Ref. Ares(2019)7354788 - 29/11/2019

PaNOSC

Photon and Neutron Open Science Cloud

H2020-INFRAEOSC-04-2018

Grant Agreement Number: 823852

Deliverable:

Report on the current technical elements of data analysis at each partner site (4.1)

Project Deliverable Information

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Sheet

Project Reference No.	823852
Project acronym:	PaNOSC
Project full name:	Photon and Neutron Open Science Cloud
H2020 Call:	INFRAEOSC-04-2018
Project Coordinator	Andy Götz (andy.gotz@esrf.fr)
Coordinating Organization:	ESRF
Project Website:	www.panosc.eu
Deliverable No:	D4.1
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	Month 12
Actual Delivery Date:	November 30th 2019
EC project Officer:	Rene Martins

Document Control Sheet

Document	Title: Report on the current technical elements of data analysis at each partner site
	Version: 1.0
	Available at:
	Files: N/A
Authorship	Written by: FANGOHR Hans (EuXFEL), GALAL Kareem (ESS), HALL Jamie (ILL), KLUYVER Thomas (EuXFEL), REIS Carlos (CERIC-ERIC), ROSCA Robert (EuXFEL), TURNER William (ILL), VINCENT Thomas (ESRF)
	Contributors: CAUNT Stuart (ILL), DIMPER Rudolf (ESRF), FANGOHR Hans (EuXFEL), GAIZER Tamás (ELI-ALPS), GALAL Kareem (ESS), GÖTZ Andy (ESRF), GREENWOOD Lottie (ESS), GROSZ Jakub (ELI), HALL Jamie (ILL), KLUYVER Thomas (EuXFEL), MARKVARDSEN Anders (STFC), PLÖTZENEDER Birgit (ELI-ALPS), PUGLIESE Roberto (CERIC-ERIC), REIS Carlos (CERIC-ERIC), ROSCA Robert (EuXFEL), SCHRETTNER Lajos (ELI-ALPS), TURNER William (ILL), VINCENT Thomas (ESRF)
	Reviewed by: Hans Fangohr (euXFEL)
	Approved: Andy Götz (ESRF)

List of participants

Participant No.	Participant organisation name	Country
1	European Synchrotron Radiation Facility (ESRF)	France
2	Institut Laue-Langevin (ILL)	France
3	European XFEL (XFEL.EU)	Germany
4	The European Spallation Source (ESS)	Sweden
5	Extreme Light Infrastructure Delivery Consortium (ELI-DC)	Belgium
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	The Netherlands

Table of Contents

The Survey

We have carried out a survey of existing data analysis infrastructure. The survey was designed by and distributed to all PaNOSC partner facilities (CERIC-ERIC, ILL, XFEL, ESS, ELI and ESRF) with the aim of building up a view of data analysis needs and services at each facility and to provide a basis for developing a set of requirements for the development of services for WP4. We also include the responses from the ISIS Neutron and Muon Facility, which is a partner in the ExPaNDS project and offered to participate in the survey.

The survey was split into five sections:

Scientists and Data: Data generation, user-community and scientific nature of each facility for today's situation and forecast for 2023.

Data analysis and reduction: Tools and services concerning data analysis and reduction.

Technology: Determines the IT infrastructure available for data analysis purposes.

Security: Provides a global view of the security requirements and solutions at each facility.

Other information: Additional information that could be useful for the development of analysis services for PaNOSC.

Survey results

Data Analysis workflow introduction

The data analysis starts with data being recorded during experiments. Already at this point so-called "online" data analysis is carried out to guide the experiment and estimate how much further data needs to be taken for the given sample and parameter settings. For this, data can be processed from saved files, or streamed over the network. For some institutions, the data rate is so high that only a fraction of the data can be processed in this streaming manner.

The data is saved to file (see *Data formats* below), and then processed, both while the experiment is carried out and the users may be present at the research facility, and remotely after the experiment, often months or even years after the data taking.
The data analysis may involve the use of high performance computing clusters where the computational requirements exceed single workstations. For some facilities and experiments, the data could be taken home, at others the size of data from a 5 day beamtime, which can easily be 100s of TB, makes this impossible.

Below, we detail various aspects of these workflows and the respective technologies used, comment on good practice, and discuss challenges within the PaNOSC project and the aim of EOSC integration. We conclude with a summary and priorities.

Persistent identifiers and data access

All partner sites can allocate Digital Object Identifiers (DOIs) to data sets, and are using or planning to use DataCite as the registration agency for persistent identifiers.
Some facilities have a data portal to access scientific data. However most of these only enable users to search and download their data through a web interface and do not offer an API that can be consumed

programmatically. WP3 is currently working on developing a common data API that will enable searching for data at each PaNOSC facility. WP4 will use this API in the common portal.

Data formats

The rate of data production at each facility ranges from tens of terabytes (TB) to about 15 petabytes (PB) per year today, and is foreseen to increase to between 0.5 PB and 100 PB per year by 2023. Data storage and data formats are of paramount importance.

The HDF5 file format (https://www.hdfgroup.org/solutions/hdf5/) is a common ground for all partners as they are either using it, moving towards it or considering to do so. HDF5 is a data file format for archiving hierarchical, multi-dimensional data with features such as compression and links to data stored in different files. This format copes with the needs of different experiments producing data where size per experiment ranges from a few kilobytes (KB) up to a record of 856 TB so far.

Five out of the seven PaNOSC institutes are also using the NeXus (https://www.nexusformat.org/) convention for metadata. NeXus provides a common ontology based on HDF5 for neutron, x-ray, and muon science, as a convention for naming and structuring the data. NeXus files support storing all additional metadata, such as the beam energy or detector geometry, aiming to allow data analysis without requiring external data.

For experiments where NeXus is not being used, data is stored in various institute specific or experiment specific formats.

The general use of HDF5 should ease re-use of tools across site by providing a common way to access the data stored in files. However, the structure of the data inside the HDF5 file is generally determined by each facility, limiting the immediate re-usability of existing analysis tools across sites. NeXus files may help to improve the situation.

WP3 is working towards improvement of metadata capture alongside the detector data, but it is well beyond the scope of this project to solve all challenges associated with that, and to achieve the efficient encoding of the metadata in machine readable files (such as NeXus).

