Data Analysis Services

Document Control Information

Settings	Value
Document Identifier:	D4.4
Project Title:	ExPaNDS
Work Package:	WP4
Document Author(s):	A. Barty (DESY), A. Buteau (SOLEIL), O. Knodel (HZDR), U. Konrad (HZDR), A. Manzi (EGI), P. Millar (DESY), C. Minotti (PSI), E. Moge (SOLEIL), M. Ounsy (SOLEIL), R. Persee (SOLEIL), K. Pozsa (PSI), Z. Matej (MAX IV), C. Reynolds (DLS), S. Schoen (DESY), N. Soler (ALBA), G. Jover-Manas (ALBA), P. Fuhrmann (DESY)
Document Reviewer(s):	P. Fuhrmann (DESY), Paul Millar (DESY), Alun Ashton (PSI)
Responsible Partner:	SOLEIL
Doc. Issue:	1.0
Dissemination Level:	Public
Date:	28/03/2022

Abstract

We present the deliverables achieved for a summarization of the *prototyped remote data analysis services* and the derived outcomes that have been realised in the process at each ExPaNDS partner facility to align with the PanOSC project and to proceed towards onboarding services into the EOSC portal. The work represents the achievement of deliverable D4.4 of the Horizon 2020 ExPaNDS project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Executive Summary

We report on the status of the prototyped "*remote data analysis services*" that have been set up at the ExPaNDS partner facilities to provide users with software, tools and workflows to analyse data associated with five out of eleven selected use cases described in deliverable D4.2 [2] on reference datasets.

This document details state-of-the-art mechanisms to ease the exchange of data analysis solutions between our national RIs for running different data analysis pipelines representing particular experimental techniques. A special section (chapter 2) is dedicated to describe dependencies on core services shared with our partner project PaNOSC.

We have demonstrated that, for the selected use cases, we were able to build analysis pipelines based on state-of-the-art cloud technologies (OpenStack, Kubernetes, Proxmox, etc.) which we deployed and ran at multiple computing infrastructures across our partners.

Furthermore, we describe the progress of our partners in registering their facility as a service provider at the EOSC Marketplace and in a second step registering individual data analysis pipelines as services for the wider EOSC community.

In addition to providing those individual services, ExPaNDS has adopted VISA as an analysis portal that we offer to those facilities which are in need of such a service. Unfortunately, the PaNOSC VISA team does not have sufficient effort to support the wide variety of computing infrastructures at our facilities. As we do not have the remit to work on the VISA software itself, our partners are in the process of bridging this gap at the facility level. Nevertheless, the success of this activity should not be measured by the adoption of VISA, but rather by the adoption of standard interfaces in whichever portal technology makes the most sense for each individual partner.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Introduction

One high-level objective of the ExPaNDS project is to provide the Photon and Neutron scientific communities with ready-to-use software tools, which are available remotely and in a uniform way between all the partner facilities. In order to build data analysis pipelines, independently of specific experimental techniques.
This implies:

- unifying access to core and scientific computing resources for PaN communities and ideally scientists working in related fields, and
- adopting the ExPaNDS architecture (see Deliverable D1.5 [1] and upcoming deliverable D1.6) in order to satisfy the anticipated increase in demand when servicing a larger user community.

To this end, the following prototype data analysis services have been developed, taking into account the needs of exemplary scientific communities while supporting the diversity of the institutions' existing computing infrastructures.
To simplify the adoption of our services and to increase the likelihood of achieving sustainable solutions, all project partners based their implementations on concepts and environments which are typical for state of the art open-source projects. (GitHub, GitLab and CI/CD or on already available core services in the EOSC Marketplace. The most prominent examples would be the EOSC monitoring and AAI mechanisms.
As there is no one-to-one relationship between the five use cases and the technical implementation implemented at the different ExPaNDS facilities, we are describing the use cases first in chapter 1 followed by the description of the technical implementation per partner and the status of the deployment of those services into the EOSC Marketplace in chapter 3.

- Chapter 1 describes the connection between the aforementioned high-level goals to the implementation of the corresponding analysis pipelines. We focus on the five most challenging use cases taken from the eleven described in deliverable D4.2. [2] (Reference Data Sets). This selection demonstrates various options for delivering software packages to be run within local infrastructures, in a facility agnostic fashion.
- Chapter 2 describes the necessary alignments between ExPaNDS solutions and our partner project PaNOSC as well as the EOSC AAI service.
- Chapter 3 describes the efforts undertaken by the different ExPaNDS partner facilities to implement standardised pipelines on their computing infrastructures. Furthermore, we report on activities to onboard the facility-provided services into the EOSC platform.

We conclude by providing perspectives for anticipated developments in order to complete the harmonisation of solutions between the project partners to succeed in providing the PaN communities with standardised software and data analysis tools that are available in the EOSC portal.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1. Use Cases

Among the eleven use cases introduced in deliverable D4.2 [2] (Reference Datasets), we selected the five most challenging ones regarding their exchangeability between the different national facilities. With those prototype implementations, we were able

- to demonstrate the feasibility and benefits of the goals of the ExPaNDS project,
- to standardise PaN pipelines and to make them interoperable and independent of the natural variety of IT infrastructures of our partners and
- to provide the necessary interfaces to publish our services in the EOSC Marketplace.

Our selected *use cases* are addressing the following PaN techniques:

- Serial Crystallography
- THz Spectroscopy
- Full-field Tomography
- Ptycho-tomography
- Small-Angle Neutron Scattering: SANS

The technical details for the provisioning of remote data analysis services to process data produced by such experiments are described in the following sections.

1.1. Serial Crystallography

1.1.1 Scientific methodology

Serial crystallography is a beamline technique for collecting information on the structure of a protein without growing large protein crystals. Instead, a large number of small protein crystals are held in a pulsed X-ray beam. In a second step, the series of produced images are used to reconstruct a precise 3-D image of the protein structure. Serial crystallography is the preferred technique for obtaining diffraction data of proteins at room temperature, where radiation damage from the X-ray beam starts rapidly. The standard software for analysing serial crystallography is “CrystFEL”.

1.1.2 The CrystFEL[1] software suite

The programs included with CrystFEL are for processing diffraction data acquired in serial crystallography experiments. They perform tasks such as indexing and integrating diffraction patterns which are often very compute-intensive and best suited for high-performance computing (HPC) clusters.

