



## SOFTWARE TOOL ARTICLE

# Norwegian e-Infrastructure for Life Sciences (NeLS) [version 1; referees: 2 approved]

Kidane M. Tekle<sup>1</sup>, Sveinung Gundersen<sup>2</sup>, Kjetil Klepper<sup>3</sup>, Lars Ailo Bongo <sup>4</sup>, Inge Alexander Raknes<sup>4</sup>, Xiayi Li<sup>1</sup>, Wei Zhang<sup>1</sup>, Christian Andreetta<sup>1</sup>, Teshome Dagne Mulugeta<sup>5</sup>, Matúš Kalaš <sup>1</sup>, Morten B. Rye<sup>3</sup>, Erik Hjerde <sup>4</sup>, Jeevan Karloss Antony Samy<sup>5</sup>, Ghislain Fornous<sup>2</sup>, Abdulrahman Azab<sup>2</sup>, Dag Inge Våge <sup>5</sup>, Eivind Hovig<sup>2</sup>, Nils Peder Willassen<sup>4</sup>, Finn Drabløs <sup>3</sup>, Ståle Nygård<sup>2</sup>, Kjell Petersen<sup>1</sup>, Inge Jonassen <sup>1</sup>

<sup>1</sup>Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

<sup>2</sup>University of Oslo, Oslo, Norway

<sup>3</sup>Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

<sup>4</sup>University of Tromsø - The Arctic University of Norway, Tromsø, Norway

<sup>5</sup>Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

**v1** First published: 29 Jun 2018, 7(ELIXIR):968 (doi: [10.12688/f1000research.15119.1](https://doi.org/10.12688/f1000research.15119.1))

Latest published: 29 Jun 2018, 7(ELIXIR):968 (doi: [10.12688/f1000research.15119.1](https://doi.org/10.12688/f1000research.15119.1))

## Abstract

The Norwegian e-Infrastructure for Life Sciences (NeLS) has been developed by ELIXIR Norway to provide its users with a system enabling data storage, sharing, and analysis in a project-oriented fashion. The system is available through easy-to-use web interfaces, including the Galaxy workbench for data analysis and workflow execution. Users confident with a command-line interface and programming may also access it through Secure Shell (SSH) and application programming interfaces (APIs).

NeLS has been in production since 2015, with training and support provided by the help desk of ELIXIR Norway. Through collaboration with NorSeq, the national consortium for high-throughput sequencing, an integrated service is offered so that sequencing data generated in a research project is provided to the involved researchers through NeLS. Sensitive data, such as individual genomic sequencing data, are handled using the TSD (Services for Sensitive Data) platform provided by Sigma2 and the University of Oslo. NeLS integrates national e-infrastructure storage and computing resources, and is also integrated with the SEEK platform in order to store large data files produced by experiments described in SEEK.




In this article, we outline the architecture of NeLS and discuss possible directions for further development.

## Keywords

Data management and sharing, compute and storage infrastructure, microservices, federated authentication, integration API, Galaxy, ELIXIR Norway

## Open Peer Review

Referee Status:  

	Invited Referees	
	1	2
version 1 published 29 Jun 2018	 report	 report
1 <b>Olivier Collin</b>  , CNRS, IRISA F-35000, France		
2 <b>Wolfgang Mueller</b> , Heidelberg Institute of Theoretical Studies gGmbH, Germany		

## Discuss this article

Comments (0)



This article is included in the **International Society for Computational Biology Community Journal** gateway.



This article is included in the **ELIXIR** gateway.

**Corresponding authors:** Kjell Petersen ([Kjell.Petersen@uib.no](mailto:Kjell.Petersen@uib.no)), Inge Jonassen ([inge.jonassen@uib.no](mailto:inge.jonassen@uib.no))

**Author roles:** **Tekle KM:** Methodology, Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Gundersen S:** Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Klepper K:** Methodology, Resources, Software; **Bongo LA:** Methodology, Resources, Writing – Review & Editing; **Raknes IA:** Resources, Software; **Li X:** Resources, Software; **Zhang W:** Resources, Software; **Andreetta C:** Methodology, Software; **Mulugeta TD:** Resources, Software; **Kalaš M:** Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Rye MB:** Resources, Writing – Original Draft Preparation; **Hjerde E:** Resources, Writing – Original Draft Preparation; **Antony Samy JK:** Resources, Writing – Original Draft Preparation; **Fornous G:** Resources, Software; **Azab A:** Resources, Software; **Våge DI:** Conceptualization, Funding Acquisition, Project Administration; **Hovig E:** Conceptualization, Funding Acquisition, Project Administration, Writing – Original Draft Preparation; **Willassen NP:** Conceptualization, Funding Acquisition, Project Administration; **Drabløs F:** Conceptualization, Funding Acquisition, Project Administration, Writing – Review & Editing; **Nygård S:** Project Administration, Resources, Writing – Original Draft Preparation; **Petersen K:** Conceptualization, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Jonassen I:** Conceptualization, Funding Acquisition, Project Administration, Writing – Original Draft Preparation

**Competing interests:** No competing interests were disclosed.

**Grant information:** The work was funded by the ELIXIR.NO (208481/F50) and ELIXIR2 (270068) infrastructure grants from the Research Council of Norway, as well as the Tryggve and Tryggve2 projects from the Nordic e-Infrastructure Collaboration (NeIC).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2018 Tekle KM *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution Licence**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Tekle KM, Gundersen S, Klepper K *et al.* **Norwegian e-Infrastructure for Life Sciences (NeLS) [version 1; referees: 2 approved]** *F1000Research* 2018, 7(ELIXIR):968 (doi: [10.12688/f1000research.15119.1](https://doi.org/10.12688/f1000research.15119.1))

**First published:** 29 Jun 2018, 7(ELIXIR):968 (doi: [10.12688/f1000research.15119.1](https://doi.org/10.12688/f1000research.15119.1))

## 1. Introduction

The **Norwegian ELIXIR node** is coordinated by the University of Bergen (UiB) and comprises the University of Oslo (UiO), The Arctic University of Norway (UiT), the Norwegian University of Science and Technology (NTNU), and the Norwegian University of Life Sciences (NMBU). The node provides services, training, and support to a broad range of national users, largely life-science researchers and students<sup>1</sup>. These scientists usually work in collaborative projects and need to store, analyze, and share data sets, often large in size, throughout all stages of the project, and between various platforms and computational resources. However, many of these users do not feel comfortable using a command-line interface, and have limited programming, system administration, or data management skills.