Remote data analysis

All facilities provide remote access through which data can be analysed remotely. The particular technologies for this vary and include ssh access (with or without X forwarding), remote desktop applications such as VISA and FastX, and Jupyter notebooks.

Jupyter notebook and Jupyter notebook services

Jupyter notebooks are used across the sites, and within the last year the majority of partners have set up a JupyterHub instance. Some of these are in production, others are installations for a smaller part of the facility, opened only to test users or in development. There is agreement that users welcome a service that allows remote access to Jupyter notebooks that can execute on centralised compute resources at the facilities. Our vision is to extend this capability to be accessible through the EOSC as well.

The use of Jupyter notebooks is considered good practice as the notebooks can combine analysis code, outputs from the code and interpretation in one document. This is a step towards reproducible and re-usable data analysis.

We have found (and support) an emerging practice of collecting notebook templates for common data analysis procedures, allowing a user to analyse their data following a recipe. This approach combines the advantages of a facility providing pre-tested code for a specific experimental setup with the ability for users to readily modify it to suit their needs or to react to unforeseen changes.

Training and support

All partners offer training and support for their users, but the level of the training and content is different at each site. Within PaNOSC we hope to develop training materials to support best-practice technologies and methodologies that help to move closer to FAIR data and data analysis. WP4 will be working closely with WP8 to agree common training plans and materials.

Computing infrastructure

All partners have High Performance Computing (HPC) installations of varying size. These allow users to efficiently analyse and process experimental data – it is increasingly unrealistic to analyse the volumes of data produced on desktop systems or individual work stations. For the largest data sets it is also a requirement to have physical co-location of data storage and processing computing power as it is unrealistic to move data sets of the order of petabytes over networks on demand. This is a challenge for the EOSC ambition of allowing central access to data and data processing: for the largest data sets this is likely to be only possible if computer resources close to the data storage can be found and used.

All of the partners, with the exception of the ILL which is using Torque, are using Slurm for their HPC job submission system.

Authentication and authorisation (AAI)

Across the sites, different frameworks for authentication and authorisation are used. Four of the seven facilities are not currently using single-sign-on (SSO), however there is a strong push to deploy SSO solutions at these facilities. Single sign on avoids users having to enter their username and password repeatedly, and is generally seen as desirable to improve the user experience.

Establishing a common authentication workflow using the same AAI protocols (i.e SAML, OpenID Connect) across each partner site will be a challenge when moving towards EOSC integration. Following best-practices and implementation details from WP6 will help solve this.

Software deployment and containers

To re-use and reproduce a given analysis – as part of moving towards FAIR data and data analysis – one should be able to repeat the analysis carried out on a particular data set in the past. We mentioned above how version control can help with archiving the software that was used to carry out this analysis. However, research software typically have dependencies ranging from compression to numerical libraries, and these can affect the processing results. It is thus important to archive these dependencies as well. Furthermore, the installation of the software may be non-trivial, and may rely on third parties continuing to provide software downloads in a consistent way. One solution to these challenges might be the use of containers to archive software environments.

The survey results show that all facilities have strong interests in container technologies, and storing/developing software in containers – either for convenient software deployment and versioning, or for the purposes of returning to an older environment to reproduce previously run analyses.

Most facilities already use Docker, with some using Singularity as well. Despite the ubiquity of Docker, it has limitations in the context of HPC which might lead us to focus on another option. The number of container runtimes has exploded in recent years (containerd, rkt, podman, cri-o, singularity, etc…). We should aim to not tie ourselves too closely to a specific technology, in case we later discover a need to migrate to an alternative. Standards work such as the Open Container Initiative could help with this.

Source code versioning

Most facilities are using Git as a version control system (storing their repositories on GitHub or internal GitLab servers). Using version control to keep track of changes in software and thus in analysis tools and infrastructure is good practice, and recommended. Git is the most widely used version control tool, and therefore well supported.
For PaNOSC we propose to move the source code for tools and services developed throughout the PaNOSC project onto the GitHub platform as open source software. This will facilitate cross-site collaboration and will serve as a single point of access for development activities.

Summary

We have identified common practice in provision of data analysis infrastructure at photon and neutron facilities in this project. Based on the requirements and existing services from partners, we suggest to focus further work on remote data analysis through Jupyter notebooks and remote desktop services: The Jupyter Notebook approach shows great potential for reproducibility, user convenience and our efforts to move towards FAIR data but is not applicable for all analysis requirements: in particular the notebooks cannot be used for existing analysis tools that are based on a graphical user interface. The remote desktop services can cater for all existing software, as they make a graphical desktop available in a web browser. We aim to put together demonstrator use cases that show how facility data can be analysed remotely using analysis templates. Together with WP3, a portal is under development that aims to serve as an entry point to remotely start an analysis session for a given data set.

The development and deployment of this portal across multiple sites is likely to be challenging due to the different AAI, data and computing infrastructure. Another challenge for data access through EOSC is that for data analysis of large data sets the only realistic option is to carry out the data analysis close to the data, and that each facility uses somewhat unique hardware and infrastructure. It may thus be necessary to exclude some data sets of largest size from the demonstrator. Finally, we note that it is non-trivial to automatically propose appropriate data analysis templates for given data sets, which will be a challenge for EOSC and the community for a long time. The efforts on better capture of metadata (WP3) will help to move closer to this goal, but we expect to be limited to a set of example data sets and associated data analysis processes within this project.

The use of containers as a way to ship software environments and enhance reproducibility shows potential and will be explored further. To ensure that the project is sustainable in the future, we must keep an eye on how the container ecosystem evolves over time. We also need to monitor, and if appropriate engage with, other developments in the field of technologies relevant to the EOSC vision, including advances in the HDF5 and Jupyter ecosystem.