[1] https://www.desy.de/~twhite/crystfel/index.html

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

1.1.3 Conventional Workflow

CrystFEL can be installed facility-wide by local infrastructure staff or by the scientists themselves on systems they have access to.

Where the facility maintains an installation of CrystFEL, access for the user is very easy. However, such installations come with a certain cost to the facility and do not give the user control over which software versions are available.

Even though the installation of CrystFEL is extremely well documented, a manual installation comes with challenges for any inexperienced user. CrystFEL depends on a wide range of packages that vary slightly according to the operating system. Installing it from source might be problematic for users, especially as they need to work in environments with restricted access rights as typical with HPC clusters. Bugs can be introduced by new versions of third-party packages that may be hard for the user to troubleshoot.

1.1.4 FAIR Workflow

To align with the FAIR principles the accessibility, installation and usage of the CrystFEL software on the different compute infrastructures of our partners was significantly simplified. This was achieved by containerizing the application and providing an installation script. A docker container image with a full CrystFEL installation is built and maintained by a CI/CD workflow on gitlab.desy.de. The image can be pulled via Docker[2] or Singularity[3] for direct use. No prior installation beyond a container management system (Docker or Singularity) is required.

To make the container deployment more user friendly, a bash installation script is also provided. As Docker is generally considered to be unfeasible for HPC systems, the script relies on Singularity instead. It pulls the Docker image and builds a Singularity image from it. Thus the only requirement is a working Singularity installation on the target system, which is usually present on HPC clusters.

Additionally, wrapper scripts for the CrystFEL commands are provided through the installation script. These scripts mimic the behaviour of the CrystFEL commands but are processed within the container. Thus, they make it possible to hide the containerization from the user, as the containerized software suite behaves and works the same way as if installed on the system. This applies to both the graphical user interface (GUI) and the command-line interface (CLI).

The workflow gives the user freedom to choose a specific CrystFEL version by selecting a specific docker image. Also, traceability of the exact versions of software and packages becomes easy, which is important for a reproducible data analysis workflow.

[2] https://www.docker.com/

[3] https://sylabs.io/singularity

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1.1.5 Slurm Integration

Slurm[4] is an important part of the CrystFEL software when running on HPC environments, which makes it necessary to interact with alternative application deployment methods.

Slurm supports the use of Singularity containers for jobs. The wrappers for the CrystFEL commands supplied by the installation script also work when used within Slurm jobs.
As Slurm needs to be set up on the host, it is not directly accessible within the container. This means Slurm is not directly supported from within the container, which is the case when starting a shell or a GUI from the container image. To make the Slurm service of the host accessible in these cases, additional wrappers for the Slurm commands are provided through the installation script. These Slurm wrapper scripts send the desired Slurm command from the container to the host where it is executed. At the time of writing, the current CrystFEL GUI (0.10.1) relies on the Slurm API, which is not accessible within the container. A new CrystFEL version is currently under development that uses the Slurm wrapper scripts.

1.1.6 Application Test

The container was successfully deployed at DESY's Maxwell HPC cluster as well as at MAX IV. In addition to the CrystFEL GUI and commands, Slurm was also accessible from within the container shell.

1.1.7 Transferability to other Applications

The proposed workflow was rendered in a standard fashion to allow it to be easily adopted by arbitrary applications. The only requirements are a Docker or Singularity image of the containerised application and a Singularity installation on the target system.
To apply the workflow to a different application other than CrystFEL, only the install script needs to be updated. The required adjustments consist of the image registry, image name and tag as well as the names of the application commands, which are meant to be supported by the containerised deployment. The most recent version of the installation script is provided at DESY GitLab[5].

[4] https://slurm.schedmd.com/documentation.html
[5] https://gitlab.desy.de/silvan.schoen/containerizedapplication

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1.2 THz Spectroscopy

The THz spectroscopy beamline is a typical user facility at the HZDR and therefore, it is necessary to make both data and analysis scripts accessible to external users. In the following section, the methodology is described as well as the conventional workflow. Following on from this, a containerized workflow is introduced that will support FAIR use. Additionally, EOSC services such as B2FIND and the future PaN-training service are described.

1.2.1 Scientific methodology

The radiation source ELBE (Electron Linac for beams with high Brilliance and low Emittance) at the Helmholtz Centre Dresden Rossendorf (HZDR) can produce several kinds of secondary radiation (see Figure 1). THz radiation is one such radiation. It typically has a pulse frequency of 100 kHz and is used to stimulate elementary low-energy degrees of freedom in matter. TELBE is open as a user facility and, in order to enable comprehensibility and reuse, the underlying raw data analysis steps are provided. The available laser infrastructure allows THz-pump laser-probe experiments within a wide range of experimental schemes and probe beam properties.

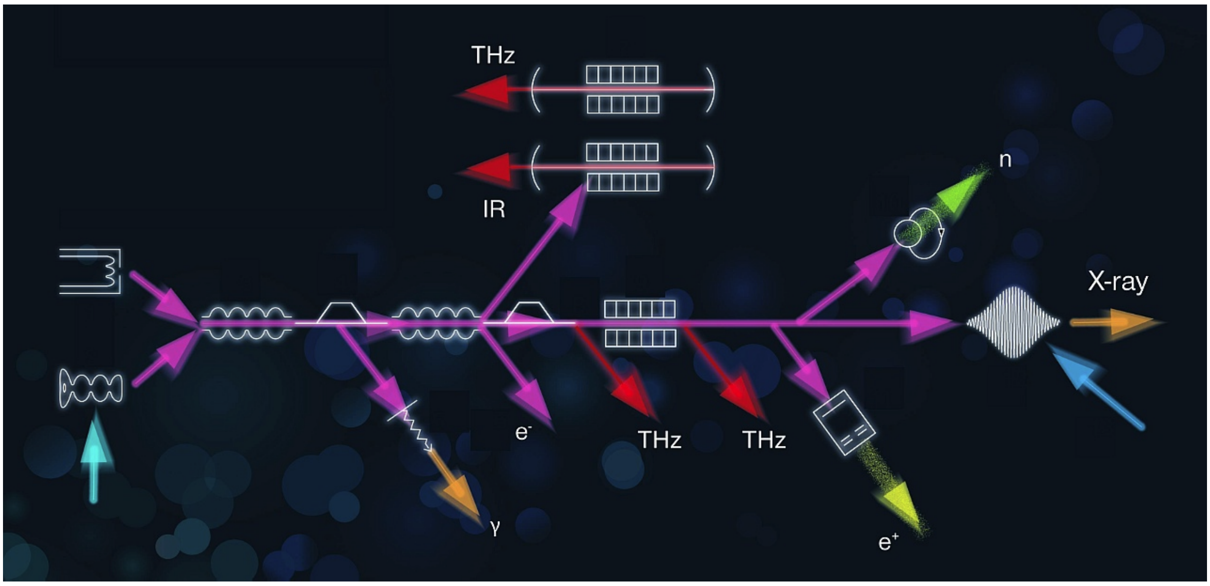

Figure 1: *Schematics of secondary radiation at the ELBE Electron Linac (Courtesy HZDR)*