Commercial workbenches such as the BaseSpace Sequence Hub<sup>a</sup> and Geneious<sup>b</sup> aim at user accessibility, but offer computation and data sharing only within their closed and expensive platform setups. On the other hand, the open-source SEEK serves as a platform for sharing systems biology project data, transparently and for free<sup>2,3</sup>. Notably, the integrative open-source GenomeSpace enables organizing and sharing data not only between users, but also between various workbenches and computational resources<sup>4-6</sup>. Although powerful, its setup could not be adapted to integrate with the available and required e-infrastructure resources in Norway.

The **Norwegian e-Infrastructure for Life Sciences (NeLS)** was built upon the previous experiences with developing and

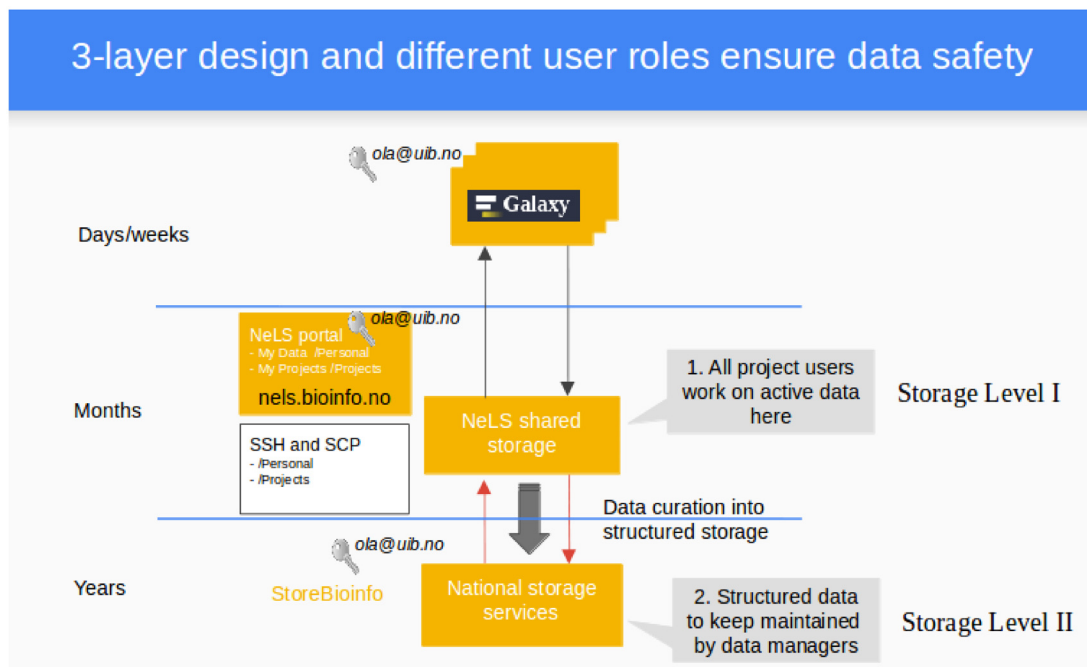
using bioinformatics workbenches in Norway, for example: the Genomic HyperBrowser<sup>7-9</sup>, an extension of Galaxy<sup>10,11</sup>; the easy-to-use UiO Bioportal<sup>12</sup>, later replaced by a Galaxy-based Lifeportal<sup>13</sup>; or eSysbio, a workbench prototype for data sharing and systems biology workflows<sup>6</sup>. NeLS provides users in Norway and their collaborators abroad an integrated system for data storage and sharing, as well as data processing and analysis. NeLS allows users to efficiently and safely store, analyze, and share their genomics-scale data and analyses, all through the use of web interfaces. Most Norwegian users can log in using the credentials – user name and password – they use at their home institution, other users need to register.

NeLS has a three-layered architecture (Figure 1). The intermediate layer (Storage Level I) provides data storage intended to be used by projects in an active analysis phase (with data being kept in this storage layer for months). Data can be accessed (and up- and downloaded) through a web portal, as well as through Secure Shell (SSH) and application programming interfaces (APIs). The latter two provide command-line-confident users with a more efficient way to work with data. The top level constitutes the data analysis workbench of NeLS. For this, we have chosen the popular Galaxy, an open-source, web-based workbench for accessible, reproducible, and transparent computational omics. Galaxy allows computational workflows to be set up and used without the need of programming skills. Our Galaxy instances have limited storage capacity and it is therefore intended that data resides on this level only for short periods of time, in the range of weeks. The bottom layer (Storage Level II)

<sup>a</sup><https://basespace.illumina.com>

<sup>b</sup><https://www.geneious.com>

<sup>c</sup> 14, pp. 53–56, 61–64, <https://bora.uib.no/bitstream/handle/1956/10658/thesis.pdf#page=61>



**Figure 1.** Overview over short-term and long-term storage in NeLS.

offers long-term storage, provided by the National Infrastructure for Research Data (NIRD), a generic e-infrastructure operated by Sigma2<sup>d</sup>. Here, data can be stored for years, requiring a more strictly organized structure with defined types and metadata.

NeLS includes Galaxy instances hosted at the five universities constituting the Norwegian ELIXIR node. These instances have a basic catalogue of tools and workflows that are relevant for researchers in life sciences, as well as more specialized ones depending on the focus of the hosting institution. NeLS provides tools integrated into Galaxy to easily push and pull data from the persistent Storage Level I. Some of the Galaxy servers are integrated with high-performance computing (HPC) resources – provided by Sigma2 – for transparent execution of computationally intensive tools.

In this article, we describe the architecture of this integrated e-infrastructure and examples of its usage, and outline the possible directions of future developments.

## 2. The architecture of NeLS

The Norwegian e-Infrastructure for Life Sciences was not built as a top-down, grand design and implementation exercise.

Rather, it was implemented through time by focusing on different parts of the problem at a time and always striving to make a functional whole. It was decided early on to avoid re-inventing the wheel and rather base the system on proven solutions and practices whenever possible. In addition, addressing different concerns in isolation while keeping the big picture in mind has proven to be an effective way for constructing the NeLS system. In the end, there were many components of different flavors: off-the-shelf systems, adaptations of available open-source packages, and also custom in-house developed systems. Figure 2 shows a componentized architecture of the NeLS system. In the following subsections, we describe the components of the NeLS system.