Detailed Survey responses

Scientists and Data

1. How many scientific visitors do you receive per year?

Institute	Current situation	Forecast 2023
CERIC-ERIC	400	500
ELI	The ELI Facilities are currently performing user-assisted commissioning experiments. They will gradually open peer-reviewed access to the experimental facilities, the expectation being to have steady-state operations in 2023. In 2019, ELI Beamlines (ELI-BL) had a number of scientific visitors for both commissioning calls (not very relevant regarding data: E1/TERESA, anticipated E3 “first light”) and a pilot call (L1-E1) with ca 4-5 different groups. The number of external users at ELI-ALPS is of the same order.	ELI-ALPS: circa 300 visitors per year, being understood that groups behind these experiments include many more scientists than the 300 indicated here; ELI-BL: 280 users
ESS	0 (now)	ESS is scheduled to start in 2023. Visitor numbers not known for now.
ESRF	6548 (2018). Some users are coming several times. The estimation of unique individuals visiting ESRF in 2018 is about 3000.	10,000
ILL	2000 visitors last year.	2200 visitors.

XFEL	Variable – 650+ (local) and 100+ (remote) in 2018, instrument commissioning is still ongoing,	- extrapolating from that estimates are 3,000+ users a year during steady operation.
ISIS	In 2018 ISIS had 2512 scientific visitors.	- There are more beamlines being built to fill out remaining available space within the target station 2 building, hence the number is likely to be higher than 2512 in 2023.

2. How many experiments are performed per year?

Institute	Current situation	Forecast 2023
CERIC-ERIC	There were 187 experiments approved in 2018, a 16% increase in comparison to 2017 (CERIC annual report 2018).	Expected increase of 20% (220 experiments).
ELI	-	ELI-ALPS: 95 experiments ELI-BL: 60 experiments
ESS	0	~100
ESRF	924	1200
ILL	600	700

XFEL	Currently ~3,000 hours of beam time available a year (per experiment) so it depends on how long the experiments run for. So around 50 to 150 experiments a year.	
ISIS	In 2018 ISIS had 1085 proposals from 32 countries and 264 Xpress proposals.	

3. How much data is produced per year?

Institute	Current situation	Forecast 2023
CERIC-ERIC	1 PB per year.	Expected to be around 50 pb per year
ELI	ELI-ALPS: 50 TB total ELI-BL: 400-500TB/year	10+ PB per year
ESS	0	< 500 TB
ESRF	9 PB	50 PB. The error bar on this number is large. It could be much more if our IT infrastructure is capable of dealing with it.
ILL	250 TB.	500 TB
XFEL	This year around 10PB	100 PB estimated once in full operation.

ISIS	About ~15 TB of raw data from April 2017 to April 2018.	With the neutron image technique beamline in full operation and improvement in detector technology this will increase significantly. A guess is ~30 TB by 2023

4. What are the mean and maximum data file sizes for an experiment?

Institute	Current situation	Forecast 2023
CERIC-ERIC	***Max***: dozens of GB (Tomography experiment at the Syrmep beamline); ***Min***: dozens of KB (experiments like Scanning Transmission X-ray Microscopy at the TwinMic beamline). ***Mean***: dozens of MB.	***Max***: Dozens of GB ; ***Min***: dozens of KB ; ***Mean***: dozens of MB
ELI	ELI-ALPS: - Min: 2.2 GB - Max: 9 TB - Mean: 1.04 TB ELI-BL: - Min: 2-4MB - Max: 2GB	
ESS	Currently undetermined.	

ESRF	Data file size is typically from 1MB to 16MB. Yet an experiment consists of many datasets, and a dataset consists of many files. For instance, mean and range for the (few) already publicly available experimental data: Mean: 500GB, Range: [2.4-3902] GB But that can be much more.	
ILL	Max: 70TB - Mean: dozens of GB	Max: 100TB - Mean: dozens of GB
XFEL	Largest proposal is 856TB, mean 70TB, standard deviation 137TB, proposals are made up of runs (largest directory is 30TB, mean is 350GB, standard is 1TB), and runs are made up of multiple data files (12GB each).	
ISIS	Range from hundreds KB to ~63 GB currently.	

5. Do you have any open data (freely accessible by anyone)?

Institute	Current situation	Forecast 2023
CERIC-ERIC	No, we don't have any open data available.	We aim at having open data according to the implementation of the CERIC Data Policy under development through the collaboration of PaNOSC WP2 activities.

ELI	No	It is something ELI-BL want to start as a pilot project in late 2020, but it depends on resource availability and prioritization.
ESS	Yes.	
ESRF	Yes, see https://data.esrf.fr 2019: 9 experiments comprised of 100 datasets	3000 experiments with 100s of thousands of datasets
ILL	Yes, see https://data.ill.eu 2019: 1300 datasets	The data policy has been in place since 2012. Data that is no longer embargoed is made available as Open Data. 2023: 3500 datasets
XFEL	Currently all data is under embargo, but some users have uploaded parts of their data to CXIDB where it can be downloaded by the public.	
ISIS	Yes, see https://www.isis.stfc.ac.uk/Pages/Data-Policy.aspx	

6. How long do you keep data for after an experiment?

Institute	Current situation	Forecast 2023
CERIC-ERIC	The new scientific Data Policy states we will keep data for at most 10 years.	Same as of today.
ELI	To be decided.	As long as possible on a best effort basis (data policy to be approved in the course of 2020)

ESS	Current ESS Data Policy states at least 5 years.	
ESRF	50 days on disk after the end of the experiment, one year on backup tapes, part of the data for 10 years in a tape archive.	30 days on disk after the end of the experiment, all data for 10 years in a tape archive.
ILL	Forever.	Forever.
XFEL	Anticipated minimum of 5 years, aim for 10 years for key data.	-
ISIS	Perpetuity.	Perpetuity.