1.2.2 Conventional Workflow

The facility's data acquisition system combines the pulse-resolved experimental data with the associated timing information. The automated logging of relevant metadata takes place in an internal electronic laboratory book and the actual measurement raw data is forwarded to a storage system at a Slurm-based HPC Cluster.

On the cluster, Slurm jobs are automatically submitted during the experiment by the data acquisition system. The cluster jobs are responsible for sorting and binning the raw data to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

reduce the size of the dataset, and executing the first analysis steps that generate metadata, which is then stored in the electronic laboratory book.

1.2.3 FAIR Workflow

The raw data stored by the data acquisition system on the HPC file system is the earliest point where it is possible to start a containerized workflow. Together with the metadata generated in the automatic processing and analysis steps, everything is packaged together in a single dataset. With additional metadata describing access rights, licence and involved scientists, the dataset is published in the data repository RODARE. The entry is then also available in the B2FIND community[6] RODARE. We include an example DOI[7] that demonstrates this. A reference to a scientific publication (Nature Communications DOI[8]) is included in the dataset's metadata.
The analysis software used to process the data on HZDR's HPC cluster is transferred into a **Jupyter notebook** to demonstrate the interaction with the data and introduce common analysis steps.
The notebooks are available on **GitHub[9]** and will first download the open available dataset. The notebooks can be executed on a federated or local HZDR Jupyter instance or in myBinder[10]. Based on the provided example a user can understand the analysis and modify parameters to generate additional results.
A description of the experiment, dataset and tools used for the analysis, and a training workflow[11] on PaN-training.eu brings all these different parts together and provides an explanation of the workflow with further information and references to the scientific background.

1.2.4 Transferability to other Applications

The beamline scientists are trained to provide the data generated in user experiments in a similar way with the tools and methods we introduced in our exemplary THz Spectroscopy experiment.

1.3 Full-field Tomography at PSI

The *TOmographic Microscopy and Coherent rAdiology experimentTs* (TOMCAT) beamline is a typical user facility at PSI and therefore, it is necessary to make both data and analysis scripts accessible to external users. In the following section, we describe the scientific methodology together with the analysis workflow.

---

6 http://b2find.eudat.eu/group/rodare
7 10.14278/rodare.277
8 10.1038/s41467-020-15613-1
9 https://github.com/hzdr/TELBE-raw-data-evaluation/blob/main/sorting_binning.ipynb
10 https://mybinder.org/
11 https://pan-training.hzdr.de/workflows/telbe

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1.3.1 Scientific Methodology

The TOMCAT beamline at PSI is operated by the X-ray Tomography Group and offers cutting-edge technology and scientific expertise for exploiting the benefits of synchrotron radiation for fast, non-destructive, high resolution, quantitative investigations on a large variety of samples. Absorption-based and phase-contrast imaging are routinely performed with isotropic voxel sizes ranging from 0.16 to 11 µm (fields-of-view (h x v) of 0.4 x 0.3 mm^2 and 22 x 3–7 mm^2, respectively) in an energy range of 8–45 keV. Phase contrast is obtained with simple edge-enhancement, propagation-based techniques or through grating interferometry.

Typical acquisition times are on the order of seconds to a few minutes. However, dynamic processes can be followed in 4D (3D space + time) using the ultra-fast endstation, which provides sub-second temporal resolution for extended time periods thanks to the in-house developed GigaFRoST system. A laser-based heating system and a cryojet and cryo-chamber are available as standard installations and are compatible with both the standard and ultra-fast endstations. It is also possible to bring specialized, user-defined instrumentation to TOMCAT.

A temporal resolution of a few (< 5) minutes can also be achieved with the hard X-ray full-field microscope setup delivering a pixel size of 65 nm for microscopic samples (~75x75 μm^2 field-of-view).

1.3.2 Conventional Workflow

3D tomographic datasets are reconstructed from 2D projections using optimised software based on Fourier methods and a user-friendly interface (i.e., an ImageJ plug-in). Remote access to a flexible HPC facility is available for subsequent, advanced post-processing and data quantification. A suite of analytical and iterative reconstruction routines are provided; additional ad-hoc tools can be installed by a user.
Raw and derived datasets are stored at PSI, along with the experiment metadata, and can later be accessed using the data catalogue. The principal investigator can trigger an action to store the datasets in our long term storage tape archive system at CSCS.

The raw data is stored by the data acquisition system on our file system, together with the metadata generated in the automatic processing and analysis steps. Everything is packaged together in one dataset. During the creation of the experiment proposal, our user office creates an access group that includes the members of the research group, which is later used to manage the authorisation of data and dataset. The members of this group can modify the metadata related to the experiment.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1.3.3 FAIR Workflow

Following the data acquisition step, group members will analyse the data. Increasingly this is done with Jupyter notebooks provided through a JupyterHub running on the PSI HPC cluster.

To achieve full reproducibility, an external user needs access to the analysis environment, workflow and corresponding data. At PSI, data is made available through a Public Data Repository after the individual embargo period has passed, which can be up to three years. The repository allows B2FIND and openAIRE to harvest the corresponding metadata in order to make the PSI data findable by the global community. This entire process is illustrated with a typical experiment [12].

At the time of writing, PSI uses modules to manage the analysis environments. While it's possible to reproduce these environments at other facilities it is time-consuming and requires the involvement of IT teams. We are assessing several containerization strategies, their impact on our current infrastructure and what a roadmap would look like to introduce this at PSI.

