PaNOSC
Photon and Neutron Open Science Cloud
H2020-INFRAEOSC-04-2018
Grant Agreement Number: 823852

Deliverable: 6.2 Integration of local compute resources into EOSC portal

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

Project Deliverable Information 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:	6.2
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	30/11/2021
Actual Delivery Date:	18/01/2022
EC Project Officer:	Flavius Pana

Document Control Sheet

Document	Title: Deliverable 6.2 PaN EOSC Compute Cloud
	Version: 3
	Available at: https://github.com/panosc-eu/panosc
	Files: 1
Authorship	Written by: Teodor Ivănoaica (ELI-ERIC), Lajos Schrettner (ELI ALPS), Martin Dostál(ELI ERIC), Jiri Majer (ELI Beamlines), Balázs Bagó (ELI ALPS)
	Contributors: Jayesh Wagh (ESRF), F. Dall'Antonia (XFEL.EU), R. Dimper (ESRF), Erwan Le Gall (ILL), J.F. Perrin (ESRF), R. Pugliese (CERIC-ERIC), D. Roccella, Andy Götz (ESRF), Thomas H. Rod (ESS),
	Reviewed by: Thomas H. Rod (ESS), Andy Götz (ESRF)
	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 ERIC (ESS)	Sweden
5	Extreme Light Infrastructure ERIC (ELI-ERIC)	Czech Republic
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	The Netherlands

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

Table of Content

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

1. Introduction

Description of the deliverable

As PaNOSC services are evolving, partner facilities are adopting or preparing the adoption of the FAIR tools and services. By preparing the key processes supporting the adoption of FAIR and open science, each RI is facing both technical implementation and integration challenges as some partners, already dealing with Open Data, have to change parts of their Data Management operations, as well as adapting for the organisational challenges and changes that come together with a change which is impacting a part of their organisation data governance models.

In the following report, we present the challenges in adopting, implementing and promoting the FAIR set of tools and services supporting PaN (but not limited to PaN) partners to integrate their services and resources into EOSC.

Aim, objectives and scope of the activity

The following report focuses on describing the technical work required to integrate partner compute resources into the EOSC environment.

2. Methodology

The activity supporting this deliverable was started from the beginning of the PaNOSC project in each work package (1) by developing guidelines and policy frameworks helping the community understand the FAIR challenge and thus giving the necessary support for the management teams, and/or (2), by developing tools and services, that help every partner adopt FAIR data standards and integrate into EOSC.

During the entire project, each work package has continuously focused on developing core FAIR services that will further support the integration of local resources into the EOSC portal.

In this report, facilities already running the tools and services are presenting the details about the implementation process, support materials and details, while the new facilities are presenting the integration risks challenges faced during the adoption process, security risks and some mitigation plans used to address those risks.

The current document collects the status of the current developments, as provided by each work package leader, together with the challenges addressed by teams working on the development of the FAIR tools and services. We have included details about the effort required to adopt and

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

implement FAIR services by each PaNOSC partner RI. By aggregating the input, we aim to present a complete and comprehensive view of the activities needed to integrate each RI's compute resources into EOSC.

3. Compute resources integration challenges

The PaNOSC project, Photon and Neutron Open Science Cloud, brings together six strategic European research infrastructures (ESRF, CERIC-ERIC, ELI Delivery Consortium, the European Spallation Source ERIC, European XFEL and the Institut Laue-Langevin – ILL, and the e-infrastructures EGI and GEANT, aiming to support the construction and development of the EOSC, an ecosystem allowing universal and cross-disciplinary access to open data through a single access point, for researchers in all scientific fields.

Having a clear role to support the construction of EOSC, the PaNOSC community is supporting the definition and development of a data commons for the Photon and Neutron Science by developing core services and tools that will allow its partners to adopt Open Science and eventually join EOSC.

Compute resources are a common underlying infrastructure for each PaN member facility, tailored for supporting the needs of each RI and their users. The computing resources of each research facility are a common element facilitating science and users' support. PaNOSC activities aim at supporting the evolution of these computing resources supporting different experiments and scientific research, which spread across various organizations serving specific experimental setups and user experiments, that can now rely on a common set of standards and data commons.

Bearing in mind that compute resources are facility-specific, and the implementation addresses specific challenges, each PaNOSC partner contributes to setting and implementing data commons to facilitate the adoption of common FAIR tools and services that would further support the EOSC integration. The successful adoption of new standards and services requires a critical mass of users and scientists to test and validate functionalities and support the IT and computing teams by providing feedback and use cases, and this is also the case of PaNOSC where scientists are encouraged to use the tools and services in their research activities.

Considering the particularities of each PaNOSC member, EOSC integration becomes possible at the level of the data commons that can support different scientific communities. In this case, data commons promoted by the PaNOSC are aggregating open data, a set of tools, services built to support FAIR standards across various different organisations. Following this strategy, each work package offers the support tools and guidelines facilitating the adoption of policies, offering the necessary organisational support, or offering the essential basic services that could facilitate the implementation of such policies and the EOSC Integration.

WP2 Policies and guidelines activities

The PaNOSC work package two (WP2), focuses on delivering the necessary policies to support the adoption of the FAIR Data Policies at PaN facilities. The WP2 activities started with a survey

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

of the existing Data Policies of the PaNOSC sites – ILL, ESRF, EuXFEL, CERIC-ERIC, ESS and Observers (PSI, ALBA, SOLEIL, Diamond Light Source), and compiled a document including commonalities and differences. Based on this analysis of the existing policies and on the analysis of the particularities of each PaN partner and, after performing an in-depth analysis of the FAIR principles (also in collaboration with a specialised organisation in implementing FAIR principles and GDPR, like GO-FAIR or FAIRsFAIR), the WP2 members have published a new Data Policy Framework[1] which offers the support for adopting a Data Policy (for new RI partners, such as CERIC or ELI ERIC) or for updating the existing Data Policies to meet the FAIR standards requirements.