7. How long is the embargo period for data produced at your facility?

Institute	Current situation	Forecast 2023
CERIC-ERIC	CERIC-ERIC Data Policy is under development and when it will be ready the embargo period will be 3 years.	3 years.
ELI	N/A	3 years (TBC).
ESS	Current ESS Data Policy states 3 years.	-
ESRF	3 years, extendable upon request.	-
ILL	3 years. Extended to 5 if a request has been made.	-
XFEL	3 years.	-

ISIS	3 years.	-

8. What data formats are used at your facility (*NeXus*, institute-specific…)?

Institute	Current situation	Forecast 2023
CERIC-ERIC	HDF5 based formats (including NeXus), csv, ssv, txt, tiff.	We are working to make *NeXus* format the standard data container for all instruments.
ELI	ELI-ALPS: data formats are experiment specific. ELI-BL: no real strategy has been developed yet.	We see an interest in HDF5 and it is something we use (and like). It is highly unlikely we will develop our own data formats.
ESS	NeXus.	-
ESRF	EDF, Specfile, HDF5, NeXus+HDF5, other data formats depending on detectors.	Majority of HDF5, NeXus+HDF5, and other data formats depending on detectors.
ILL	NeXus (for all instruments except nuclear physics and ROOT (for nuclear physics).	-
XFEL	Institute-specific HDF5.	-
ISIS	NeXus and for those instruments who are not yet using this format they use an ISIS binary .raw format.	-

9. Does your facility generate DOIs for your experimental data? If yes, when are they generated?

Institute	Current situation	Forecast 2023

CERIC-ERIC	DataCite has been chosen as our DOI provider. We are under testing phase with their API. We should start to generate DOIs by 2020.	We will generate DOIs for all experimental data through DataCite.
ELI	Not yet, but this is of course our intention.	
ESS	During data acquisition, DOIs will be assigned as soon as experiment data is first written to disk.	
ESRF	Yes. DOIs are generated at the beginning of the acquisition. 2018: Partially deployed on the beamlines.	Implemented on all beamlines.
ILL	Yes. After an experiment has been carried out.	For the moment we are only minting DOIs for proposals. By 2023, we should have DOIs for subsets of data.
XFEL	Yes. Generated along with metadata catalogue entry, only made public once the embargo is lifted.	-
ISIS	Yes currently for all except Xpress experiments. They are generated before cycles start.	-

10. Does your facility have a User Portal (e.g. to submit proposals)?

Institute	Current situation	Forecast 2023
CERIC-ERIC	Yes. CERIC-ERIC uses the Virtual Unified Office (VUO) portal to manage the whole experiment life cycle.	By 2023 we hope to integrate our User Portal with the partner facilities of PaNOSC.

ELI	The ELI Facilities are currently using ad-hoc solutions for proposal submission in the framework of commissioning calls.	A user portal will be developed in the coming years, with a first version anticipated in the second half of 2020 once ELI ERIC is established and launches its first call for peer-reviewed access.
ESS	Not yet, but this is planned to exist by 2023.	-
ESRF	Yes, https://smis.esrf.fr/.	-
ILL	Yes, https://userclub.ill.eu/userclub/.	-
XFEL	Yes.	-
ISIS	Yes.	-

11. How would a user download all of their data at your facility?

Institute	Current situation	Forecast 2023
CERIC-ERIC	The VUO portal allows for remote data browsing and downloading when data is still under embargo period. After this period data can be downloaded by contacting the IT group.	We will keep data downloading through VUO and will integrate it with the common portal under development in PaNOSC.
ELI	ELI-ALPS: FTP ELI-BL: *Catapult* (https://www.catapultsoft.com/)	To be decided (e.g. Globus).
ESS	Currently undetermined, but likely through a GridFTP-like solution or similar.	

ESRF	2018: Via the Internet at https://data.esrf.fr/ or through ICAT API for ~10% of the data. Portable media (disks) for ~90% of the data.	globus-online
ILL	Downloading using HTTP (https://data.ill.eu). SFTP transfer (dt.ill.eu).	Perhaps using OneData or Globus (undecided for the moment).
XFEL	In theory they wouldn't. However if they want to then we offer Globus for large downloads, alternatively for a small subset of files rsync or sftp would be an option.	-
ISIS	Through the ISIS Data catalogue.	-

Data Analysis and Reduction

1. Do you provide Jupyter Notebook/JupyterLab services to visitors?

Institute	Answer
CERIC-ERIC	**Today:** We have a prototype installation of JupyterHub which has been used for use cases testing. **Forecast 2023:** We plan to offer a production environment of the JupyterHub system (including sample Jupyter Notebooks, tutorials, GPUs) to visitors.
ELI	Not at present. ELI-ALPS have Jupyter notebook set up and internal staff is testing it. The plan, within the context of PaNOSC, is to integrate this technology to our offering and promote its use internally and to external users.
ESS	Jupyterhub service is currently in testing stage, expected to be in production by end of 2020.
ESRF	Yes: some beamlines have some JupyterHub servers running and JupyterHub+Kubernetes, BinderHub and JupyterHub+Slurm are also been made available recently.

ILL	Yes. We have recently deployed JupyterHub. It is currently in a pilot phase.
XFEL	Yes. JupyterHub is provided as a service to the users.
ISIS	No, not directly. But we provide machine with Python on where users could setup Jupyter Notebooks.

2. How many staff and visiting scientists/end-users currently use Jupyter Notebook/JupyterLab at your facility?

Institute	Answer
CERIC-ERIC	**Today:** About 5 people from our staff are currently in use of such technologies. **Forecast 2023:** We expect to increase those numbers as the infrastructure to offer such services in a production environment will be ready.
ELI	ELI-ALPS: 2 staff members ELI-BL : 2 staff members
ESS	See previous answer.
ESRF	No idea but only a small number (10+) up to now, yet growing.
ILL	Not many for the moment but hopefully we will see this number grow due to making notebooks available to ILL users.
XFEL	~150 during one week on the JupyterHub, plus an unaccounted number of users running their own Jupyter Notebook instances.
ISIS	This is not recorded.

3. Do external users currently have remote access to a data analysis/reduction services (remote desktops, grid computing, shell...)?

Institute	Answer
CERIC-ERIC	**Today:** Yes. A remote data browsing and visualization web application allows users to access and visualize HDF5 files through the VUO portal. Remote shell access is also available through the same portal. **Forecast 2023:** We aim at providing Jupyter Notebook data analysis/reduction services.
ELI	Not at the moment.
ESS	External users of the cluster can access it via SSH.
ESRF	2018: Remote desktop provided through NoMachine (nx.esrf.fr), else connection using SSH.
ILL	Yes. They have access to remote desktops (VISA) externally and the HPC cluster.
XFEL	Yes. SSH and remote desktop access to registered users.
ISIS	Only for those instruments that are using the new- ISIS Data Analysis as a Service platform. By 2023 this will have increased.