Workflows are very diverse and often modified by scientists to suit their needs. Hence, the responsibility of versioning and sharing them is left to the individual scientist. We are evaluating different solutions (e.g. renku, aiida, etc.) to introduce standardisation and common rules to be followed to ensure the FAIRness of workflows.

1.4 Ptycho-tomography

Ptychography and ptychographic X-ray computed tomography (PXCT) are synchrotron-based imaging techniques that are taking advantage of the increased coherence of the new generation diffraction-limited storage rings. PXCT allows reconstruction of the sample electron-density in space and imaging 3D samples with resolution down to nanometers (see Figure 2). NanoMAX and SoftiMAX beamlines at the MAX IV Laboratory provide cutting edge coherent nano-beams very suitable for PXCT experiments in hard and soft X-ray regimes. PXCT experiments produce large data volumes and are demanding on computing resources.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

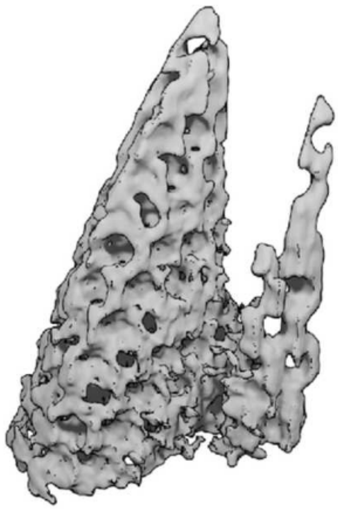

Figure 2: A three-dimensional reconstructed electron-density of an artificial nickel-based inverse opal material. Image, experiment and data processing by M. Kahnt et al., DOI: [10.1107/S160057672001211X](https://doi.org/10.1107/S160057672001211X)

1.4.1 Scientific methodology

A single spatial point of the PXCT measurement is a 2D scattering image recorded by an X-ray detector. The sample is 2D scanned with an X-ray nano-probe. Such a set of 2D spatial points is forming a single projection measurement. In the next step, the sample is rotated and 2D scanned again. Similar to conventional tomography, the full dataset consists of several hundreds of angular projections.

The 3D electron density in the sample is then reconstructed in the following three steps. A 2D ptychographic-like reconstruction is obtained from each projection sub-dataset. In the next step, the resulting projections are aligned and, in the final step, a tomography-like reconstruction is applied to the aligned projections.

1.4.2 Conventional Workflow

The recorded detector data and the relevant experimental metadata are stored in shared network storage and available for processing on an HPC cluster. The researcher has to provide (from a template) a reconstruction script describing the procedure and relevant parameters for the first ptychographic reconstruction step. A ptypy software [6] is used. The reconstruction script for a series of projection datasets is submitted to the HPC cluster using Slurm. This results in text logs, reconstructed data in the HDF5 format, and standardised images illustrating the quality of the reconstruction procedure and other relevant information.

1.4.3 The *ptypy* software suite.

The ptypy[12] software is published under the *GNU General Public Licence*, version 2. However, there is a company (Phase Focus Ltd, UK) that holds a portfolio of patents related to ptychography. Phase Focus claims that, by using the ptypy software, a scientist may be

[12] ptypy

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

using one of their patented inventions, for which the scientist would require a license. The company offers a free license for non-commercial, academic research.
Such situations are solved at the HPC centres and synchrotron facilities in two ways:

- by allowing the scientist to use the software only after they agreed to follow the Code of Conduct and the specific licence agreements from the software authors and vendors
- by providing an environment in which the scientist may upload the required software. In this way, the user takes full legal responsibility for the software they use.

Given the availability of a free license for academic use, we consider this procedure is only a minor inconvenience, particularly as the issue of whether patents apply to software is not completely resolved within the EU.

1.4.4 The FAIR Workflow

The increasing popularity of Jupyter-notebook environments and their expanding possibilities allows for PXCT analysis to be supported by adopting a Jupyter-like service. Merging data access patterns, software initialisation, reconstruction scripts, data processing logs and visualisation of the most relevant results into a single frontend simplifies the workflow for users and puts lower requirements on the complexity of IT infrastructure. Data will be accessible from the open data partition at the MAX IV Laboratory [7]. As indicated above, the ptypy software Code of Conduct and good IT security practice as requirements for accessing the ptypy software presents a small challenge. A tested software environment in which the scientist can easily install the *ptypy* software will be provided. This places the legal burden of complying with the patent licensing agreement on the scientist. The baseline is a well-documented procedure using the Jupyter terminal and a GitHub personal access token. Ideally, a more straightforward and user-friendly authentication can be implemented or developed using e.g. GitHub REST API. The Jupyter environment definition and the reconstruction scripts will be provided via a standard git repository. The software environment can be based both on Conda or environment modules as both are used at MAX IV Laboratory. The testing framework (see D4.3 [8]) introduced by ExPaNDS will be used for continuous testing of the PXCT workflow.

1.5 Small-Angle Neutron Scattering, SANS

Small-Angle Neutron Scattering, SANS is a special form of elastic neutron scattering, providing information on the size and shape of particles on length scales from ~1 nm to ~1000 nm. The probe, at which the neutron beam is directed, can be an aqueous solution, a solid, a powder or a crystal. For more information, please find a detailed description and discussion in our PaN teaching portal[13].