FAIR Policies at European Synchrotron Radiation Facility (ESRF)

The current ESRF Data policy is based on the outcomes of the PaN data framework and was established in 2015. While adapting the PaNOSC data policy framework, we noticed the following needs for an update in the ESRF data policy.

ESRF Data Policy (2015)	**Proposed Updated ESRF Data Policy (2021)**
FAIR concepts not mentioned	Explicitly mention FAIR and its objectives
Reduced or compressed data not mentioned	Explicitly mention the possibility of storing reduced/compressed data as raw data
Processed data are not included (mention is made of curation of results on a best effort)	Explicitly open the possibility of storing processed data covered by the data policy
Auxiliary data not mentioned	Define and mention auxiliary data
Electronic logbook not mentioned	Explicitly mention the electronic logbook as part of the metadata capture
ORCID not mentioned	Explicitly mention ORCID as a means of linking users to data DOIs
Termination of data custodianship not mentioned	Explicitly include data custodianship termination clause
The data format is not explicitly mentioned	Explicitly mention that the preferred data format is HDF5
Only metadata from ESRF software accepted	Allow the possibility of metadata from non-ESRF software especially for processed data

[1] https://zenodo.org/record/3862701#.YYIE22DMKUk

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

Persistent identifiers (PIDs) mentioned in general	Explicitly mention the use of DOIs as PIDs
The granularity of PID is experiment and dataset	Explicitly mention PID can refer to a bespoke collection of datasets in addition to experiment and dataset automatically generated
No explicit mention of data under embargo being available for AI/ML use by the facility	Explicitly mention that data under embargo can be used for AI/ML by the facility
No formal process for making changes to the Data Policy	Define an internal process for making minor changes to the Data Policy

A formal process for adopting the changes has been started, an initial step involving a discussion with scientific representatives has been done in 2021, the process should now continue with directors, formal approval is expected by the end of 2022.

The introduction to the Data Management Plan for ESRF users was initially discussed with beamline scientists, the implementation is an ongoing process that will need to select relevant questions for ESRF experimenters in order to add value to the data collected without impacting the efficiency of the scientific workflow.

The European Spallation Source (ESS)

The adoption of the new data policy requires the review and approval of several stakeholders at ESS. The process starts at the Data Management Software Centre, who with input from the Scientific Activities Division prepares an initial draft of the policy document in line with the Statutes of the ESS and the insight to best practice gained from PaNOSC and ExPaNDS. This preparatory stage requires several rounds to ensure precise and accurate language is used, clearly conveying each point and ensuring uniformity between policies in terms of acronyms used and defined phrases. After the preparatory stage, the scientific management team reviews the document, which is then sent to the council for final approval.

At the moment, the status of integration of the data policy in ESS can be described as follows:

- The current ESS Policy for Scientific Data has been used to guide the first draft of an updated Policy for Scientific Data.
- A team from the DMSC and Scientific Activities Division have undertaken an iterative cycle of revisions to converge on a document with the clarity and precision that is necessary for such a policy and has now converged on a draft.
- The Science Management Team is primed to review the policy.

ESS has installed and customized a tool for creating data management plans and integrated it with the IT landscape. The integration has so far involved the user office software as well as the metadata catalogue with more in-depth integrations to follow. This enables all users to get a partially filled DMP upon proposal submission. It is then optional for the user to enhance this

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

DMP by providing additional details.

Institut Laue-Langevin (ILL)

The Data Policy draft is currently in the two-step validation process. Once the changes are integrated, the DPP (Data Protection and Processing group) will review the draft. After their validation, the document will be submitted to the Management Board of the ILL, which will decide on its adoption. Then the new DP will be published on ILL's website.

Data Management Plan is defined in the new Draft of the Data Policy. Since the ILL's proposal portal will be redesigned, the integration of the DMP form generation is scheduled. Nevertheless, third-party tools like Data Stewardship Wizard will be considered. Also, the data portal that indexes the experiment's data will be modified to allow DMP's update all along with the experiment's progression.

Central European Research Infrastructure Consortium (CERIC-ERIC)

The CERIC-ERIC General Assembly has approved the Data Policy, and one of the partner facilities (i.e., Elettra) has an implementation in place which is very similar and as FAIR as the CERIC one; CERIC is taking steps forward to make other partners benefit from the FAIR services offered by Elettra to make DP real for as many partners as possible, and is also evaluating cloud solutions for a wider adoption among the partners.

Regarding the Data Management Plan, CERIC-ERIC is evaluating DMP management tools such as Data Stewardship Wizard to decide how to integrate DMPs into the user office system. CERIC is designing surveys for specific stakeholders (e.g., users, beamline scientists, researchers, etc.) to collect their respective viewpoints. When strategic decisions are taken, CERIC will design a DMP template for each instrument as compatible as possible with that proposed by the PaNOSC project.

European XFEL (XFEL.EU)

European XFEL has received a positive recommendation from the external committee appointed by the Management of the European XFEL concerning the adoption of the PaNOSC FAIR data policy framework. During the annual European XFEL User Meeting,
the user community consulted the topic in a dedicated session on "FAIR Data Management".

A new version of the Data Policy has been drafted based on the received feedback. Currently, the updated draft is under consultation with internal stakeholders, starting from Data Department groups, and then further with Instrument Scientists, User Office, and Legal group.

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

After subsequent analysis and incorporation of this feedback, an updated version of the Scientific Data Policy will be presented to the Management Board for approval and subsequent presentation to the Scientific Advisory Committee to obtain its recommendation.
The final document will be then proposed to the European XFEL Council for final approval.

The introduction of the Data Management Plan is proposed in the updated Data Policy draft and its integration with facility tools like User Portal and Metadata Catalogue is being evaluated.