### 2.1 Authentication

Authentication is a process by which the user’s credentials are verified against a user-information catalogue in order to determine whether the user is who they claim to be, before granting access to resources. NeLS supports multiple identity sources based on the Security Assertion Markup Language 2.0 (SAML 2.0) standard<sup>e</sup>. Currently, it supports the Norwegian Federated Electronic Identity service (FEiDE<sup>f</sup>) and NeLS’s own

<sup>d</sup><https://www.sigma2.no>

<sup>e</sup><https://wiki.oasis-open.org/security/FrontPage>

<sup>f</sup><https://www.feide.no/introducing-feide>

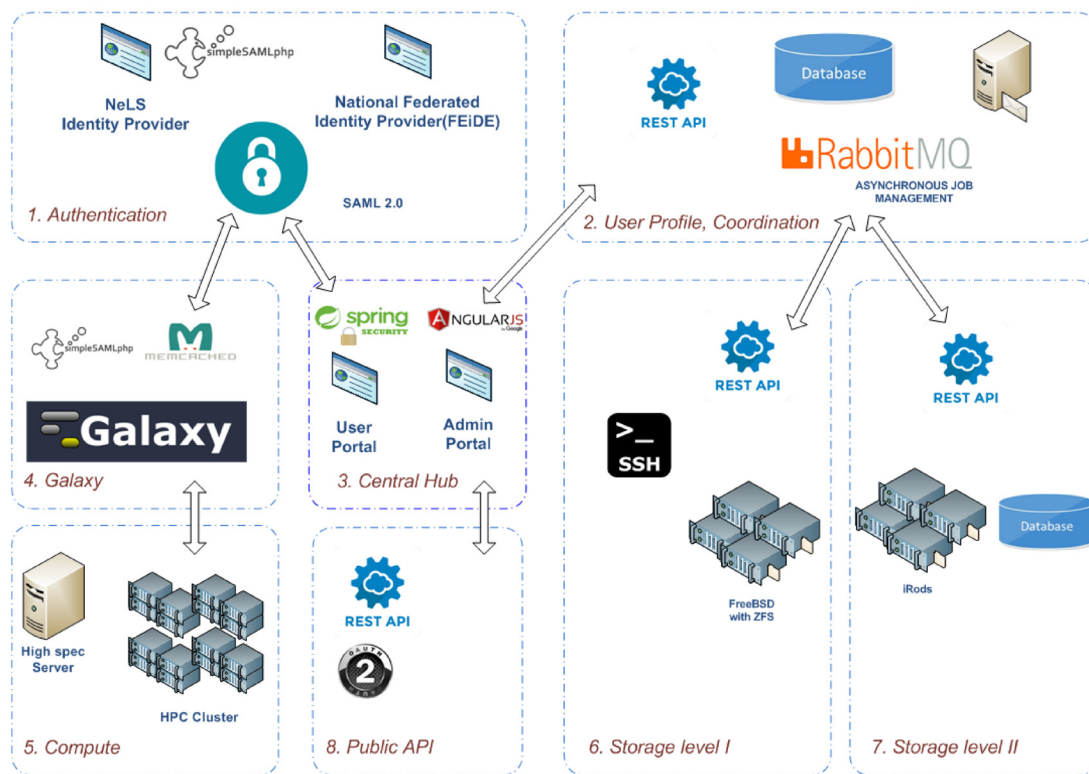


Figure 2. NeLS Architecture.

identity provider. NeLS's identity provider was constructed by configuring Simple-SAMLphp<sup>g</sup>, an open-source security software system. Integration with the ELIXIR Authentication and Authorisation Infrastructure (AAI)<sup>15</sup> as an identity provider has been tested technically, and can in the future be used to differentiate resource allocations further in the NeLS network of services.

## 2.2 User-profile management and coordination

The first time a user logs in using any of the supported identity providers, NeLS creates a user profile and subsequent derived identities. Secure Shell (SSH) access credentials are generated for the user and can be fetched from the NeLS portal (central hub). This coordination layer holds metadata about projects and associated membership of users. The user-profile management and coordination block provides a Representational State Transfer (REST, "RESTful") web application programming interface (API)<sup>16-18</sup> for other units, and enables asynchronous job management by leveraging an off-the-shelf message-queuing system, RabbitMQ<sup>h</sup>. Structured logging, e-mail communication, and related management tasks are supported in this block.

## 2.3 The NeLS portal (central hub)

The NeLS portal is the central hub of the whole system. It is a Java-based web application with multiple responsibilities and uses the Spring Security package<sup>i</sup> to interface with the different SAML 2.0 identity providers. Upon first login, NeLS creates a profile for the user and initializes all necessary components. Following are the four distinct responsibilities of the NeLS portal:

- (a) Web-based file-system browser to the NeLS Storage Level I (see subsection 2.6).
- (b) Initiate and monitor asynchronous jobs for copying, moving, and sharing files within the Level I storage layer, as well as transfer across storage Levels I and II.
- (c) Facilitate OAuth<sup>j</sup> token provision. The NeLS portal acts as a bridge towards the identity providers and avails NeLS metadata for the OAuth service (see subsection 2.4).
- (d) Interface with external systems. The NeLS portal has been successfully integrated with a national solution for sensitive data, the TSD<sup>k</sup><sup>19</sup>. NeLS makes two-factor authentication of the TSD easier for users, and provides an easy web-based way for initiating data-transfer jobs between the NeLS and TSD infrastructures.

## 2.4 Galaxy

The Galaxy block is the workhorse of the NeLS system. It gives the user a curated set of tools and workflows supported by the ELIXIR Norway help desk<sup>l</sup>. Technically, the Galaxy block

comprises Galaxy in a remote-user configuration, and an authentication layer in front to interface with the same set of identity providers as the NeLS portal. SimpleSAMLphp in service provider (SP) configuration with its AuthMemCookie<sup>l</sup> solution on top of Memcached<sup>m</sup> is used to interface with the Apache web server, transforming the SAML 2.0 authorization information into a Galaxy-compatible format. NeLS also provides Galaxy tools for data import and export, which work in tandem with the NeLS portal to give the user the possibility of pulling data from the Storage Level I in NeLS into a Galaxy history, and also be able to push results of Galaxy jobs into the NeLS Storage Level I.

NeLS provides different Galaxy instances hosted by different institutions.

## 2.5 Compute block

The NeLS compute block executes NeLS Galaxy jobs. The jobs are either executed on the same high-spec (fat) servers as the Galaxy server, or they can be submitted to a high-performance computing (HPC) cluster for parallel execution. The job execution details, including HPC job management, are hidden for the user. The HPC jobs are run using a pre-allocated compute quota.