4. Do you offer data analysis/reduction training at your facility?

Institute	Answer
CERIC-ERIC	**Today:** We offer on-site support for data analysis/reduction during the experiment but not a formal training. **Forecast 2023:** With the developments of PaNOSC we will be able to offer in an easier way for local and remote training.
ELI	Some internal training on image processing has been initiated at ELI-BL and will repeat in the upcoming year. However, this is more intended to enable scientists to automate processes (alignment) without CS help.

ESS	The data analysis, reduction and modelling group offer training on specific software applications they develop.
ESRF	2018: We provide some Python trainings oriented towards data analysis (Training material: https://github.com/silx-kit/silx-training). 2023: Training material should be made available and possibly training should be complemented by tutoring services.
ILL	Not specifically at ILL, but we propose training as part of the software development collaboration.
XFEL	Occasional in person training for groups, one-to-one training on request, written documentation.
ISIS	Yes, through annual Neutron training school and Muon training school. Plus e-learning material, documentation of software used on beamlines and through training by instrument scientists.

5. Do your scientists support users with their data analysis/reduction after their experiments?

Institute	Answer
CERIC-ERIC	**Today:** Yes. Most of the times the scientific computing team and the beamline responsible scientists offer support for the data analysis/reduction required by the users. **Forecast 2023:** We will keep this support.
ELI	At ELI-BL, we already see that informally in E1 and partially with the commissioning experiments. There is a massive effort for reconstruction of experimental TOF data and scientists seem to be doing that collaboratively with their users.
ESS	We do not yet run experiments. We expect this to be the case in the future.
ESRF	2018: Yes, but not officially. This is done as an additional informal service by many of the local contacts on the beamlines. 2023: A clear mandate for DaaS should be in place and define the services provided by the ESRF.

ILL	Yes but not officially, this is often done by the local contact.
XFEL	Yes. Support and advice is provided for our tools and software.
ISIS	In some cases if required.

6. Are data analysis/reduction results preserved at your facility?

Institute	Answer
CERIC-ERIC	In some cases, during the experiment the user has available a data analysis directory in which one can do and preserve analysis results. Such data is also accessible through the VUO portal. Still, there is no common policy defined as of today.
ELI	Not at the moment, because data analysis is mostly handled by users and not yet integrated into a processing pipeline. However, we anticipate storing results of some of those reconstructions (especially at ELI-BL, where we will provide analysis as part of the DAQ chain for massive amounts of raw data that we might not be interested in saving).
ESS	We do not yet have a system set up for this. We expect that this will be done to some extent, at least for auto-pipelined data products.
ESRF	Yes for in-house experiments, because these storage areas are still covered by a backup. And no for external users, this is currently not part of the data policy. 2023: It may be necessary to archive also processed data, if economically feasible.
ILL	Yes. Analysis and reduction results are stored on the central storage along with the raw data.
XFEL	Yes. Processed/calibrated data is stored, calibrated data is not always stored but can be re-created on demand.
ISIS	Not in perpetuity currently.

7. How long do you keep the analysis results?

Institute	Answer
CERIC-ERIC	Again, there is no data policy defined for this yet. Data is kept according to each beamline and research team need.
ELI	To be defined in future ELI ERIC data policy - we would consider and treat analysis results like user data in the data policy.
ESS	Currently undetermined.
ESRF	One year for in-house experiments.
ILL	Forever.
XFEL	6 months for calibrated data, user files are saved for 24 months.
ISIS	A policy for how long to keep analysed results created through the ISIS Data Analysis as a Service platform is currently being developed.

8. How much data is produced after data reduction / analysis?

Institute	Answer
CERIC-ERIC	In some cases, the same amount as raw data.
ELI	The steady-state estimate for the overall amount of data archived at ELI is 10+ PB per year. Significant reduction will be required for some high-repetition-rate experiments (for example, E1 at ELI-BL, where reduction is from 20-30Gb/s to 1Gb/s).

ESS	Currently undetermined. For 2023, <500TB is expected.
ESRF	N/A, this highly depends on the kind of experiment/analysis.
ILL	This highly depends on the kind of experiment/analysis - it could represent more than the raw data.
XFEL	We do not currently have data reduction set up, volume of analysis results vary from a few KB to TB.
ISIS	It can vary a lot from instrument to instrument, but to a first approximation, about the same as the raw data volume.

9. Can external users remotely access experimental logs after the experiment?

Institute	Answer
CERIC-ERIC	**Today:** An electronic logbook is available for users during the experiments. The user can place this e-logbook in the experiment results directory and then access it through the user portal (VUO). **Forecast 2023:** We aim at have interactive log books using Jupyter Notebook technologies to allow users to reproduce the results described in the log book directly from it.
ELI	ELI-ALPS: yes (custom logbook application with external access). ELI-BL: Not yet.
ESS	Yes. Through an interactive chat interface (SciChat).
ESRF	An electronic logbook is being implemented for all beamlines as part of the data policy .
ILL	Yes. All logs are collected from the instrument control machines and persisted to a database. They are made available via web interface (https://logs.ill.fr).

XFEL	Yes. Metadata catalogue and elogs can be accessed as long as the account is active.
ISIS	The experimental logs are included in Nexus files.

Technology

1. What operating systems for data analysis does your facility support (please specify any Linux distributions)?

Institute	Answer
CERIC-ERIC	Ubuntu (16.04, 18.04), CentOS 7, Windows.
ELI	CentOS 7, 8 (2020).
ESS	CentOS7.3 is provided on the HPC Cluster. CentOS7.6 and Ubuntu18.04 are used for virtual machines.
ESRF	Debian 8, 9, 10.
ILL	Ubuntu 16.04 and 18.04. We also support OS X and Windows for some software.
XFEL	CentOS 7, Ubuntu 18.
ISIS	ISIS's support includes: Windows, Redhat, CentOS 7 and Scientific Linux.

2. What is the main Linux distribution used for the server infrastructure at your facility?

Institute	Answer
CERIC-ERIC	CentOS 7.

ELI	ELI-ALPS: Ubuntu ELI BL: CentOS 7 and some NI Linux RT (for control applications).
ESS	CentOS7.x.
ESRF	Debian.
ILL	Debian.
XFEL	CentOS 7.
ISIS	Redhat and varieties.