Extreme Light Infrastructure ERIC (ELI-ERIC)

With the FAIR Data Policy approved by the General Assembly, ELI ERIC is now able to leverage the experience offered by the PaNOSC collaboration in implementing FAIR tools and services that are facilitating the implementation of the Data Policy.

ELI's Data Policy and Access Policy are forming the cornerstone of the ELI Scientific Data Management System, providing the management and organisational support to implement data commons for ELI facilities. The policies are stating the commitment of ELI to FAIR, open science and open-source software considered, as much as possible, for the software developed at ELI.

Following the PaNOSC activities related to PaNOSC PaNOSC D2.2 DMP template for facility users[2], the ELI ERIC DMP is now under development and will be implemented through the proposal system, all preparation activities are ongoing and evaluated by the ELI Users' Office team and will be soon part of the Data Policy implementation plan.

As a next step, as part of the Data Policy implementation roadmap, ELI teams are now defining and implementing Data Stewardship and Data Custodian roles and processes, thus facilitating the adoption of the common data strategy for ELI. In this context, based on the recommendations of the ELI International and Scientific Advisory Board, ELI Scientific Computing and Data Management Group will receive the support to promote the new standards and attract the necessary critical mass of scientists that will further contribute at improving and customizing the ELI FAIR data tools and services forming the ELI Scientific Data Management System.

WP3 Data Catalogue Services

In work package 3 (WP3), the partners create means for users and third parties to find datasets from photon and neutron source domain-specific search terms. The core activity of this activity is focusing on developing/adopting data commons and features supporting the PaN members (but not limited to) to make the data findable for their users.

A user-facing frontend to the federated search service is being developed in close collaboration with WP4 and technically led by ELI. In this process, a number of practical use cases have been explored which, together with testing at partner sites, provided useful feedback that will result in revisions to the search API and clarifications in the documentation. For example, a common set of shared experimental parameters has been defined and the roles for personnel associated

[2] https://zenodo.org/record/5639428#.Yd7c81jMLvU

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

with datasets have been agreed upon. To map and curate these definitions onto the locally held datasets will be an ongoing task.

Significant progress has been made on how to appropriately label and query the experimental technique that forms the basis of a dataset. PSI (ExPaNDS) and ESS are jointly leading the effort on this. A reference specification and implementation are in preparation, following the submission of a related ontology deliverable in ExPaNDS.

In addition to these filtering options (by technique, by person, by presence or value of a specific parameter) the search must also rank the results by relevance based on the text search terms. Since the federated service does not (necessarily) have all the information that underpins a match, this ranking metric has to be calculated locally and in an agreed way. A reference implementation of how this has to be done has been developed by ESS and been demonstrated to partners. In the following period, the same ranking is expected to be tested and further implemented by the PaNOSC project partners.

European Synchrotron Radiation Facility (ESRF)

The ESRF is opening access to the data generated by the experiments through its data portal . In addition to the web frontend, the data portal exposes the catalogue of data (currently 4000 data sets) through OAI-PMH, the PaNOSC search API has also been implemented in the course of WP3 activities. Data are made public after an embargo period of a maximum of 3 years. The data portal has been recently registered to the EOSC Portal.

European Spallation Source ERIC (ESS)

The ESS is developing and has provisioned SciCat (https://scicat.ess.eu) as its metadata catalogue service. The service is being developed in collaboration with MAX IV (SE) and the Paul Scherrer Institute (PSI). SciCat will be used to make data findable and accessible according to both the existing and the updated data policy (see Section WP2 Policies and guidelines activities. Besides SciCat, data are exposed through the PaNOSC search-API. Data can also be found through B2FIND and OPENAIRE by means of OAI-PMH.

ESS is also hosting the federated search that enables a user to search across the partner facilities. The next step is to implement uniformly scoring of datasets across the involved facilities so that a user receives the most relevant result at the top for a given query. To this end, ESS has created a detailed description of how to calculate scores as well as a reference implementation that will be sent out to the partner facilities

Central European Research Infrastructure Consortium (CERIC-ERIC)

After evaluating the alternative solutions, CERIC has decided to adopt the ICAT metadata

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

catalogue.
The reasons that led to this choice were:

- the amount of work required to complete integration,
- the adoption of a solution already in use in the PaN community,
- the possibility to have out-of-the-box integration with the EOSC.

CERIC is a distributed research infrastructure. Consequently, this has affected the chosen setup which consists of a centralized ICAT installation populated by software tools that ingest data from partner facilities dataset and the user office database.

The need to collect experimental data that is the outcome of experiments carried out at CERIC partner facilities will be satisfied through tools already developed that allow saving data on storage provided by Elettra and cloud storage.

Regarding services and interfaces accessible via EOSC, ICAT comes with a ready-made OAI-PMH plugin and CERIC is registered to re3data.org and DataCite. The ICAT community works on an ICAT plugin implementing the PaNOSC wp3 search-API. The release of this plugin, and the consequent installation at CERIC, is foreseen at the end of 2021.
At the moment, we have identified no risks or IT infrastructure challenges.
Although ICAT is the current catalogue system, CERIC is evaluating the capabilities of INVENIO too. Furthermore, Elettra will extend its User Office database (VUO) to store experimental metadata to be ambidextrous and ensure sustainability in the long term.

European XFEL (XFEL.EU)

At the start of the project, XFEL had already in place myMdC as its metadata catalogue service.

To comply with the PANOSC goal and deliveries, myMdC had to be enhanced with the necessary features and integration endpoints.

In PANOSC, techniques metadata and necessary RESTful wp3 search APIs were developed into myMdC, which provided real-time production metadata to the registered partner services and users.

Regarding other services and interfaces accessible via EOSC, myMdC is registered at re3data.org and DataCite. DOIs are automatically generated for all proposals after their beamtime, and the OAI-PMH plugin will become available in production during the first half of 2022.