[13] https://e-learning.pan-training.eu/wiki/index.php/Small_angle_neutron_scattering,_SANS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

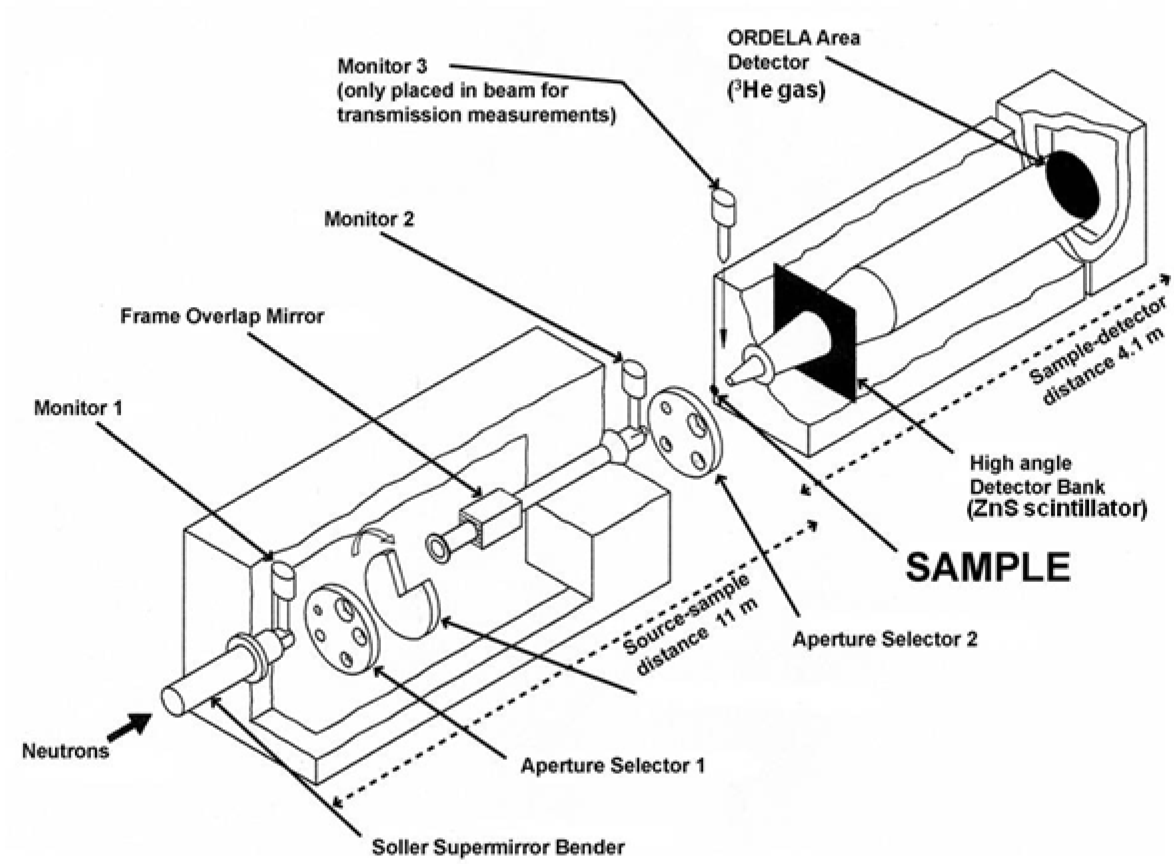

Figure 3: Schematic of the ISIS LOQ beamline[14].

1.5.1 Scientific methodology

This use case employs the data reduction software Mantid[15], the most used data reduction software at ISIS and adopted by a number of other facilities. Furthermore, it demonstrates a typical workflow to turn raw data into reduced data, which can then be used by subsequent analysis. The specific example here is of ‘protein standard’ data required for calibration purposes on the LOQ beamline and it has been published [9].

The LOQ beamline is used to investigate the shape and size of large molecules, small particles or porous materials with dimensions in the range of 1–100nm. Length scales of up to 400 nm can be probed in highly anisotropic systems. The experimental technique leveraged for this is called small-angle neutron scattering, whereby neutrons are scattered elastically (without energy loss) by interacting with the nuclei of the atoms and with the magnetic moment due to unpaired electrons.

The main purpose of this use case is to highlight a bread-and-butter analysis workflow for analysing neutron data using the *Mantid* software. At a high level this workflow is similar for analysing data from a whole range of other neutron experimental techniques and is the most common analysis workflow at the ISIS Neutron and Muon Source.

[14] https://www.isis.stfc.ac.uk/Pages/Loq-technical-information.aspx
[15] Mantid

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

1.5.2 The Conventional and the FAIR Workflow

Data Access

All the associated files with this use case are published at GitHub[16]. This includes documentation about the setup and a log of the experiment, together with the raw and reduced data (for comparing), the *Mantid* files needed to run the reduction and a Jupyter notebook to run the reduction interactively.

Running the workflow

This use case can be run out-of-the-box on `SANS` workspaces within the ISIS Data Analysis as a Service. If used at a different facility then the following are required:

- Mantid (GPLv3)
- Jupyter (3-clause BSD)
- internet access

and the availability of 'git' is optional.

Once these requirements are fulfilled, the path to the *Mantid* installation needs to be updated in the Jupyter notebook, together with the directory paths to be used.

2. Alignment with EOSC and PaNOSC Services

As described in the high-level objectives of ExPaNDS and particularly for WP4, commonly used PaN analysis services should be based on core services provided by other research institutes in the context of EOSC, whenever possible. This is particularly true for the AAI system, which must be common to ESFRI and national PaN facilities. In subsection 2.1 we'll briefly describe the AAI mechanisms against which ExPaNDS will align, for authentication. Furthermore, as the EFRI PaN cluster project PaNOSC started before ExPaNDS, our proposal already anticipated an analysis portal framework delivered by PaNOSC to which we agreed to contribute in terms of validation and deployment. Subsection 2.2 provides the necessary background on that initiative. Finally, in order to provide training and education to the PaN community, ExPaNDS and PaNOSC are working on a common training platform. This is mentioned here for completeness only, as it is worked on within WP5, but requires input from our work package.

[16] https://github.com/DAaaS-reference-data/SANS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

2.1 UmbrellaID as the common PaN AAI solution

Since its inception in 2012, UmbrellaID[17] has been serving the AAI needs of the PaN community. Within the PaNOSC project, WP6 "EOSC integration" aims to integrate the PaNOSC cluster services within EOSC and one of the results has been the integration of the UmbrellaID AAI within the EOSC AAI general architecture[18]. The integration process is detailed in the PaNOSC deliverable D6.3 [13].

One of the first achievements has been the adoption of the eduTEAMS service, operated by GÉANT, as the core infrastructure for the UmbrellaID AAI. The eduTEAMS infrastructure for UmbrellaID has been established in Autumn 2020 and a series of technical workshops have been organised to allow the migration of all *PaN Service Providers* by March 2021. A second workshop scheduled for March 2022[19] is organised to further discuss the integration of the new UmbrellaID AAI into infrastructures of ExPaNDS facilities.