We use the Light-weight Runner (LWR, now renamed to Pulsar<sup>n</sup>) Galaxy services to submit computationally intensive Galaxy jobs to an HPC system (such as the Stallo supercomputer in the UiT Galaxy<sup>o</sup>). LWR communicates with Galaxy via the Galaxy API and a RabbitMQ Advanced Message Queuing Protocol (AMQP) message queue. It specifies the required parameters for the tool and executes a wrapper script for the tool. The wrapper creates temporary directories, submits tool jobs to the HPC scheduler (PBS/Torque<sup>p</sup>) with selected parameters, saves results, and deletes temporary files. Once the jobs are completed, LWR transfers the data back to Galaxy for the user to inspect.

## 2.6 Storage Level I

This layer of NeLS provides flexible storage with advanced access control list to allow appropriate sharing and data protection in scientific projects.

The NeLS Storage Level I layer features a dedicated private directory for each user's personal data as well as a project-based shared storage area for collaboration and sharing. A user can be added to a NeLS project with three possible roles: member, power-user, or principal investigator (PI). Each role has a predefined set of permissions allowed in the project area. Technically, the NeLS Storage Level I is built using FreeBSD<sup>q</sup> with its support for the ZFS<sup>r</sup> file system. It employs advanced access

<sup>g</sup><https://simplesamlphp.org>

<sup>h</sup><https://www.rabbitmq.com>

<sup>i</sup><https://projects.spring.io/spring-security/>

<sup>j</sup><https://tools.ietf.org/html/rfc5849>

<sup>k</sup>Services for Sensitive Data, in Norwegian *Tjenester for Sensitive Data*

<sup>l</sup><https://zenprojects.github.io/Apache-Authmemcookie-Module/>

<sup>m</sup><https://memcached.org>

<sup>n</sup><https://galaxyproject.org/admin/config/pulsar/>

<sup>o</sup><https://galaxy-uit.bioinfo.no>

<sup>p</sup><https://www.adaptivecomputing.com/products/open-source/torque>

<sup>q</sup><https://www.freebsd.org/>

<sup>r</sup><https://en.wikipedia.org/wiki/ZFS>



control lists and also provides SSH access to more tech-savvy users. It provides a RESTful web API (Java) and command-line management tools (Python).

## 2.7 Storage Level II

Level II of the NeLS storage, also known as the StoreBioinfo layer, enforces more strict organization of datasets, and is facilitated through integration with national storage resources provided by the Norwegian Infrastructure for Research Data (NIRD). Its purpose is to act as a long-term storage and data warehouse, with capacity to hold all of a project's generated data, from raw to final results, including any data replicated to the Storage Level I. It has a metadata database and interfaces with the data-warehousing system, iRods<sup>5</sup>, via specialized server-side scripts. The NeLS Storage Level II provides a RESTful web API and is orchestrated via the NeLS central coordination block (3. Central Hub in Figure 2).

## 2.8 Public API

The NeLS public API is a RESTful web service targeted towards external systems. It supports implicit and authorization-code OAuth<sup>1</sup> grant profiles. It exposes a well-defined navigation and linking mechanism into the structured data of NeLS Storage Level II. The OAuth service is an in-house built Python-based system that uses Tornado<sup>6</sup> and python-oauth2<sup>7</sup> libraries by interfacing with the NeLS portal. In collaboration with Digital Life Norway<sup>8</sup>, the public API is developed to support integration with the SEEK data management system<sup>2,3</sup>, to allow a resolvable URL to a dataset in NeLS be referenced in SEEK.

## 3. Operation

NeLS is inherently a distributed infrastructure of multiple microservices which naturally would be deployed on different servers. The scale and availability of the different resources to be integrated – such as compute, storage, identity providers, databases, *etc.* – heavily influences its deployment. In Norway, the NeLS production instance is deployed on 7 different servers, including 2 storage master systems (additional slave storage nodes – sub-systems – are not counted).

For testing or a proof-of-concept setup, all microservices and web components are possible to run on a single host with 2–4 cores and 64GB of memory, while the storage levels would naturally require their own setups.

## 4 Workflows

To cover the most prominent NeLS user needs, Galaxy workflows for analyzing RNA sequence data (prokaryotic and eukaryotic), and workflows for both the taxonomic and functional

profiling of metagenomic data have been developed. In addition, workflows for the analysis of miRNAs and ChIP-seq analysis are available, see Table 1. All workflows are maintained in order to provide the state-of-the-art tools for the analysis to the users. Upon demand, each work-flow can be modified to accommodate specific user needs, *e.g.* that a tool is replaced by another tool, or version. A complete overview of the NeLS workflows and links to the Galaxy instance in which a workflow is available can be found in the NeLS portal.

To ensure data reproducibility and to reduce the compute time for the user, a common data repository with pre-indexed reference genomes has been built. The repository is available across all five Galaxy instances.

For first-time users of the NeLS Galaxy, a quick start guide that contains information on the Galaxy basics is available on each NeLS Galaxy start page, and more detailed documentation and tutorials on the NeLS workflows are also available there. Finally, the user can contact the national help desk or access a Q&A forum directly from the NeLS Galaxy.

## 5 Use cases

### 5.1 Main steps in an ordinary NeLS project

In a project with non-sensitive data, a user will perform the following steps (See Figure 3e (a)-(e))

- (a) Log in to the NeLS portal. If the user does not have an account, and neither has a FEiDE account, they can easily apply for a NeLS account.
- (b) Upload data to NeLS repository using SSH or Filezilla<sup>9</sup>. If the data are generated by a Norwegian high-throughput sequencing core facility, *e.g.* one linked with the Norwegian Consortium for Sequencing and Personalized Medicine (NorSeq<sup>10</sup>), the user is offered to have the data uploaded directly from the core facility.
- (c) (Recommended) Synchronize data to Storage Level II (StoreBioinfo) for annotation and long-term storage.
- (d) Log in to one of the Galaxy instances, *e.g.* <https://galaxy-ntnu.bioinfo.no>.
- (e) Get data from NeLS Storage Level I to Galaxy, and run Galaxy tools and/or workflows. Share or publish Galaxy history.
- (f) Copy new results to Storage Level I and synchronize to Storage Level II (recommended).