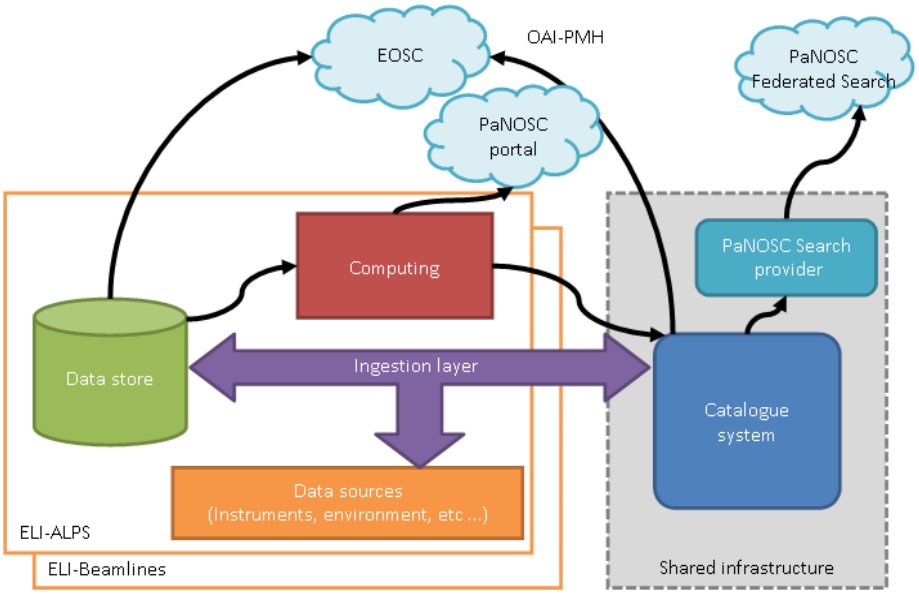
Concerning integration with wp4, we are currently developing the necessary interfaces to guarantee that myMdC and XFEL VISA portal are fully integrated using RESTful APIs.

At the moment, we have identified two organizational challenges: a) understand implications of anonymous data access concerning the legal regulations; b) prepare and build IT infrastructure and procedures, allowing users secure access to raw data.

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

Extreme Light Infrastructure ERIC (ELI-ERIC)

ELI is progressing towards designing and implementing a data management system outlined in the drawing below.



This process is connected to the WP3 activities at several points. ELI is in the second phase of evaluation of catalogue systems and is gathering experience in setting up and operating an ICAT and an Invenio RDM instance.

ELI does not have publicly available datasets at this moment, but example datasets have been prepared and made available through a PaNOSC search API provider (reported in D3.2). The first version of the search provider had direct access to the example datasets; since then, work has concentrated on putting the datasets behind the evaluated catalogue systems and creating a link between the catalogue system and the search provider. The necessary components have been developed for Invenio RDM. The same data path will be available for ICAT because of an ongoing development effort by the ICAT development team. As a result, ELI is technologically prepared to support a PaNOSC search provider instance, and hence to contribute search results to the federated search instance.

The evaluated data catalogue solutions support OpenID Connect, through which successful UmbrellaID authentication tests have been carried out, and thus future support for UmbrellaID can be ensured.
Work has started towards establishing a shared infrastructure for the deployment and operation of an ELI wide catalogue system and search provider. However, there are still numerous challenges in this area.

WP4 Data Analysis

Work Package 4 focuses on creating and providing data analysis services locally at the partner sites, and eventually through the EOSC. The work package has significant interactions with and dependencies on other work packages.

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

Core activities of WP4 are centred around the provisioning of a PaN portal for remote data analysis. Currently, two major services are envisaged within this context:

- The VISA platform, as developed by colleagues at ILL, provides a complete cloud service with web user-interface, and API to spawn virtual machines as data analysis instances giving access to resources on OpenStack (or other) RI-local compute infrastructure providers, and two flavours of workspaces: a remote desktop for access to GUI software, and a Jupyter notebook service for Python code including - among other - packages developed in the PaNOSC context. This resource was originally designed with the respective facility's users in mind, but there are perspectives to extend/adapt this to allow also for 'external' PaN community users. Currently, the deployment of VISA at partner RIs is ongoing.

- A PaN search portal for open data, combining a web user interface contributed by WP4 developers (ELI) with the federated search API developed in WP3. The aim of such service is the retrieval of aggregated results from metadata catalogues of all PaNOSC and ExPaNDS partners, and re-directing the user to the website of the RI where the data is hosted, ideally by DOI registration of open data sets, and optionally a selective forwarding to existing local download and/or computing services.

Other activities of the WP4 have produced initially local services which are partly already shared among PaNOSC RIs. These are the h5Web tool (as stand-alone web service or part of a JupyterHub service via JupyterLab plug-in) by ESRF, domain-specific data viewers that will be based on a refactored h5nuvola framework by CERIC, or h5glance to view HDF5 structures in a terminal or Jupyter notebooks.

Generally, data analysis solutions developed in the scope of PaNOSC-WP4 are registered in the PaNdata software catalogue (software.pan-data.eu), including also more facility-specific tools/frameworks like EXtra-data and EXtra-geom from EuXFEL. The PaNdata software catalogue itself is - as "PaNOSC Software Catalogue" - registered in EOSC. The direct registration of WP4-developed tools in EOSC could be on the RI's level or the federated level. There are plans to host a VISA instance as a federated service by the EGI, and such an instance would be an obvious candidate for EOSC registration.

Each PaN Partner should add details about how they see the support for (remote) users, accessing computing resources and data at each facility

European Synchrotron Radiation Facility (ESRF)

In January 2021, ESRF rolled out a Jupyter notebook service (https://jupyter-slurm.esrf.fr/) connected to the ESRF HPC cluster primarily intended for remote users. This service allows

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

scientists to perform analysis on the ESRF data without having to download them, keep track and ultimately publish the processing sequence leading to the scientific results.