The migration to eduTEAMS will simplify the integration of UmbrellaID with EOSC AAI, where eduTEAMS is one of the Community AAI proxy, by

- enabling UmbrellaID users to access EOSC services, which are integrated with the EOSC AAI
- enabling access to ExPaNDS services via EOSC AAI.

2.2 The VISA[20] platform as the common PaN portal

As mentioned above and detailed in the ExPaNDS deliverables D4.1 [14] and the PaNOSC deliverable D4.1 [15], PaNOSC is developing a *Common Portal* for *Data Analysis Services* enabling users to remotely analyse data from PaN facilities. Within ExPaNDS WP4, we support PaNOSC in deploying their solution on the infrastructures of our partners and in adapting the portal system to local conditions and requirements, as needed.

With long term sustainability in mind, the PaNOSC project decided to base the portal system upon an established software stack (VISA), developed and used in production at the ILL facility in Grenoble. To allow external contribution, ILL moved the VISA software to the open-source repository[21].

The core functionality of the VISA platform allows launching predefined PaN analysis pipelines on PaN facility and horizontal cloud infrastructures, enabling scientific users to analyse their experimental data remotely using a web browser only.

When using VISA, the user can choose between two types of application. A remote desktop to access the software via its native graphical interface or a Jupyter notebook service.

As VISA is the official ILL portal service, the portal is continuously improving to match new technologies in hardware and software. Examples are the ability to access GPUs which are

[17] https://umbrellaid.org/what.html
[18] https://op.europa.eu/en/publication-detail/-/publication/d1bc3702-61e5-11eb-aeb5-01aa75ed71a1
[19] https://indico.psi.ch/event/12701/
[20] https://visa.ill.eu/login
[21] https://github.com/ILLGrenoble/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

shared between multiple analysis virtual machines and the distribution and versioning of analysis software made available using CVMFS[22].
The deployment and integration of the VISA platform is a common activity ongoing both at PaNOSC and ExPANDS facilities, described in the PaNOSC deliverable D6.2 [10] and for ExPaNDS in Section 3 of this document.

2.3 The common PaN training platform

As ExPaNDS and PaNOSC essentially share the same community of scientists, we agreed to build a single portal for training resources. In this context, PaNOSC offers an e-learning platform with Jupyter notebooks integrated into online courses[23]. The PaN portal[24] provides a catalogue with training material and events, which is developed and provided within the framework of the ExPaNDS project. In addition, the portal provides an environment that describes training workflows where the different steps can be understood via illustrative examples. The references to the material (e.g. scientific publication, dataset, analysis tools) are provided in the workflows as well as further information. The workflows can be associated with different research institutes.

3. Enabling Facilities Computing Infrastructures for common analysis services

In this chapter we describe the efforts undertaken by each facility to prepare the local infrastructure for implementing the use cases described in chapter 1, as well as the current status of the EOSC onboarding process.

[22] https://cernvm.cern.ch/fs/
[23] e-learning.PaN-training.eu
[24] PaN-training.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.1 PSI

3.1.1 Technical activities

The following is a list of changes that have been made or initiated at PSI. These changes have been triggered or influenced by ExPaNDS.

- We jointly developed a facility-independent Jupyter notebook that supports tomographic reconstruction. This notebook has been demonstrated to work correctly at DESY, MAX IV and PSI.
  The repository includes the Jupyter notebook containing the reconstruction itself, the Conda environment files to reproduce the Jupyter kernels necessary to execute the notebook and references to a public 'test' dataset (16GB) that can be used with the workflow. For more information please see the README.
  In the future, files to create containers capable of executing the workflow will also be part of the repository.
- The Paul Scherrer Institute organisation on GitHub will be used to publish common development efforts in the future.
- PSI's base infrastructure currently does not support executing containers. Work has begun to evaluate different container execution engines and introduce this capability at PSI.

3.1.2 The EOSC Onboarding Process

- PSI has registered as an EOSC provider.
- PSI's only data analysis service in production (that is available without VPN access) has been registered as an EOSC resource.

3.2 Diamond

3.2.1 Technical activities

Diamond has developed the following set of computing facilities for shared data analysis:

- A JupyterHub installation has been deployed on Diamond's Kubernetes platform. The configuration is such that users can supply their own Notebook container, complete with dependencies, by providing a URL to a container hosted in any container registry (e.g. dockerhub). Diamond is working with other facilities to demonstrate Notebook portability between our JupyterHub and others. Data portability is a significant obstacle. The initial prototypes anticipate using a bucket on a globally accessible object storage system provided by STFC.
- Diamond has no plans to deploy VISA due to the significant dependency of an OpenStack IaaS. Diamond's on-premises Cloud is based on Kubernetes, since this

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

supports the popular use case of self-service container orchestration, as well as the potential orchestration of legacy VMs using tools like KubeVirt. Kubernetes is likely to be at the heart of the planned Diamond II[25] upgrade, running Data Acquisition and Analysis services, as well as control systems.[26]

3.2.2 The EOSC Onboarding Process

Diamond has been registered as an EOSC provider. It is in the process of adding the Nomachine/NX based remote desktop system as an EOSC Marketplace service. The NX system acts as a gateway and allows remote desktop access to a Diamond Desktop Server, which in turn allows access to data analysis tools. This includes access to HPC clusters as well as a JupyterHub installation.

3.3 UKRI

3.3.1 Technical activities

The following work was done at UKRI as part of ExPaNDS WP4:

- sharing of our experience implementing a VM-based Data Analysis as a Service platform at the ISIS Neutron and Muon Source. This service called IDAaaS has been in production since 2019
- gathered, prepared, tested and published reference datasets and the helped create the template for the related deliverable

3.3.2 The EOSC Onboarding Process

UKRI has already been registered as a service provider on the EOSC Marketplace. The enrollment process for *IDAaaS as a service* on the EOSC Marketplace is currently blocked until after IDAaaS is updated so it can restrict access to users of the facility.

3.4 MAX IV

3.4.1 Technical activities

The following points summarise the achievements of the MAX IV within ExPaNDS WP4:

- A continuous testing framework for Jupyter-like services has been developed by ExPaNDS [11] and is in production on several types (containers and HPC Slurm) instances of JupyterHub services at MAX IV. This allows the rapid development of specialised scientific Jupyter kernel environments[27] and supports their sustainable operation and maintenance.