3. What compute infrastructure is dedicated to scientific computation at your facility?

Institute	Answer
CERIC-ERIC	Tesla units (GPU), HPC cluster, local cloud
ELI	ECLIPSE HPC Cluster, Upcoming (2020) HPC Cluster; DAQ servers with 14 blades/24 cores each + 768GB RAM for memory buffer.
ESS	For scientific computation we have an HPC cluster (see below), as well as a virtualisation environment used for infrastructure VMs as well as prototyping the data analysis environment - this consists of three virtualisation hosts (32 cores, 384 GB RAM each).
ESRF	Compute clusters & dedicated machines for online data analysis.
ILL	Compute cluster & dedicated machines for online data analysis.

XFEL	General purpose HPC cluster (Maxwell Cluster).
ISIS	SCARF and more recently the SCD Cloud as well as dedicated individual machines

4. Do you have an HPC cluster? If yes, what is the size of the cluster? I.e. number of cores, memory etc.

Institute	Answer
CERIC-ERIC	Yes. It has 16 work nodes (8 nodes of 2.50GHz, 8 nodes of 3.40GHz) interconnected 10 Gb Ethernet, 132 GB RAM, CentOS 7, Slurm
ELI	ELI-ALPS: - CPU computing cluster with 5 nodes, 36 cores and 768 GB RAM each. - GPU computing cluster with 5 nodes, 20 cores, 128 GB RAM and 2 NVIDIA Tesla K80 card each. ELI-BL: ECLIPSE HPC Cluster (1344 cores and 10.75 TB RAM in 84 nodes), Upcoming (2020) HPC Cluster (~8000 cores, ~50 TB RAM in ~324 nodes).
ESS	Modest HPC cluster – 1400 cores, 5GB/core on average, QDR IB (40 Gbps), no GPUs.
ESRF	Yes, 2018: we have currently 3200 Intel cores and 28 GPUs. 2023: a substantial upgrade is required, ideally to 20k Intel (or AMD) cores and 150 GPUs.
ILL	Yes. HPC cluster has ~800 cores with each node having between 32gb to 256gb of RAM and access to 30TB of storage. The HPC cluster is mostly used for data simulation. For data analysis and reduction, dedicated servers are available.
XFEL	EuXFEL owns a partition of the Maxwell cluster. Memory 256GB-1.5TB per node, 40-80 cores per node, HDR/EDR IB Backbone. Currently around 300 nodes, ~18,000 cores, ~150TB RAM.

ISIS	SCARF. See https://www.scarf.rl.ac.uk/hardware.

5. How is your scientific software distributed? (CERN-VMFS, custom Linux repositories, release downloads, artefact repositories...)

Institute	Answer
CERIC-ERIC	Internal GitLab, GitHub, Custom Linux repositories
ELI	No distribution.
ESS	Conda, github, dedicated websites, custom linux repositories.
ESRF	As Debian packages and Python packages in pypi and eventually in conda (when other persons are packaging them).
ILL	On git repositories, web site and NFS shares.
XFEL	Custom repositories, shared file system installations (GPFS, AFS), CVMFS available but currently not used for photon science.
ISIS	Some as cross platform binary distributions such as Mantid https://www.mantidproject.org/Main_Page). The forward looking plan is to distribute through cloud services more.

6. What emerging tools are you currently looking into or are interested in using?

Institute	Answer
CERIC-ERIC	JupyterHub, containers technologies (Docker, Singularity), Data Analysis as a Service, remote desktop technologies, K8S.

ELI	We are looking a bit into containerization instead of classical virtualizations - but this is more to help with CS development and deployment on diverse hardware rather than provided as a service.
ESS	DASK (Distributed NumPy Arrays), Py-bind-11, Singularity.
ESRF	Singularity.
ILL	Kubernetes, Singularity and Ansible.
XFEL	Hard to answer due to how closely linked our infrastructure is to DESY. At XFEL we are interested in BinderHub, Singularity, FPGA, ML/AI for data reduction, and Common Workflow Language. From the DESY side: Cpack on CVMFT. nextflow, airflow, airavata, scicat. k8s on bare (HPC) metal. Podman. Qiskit and related. Keycloak, Slurm cloud bindings, ...
ISIS	Use of the cloud services for users.

7. Do you currently provide any remote desktop services to users internally at the institute and to users after their experiments?

Institute	Answer
CERIC-ERIC	VNC for users during the experiment.
ELI	No. Currently nothing anticipated; however not excluded if needs emerge.
ESS	Currently developers can use VNC to access build machines or test software on VMs.
ESRF	NoMachine (nx.esrf.fr).

ILL	Yes. Using https://visa.ill.eu which gives users access to compute resources and remote desktops.
XFEL	Full graphical remote access for all users via FastX.
ISIS	ISIS Data Analysis as a Service (isis.analysis.stfc.ac.uk).

8. Are you currently running any Data Analysis as a Service pilot projects (such as VISA/CalipsoPlus at the ILL)?

Institute	Answer
CERIC-ERIC	Yes, CalipsoPlus.
ELI	No.
ESS	No, we are investigating this currently – considering something integrated with SciCat and Jupyter notebooks.
ESRF	jupyterhub + kubernetes, binderhub and jupyterhub+slurm.
ILL	Currently using VISA.
XFEL	None at EuXFEL, CalypsoPlus hosted at DESY.
ISIS	ISIS Data Analysis as a Service

9. Do you have a cloud infrastructure in place (OpenStack, Nebula, Orchestration etc.)?

Institute	Answer

CERIC-ERIC	Yes
ELI	Not at the moment, some interest - but no concrete plans.
ESS	We use a combination of Ovirt and foreman to deploy VMs for users/infrastructure.
ESRF	We are currently strongly limited by our human resources. The initial idea was to offer OpenStack, but this has been put on hold for the time being.
ILL	OpenStack, VMWare and Proxmox.
XFEL	EuXFEL and DESY share some infrastructure, OpenStack is available via DESY and can technically be used at EuXFEL.
ISIS	OpenStack.

10. What are you using for machine virtualisation (VMWare, KVM…)?

Institute	Answer
CERIC-ERIC	Proxmox (KVM).
ELI	VMWare; No virtualisation on HPC.
ESS	Ovirt.
ESRF	N/A.
ILL	KVM and VMWare.

