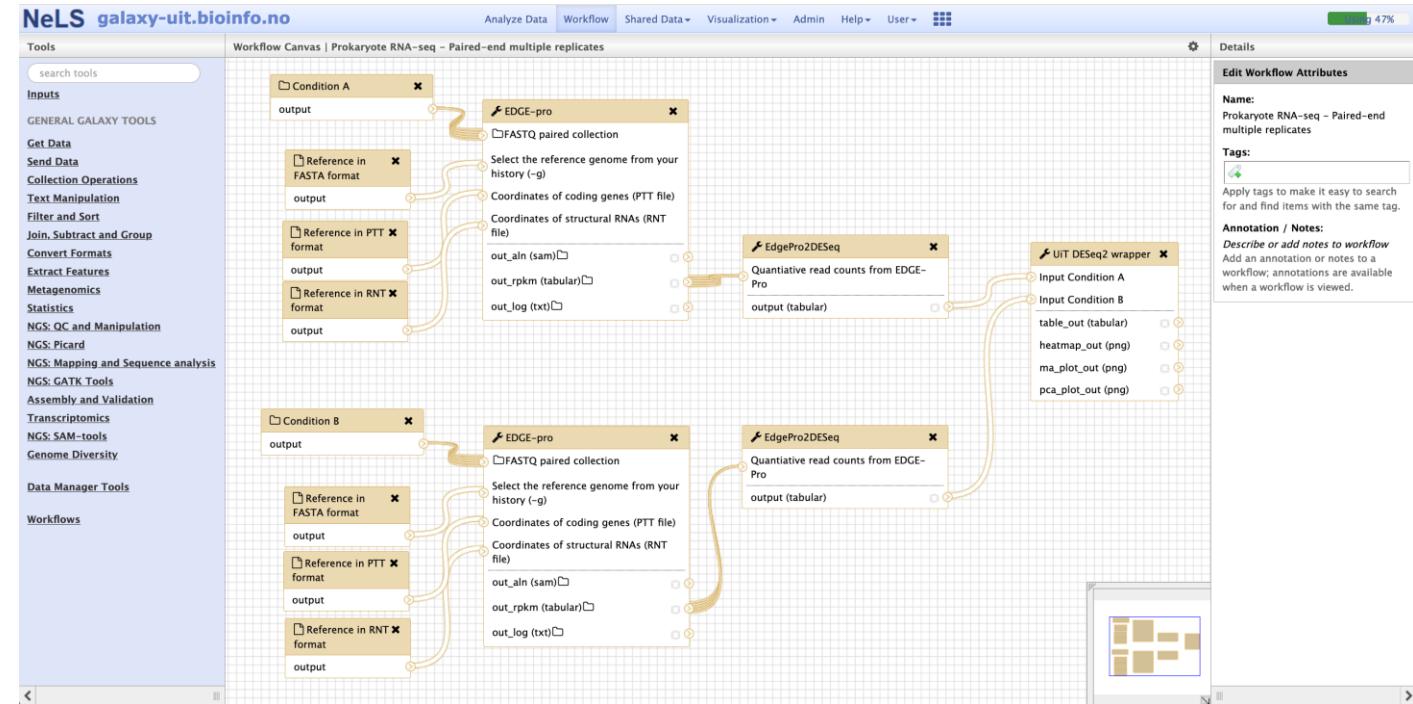
The “nODULES” indicate which output file from one acts as input for the next tool



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

The screenshot shows the Galaxy web interface with the following details:

- Header:** NeLS galaxy-uit.bioinfo.no, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Workflow Title:** Workflow: Taxonomic classification of metagenomic sequences – Paired-end
- Workflow History Options:** Send results to a new history (Yes/No). The history is currently empty.
- Workflow Steps:** A list of 20 steps, each with a small icon and a brief description:
 - 1: Input dataset – FASTQ/FASTQSANGER format
 - 2: Input dataset – FASTQ/FASTQSANGER format
 - 3: Concatenate datasets (Galaxy Version 1.0.0)
 - 4: Filter FASTQ (Galaxy Version 1.0.0)
 - 5: FASTQ to FASTA (Galaxy Version 1.0.0)
 - 6: Predict 16S rRNA Reads (Galaxy Version 1.0.0)
 - 7: UIT Megablast Wrapper (Galaxy Version 1.0.0)
 - 8: LCAClassifier (Galaxy Version 1.0.0)
 - 9: Krona pie chart (Galaxy Version 2.6.1)
 - 10: DESeq on data 16 and data 15: PCA plot
 - 11: DESeq on data 16 and data 15: MA-plot
 - 12: DESeq on data 16 and data 15: Heatmap
 - 13: DESeq on data 16 and data 15: Results
 - 14: EdgePro2DESeq on data 13
 - 15: EdgePro2DESeq on data 10
 - 16: EdgePro2DESeq on data 6, data 5, and others: log
 - 17: EdgePro2DESeq on data 6, data 5, and others: rpk
 - 18: EdgePro2DESeq on data 6, data 5, and others: alignment
 - 19: EdgePro2DESeq on data 6, data 5, and others: log
 - 20: EdgePro2DESeq on data 6, data 5, and others: rpk
- History Panel:** Shows a list of 20 datasets, each with a preview icon, edit icon, and delete icon. The datasets are:
 - test_run RNAseq
 - 5.6 GB
 - 20: DESeq on data 16 and data 15: PCA plot
 - 19: DESeq on data 16 and data 15: MA-plot
 - 18: DESeq on data 16 and data 15: Heatmap
 - 17: DESeq on data 16 and data 15: Results
 - 16: EdgePro2DESeq on data 13
 - 15: EdgePro2DESeq on data 10
 - 14: EdgePro2DESeq on data 6, data 5, and others: log
 - 13: EdgePro2DESeq on data 6, data 5, and others: rpk
 - 12: EdgePro2DESeq on data 6, data 5, and others: alignment
 - 11: EdgePro2DESeq on data 6, data 5, and others: log
 - 10: EdgePro2DESeq on data 6, data 5, and others: rpk
 - 9: EdgePro2DESeq on data 6, data 5, and others: alignment
 - 8: Treated_R2.fastq
 - 7: Treated_R1.fastq
 - 6: reference.rnt
 - 5: reference.ptt