[25] Diamond II
[26] KUBERNETES FOR EPICS IOCS
[27] jupyter-docker-stacks

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

- MAX IV has contributed to the full-field tomography ExPaNDS scientific case jointly developed with PSI using the PSI tomography dataset.
- MAX IV cloud services are running mostly on Kubernetes and VMware infrastructure
- MAX IV is currently developing a public JupyterHub instance as a MAX IV EOSC prototype service. The service is supported by the production MAX IV Kubernetes infrastructure, utilising EGI heck-in and Keycloak for federated authentication.
- MAX IV open data are accessible as MAX IV - Open Access Data[28] Collection via Globus Connect.[29]
- MAX IV scientific cases for the ExPaNDS data analysis services are accessible at gitlab.com/MAXIV-SCISW.[30]

3.4.2 The EOSC Onboarding Process

The MAX IV laboratory is not registered as an EOSC provider because of recent updates to EOSC provider inclusion criteria. Lund University, a MAX IV Laboratory legal entity, would need to be registered first according to the new criteria. This was not foreseen and cannot be committed within the ExPaNDS project. MAX IV EGI Jupyter-notebook service will be available to ExPaNDS users via PaN-portal.

3.5 ALBA

3.5.1 Technical activities

The following actions were taken in ALBA under the impulse of WP4:

- A change of technical specifications for our DAaaS platform. In order to be aligned with most PaN facilities present in the project, the initially planned implementation of the CALIPSOplus service was aborted in favour of the VISA portal. Alba does not have OpenStack implemented so our teams are investigating using our CITRIX infrastructure instead, while not excluding Proxmox as an alternative in case of any major setback. Our goal for the end of 2022 is to have VISA successfully working internally on a use-case from our SAXS/WAXS beamline (NCD-SWEET) for which iCAT is already installed and running.
- The installation of a JupyterHub service (internal, in test phase) that can launch instances of JupyterLab using our local HPC cluster (currently in the demo phase, JupyterHub launches the JupyterLab instances by submitting jobs to our Slurm cluster). This service has been successfully tested running Serial Crystallography use case (1.2.3 FAIR Workflow).

[28] MAX IV - Open Access Data
[29] Globus Connect
[30] gitlab.com/MAXIV-SCISW

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.5.2 The EOSC Onboarding Process

The registration to the EOSC as a service provider was accepted recently. There are ongoing discussions on which of the above services to register.

3.6 Soleil

3.6.1 Technical activities

As in the task of prototyping the remote data analysis services, we have been able to set up two OpenStack Infrastructures in Docker containers, deployed using Kolla-Ansible:

- the first as an "All-in-One" (AIO) OpenStack dockerized components running on a single physical server for development purpose;
- the second as an OpenStack distributed dockerized components platform on physical servers, for a scalable production-grade mock-up, including GPU nodes.

- To provide data analysis services, we have installed the VISA platform (see subsection 2.2) on a Docker-Compose virtual machine hosted on a *Proxmox* infrastructure. VISA is interfaced with our "All-in-One" OpenStack infrastructure and can instantiate virtual machines[31] on the OpenStack backend.
- VISA user authentication is done through a *Keycloak* open source identity and access management solution, using our local "Experiences Network Directory" (LDAP-RE) for account/password challenge.
- This Keycloak instance will be connected to the future UmbrellaID once the latter is available and access is granted.
- External user and project accounts used for experiments are provided by the SOLEIL User Office web portal SUNset[32]. Some of the required information for the VISA data model[33] are only available in the SUNset portal. At this stage, there is no 'ready to use' solution available to interface VISA with the SUNset application. As a workaround, we are developing a REST API above our LDAP-RE directory, and an ETL to extract data from the directory using these APIs. The ETL is developed to be reusable with the SUNset portal once a REST API will be provided.
- Data analysis applications are provided in virtual machines hosted on the OpenStack infrastructure. The virtual machine images are built using Packer by following the procedure described in the VISA documentation by the ILL.
- The build process is integrated into a CI/CD system using our internal GitLab infrastructure and the built images are automatically uploaded to OpenStack using the OpenStack CLI.
- We started working with the SOLEIL data analysis team to define virtual machine requirements and a workflow for data analysis application integration into the virtual machine images. The data analysis team describes their application requirements in

[31] https://www.keycloak.org
[32] https://sun.synchrotron-soleil.fr/sunset/bridge/sunset/
[33] https://visa.readthedocs.io/en/latest/development/development-etl.html

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Packer manifest files that are versioned in our GitLab repository. Once changes are approved, the changes are merged with the main repo for testing and automated image building.

3.6.2 The EOSC Onboarding Process

- Soleil is assessing the FAIRness of its data repository, before making it available to EOSC B2FIND.
- Soleil is in the process of registering as an EOSC service provider. The initial EOSC services will be the Soleil SciCat and the VISA platform.

3.7 HZDR

3.7.1 Technical activities

To enable the FAIR workflow described in 1.2 the infrastructure at HZDR has been improved by the following measures:

- Installation of an OpenStack infrastructure in order to provide the virtual machines for the data analysis workflow.
- Based on that infrastructure container instances are generated that run the Jupyter Notebooks.
- User authentication is provided by federated AAI (Helmholtz AAI and future UmbrellaID).
- The FAIR data repository RODARE was made available to EOSC B2FIND with an own community as EOSC service.
- Our federated GitLab instance is used for common development and testing of all components of the analysis workflows.

3.7.2 The EOSC Onboarding Process

- HZDR provides its FAIR data repository "RODARE", which is already harvested by the EOSC and all datasets are available at EOSC B2FIND.
- HZDR is in the process of being registered as an EOSC service provider. The initial EOSC service will be RODARE .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.8 Desy

3.8.1 Technical activities

ExPaNDS had a significant impact on efforts at DESY, which were conducted in order to provide analysis services.

