Implementing the national Photon and Neutron RI's analysis services within the EOSC

Document Control Information

Settings	Value
Document Identifier:	D4.1
Project Title:	ExPaNDS
Work Package:	WP4
Document Author(s):	Andrea Manzi (EGI), Krisztian Pozsa (PSI), Alun Ashton (PSI), Daniel Salvat (ALBA), Uwe Konrad (HZDR), Franz Lang (UKRI), Zdenek Matej (Max IV), Majid Ounsy (SOLEIL), Christopher Reynolds (Diamond), Anton Barty (DESY)
Internal Reviewer(s):	Brian Matthews (UKRI), Daniel Salvat (ALBA), Patrick Fuhrmann (DESY), Sophie Servan (DESY)
Responsible Partner:	DESY
Doc. Issue:	1.0
Dissemination level:	Public
Date:	29/03/2021

Abstract

An operational framework document describing the sustainable, coordinating organisational structure spanning national Photon and Neutron Research Infrastructures and EOSC. Common rules and best practices for implementing the national RI's analysis services, and future data analysis services, within the EOSC.

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

Table of Contents

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

Executive Summary

This document describes the organisational structure to be implemented for a sustainable integration of ExPaNDS participants into the EOSC both during the project and beyond. It relates to the work within Task 4.1, which aims to implement roles and processes for coordination between the national RIs, the EOSC, and partner ESFRI institutions. The document describes the core services and policies required to contribute to the EOSC, either through the common PaN portal developed by PaNOSC or through our project partner EGI as one example of a horizontal infrastructure to support the PaN community. We discuss the goal of enabling implementation of FAIR data policies and the onboarding of PaN community services into the EOSC as a means of achieving this. On the technical side we describe the requirements for common software catalogues, common authentication mechanisms, and the implementation of the resulting services through the PaN portal developed by PaNOSC into the EOSC service infrastructure.

Three high level areas are identified to work on:

1. Coordination between the partners: We describe a structure for coordinating both the national research institutions in ExPaNDS and the ESFRI research institutions in PaNOSC so that they present a unified approach with respect to EOSC integration both with respect to each other and with our partners in PaNOSC.

2. Integration with EOSC: We describe the presentation of a common, standardised Photon and Neutron community "front window" interface to data and data analysis services for integration into EOSC, providing access to both participating national RIs in ExPaNDS and partner ESFRI institutions in PaNOSC. This will not only be beneficial for EOSC integration efforts of the wider PaN community, but also of direct benefit for users of the PaN national RIs and ESFRI institutions as they will benefit from the development of standardised data analysis services and access mechanisms across institutions.

3. Sustainability beyond the project: Although the task of sustaining ExPaNDS results and services is primarily contained within WP1, the efforts and choices made within WP4 will have a large impact on sustainability beyond the end of the project. A fundamental prerequisite for ongoing sustainability of ExPaNDS analysis services after the project concludes is to have them integrated into the RI's core data analysis service portfolio without unnecessary delay. Moreover, the selection from day one of ExPaNDS data services that are aligned with the users workflow will create a desire by facility users to keep those services maintained after the ExPaNDS project funding period ends. Furthermore, the ability to improve software portability between facilities will directly benefit the PaN user community and facilitate migration towards horizontal infrastructures, as well as lowering the cost of integration by consistently applying industry standards and interfacing to common software stacks across facilities.

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

1. Structure of this document

This document is set out as follows:

Section 2 describes the organisational framework implemented for coordination with PaNOSC[1] during the course of both the projects, as well as a sustainable structure to be adopted once the projects end. We also describe the planned structure for integration into EOSC through a common Photon and Neutron community portal. This section addresses the description of "An operational framework document describing the sustainable, coordinating organisational structure spanning national Photon and Neutron Research Infrastructures and EOSC" part of D4.1.

Section 3 describes the rules, practices and principles to be adopted by ExPaNDS participants with regard to integrating existing data analysis services into EOSC. This section directly addresses the description of "Common rules and best practices for implementing the national RI's analysis services, and future data analysis services, within the EOSC" part of D4.1.

Finally, the appendix contains a lengthier description of the operation of horizontal infrastructures within EOSC. This section describes how tighter PaN integration with EOSC can be implemented but strictly speaking goes beyond the scope of work within ExPaNDS WP4.

2. Organisational framework

This section addresses the description of "An operational framework document describing the sustainable, coordinating organisational structure spanning national Photon and Neutron Research Infrastructures and EOSC" part of D4.1. It describes the organisational framework implemented for coordination with PaNOSC during the course of both the projects, as well as a sustainable structure to be adopted once the projects end. We also describe the planned structure for integration into EOSC through a common Photon and Neutron community portal.