### 5.2 META-pipe

META-pipe<sup>23–25</sup>, developed within the ELIXIR Marine metagenomics project<sup>26,27</sup>, efficiently produces full-length annotated genes from metagenomic assemblies, and offers the extensive annotation options, flexibility, and visualization needed to

<sup>5</sup><https://irods.org/>

<sup>6</sup><https://tools.ietf.org/html/rfc5849>

<sup>7</sup><https://pypi.org/project/tornado/>

<sup>8</sup><https://github.com/joestump/python-oauth2>

<sup>9</sup><https://digitallifenorway.org/>

<sup>9</sup><https://filezilla-project.org/>

<sup>10</sup><https://www.norseq.org>

**Table 1. Current NeLS workflows.**

Category	Name	Description	Node
DNA-seq	Germline variant calling	Discovery of germline variation in DNA-seq samples	UiO
	Somatic variant calling	Discovery of somatic variation based on a sample pair	UiO
	LiceBase <sup>20</sup> z cDNA mapping	cDNA to genome mapping workflow for sea lice samples	UiB
RNA-seq	Eukaryote RNA-Seq	DE analysis between two collections of eukaryote samples	NTNU/UiB
	Prokaryote RNA-Seq	DE analysis between two collections of prokaryote samples	UiT
	LiceBase RNA-Seq	Alignment and count workflow for multiple Sea Lice samples	UiB
	RNA-seq counts - STAR <sup>21</sup>	Create RNA count matrix from RNA-seq FASTQ files	UiB
	RNA-seq counts - HISAT2 <sup>22</sup>	Create RNA count matrix from RNA-seq FASTQ files	UiB
miRNA-seq	miRNA prediction	Prediction of miRNA	UiO
	miRNA processing	Alignment and DE analysis between two collections of samples	NTNU
ChIP-Seq	ChIP-Seq analysis	<i>workflows in testing, to be released soon</i>	NTNU/UiO
Metagenomes	Taxonomic classification	Taxonomic profiling of 16S rRNA reads from shotgun reads	UiT
	META-pipe <sup>23,24</sup>	Functional annotation of assembled metagenomic shotgun samples	UiT

pick interesting targets for further investigation. The NeLS version of META-pipe provides taxonomic and functional analysis from whole-genome shotgun sequence data. It supports high-throughput sequencing data and provides assembly, focusing the analysis on full-length genes. The pipeline consists of three major modules: preprocessing, taxonomic classification, and functional analysis. All modules are available as individual workflows, except for assembly in pre-processing, which is run manually on either a high-memory computer or our supercomputer. Workflows can be tailored to the specific needs for the analysis of a sample and it is also possible to add additional steps or to omit some of the steps.

To use META-pipe, the user follows the generic steps for login and data upload above, using the NeLS Galaxy instance of UiT<sup>2a</sup> and the NeLS portal to administrate the data of the metagenomics project. Input files are transferred from the NeLS project in Storage Level I, to the Galaxy history of the user using the provided NeLS data transfer tools in Galaxy tool menu.

To analyze the data, the user selects the META-pipe tool in Galaxy and then configures the pipeline parameters such as which tools to run, input files, reference database versions, and output formats. Once the workflow is configured, the user presses the execute button in Galaxy to execute the pipeline in the background. This will create a history element in Galaxy where the user can view the current status of the job. Currently queued or running jobs are colored yellow, and completed jobs are colored green. When the job is done, the user can examine the output data in the Galaxy view panel, transfer results to the NeLS project in Storage Level II, or download the files to their own computer.

### 5.3 Sensitive data

NeLS was not designed for hosting sensitive data such as human genome data from Norwegian patients. ELIXIR Norway is collaborating with the TSD project and infrastructure<sup>ab 19</sup>, created by USIT (The University Centre of Information Technology) at the University of Oslo, to offer a service to researchers in Norway for storing and processing sensitive data, including health data.

NeLS allows for seamless data transfer from the NeLS storage services (Layer I) to the TSD File Lock servers for import of supplementary non-sensitive data that user projects would need available inside TSD to interpret their sensitive data.

Workflow development for sequencing data analysis *etc.* performed in ELIXIR Norway is implemented either as tools and workflow definitions for Galaxy or as software containers, such that workflows can be deployed in the appropriate compute environment to facilitate analysis of both sensitive and non-sensitive data. ELIXIR Norway and TSD are working towards a Galaxy service in TSD, and the aspect of workflow mobility is also a key aspect in the Nordic project Tryggve2<sup>ac</sup>, with ELIXIR partners from Norway, Finland, Sweden, and Denmark, and respective national infrastructures for sensitive data.

### 5.4 Example bioinformatics analysis project

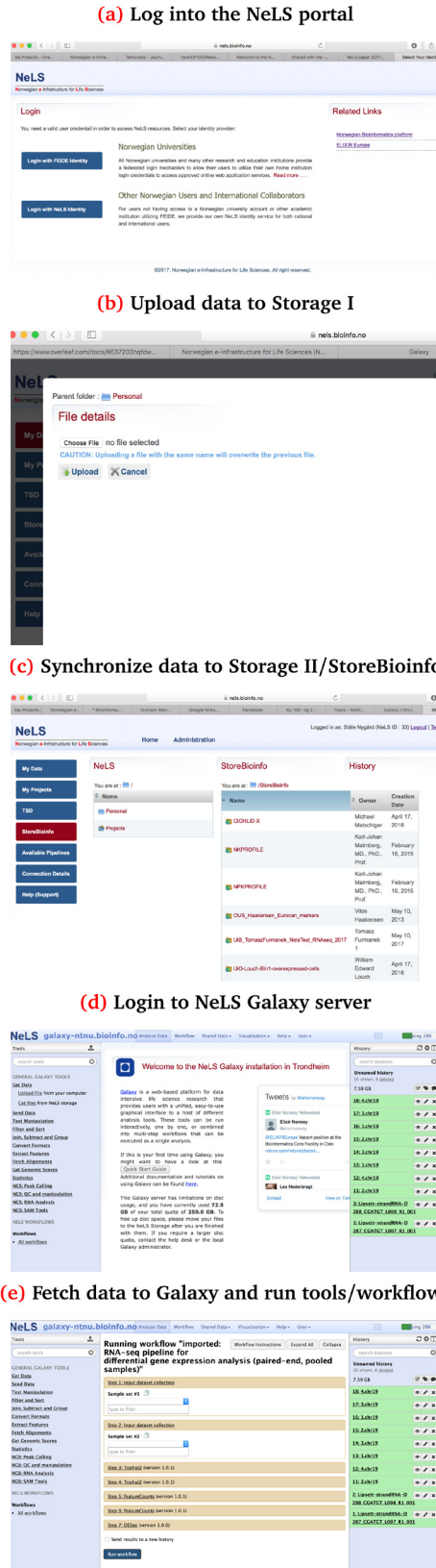
To illustrate how NeLS is used in daily operation of the help desk of ELIXIR Norway, we include an example. The help desk was contacted by a researcher wanting to analyze 15 RNA-Seq