- DESY is one of the driving forces to adapt the VISA platform for our ExPaNDS partners. It has contributed Helm charts to deploy VISA on Kubernetes, and has automated pipelines for production, staging and test environments that are needed to test new features and integrations. Building and releasing the Helm Charts is automated using GitLab CI/CD pipelines.
- DESY built a workflow for virtual machine images for VISA using Packer with CI/CD on desy.gitlab.de was established. It directly uploads to OpenStack via the OpenStack CLI. The containerization software Singularity was integrated into these images to provide a secure and well-adopted containerization option. In this context, CVMFS was discussed and tested for the distribution of Singularity container images. Further, the Slurm API was enabled for VISA instances and its usability was evaluated in collaboration with users. It made the transfer of heavy compute loads and parallelized processing (MPI) to the Maxwell cluster available from VISA instances. Also, a SciCAT data catalogue prototype is being deployed and evaluated to manage metadata for the VISA ETL process.
- At DESY, VISA authentication is done by Keycloak, which is connected to UmbrellaID. Thus UmbrellaID can be used to authenticate with the DESY Keycloak either using EGI Check-in or via the German federated Helmholtz AAI. A prototype to enhance the local user registry was developed to ensure seamless integration of federated accounts to DESY's local infrastructure and personalised storage resources.
- DESY implemented a GitLab instance[34] including a container registry running on an OpenStack infrastructure. The GitLab Runner infrastructure is realised by a Kubernetes container orchestration system.

3.8.2 The EOSC Onboarding Process

DESY registered early as an EOSC provider, with multiple pilot services. From these JupyterHub[35], GitLab[36] and a demo instance of the dCache[37] storage system is still used and further developed for ExPaNDS.

[34] https://gitlab.desy.de
[35] JupyterHub
[36] GitLab
[37] dCache

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

4. Conclusion and next steps

This document has demonstrated the efforts coordinated by ExPaNDS at its partner-facilities to pave the road to the final goal of providing PaN scientists in Europe with standardised data analysis pipelines available through the EOSC portal.
The approach is based on an overall modular architecture (see Deliverable D1.5 [1] and upcoming deliverable D1.6) conceived around generic APIs letting the room open for each facility to adapt to whatever future web portal they may choose regarding their local infrastructure specificity. Prototype analysis services have been set up and tested in different configurations (through Jupyter notebooks or ready to use remote desktops) for the end-user to use them in the same way in any participating facility.

In the remaining time of the project, we will complete the work depicted in our DoW, to

- ease the link to the outcomes of WP3 on standardised data catalogues, API searches, common vocabularies and ontologies targeted to the end-users (see deliverable D3.3 [3])
- assess the FAIRness of WP4 outcomes particularly regarding the reusability of the proposed software solutions (see deliverable D2.2 [4])
- enlarge the audience of facilities using the VISA portal giving our scientific community a unified endpoint via the EOSC portal to all our current and future outcomes
- enrich the training platform with more material to ease access for the users to WP4 outcomes (see D5.4 [5])

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

References

[1] Scardaci, Diego; Salvat, Daniel; Fuhrmann, Patrick; Barty, Anton; Ashton, Alun; Servan, Sophie. (2020). ExPaNDS General Architecture description in relation to the EOSC services. DOI: 10.5281/zenodo.3697704

[2] Ashton, Alun; Barty, Anton; Fuhrmann, Patrick; Konrad, Uwe; Lang, Franz; Matej, Zdenek; Ounsy, Majid; Reynolds, Christopher; Servan, Sophie. (2021). Photon and Neutron reference data sets. DOI: 10.5281/zenodo4558708.

[3] Alejandra Gonzalez-Beltran; Carlo Minotti; Louise Davies; Marco Leorato; Matthew Richards; Rolf Krahl; Sudha Padmanabhan Viktor Bozhinov (2022). Demonstrate ICAT and SciCat released with APIs compatible to ExPaNDS federated EOSC. DOI: 10.5281/zenodo.6363591

[4] Salvat, Daniel; Gonzalez-Beltran, Alejandra; Görzig, Heike; Matthews, Brian; McBirnie, Abigail; Ounsy, Majid; Da Graca Ramos, Sylvie; Vukolov, Andrei (2020). Draft recommendations for FAIR Photon and Neutron Data Management. DOI: 10.5281/zenodo.4312825

[5] Knodel, Oliver; Konrad, Uwe ;Valcarcel-Orti, Ana; Fuhrmann, Patrick; Servan, Sophie; Roarty, Kat. (2021). Demonstrator for using e-learning platforms for PaN. DOI: 10.5281/zenodo.5171766

[6] Ender, B.; Thibault, P. (2016). A computational framework for ptychographic reconstructions. DOI: 10.1098/rspa.2016.0640

[7] Kahnt, Maik; Sala, Simone; Johansson, Ulf; Björling, Alexander; Jiang, Zhimin; Kalbfleisch, Sebastian; Lenrick, Filip; Pikul, James H.; Thånell, Karina (2020). Raw data for "First ptychographic X-ray computed tomography experiment at the NanoMAX beamline". DOI: 10.1107/S160057672001211X

[8] Brudvik, Jason; Schoen, Silvan; Matej, Zdenek; Barty, Anton (2021). ExPaNDS Testing and Validation Framework. DOI: 10.5281/zenodo.5718671

[9] Myatt, Daniel P.; Hatter, Louise; Rogers, Sarah E.; Terry, Ann E.; Clifton, Luke A. (2017). Monomeric green fluorescent protein as a protein standard for small angle scattering. Biomedical Spectroscopy and Imaging **6** (2017) 123–134 123, DOI: 10.3233/BSI-170167

[10] Ivănoaica, Teodor; Schrettner, Lajos; Dostál, Martin; Majer, Jiri; Bagó, Balázs (2022). Integration of local compute resources into EOSC portal, DOI: 10.5281/zenodo.5913423

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

[11] Brudvik, Jason; Schoen, Silvan; Matej, Zdenek; Barty, Anton (2021). ExPaNDS Testing and Validation Framework. DOI: 10.5281/zenodo.5718671

[12] Tran-Lundmark, Karin (2021). Synchrotron Imaging of Complex Vascular Lesions in Human Pulmonary Hypertension: Pathology Distribution in 3D Space. DOI: 10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7

[13] Perrin, Jean-François (2022). Integration of the PaN AAI into the EOSC. DOI: 10.5281/zenodo.5913472

[14] Manzi, Andrea; et al. (2021). Guidelines for implementing the national Photon and Neutron RI's analysis services within the EOSC. DOI: 10.5281/zenodo.4569421

[15] Fangohr, Hans et al. (2019). Deliverable: Report on the current technical elements of data analysis at each partner site. DOI: 10.5281/zenodo.5905387

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.