Jupyter notebooks represent a natural step forward as they allow to perform data analysis remotely and publish these analyses in a reusable manner, but not all software and processes are currently usable inside a notebook. ESRF has decided to implement and deploy the VISA solution to bridge the gap between the variety of scientific software and the need to provide remote access to data analyses services. As of December 2021, ESRF has deployed VISA on a production-grade OpenStack infrastructure. In early 2022 we plan to meet the beamline scientists to prepare the environment for their users and progressively open the service to all users.

While these services are preliminary intended for ESRF users, we expect in 2022 to dedicate some resources for authorizing data scientists who are not considered historically ESRF users to benefit from these analyses services for processing the ESRF open data.

The European Spallation Source (ESS)

The VISA service for ESS is currently deployed in a test setup with a prototype OpenStack backend. While this interim setup is not yet production-ready, it is sufficiently capable for gaining experiences from ESS internal users in collaboration with ESS infrastructure personnel in order to iron out any caveats in the process of rolling out VISA and OpenStack in production. At ESS the Data Management and Software Centre (DMSC) has Keycloak running providing federated authentication against local identity sources as well as against UmbrellaID. The plan is to integrate OpenStack and VISA with Keycloak for federated authentication of data analysis services.
Work is planned for the remaining PaNOSC project to move towards a production-grade setup and to provide VISA images containing the relevant software packages for data analysis for the beamlines planned for ESS. Also, integration with the ESS user office software and data catalogue is planned activities in that period.

VISA is currently seen as a very strong contender for the future data analysis service for ESS.

The integration of VISA against federated authentication and not least authorization sources is seen as a challenge. While important experience has already been made with UmbrellaID in work package 8, that experience also indicates that the authorization part with federated identity sources does present both technical and security challenges when integrating with any specific system. At ESS these challenges are being tackled by incorporating federated authentication and authorization considerations early on in the design and implementation work of the VISA solution at ESS.

Institut Laue-Langevin (ILL)

VISA has been running at the ILL in production since January 2020 providing remote analysis services as well as a platform to perform remote experiments. Initially concentrating on Remote Desktop access, JupyterLab was integrated in September 2020. Since June 2021 VISA has been

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

open-sourced and the ILL is now providing support to deploy VISA at other facilities. A pilot project of running VISA on EGI infrastructure is also in progress, again with the support of the ILL.

Current IT infrastructure challenges involve providing access to GPUs shared between multiple analysis VMs and the distribution and versioning of analysis software made available using CernVM-FS.

Authentication is achieved using OpenID connect and the ILL uses Keycloak as the service provider. Currently, authentication is provided against local identity sources however in the near future, UmbrellaID will be supported. However, access to VISA services is uniquely available to ILL users.

The ILL has no immediate plans to provide VISA as a public service within EOSC. In the context of providing FAIR access to data and services, VISA could be adapted to provide access to Open Data and compute services. There are however a number of hurdles to achieve this including dataset ACLs, compute quotas, infrastructure capacity and long-term sustainability.

Central European Research Infrastructure Consortium (CERIC-ERIC)

CERIC has developed a set of singularity containers with GUI applications to use QUASAR, PyMCA and other python-based data analysis pipelines, and jupyter-lab notebooks; these tools are available also remotely to users via the CERIC VUO. Two specific purpose web tools (XrfFitVis and FidViewer) have been developed as two PaNOSC use cases and they are publicly available. Meanwhile, a general web-based hdf5 explorer, h5nuvola based on Django Python framework, has been deployed locally as a modular application that can be easily extended:

- Different tools (including XrfFitVis), are available as customizable plugins.
- Data access has been generalized, local storage or NFS shares are supported, read and write operations are performed with the correct uid/gid.
- New tools and visualizers can be added to the application using pure python and using Dash/Plotly visualization libraries.

Moreover, CERIC is porting the VISA Portal to Proxmox in order to install and take advantage of the VISA Portal in the currently available infrastructure.

European XFEL (XFEL.EU)

The aforementioned tools and frameworks: h5glance, EXtra-data and EXtra-geom, have been developed in the data analysis group at EuXFEL. There has also been a contribution to h5py, a package to wrap the HDF5 back-and into Python. There are currently no concrete plans to register either of these in EOSC, in particular, because they are not stand-alone web services (but can currently be run by EuXFEL users in Jupyter notebooks on the DESY Max-Jhub service).

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

The VISA service for EuXFEL is currently being deployed on the OpenStack production infrastructure at DESY. The actual process has been successfully accomplished, but in this phase, the VISA usage itself is in a test state, until the ETL process has been adapted to pull user data from the metadata catalogue service myMDC. The environment of EuXFEL/DESY has Keycloak working to support OpenIDConnect tokens as an authentication method; this technology will be suited to work with the Umbrella-AAI, once this will be in the mature state.

The EuXFEL strategy foresees distinct usage phases for VISA. It is planned to start with access to the subset of open (post-embargo) data, to which all EuXFEL users, but also new PaN community users would have access. Technical challenges in this context are:

- transfer of LDAP information to OpenIDConnect for existing users while at the same time providing a registration (e.g., via Umbrella) for first-time PaN users.
- separation of open data from embargoed data, either physically by storage volume instance, or virtually by file tagging, allowing data mounts from VISA/OpenStack VM instances - by NFS or WebDav - only for the open set.

Eventually, VISA would also become a service for EuXFEL users with respect to their own (still embargoed) data, which poses additional challenges

- differential/dual data access: PaN users can access only open data, EuXFEL users both open and own data
- alignment of authentication (token information) with file access authorization to allow for these selective access levels.

Extreme Light Infrastructure ERIC (ELI-ERIC)

Having the FAIR ELI Data Policy approved in December 2021, is now giving the teams the necessary support to start working on the implementation and integration of the services facilitating the integration and implementation of the policy for the two facilities of ELI ERIC. The adopted data policy is an expression of the management support and commitment for the implementation and integration of the ELI Data Policy, following the FAIR standards and tools promoted by the PaNOSC community (e.g., file cataloguing, new rich metadata standards, data management and data transfer solutions).