XFEL	Varies, both used at EuXFEL (online cluster/remote desktop); additional software at DESY.
ISIS	-

11. What are you using for containerisation (Docker, Singularity…)?

Institute	Answer
CERIC-ERIC	We are mainly using Docker but also investigating Singularity approach.
ELI	ELI-ALPS: Docker ELI-BL: Singularity, Docker for early experiments.
ESS	Currently Docker containers orchestrated by Kubernetes.
ESRF	Docker.
ILL	Docker.
XFEL	Docker, Singularity. Considering Podman.
ISIS	Docker and Singularity.

12. Which protocol(s) are you using to access experimental data (NFS, SMBFS, SFTP....)?

Institute	Answer
CERIC-ERIC	CEPH, NFS, SFTP, SMBFS.

ELI	ELI-ALPS: SMBFS ELI-BL: NFS/SMBFS and a custom ZMQ based streaming solution; however we've just tendered a software-defined storage (commissioning in 2020) and plan to support a number of protocols depending on scientific needs and capacity to integrate.
ESS	NFS.
ESRF	NFS, SMBFS, GPFS.
ILL	NFS, SMB, GPFS, HTTPand SFTP.
XFEL	GPFS, BeeGFS locally. NFS exports, FTPS for remote copying, Globus online, ...
ISIS	-

13. Do you provide functions-as-a-service?

Institute	Answer
CERIC-ERIC	No.
ELI	No.
ESS	No, we have previously investigated them as part of the EOSC pilot.
ESRF	No.
ILL	No but we are currently looking into using something like OpenWhisk or Kubeless.

XFEL	Partially. EuXFEL does not, however some FaaS from DESY can be used.
ISIS	-

14. What is the incoming and outgoing bandwidth at your facility?

Institute	Answer
CERIC-ERIC	Incoming and outgoing are 10Gbps, moving to 100Gbps next year (2020).
ELI	ELI-ALPS: 10G redundant currently, with an option to extend 100G. ELI-BL: 10G dedicated to user data; anticipated extension to 100G.
ESS	Currently 10G (DMSC/Copenhagen – Danish Research Network).
ESRF	2018: A shared 10Gbps Internet connection (to the metropolitan network TIGRE, then RENATER, then GEANT). 2023: 100Gbps.
ILL	2018: A shared 10Gbps Internet connection (to the metropolitan network TIGRE, then RENATER, then GEANT). 2023: 100Gbps.
XFEL	Incoming: 2x50Gb. Outgoing: 2x50Gb.
ISIS	-

15. Do you use cloud providers for hosting external services or is everything hosted internally?

Institute	Answer

CERIC-ERIC	Everything is hosted internally.
ELI	ELI-ALPS: Currently everything is on-prem, there are plans to move non-core services to the cloud. ELI-BL: Hosting is currently 99% internal, but there are no general reservations against cloud providers.
ESS	Everything is provided internally.
ESRF	2018: everything is hosted on-site. 2023: scale-out with commercial cloud providers.
ILL	2018: Everything related to scientific data is hosted internally.
XFEL	Internally.
ISIS	Currently internally but plans to expand to use external services also.

16. If you were to give users access to remote virtual machines, would you give them root access?

Institute	Answer
CERIC-ERIC	No.
ELI	ELI-ALPS: Only under special circumstances ELI-BL: Possible, not required up to now.
ESS	For developers we currently give them root access on dev machines, for scientific users, this is currently undetermined (but likely not (if possible)).

ESRF	No.
ILL	No.
XFEL	Depends on the user, likely not. Restrictions apply (e.g. no direct access to experimental data).
ISIS	In general no, only for specific testing purposes such as for example software testing of specific critical facility software.

17. How would you govern the computational resources required (i.e. CPUS, memory) for a given experiment?

Institute	Answer
CERIC-ERIC	The computational resources are managed by the IT Group according to each experiment need. In future, with a JupyterHub installation for instance, we will be able to predefine available resources so users can choose the amount accordingly.
ELI	No established governance yet.
ESS	Currently undetermined – likely through VM environment allocation rules.
ESRF	2018: free access first come first served. 2023: data intensive experiments. Data Management Plan and accordingly resource allocation
ILL	2019 (pilot phase): free access first come first served.
XFEL	Not governed. Beamline staff makes a rough prediction, which can be adjusted on demand.
ISIS	Depends on the complexity of the instrument and the experiments

18. Can users choose the amount of resources required for data analysis at the facility?

Institute	Answer
CERIC-ERIC	Not nowadays. See previous question for explanation.
ELI	No constraints via governance, limitations by the hardware/capacity taken into account as part of the feasibility assessment of the user proposal.
ESS	Currently undetermined, but, likely yes. Where the user will be provided with sufficient resources during the active experiment, post-experiment analysis will likely be provided with a selection of configurations (hopefully the amount of allocated resources can be coupled to the proposal, but, currently undetermined).
ESRF	See previous question.
ILL	Yes, using VISA.
XFEL	Not really. Depends also on the type of compute pipelines. In principle all users have access to the entire HPC cluster.
ISIS	Yes, but only a limited number of choices are given

19. Do you have any particular quotas in place related to computation processing (i.e. number of hours, cores, ...)?

Institute	Answer
CERIC-ERIC	No.
ELI	No established policy yet.

ESS	Currently, we have Slurm fair-share allocation implemented on our HPC.
ESRF	2018: no, free access, but monitoring and complaints management. 2023: resource allocation methods have to be found which allow us to regulate the usage and to make sure everybody can work satisfactorily.
ILL	Maximum of three running jobs.
XFEL	Varying quota on number of concurrent jobs, number of nodes per job. Jobs have mostly exclusive access to nodes. A small number of nodes are reserved for each experiment during the beam time, plus dedicated compute hardware close to the experimental hutch ("online cluster").
ISIS	No

20. Do you have a job submission system for scientific computing resources (Sungrid, Torque, Slurm, Oar...)?

Institute	Answer
CERIC-ERIC	We do have Slurm as the job submission system for the HPC cluster.
ELI	Slurm.
ESS	Slurm.
ESRF	Currently changing from OAR to Slurm.
ILL	Torque. Currently looking into using Slurm.
XFEL	Slurm.

ISIS	SCARF uses Slurm.