The ExPaNDS project aims to integrate large scale research infrastructures of European national Photon and Neutron facilities and support horizontal services providers into the EOSC. This requires to design and implement unified and consistent interfaces to local resources, particularly important for this work package, to interface analyze pipelines namely traditional batch systems as well as modern cloud access mechanisms like virtual machines, containers and notebooks. In order to cover not only national RIs but similarly ESFRI[2] facilities, we closely collaborate with our sister project PaNOSC. Within this context, ExPaNDS WP4 is responsible for the coordination, adaption, and alignment of existing data analysis services at national RIs with the new EOSC services available via the EOSC Service Catalogue and EOSC Portal.

[1] https://www.panosc.eu/
[2] European Strategy Forum on Research Infrastructures

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

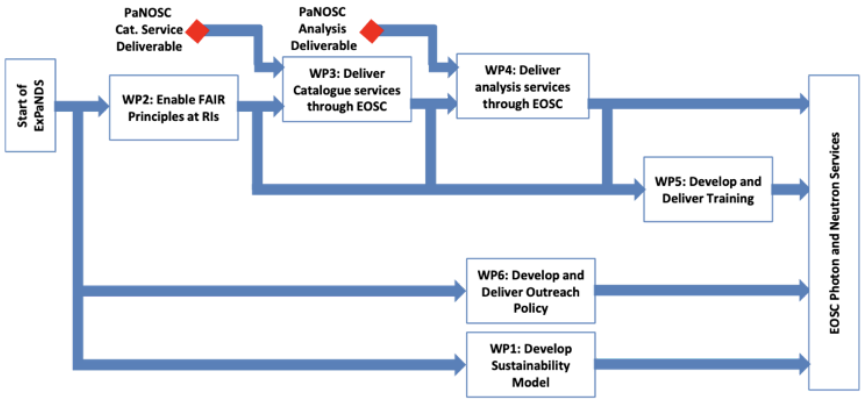
2.1. Background

ExPaNDS is only one piece in a larger, already existing PaN ecosystem which extends beyond the scope of the project itself. As such, achieving a unified PaN interface suite towards the EOSC requires coordination across several levels:

1. Collaboration amongst national research infrastructures within the ExPaNDS project;
2. Collaboration between the ExPaNDS and PaNOSC projects to maintain common alignment of the projects, as both projects essentially cover the same user community;
3. The work on a common framework for the coherent integration of both national and European infrastructures with the EOSC.

The operational structure therefore must span these three levels and provide a framework that ensures sustainability and harmonisation of EOSC integration beyond the project duration.

An additional consideration is the interlinkage between the PaNOSC and ExPaNDS projects foreseen in the proposal document. Although the two projects run in parallel, they are in fact interlinked with the intention of presenting a common Photon and Neutron ('PaN') interface to EOSC. Indeed, the PaNOSC project started one year prior to the start of ExPaNDS, and by the time ExPaNDS started concepts for search interfaces and service portals had already been developed within PaNOSC. The intention expressed in the proposal is that ExPaNDS mirrors decisions in PaNOSC as far as practically possible, and that ExPaNDS ideally uses the same common PaN portal and APIs in order to present a unified 'Photon and Neutron' interface to EOSC. As a consequence, WP3 and WP4 of ExPaNDS have a dependency on the corresponding PaNOSC work packages. This was illustrated in the proposal (p. 44) as follows:



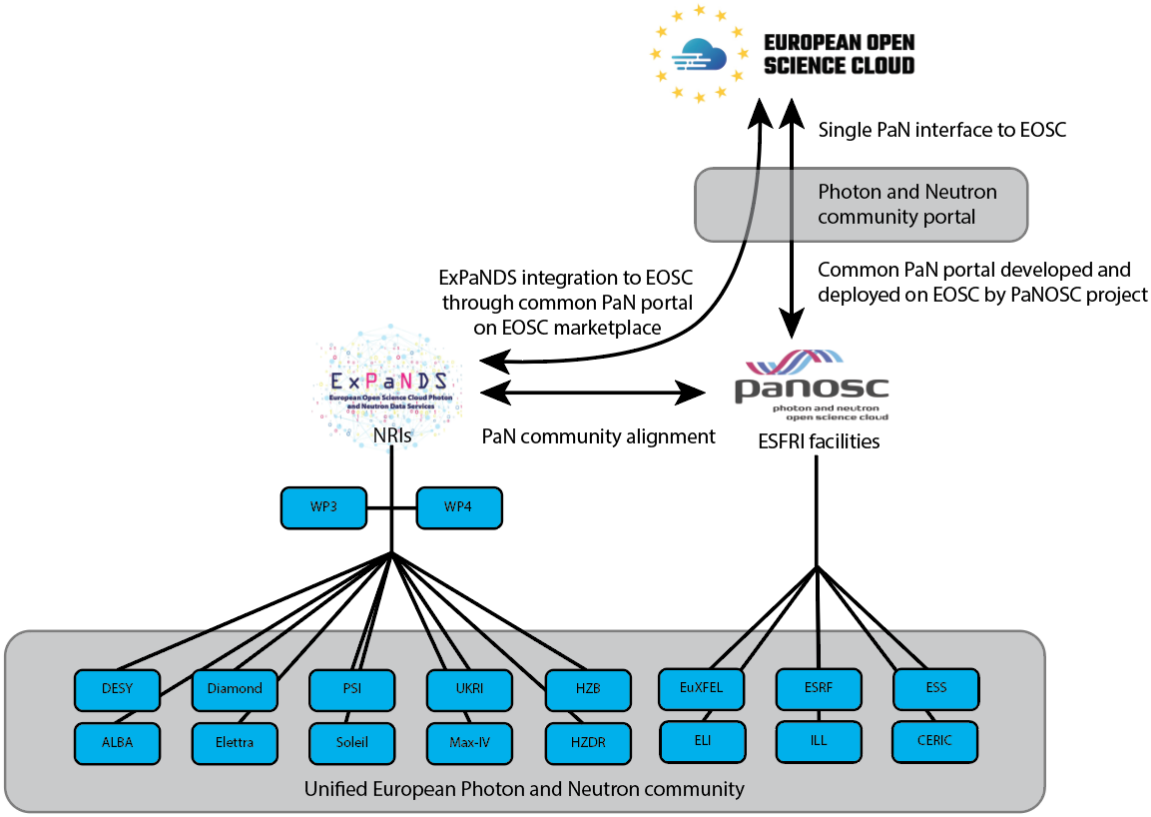
The interlinkage between the ExPaNDS and PaNOSC projects requires tight coordination between projects during project execution. However, from the perspective of the PaN community the distinction between national and European infrastructures is an artificial one. A different organisational structure which breaks down the artificial PaNOSC and ExPaNDS distinction between national and European infrastructures is required to create a sustainable coordinating organisational structure spanning National RIs and ESFRI institutions for

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

sustainability after the separate projects have ended. We therefore have different operational frameworks during the project versus after the project has ended.

2.2. Operational framework during project execution

Alignment to the PaNOSC project and EOSC deployment through the PaN portal developed by PaNOSC is maintained by adopting the following organisational structure during execution of ExPaNDS. This structure addresses the coordination, adaption, and alignment of existing data analysis services at national RIs with the new EOSC services available via the EOSC Service Catalogue and EOSC Portal, using PaNOSC services. This ensures that we do not artificially fragment the PaN community based on grant funding nor do we attempt to duplicate the ongoing PaNOSC effort on the PaN portal.



Key aspects are:

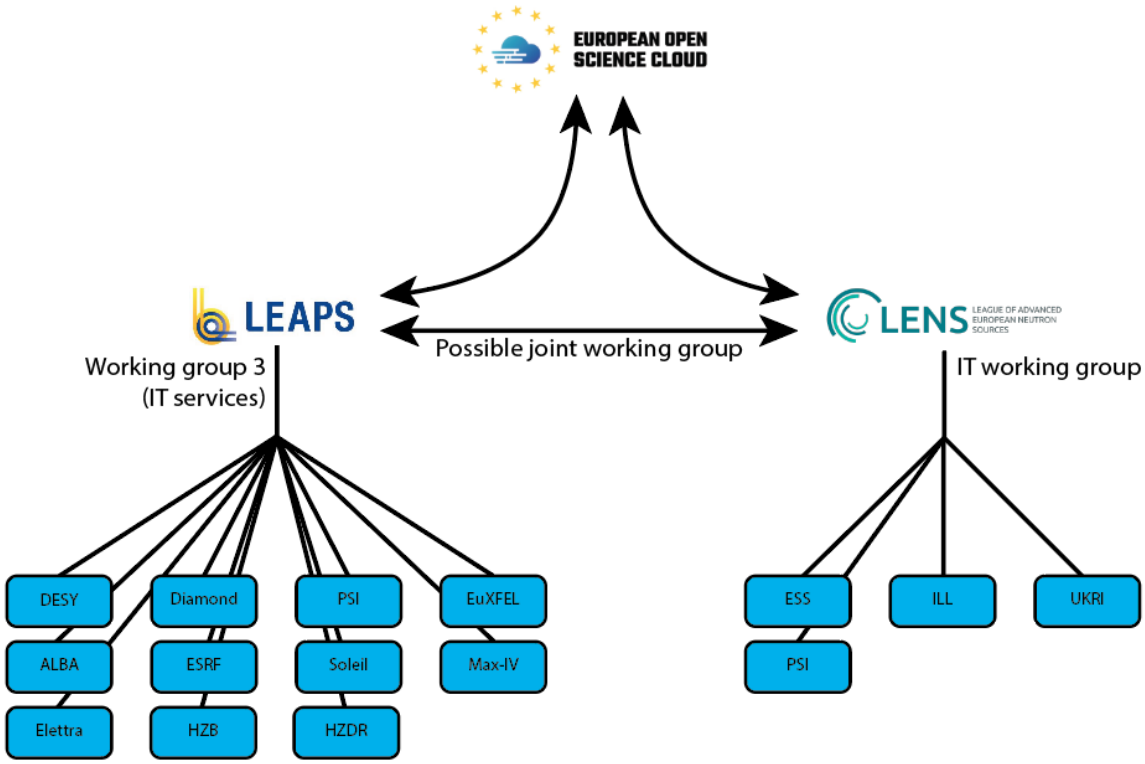
1. Activities within each of WP3 and WP4 are coordinated within ExPaNDS by the respective lead institutions (PSI and DESY, respectively).
2. Alignment to PaNOSC service development is maintained through sharing meeting participation of associated work packages WP3 and WP4 of PaNOSC and ExPaNDS.
3. Harmonised deployment of the PaN community (both NRIs and ESFRI institutions) interface within EOSC is achieved by deploying ExPaNDS services through PaN portal developed as a part of PaNOSC.