<sup>2</sup>[licebase.org](http://licebase.org)

<sup>2a</sup><https://galaxy-uit.bioinfo.no>

<sup>ab</sup>Services for Sensitive Data, in Norwegian *Tjenester for Sensitive Data*

<sup>ac</sup><https://neic.no/tryggve2/>



samples from the Atlantic salmon. We decided that the already prepared RNA-Seq workflow in our Galaxy instance at NTNU Trondheim consisting of HISAT2 alignment<sup>22</sup>, followed by read assignment by featureCounts (subRead package)<sup>28</sup> and differential expression analysis in voom<sup>29</sup>, would be suitable for the initial analysis of the data. Atlantic salmon is not an organism with pre-processed genome and transcriptome readily available, so our help desk first had to create a HISAT2-indexed reference genome from the original Atlantic salmon genome FASTA file and transcriptome GFF file downloaded from SalmoBase<sup>30</sup>. The initial indexing (using HISAT2 indexing with 1.5TB of memory) was done by the ELIXIR Norway staff at NMBU Ås responsible for SalmoBase. This step only needs to be performed once for any reference genome, and can be reused for other users targeting the same organism. The indexed reference was made available for selection in the workflow by ELIXIR Norway staff at NTNU. To run the workflow, the NTNU help desk created a shared project in NeLS Storage Level I layer, and shared the project with the researcher. They in turn uploaded the raw sequencing data to the shared project in NeLS (then becoming available to the responsible person in the help desk), who ran the RNA-Seq workflow in the Galaxy environment. The workflow made use of the dataset collection feature in Galaxy to run alignment and read-assignment on all 15 samples in a single step. Sample group assignments for comparison in differential expression analysis can be defined at the beginning of the workflow, or by adding assignment and comparisons during a re-run of the last step in the workflow (voom analysis). In this way, the user only needs to run the computationally demanding alignment and assignment steps once, but still have the flexibility to change samples assignments and group comparisons in subsequent analysis. In total, four group comparisons were made, reporting differentially expressed genes in each comparison. The total processing and analysis were done with a minimal effort for the user who basically only had to 1) upload the data, 2) define a dataset collection, 3) select the correct organism reference genome (for alignment) and transcriptome (for counting), and 4) define the sample groups assignments and the groups to be compared (differential analysis).

### 6 Unified service toward data-generating platforms

National or other large data-generating platforms, such as the Norwegian Consortium for Sequencing and Personalized Medicine (NorSeq<sup>ad</sup>) produce user-requested sequencing for multiple purposes. The data are produced on receipt of DNA samples, and these may be of both human and non-human nature, requiring different data handling procedures. The goal is to provide a unified and seamless user experience, in which the user is provided with a resulting dataset in an environment that is suitably equipped with compute resources, relevant analytical tools, reference data, and initial analysis results. All of this should be provided and documented with no action required from the user after the initial agreement. This requires a tight collaboration between the data-generating platform, the national hardware (storage, compute) resources, as well as the ELIXIR help desk

Figure 3. Illustration of the main steps in an ordinary NeLS project.

<sup>ad</sup><https://www.norseq.org>



facilitating for the user experience in providing relevant tools, workflows, support, and documentation.

Non-sensitive data handling is coordinated through the use of NeLS and its layered architecture. NorSeq staff uploads the generated data to a NeLS project area created by ELIXIR Norway help desk on behalf of the research group ordering the sequencing. After verification of the uploaded data, ELIXIR Norway help desk assists the research group in synchronizing the data also to the Store-Bioinfo services at the Norwegian Infrastructure for Research Data (NIRD), and provides access in different roles to the different members of the research project. Users may then analyze the data utilizing the Galaxy front ends in the NeLS ecosystem, and receive support and training from the ELIXIR Norway help desk.

Sensitive data handling is achieved by utilization of Services for Sensitive Data (TSD)<sup>19</sup>. NorSeq staff uploads the generated data to a special TSD project that allows for initial data analysis using ELIXIR Norway provided workflows jointly by NorSeq and ELIXIR Norway help desk staff, before the raw and processed data are made fully available to the user's TSD project.

## 7 Discussion

We have described the NeLS system developed to serve a broad spectrum of bioinformatics users, with focus on Norwegian users and on genomics data. The system supports data storage, sharing, and data analysis in a project-oriented fashion. A strength of the system is that it utilizes a federated identity provider allowing most users to use their institutional login. Furthermore, it integrates storage and compute resources offered by the generic infrastructure Sigma2, set-up and funded to support users across all research fields in Norway. This avoids duplication of effort and caters for a more harmonized policy with respect to allocations of compute and storage between life science and other fields. An additional strength of the system is that it has interfaces adapted to both advanced users through a programmatic interface (API) and SSH, and to less computer-savvy users through a web portal. This allows different categories of users to work efficiently with the system, and to collaborate through joint projects. The system has been in production since 2015, and has been adapted according to user feedback accumulated over a series of workshops.

The system has been designed to use existing open-source solutions whenever possible. We believe this strategy produces a system that is easier to maintain and therefore more sustainable. NeLS has been developed in an iterative fashion with short agile development cycles facilitating adaptation to changing needs.

Until now, we have had one instance of NeLS running at the University of Bergen, linked with five instances of Galaxy, one at each of the partner institutions in ELIXIR Norway. For the future, we are investigating a more dynamic approach launching Galaxy instances on demand.

NeLS itself does not provide the level of security required for handling sensitive data. To support such projects, NeLS is linked with the TSD (Services for Sensitive Data) platform in Oslo<sup>19</sup>.

NeLS and TSD are integrated, allowing transfer of data and workflows between the systems, making for more resource-efficient support of both types of projects benefiting both their operation and their users.

NeLS uses the national Federated Electronic Identity provider (FEiDE) linking all Norwegian universities. The technology used is the same as that used for the ELIXIR Authentication and Authorisation Infrastructure (AAI)<sup>15</sup>. It is therefore possible to extend NeLS to also support ELIXIR AAI identity provision. NeLS has so far been designed and resourced for supporting Norwegian projects, and new policies – and ideally also new funding mechanisms – would be needed to extend the scope beyond Norwegian projects.