Security

1. Which authentication provider(s) is used (i.e. umbrella, keycloak etc.) for user access (internal/external) ?

Institute	Answer
CERIC-ERIC	Local (VUO, LDAP). We do have Umbrella authentication as well.
ELI	No decision about the solution for external users has yet been taken.
ESS	Currently, no federation AAI is implemented – all AAA is internal (AD/LDAP).
ESRF	Umbrella and Keycloak.
ILL	Umbrella and Keycloak.
XFEL	LDAP.
ISIS	Umbrella is supported.

2. Which authentication protocols do you use to authenticate your users? I.e. OpenConnect, CAS, SAML etc.

Institute	Answer
CERIC-ERIC	Local based on LDAP.
ELI	No decision about the solution for external users has yet been taken.

ESS	Currently, no federation AAI is implemented.
ESRF	SAML.
ILL	SAML, CAS, OpenIdConnect.
XFEL	LDAP/Kerberos.
ISIS	-

3. Can people who have never visited your facility access data or analysis services (i.e. to access open data)? If so how do they authenticate?

Institute	Answer
CERIC-ERIC	Yes, but they have to be registered in the user portal (VUO).
ELI	No decision about the solution for external users has yet been taken.
ESS	Currently, no federation AAI is implemented.
ESRF	No, so far available services are not visible from outside, and when made available users will have to authenticate. However, getting an account is possible if collaborating with an ESRF staff member.
ILL	Using the Keycloak SSO.
XFEL	If users are part of an experiment they can request an account, once approved this account has access to the data from any experiments that they have participated in. In the future once the embargos start lifting on past experimental data some system to request a 'public' account will be needed.

ISIS	Open data is accessible for everybody. For non-open data access is only given to users who register with the user office and only for experiment they have been participating in.

4. Do you provide a means for people to apply for access to open data and analysis services?

Institute	Answer
CERIC-ERIC	It is all connected to the proposal.
ELI	No decision about the solution for external users has yet been taken.
ESS	Currently undetermined.
ESRF	https://data.esrf.fr allows access to open data, but there is no plan to give access to analysis services publicly.
ILL	Yes. Users can create an account and access data using https://data.ill.eu , however, there is currently no way to apply for access to compute services. This is currently being looked into.
XFEL	No.
ISIS	-

5. What IT services are provided to external users (ssh, sftp, specific web applications)?

Institute	Answer
CERIC-ERIC	ssh, sftp, data browsing and downloading through user portal (VUO). Data visualisation (HDF5 files) through web application. Remote desktop via Guacamole.

ELI	ELI-ALPS: SFTP and an eLearning web application. ELI BL: data transfer for specific commissioning experiment.
ESS	Currently undetermined.
ESRF	2019: SSH 2023: JupyterHub + Remote Desktop
ILL	2019: SSH, SFTP, JupyterHub, VISA, data catalogue.
XFEL	SSH, SFTP, JupyterHub, Globus, remote desktop, metadata catalogue.
ISIS	For example the ISIS data catalogue and ISIS Data Analysis as a Service.

6. Do you have any security restrictions at your facility (i.e. is a VPN connection required before accessing services)?

Institute	Answer
CERIC-ERIC	For some services such as ssh a VPN connection is required. For data browsing, downloading and visualisation a valid VUO account is required.
ELI	ELI-ALPS: internal services (if any) can be accessed only via VPN. ELI-BL: VPN connection is required for most internal services; very limited and controlled outside interfaces (data transfer).
ESS	Not to access external resources, otherwise, ssh (or VPN for internal users).
ESRF	A moving target. IT security has to be enforced. VPN access will most likely disappear.

ILL	No. A VPN connection is not required to access data services.
XFEL	No.
ISIS	Authentication with the user office account.

7. How are users associated to their experimental data? (i.e. LDAP groups, POSIX attributes…)

Institute	Answer
CERIC-ERIC	LDAP groups.
ELI	There is no structured solution yet. This is anticipated at ELI Beamlines in 2020. Constrained via LDAP groups/network separation/localised storage to specific experiments.
ESS	LDAP/POSIX.
ESRF	ACLs.
ILL	LDAP groups and extended NFS ACL attributes.
XFEL	No.
ISIS	-

8. Do scientific visitors have a user home at your facility (Linux home, Windows home…)?

Institute	Answer

CERIC-ERIC	Yes. Each user has a Unix user id which is used to access their user home and other resources available in the IT infrastructure.
ELI	Not yet.
ESS	Yes, for cluster users (neutron facility not yet in operation, though).
ESRF	2019: No 2023: Yes
ILL	Yes.
XFEL	Yes.
ISIS	Those who use the Data Analysis as a Service have.

Other information

Institute	Answer
CERIC-ERIC	Ideally all the services should be easily accessible via web through the internal portal based on the VUO and integrated with PaNOSC common portal as well as EOSC in future.
ELI	-
ESS	Please note that the replies herein relates to ESS, and that the facility is not currently in operation, and does not yet have a user programme (test user programme starts in 2023). The replies therefore are not representative for active neutron facilities and should, in general, be used with care in the context of active facilities (e.g. care should be taken so the expected data rates are not included in averages together with other facilities etc.). In addition, values, policies and procedures are still on the planning stage, and are very much subject to change.

ESRF	Data analysis and data analysis software developed/maintained at ESRF owes a lot to open source software, including: • The Python “scientific stack”: numpy, scipy, matplotlib, scikit-*, ... • HDF5 library, h5py, h5glance • Qt/PyQt • Jupyter/IPython ecosystem • Many other python packages and other softwares (e.g., ImageJ)
ILL	-
XFEL	As mentioned a few times, a lot of EuXFEL infrastructure is closely linked to DESY, so answering questions on what nodes are available to us and what software services are run at our facility can be difficult due to the blurry line between what might count as ‘ours’ and what counts as DESY’s. Our services are underpinned by a number of publicly available libraries and tools, including: · pyFAI · HDF5/H5py · H5glance · Silx · Jupyter Notebook/Jupyter ecosystem tools · ZeroMQ · MessagePack · Many SciPy ecosystem libraries – numpy, matplotlib, xarray, pandas, etc… · PyQt5/PyQtGraph
ISIS	-