As part of the PaNOSC FAIR Data Policy implementation activities, ELI has deployed and is now testing Keycloak and UmbrellaID (https://aai.eli-laser.eu) to authenticate users accessing the two file cataloguing solutions, ICAT and Invenio RDM, which are already deployed and running as test instances. At the same time, in the development of the ELI ERIC User Portal, keycloak is considered for providing federated authentication against local identity management systems for users and staff.

Following the PaNOSC activities, the ELI Alps team has started deploying a test instance of VISA and, based on these initial tests, the team is now investigating the possibility to use Proxmox, as a virtualisation hypervisor for the ELI Computing backend. (in the long-term, ELI will further investigate and prepare a possible transition to a private OpenStack cloud solution).

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

Authentication, using OpenID connect with Keycloak, already tested for different services and will be used for ELI users, while as Identity Providers (IdPs) ELI will aim at having UmbrellaID, as well as the possibility to allow users to use their ORCID ID.

WP6

Work Package 6 prepares and ensures the integration of the PaNOSC services into EOSC by organizing the services support, preparing the Authentication and Authorisation Infrastructure, exploring the data transfer to the computer facilities and participating in the EOSC definition.

UmbrellaID (https://umbrellaid.org/what.html) has been serving the AAI needs of the PaN community since 2012. In the WP6 of PaNOSC, we are transforming UmbrellaID to make it ready for the EOSC challenges. Services need to be open to other communities, and vice versa. The PaN community should be able to benefit from other community services seamlessly. WP6 is driving the integration of UmbrellaID with EOSC AAI with support from GÉANT. Deliverable 6.3 provides in-depth information on these activities.

Data transfer - D6.1 Data Hub> definition with 3 use cases, EGI Datahub and conclusion

Data transfer concerns local to remote data access and processing. The following three use cases drive the PaNOSC activities concerning data transfer.

1. A Research Infrastructure wants to archive its experimental data in a remote data centre.
2. A user wants to access a data analysis service; data must be available “transparently”.
3. A facility user wants to transfer a large dataset from an RI’s archive to a remote compute centre or their home pc.

The first and third use cases have been addressed after conclusive pilot activities. The Rclone (https://rclone.org/) solution is now daily used in production for archiving ILL data to the STFC cloud storage. European-XFEL and ESRF users transfer large datasets to their home organisation storage using Globus Connect services.

To select the most suitable technical solution for the 2nd data transfer use case, a workshop was jointly organised by the EGI community with the ESCAPE and XDC H2020 projects in Amsterdam in July 2019. The PaNOSC data transfer use cases were presented and discussed during this workshop.

The necessity to have a local cache of the data at the analysis service location and the authorisation mechanism of the community that necessitates a translation of the User ID between the community and the RI local accounts drove us to first test the EGI DataHub service, which is based on the Onedata software stack.

The main components of Onedata are:

- Spaces, which are distributed data volumes, that users can access and organise,
- Zone, which is the federation of providers. The Onezone service has been installed by EGI and is the core of the EGI DataHub. A specific zone has been created for PaNOSC,
- Providers are the entities that support spaces and provide the data storage.

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

The outcomes of the pilot OneData framework are very promising. After being set up in the RIs it allows exposing a uniform interface for the users regardless of the location of the service and data. In the same interface, the user can access datasets archived in different RIs. It allows combining data from different facilities in a single analysis process. This is the same interface for all the data distributed across RIs' archives. Deliverable 6.1 provides in-depth information on the OneData framework implementation. As a next step, the framework will be implemented on more PaNOSC RIs and users' feedback will be collected.

CVMFS - Community software provisioning for computing resources.

Most of the PanOSC and ExPANDS RIs are deploying VISA to offer remote analysis services to the user community. These services provide data access, compute resources, specialised support and software ready to be used. These software are numerous, typically over 100 different software are in use by each RI, they need maintenance and regular installation or update. It is also quite common that the same software is in use in different RIs. Therefore, we have looked for achieving better synergy and for a technical solution that would avoid duplication of efforts in the community. Additionally, this software should also be available 'ready to use' for any organisation and especially for cross-community organisations like EGI that are deploying VISA on their infrastructure to extend the community capacity and offer more resources for data scientists.

We were looking for a solution where:

- we can deploy easily new releases on all the platforms independently of the solution the developer has chosen to package it
- a unified solution that could work for VM, cluster nodes, …
- users could access easily previous releases
- we could benefit from the work of the community and trust the repository
- we could send our users to external compute resources without having to worry that software is easily/already available (typical case is computing resources that could be made available by EGI or other providers)

CVMFS (CERN Virtual Machine FileSystem) has quickly been identified as the solution to provision software that fits our requirements.

Currently, the CVMFS integration into RIs' workflow is evaluated by ESRF, DESY and ELI while STFC is building a new, high-capacity CVMFS service for use by EGI-ACE and other customers, concentrating on efficiently serving container images and making the best use of the STFC ECHO object store to support more requests. STFC also on CVMFS authentication, including integrating with identity providers proxies such as UmbrellaID or EGI Check-In.

Once these evaluations and developments will be over, we will have all the information necessary to decide on a single CVMFS repository or for a federated approach.

Helpdesk Service definition, customized based on the PaNOSC specifics and discussed in the context of WP6 are developed and continuously updated and correlated with the particularities of each of the developed services.