The NeLS system can be used as an example of how to set up a flexible and relatively light-weight system providing bioinformatics projects with data storage, sharing, and analysis. The NeLS source code is available on GitHub and can help in the building of similar projects elsewhere, although adaptations must be expected, for example to integrate with storage and compute resources.

The modularity of NeLS allows its parts to be reused in other contexts. An example is the integration of NeLS with the SEEK platform<sup>2,3</sup>, where users can link data sets in NeLS with their metadata in SEEK. Future work may include functionality for allowing users to export annotated data from NeLS (optionally integrating linked metadata from SEEK) into public data repositories such as ArrayExpress<sup>31</sup> and PRIDE<sup>32</sup>.

## 8 Conclusions

The NeLS system is in production and serves as an important platform for the operation of ELIXIR Norway and its help desk for users in molecular life sciences. The system will therefore be maintained and supported in the foreseeable future. We benefit from sharing experiences with other similar projects within and beyond ELIXIR, through the wide adoption of Galaxy across many ELIXIR nodes.

### Data availability

All data underlying the results are available as part of the article and no additional source data are required.

### Software availability

Bio.Tools<sup>33</sup> ID: *NeLS* (<https://bio.tools/nels>)

RRID: SCR\_016301

NeLS is available at <https://nels.bioinfo.no>, without extra registration for all Norwegian academic users (via FEiDE<sup>36</sup>), and with registration upon request for all other users.

The source code of the core NeLS modules is available at <https://github.com/elixir-no-nels/nels-core>, under the Apache License 2.0<sup>37</sup>.

<sup>36</sup><https://www.feide.no/introducing-feide>

<sup>37</sup><https://www.apache.org/licenses/LICENSE-2.0.html>

Archived source code at the time of publication is available here: <http://doi.org/10.5281/zenodo.1251639> under an Apache License 2.0<sup>34</sup>.

## Use cases

The source code of the module integrating NeLS with the national Service for Sensitive Data (TSD), can be found within the above repository and archive in [elixir-no-nels/nels-core/tsd-proxy](#).

The source code of the integration module of META-pipe with NeLS, including Galaxy front end and HPC back end (Stallo supercomputer at UiT), is available under the MIT license<sup>35</sup> at <https://gitlab.com/uit-sfb/meta-pipe-galaxy-wrapper>, archived in 35.

<sup>35</sup><https://opensource.org/licenses/MIT>

## References

- Nygård S, Jonassen I: **Norwegian Bioinformatics Platform**. *NBS Nytt*. 2014; **36**(2): 32–35.  
[Reference Source](#)
- Wolstencroft K, Owen S, Krebs O, et al.: **Semantic Data and Models Sharing in Systems Biology: The Just Enough Results Model and the SEEK Platform**. In *Proceedings of the 12th International Semantic Web Conference - Part II. ISWC '13*, New York, NY, USA, Springer-Verlag New York, Inc. 2013; 212–227.  
[Publisher Full Text](#)
- Wolstencroft K, Owen S, Krebs O, et al.: **SEEK: a systems biology data and model management platform**. *BMC Syst Biol*. 2015; **9**(1): 33.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Reich M, Liefeld T, Ocana M, et al.: **GenomeSpace: an environment for frictionless bioinformatics**. *F1000Posters*. Poster. 2013; **4**: 804.  
[Reference Source](#)
- Garamszegi S, Mesirov JP, The GenomeSpace Team: **GenomeSpace: An environment for frictionless bioinformatics [v1; not peer reviewed]**. *F1000Res*. Poster. 2015; **4**(ISCB Comm J): 349  
[Publisher Full Text](#)
- Qu K, Garamszegi S, Wu F, et al.: **Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace**. *Nat Methods*. 2016; **13**(3): 245–247.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sandve GK, Gundersen S, Rydbeck H, et al.: **The Genomic HyperBrowser: inferential genomics at the sequence level**. *Genome Biol*. 2010; **11**(12): R121.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sandve GK, Gundersen S, Johansen M, et al.: **The Genomic HyperBrowser: an analysis web server for genome-scale data**. *Nucleic Acids Res*. 2013; **41**(Web Server Issue): W133–W141.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Simovski B, Vodák D, Gundersen S, et al.: **GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome**. *GigaScience*. 2017; **6**(7): 1–12.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Giardine B, Riemer C, Hardison RC, et al.: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome Res*. 2005; **15**(10): 1451–1455.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Algan E, Baker D, van den Beek M, et al.: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update**. *Nucleic Acids Res*. 2016; **44**(W1): W3–W10.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kumar S, Skjæveland A, Orr RJ, et al.: **AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses**. *BMC Bioinformatics*. 2009; **10**(1): 357.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kumar S, Krabberød AK, Neumann RS, et al.: **BIR Pipeline for Preparation of Phylogenomic Data**. *Evol Bioinform Online*. 2015; **11**: 79–83.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kalaš M: **Efforts towards accessible and reliable bioinformatics**. PhD thesis, University of Bergen, Norway, 2015.  
[Publisher Full Text](#)
- Linden M, Procházka M: **ELIXIR Authentication and Authorisation Infrastructure (AAI) [version 1; not peer reviewed]**. *F1000Res*. Poster. 2016; **5**(ELIXIR): 332.  
[Publisher Full Text](#)
- Fielding RT: **Architectural Styles and the Design of Network-based Software Architectures**. PhD thesis, University of California, Irvine, 2000.  
[Reference Source](#)
- Richardson L, Ruby S, Hansson DH: **RESTful Web Services**. O'Reilly, 2007.  
[Reference Source](#)
- Richardson L, Amundsen M, Ruby S: **RESTful Web APIs**. O'Reilly, 2013.  
[Reference Source](#)
- Azab A, Domanska D: **Software Provisioning Inside a Secure Environment as Docker Containers Using STROLL File-System**. In *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*. IEEE, 2016; 674–683.  
[Publisher Full Text](#)
- Eichner C, Dondrup M, Nilsen F: **RNA sequencing reveals distinct gene expression patterns during the development of parasitic larval stages of the salmon louse (*Lepeophtheirus salmonis*)**. *J Fish Dis*. 2018; **41**(6): 1005–1029.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dobin A, Davis CA, Schlesinger F, et al.: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics*. 2013; **29**(1): 15–21.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods*. 2015; **12**(4): 357–360.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Robertson EM, Kahlke T, Raknes IA, et al.: **META-pipe-Pipeline Annotation, Analysis and Visualization of Marine Metagenomic Sequence Data**. *CoRR*. 2016; abs/1604.04103.  
[Reference Source](#)
- Agafonov A, Mattila K, Tuan CD, et al.: **META-pipe cloud setup and execution [version 2; referees: 1 approved, 1 approved with reservations]**. *F1000Res*. 2018; **6**(ELIXIR): 2060.  
[Publisher Full Text](#)
- Raknes IA, Bongo LA: **META-pipe Authorization service [version 1; referees: 2 approved with reservations]**. *F1000Res*. 2018; **7**(ELIXIR): 32.  
[Publisher Full Text](#)
- Robertson EM, Raknes IA, Tartari G, et al.: **ELIXIR Pilot: Marine metagenomics [version 1; not peer reviewed]**. *F1000Res*. Poster. 2016; **5**(ELIXIR): 864.  
[Publisher Full Text](#)
- Robertson EM, Denise H, Mitchell A, et al.: **ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services [version 1; referees: 1 approved, 2 approved with reservations]**. *F1000Res*. 2017; **6**(ELIXIR): 70.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics*. 2014; **30**(7): 923–930.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Law CW, Chen Y, Shi W, et al.: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome Biol*. 2014; **15**(22): R29.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Samy JKA, Mulugeta TD, Nome T, et al.: **SalmoBase: an integrated molecular**