Galaxy – Workflow documentation

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

Also test data sets available for each workflow

Accessible Page | RNA-Seq pipeline: eukaryotes

RNA-Sequencing workflow

Illumina RNA Sequencing differential expression analysis pipeline between two collections of eukaryote samples. The collections represents the two conditions being compared, and the datasets in each collection is defined by the user.

The workflow use TopHat2 for alignment. TopHat2 first performs splice-aware alignment to the reference transcriptome. Reads which not aligned to the transcriptome are then aligned to the genome. Splice sites are reported. FeatureCounts from the SubRead-package is then used on the aligned reads to count transcript features from a reference file. The transcript-features do not need to be the same as the transcriptome used for alignment. Finally differentially expressed features (usually genes) are calculated by DESeq, a standard tool for differential expression of count data from sequencing experiments. DESeq corrects for the effect of overdispersion in the differential analysis.

The user performs the following steps before running the workflow

1. Upload data to Galaxy: Use "Get Data" on the menu, and upload your sequence files (paired-end illumina sequences in .fastq format)
2. Create dataset collections: Add datasets to the two collections you want to compare (condition 1 and condition 2)
3. Set parameters: See description below for which parameters to set, and which sub-workflow to select
4. Run workflow:

The workflow performs the following steps automatically

1. Perform sequence alignment to reference transcriptome/genome: The workflow uses the program TopHat2 for alignment.
2. Assign reads to features and create read-count for each feature: The workflow uses the program featureCounts from the SubRead package. Usually genes are the preferred feature.
3. Differential expression of features: The workflow uses the program DESeq for differential expression.
4. Return output files: DESeq returns a table of differentially expressed genes, and a heatmap, ma-plot and pca-plot of the results.

Galaxy Workflow | imported: RNA-seq pipeline for differential gene expression analysis (paired-end, pooled samples)

Step	Annotation
Step 1: Input dataset collection	Sequence dataset collection for condition 1 (.fastq, paired-end)
Sample set #1 select at runtime	



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

The screenshot shows the Galaxy web interface with the title "Galaxy / uit". The main content area displays a table titled "Histories shared with you by others". The table has columns for Name, Datasets, Created, Last Updated, and Shared by. The data includes:

Name	Datasets	Created	Last Updated	Shared by
Sample11 Trimmed	5 6	Nov 25, 2016	Dec 06, 2016	iro@nifes.no
Sample1 Trimmed	3 6 1	Nov 25, 2016	Dec 06, 2016	iro@nifes.no
Sample2 Trimmed	7 5	Nov 25, 2016	Dec 06, 2016	iro@nifes.no
Unnamed history	28 12	May 10, 2016	May 26, 2016	miv023@uit.no
Theme2	21	Apr 20, 2016	Apr 21, 2016	era036@uit.no
Unnamed history	2 1 5	Jan 28, 2015	Feb 02, 2015	aus023@uit.no

Below the table are buttons for "Copy" and "Unshare". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Admin", "Help", and "User". A sidebar on the left lists various tools and workflows. A vertical toolbar on the right provides options for managing datasets and histories.



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

Galaxy Tool Shed		
Repositories by Category		
Name	Description	Repositories
Assembly	Tools for working with assemblies	81
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	42
Combinatorial Selections	Tools for combinatorial selection	6
Computational chemistry	Tools for use in computational chemistry	27
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	11
Convert Formats	Tools for converting data formats	68
Data Export	Tools for exporting data to various destinations	1
Data Managers	Utilities for Managing Galaxy's built-in data cache	33
Data Source	Tools for retrieving data from external data sources	48
Epigenetics	Tools for analyzing Epigenetic/Epigenicomic datasets	4
Fasta Manipulation	Tools for manipulating fasta data	77
Fastq Manipulation	Tools for manipulating fastq data	60
Flow Cytometry Analysis	Tools for manipulating and analyzing FCS files	
Genome-Wide Association Study	Utilities to support Genome-wide association studies	20



Galaxy – Helpdesk assistance

New reference data

New tools/wrappers

New public workflows if needed by many

Hands-on workshops regularly

NeLS Reference Data Repository		
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



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