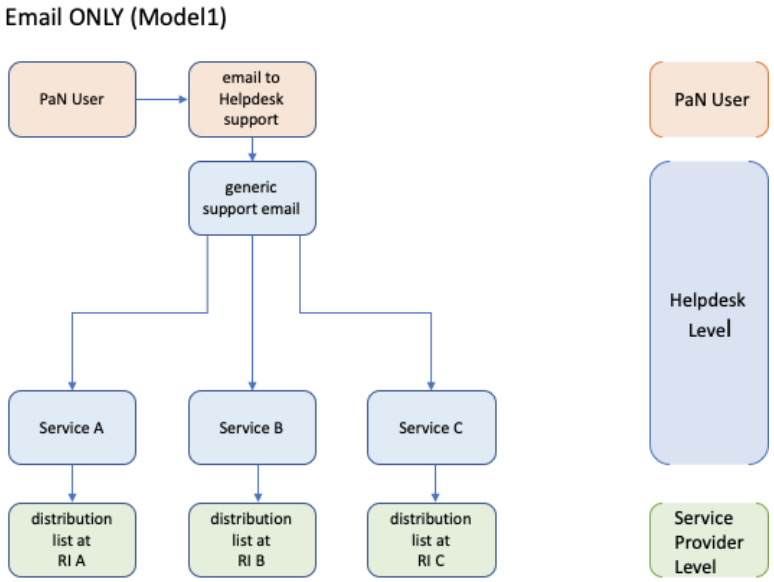
The Helpdesk system is the single resource for managing the incoming support requests for all services hosted at the PaN portal. The PaN helpdesk will ensure that all issues are

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

communicated, assigned, and resolved in a reasonable time frame.

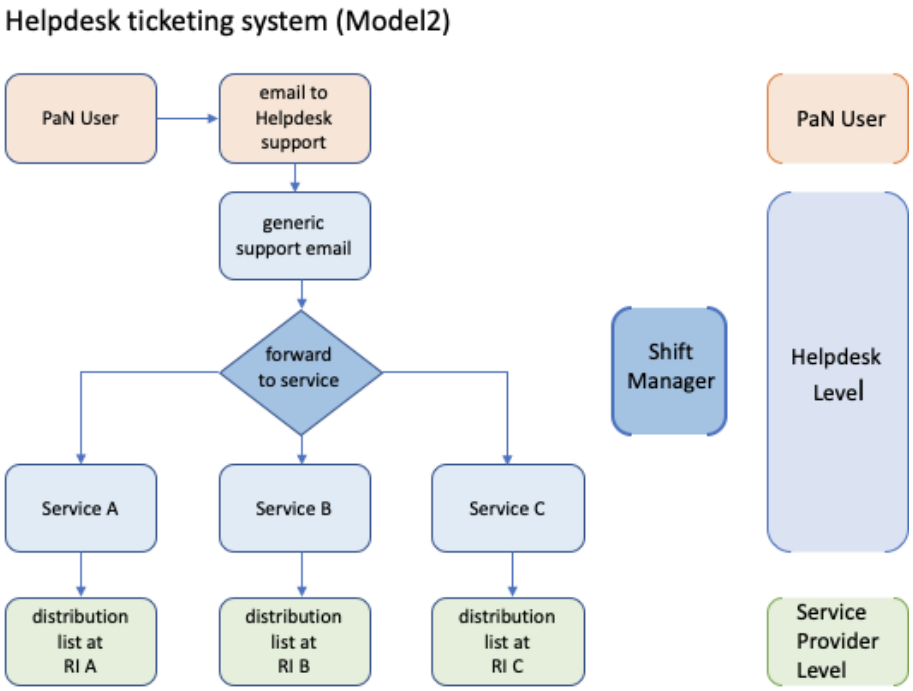
After considering the requirements of PaN RIs and individual researchers, the following two helpdesk models have been proposed.

- Email-only model: In this model, a PaN user submits a ticket to the Helpdesk based on an email ONLY. The advantages of this model are its simplicity and readiness. A user can submit through a simple email, and this model is easy to set up and configure. The downside of this model is the lack of a complete overview of progress on the submitted tickets.



- Ticketing model: In this model, a PaN user submits a ticket to the Helpdesk based on a Helpdesk ticketing system. This model provides a complete overview of progress on the submitted tickets. However, this model requires a specific configuration as well as resources for setup.

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



At present, the email model of the helpdesk is in place, and it allows tracking support requests. Nevertheless, an evolution towards ticketing systems is envisaged, and it will depend on EOSC Future project outcomes.

Out of the two helpdesk models, the PaNOSC community has chosen the email-based helpdesk system as it is easy to configure and set up. In the meantime, we are looking forward to receiving further advice from the EOSC Future project on the helpdesk system. The PaN community will adapt its helpdesk model according to the recommendations from EOSC Future.

WP8

WP8 is developing the online learning platform (currently pan-learning.org) that uses moodle and which is already a registered EOSC service. Among other things, the platform will be used to showcase and provide training for the software developed by WP5. The first example of this is the course on how to use the McStas script (found as part of the IKON 21 python training). This course makes use of the JupyterHub integration on the platform using Docker containers. Currently, the containers run on servers at the ESS, which obviously imposes a theoretical limit on the number of courses that can use this feature. More importantly, the current setup limits how many students can access and run a course at any one time. For future large training schools or conferences, the ability to use compute resources from other institutes would be useful. In addition, this would allow courses on simulation techniques that require much larger compute resources, for example, molecular dynamics and density functional theory, to be hosted on the platform.

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

EGI cloud resource provider for future PaNOSC users

VISA integration in EGI

At the time of writing of this deliverable, the VISA integration into the EOSC Cloud Compute resources is still in progress. Additional work was requested by the VISA developers (WP4) to support other OIDC providers using the new angular-oauth2-oidc module[1]. The latest release of the VISA portal (v2.1) is available in the GitHub repository. The integration of the VISA portal deployed in the EOSC Cloud Compute resources with the e-Infrastructure proxy operated by EGI (Check-in) is in progress.