- data resource for *Salmonid* species.** *BMC Genomics.* 2017; **18**(1): 482.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Parkinson H, Sarkans U, Shojatalab M, *et al.*: **ArrayExpress--a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res.* 2005; **33**(Database issue): D553–5.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  32. Vizcaíno JA, Reisinger F, Côté R, *et al.*: **PRIDE: Data Submission and Analysis.** *Curr Protoc Protein Sci.* 2010; Chapter 25: Unit 25.4.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  33. Ison J, Rapacki K, Ménager H, *et al.*: **Tools and data services registry: a community effort to document bioinformatics resources.** *Nucleic Acids Res.* 2016; **44**(D1): D38–47.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  34. Tekle KM, Li X, Zhang W, *et al.*: **elixir-no-nels/nels-core: NeLS core 2018-05-16.** *Zenodo.* 2018.  
[Data Source](#)
  35. Bongo AL, Raknes IA: **META-pipe Galaxy Wrapper.** Software version. *Zenodo.* 2018.  
[Data Source](#)

# Open Peer Review

Current Referee Status:  

---

Version 1

Referee Report 13 September 2018

doi:[10.5256/f1000research.16472.r35602](https://doi.org/10.5256/f1000research.16472.r35602)



**Wolfgang Mueller**

Heidelberg Institute of Theoretical Studies gGmbH, Heidelberg, Germany

I am a member of ELIXIR, and as member of the FAIRDOM project, I am heavily involved with the SEEK system whose link to NeLS is hinted inside the paper. However, I have been only loosely involved with that, so I was deemed to be an appropriate reviewer. In order to make things crystal clear, I hereby make my status visible.

In my view, this paper has diverse functions:

- It provides an architectural view on how a comparatively lightweight combination of a variety of pre-existing tools can yield a powerful national research infrastructure.
- It provides a short justification of many architectural decisions.
- It describes main steps in an ordinary NeLS project.
- It provides a description of typical use cases and a reference of workflows that can be run by users.

I find the paper clear and readable.

I second the major remark by Olivier Collin.

I was also surprised to find a whole paper about an infra structure that does not reference the FAIR principles explicitly. I think they could provide some point of reference regarding the handling of metadata, identifiers, licenses.

However, it is only a matter of making existing links to FAIR more explicit.

One suggestion: Please adjust the numbering in the drawings with the numbering of subsections in the paper. This would make things much easier.

Bullet point 2.3 (c) references subsection 2.4, shouldn't this be 2.8, the public API? Please clarify.

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** I am coauthor on a Nettab Abstract with Inge Jonassen. No publication, but conference contribution. Collaboration on linking SEEK and NeLS.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Referee Report 16 July 2018

doi:[10.5256/f1000research.16472.r35600](https://doi.org/10.5256/f1000research.16472.r35600)



**Olivier Collin** 

Univ Rennes, Inria, CNRS, IRISA F-35000, Rennes, France

The authors describe the architecture of NeLS, the Norwegian e-Infrastructure for Life Sciences. Use cases provide an overview on how the NeLS infrastructure operates. The infrastructure can deal with sensitive and non-sensitive data.

The microservices oriented architecture allowing several universities to propose a national service is well explicated. The integration with data-generating platforms will help users to manage efficiently their research processes.

The article is well written and provides a good overview of the infrastructure. The description of the use case improves the understanding of the operations. Additional informations and source code are available on a GitHub repository.

### Comments

*Major remark :*

The data management could be more explicitly described in the article.

The data analysis workflow is well described from the data analysis point of view but the usage of SEEK in the NeLS infrastructure is only briefly mentioned. Some additional description on SEEK usage and



interest for NeLS users could improve the article.

In the Figure 1 : there is a mention of "data curation into structured storage". This data curation is not explained in the text and this raises some questions. Is this data curation using SEEK ?

*Minor remarks :*

In the abstract, it is stated that "NeLS is also integrated with the SEEK platform in order to store large data files produced by experiments described in SEEK". This should be rephrased since it gives the impression that the SEEK platform is used to store large data files.

In the second chapter of the introduction, the flow of ideas is not clear. A comparison is made between BaseSpace, Geneious and SEEK before mentioning GenomeSpace. Why compare BaseSpace, Geneious and SEEK that are tools that do not really compare since the first two are focused on data analysis and the last on data management? Maybe there could be some rephrasing needed (this could also help to improve the data management focus mentioned in my first comment).

Concerning BaseSpace and Geneious, it could be less judgemental to say that their service is chargeable instead of expensive. And to say that their platform is private.  
It could be interesting to describe briefly why the GenomeSpace setup could not be adapted.

In the data reproducibility chapter some reference genomes are made available on the five Galaxy instances. Is it possible to describe how this is achieved and how everything is synced?

Some quantitative data about the number of users are needed in order to better estimate the community covered.

In the abstract there is a mention of a "project-oriented fashion". This expression is strange to me. Maybe replacing fashion by way or method or mode?

*Typo :*

The headlines are numbered and sometimes there is a dot after the number and sometimes not. And they are not numbered in the templates if I am not mistaken.

In 2.1 Authentication

whether the user is who he claims to be instead of whether the user is who they

Page 9 col 1 line 29 : e-infrastructure

Page 9 col 2 line 15 : some strange font glitches in the pdf file (but not online)

Page 9 col 2 line 17 : can help in instead of helpin

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**