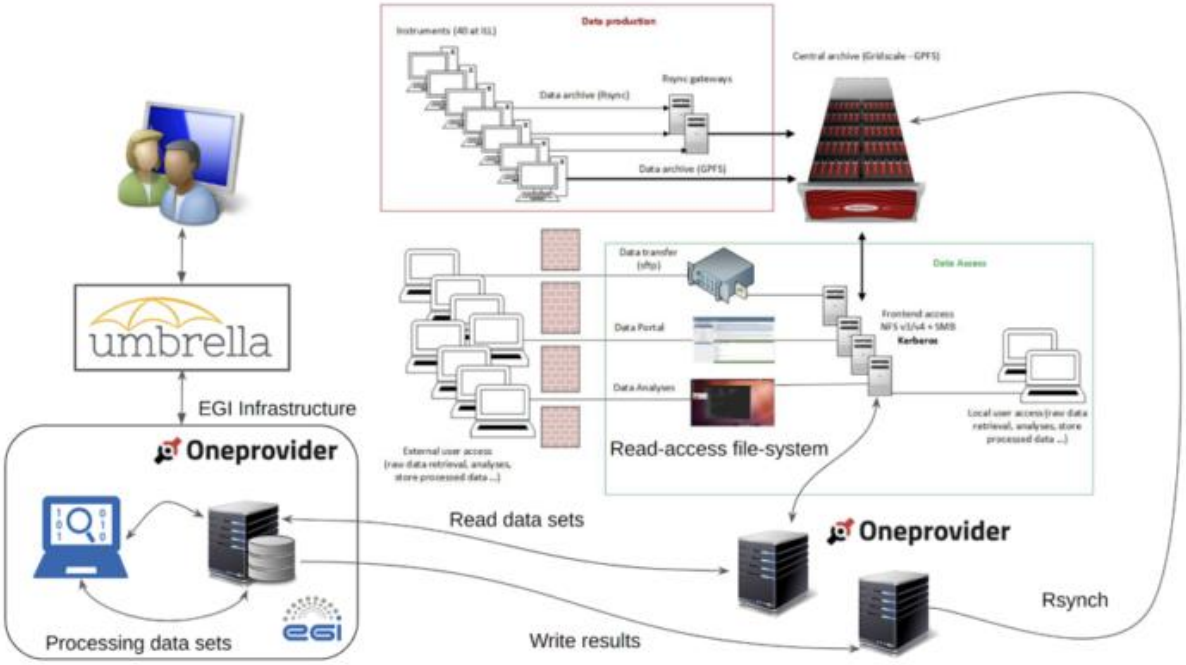
UmbrellaID integration with the e-Infrastructure proxy of EGI (EGI Check-in)

From the Authentication and Authorization perspective, the UmbrellaID community proxy has been integrated with the e-Infrastructure proxy of EGI[2] to allow PaN users with UmbrellaID accounts to seamlessly access the EOSC computing resources. A more global integration approach for the AAI proxies is developed in the EOSC Future project. The PaNOSC AAI has been used as a typical EOSC use case and was successfully demonstrated during the first intermediate review of the EOSC Future project.

Data Transfer services

To support data transfer solutions and allow PaN users to transfer data outside the RI facilities and perform analysis on a different e-Infrastructure, the following solutions were considered:

- EGI DataHub[3] to expose the RIs data using a locally deployed component (OneProvider) and implement the custom mapping between local user accounts or credentials on storage resources (e.g., POSIX user ID/group ID, LDAP DN, Ceph username, GlusterFS UID/GID, etc.) to AAI credentials. The following figure shows the pilot that has been deployed at ILL, CERIC and subsequently at ESRF.



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

- EGI Data Transfer service, in conjunction with Rucio to offer policy-based data orchestration. This is the data management solution that has been adopted by the ESCAPE project. For this particular use case, facilities need to expose their data using storage solutions or gateways providing the protocols for data transfer supported by FTS (GridFTP, HTTP/WebDav).

Onboarding new services in EOSC

The procedure for onboarding new services from PaN RIs in the EOSC Portal is the following:

- Provider registers him/herself into the EOSC Portal. Login with EGI AAI CHECKIN in the EOSC Portal - Select your identity.
    - In the add provider page, 8 blocks of information are shown, and we recommend filling up the non-mandatory fields too to help the user search for the service: Leaving optional fields blank will remove the relevant heading from the published resource/provider profile.
    - Further information and examples can be found here: https://eosc-portal.eu/providers-documentation/eosc-provider-portal-provider-profile.
- Provider onboards (and updates) the Provider information.
- Provider onboards (and updates) the Resources offered by the Provider: **"**onboard a Resource under a registered Provider".

Resources can be added after the successful registration of the Provider, according to the Resource Maturity Classification described here.

Who can onboard resources to EOSC?

- Resources must be onboarded by a legal entity (although the legal entity may do so on behalf of a project or consortium in which they participate, with the agreement of those groups).
- Providers onboarding a resource must assert that they are able to ensure the resource is delivered by them or their collaborators and agree to remove resources that are no longer operational or available.

What resources may be connected to EOSC?

- At present only services are being onboarded.
- It must be an actual service.
- It must be a specific service offered 'live' to customers. This may be an IT service, or a human service (e.g., training, consultancy).
- It may not be a research product, for instance, a document, a dataset or a piece of software.
- The Service must be discrete. It must be available and offer value on its own. It may not be only a feature of a larger service available while already using that service.

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

5. Conclusions

After three years of the PANOSC project, the participating ESFRI projects have made significant progress in adopting FAIR data policies based on the guidelines from WP2. Almost all the PaNOSC ESFRI projects are currently reviewing the FAIR data policies and are expected to make further progress in its adoption by the end of the project in 2022.

A set of services has been registered in the EOSC Portal, most of the PaNOSC ESFRI projects have already registered or planning to register soon their respective data catalogues on the EOSC portal.

Data analysis services based on the Jupyter notebook have been opened to users, the VISA remote desktop analyses service has been open at ILL for 2 years and is currently deployed by all partners to be open to users in the first quarters of 2022.

Infrastructure solutions for data transfer, software provisioning and AAI have been identified and for most of them rolled out for production.

2022 should be devoted to completing the service's rollout and supporting users.

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