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

In practice all members of the ExPaNDS WP3 and WP4 work packages are welcome at the corresponding PaNOSC technical meetings, and vice versa. Indeed, among other things we share topical mailing lists and coordinate technical workshops on common topics.

2.3. Sustainable long-term coordinating framework

Sustainable coordination of European photon and neutron facilities after the ExPaNDS and PaNOSC projects have completed requires a different organisational framework. Here, we coordinate with the European initiatives LEAPS[3] (League of European Accelerator Photon Sources) and LENS[4] (League of European Neutron Sources), both of which are umbrella organisations for the European photon and neutron sources respectively. Both LEAPS and LENS have dedicated IT working groups which have a cross-institution coordinating layer aimed at exploiting synergies between the research infrastructures of both NRI and ESFRI facilities. In addition to its other roles between research infrastructures, this existing organisational structure is ideally suited to assuming the role of coordinating research infrastructures with respect to EOSC after the ExPaNDS and PaNOSC projects have ended.



[3] https://leaps-initiative.eu
[4] https://www.lens-initiative.org

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

Of course, it remains up to LEAPS/LENS and the member infrastructures to determine the manner in which support is provided and its scope. However steps which could be taken include:

- Integration of ExPaNDS and PaNOSC work package leaders at each institution into the respective LEAPS working groups
- Enhancement of LEAPS working group activities to with respect to facility interoperability and exploitation of horizontal infrastructures for the respective user communities
- LEAPS and LENS IT working groups should align themselves with respect to each other as far as possible
- Active engagement of EOSC with LEAPS IT within wider EU research infrastructure frameworks

We have taken steps to embed this future operational framework within both ExPaNDS and PaNOSC projects. Already, both ExPaNDS and PaNOSC project leaders are members of the respective LEAPS IT working group. Additionally, many of the ExPaNDS facility representatives are already in the respective LEAPS and LENS working groups. After the end of the project, the respective LEAPS and LENS working groups could become conduits for maintaining alignment and coordination of all photon and neutron facilities' EOSC services. However, as noted above it remains up to LEAPS/LENS and the member infrastructures to determine the manner in which support is provided and its scope.

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

3. Common rules and practices

This section describes the rules, practices and principles to be adopted by ExPaNDS participants with regard to integrating existing data analysis services into EOSC, addressing the description of "Common rules and best practices for implementing the national RI's analysis services, and future data analysis services, within the EOSC" part of D4.1.

3.1. Existing analysis services

ExPaNDS aims to integrate existing data analysis services into EOSC via PaNOSC services. Existing analysis services are those services existing at the facilities at the start of the project. A brief overview of analysis and data transfer services* at each facility are:

Facility	Data analysis service	Data transfer service
DESY	FastX	Globus online
PSI	NoMachine + JupyterHub	Globus online
Diamond	NoMachine + JupyterHub	Globus Online
UKRI-STFC	VMs + JupyterHub	Globus Online (separate to analysis services)
MAX IV	JupyterHub, ThinLinc HPC	SFTP + Globus online
Alba	VMs	SFTP
Soleil	NoMachine	SFTP + Globus (in test)
HZDR	NoMachine + JupyterHub	Globus online
Elettra		

*As of start of the project

There is a notable lack of harmonisation in data analysis services between facilities. Existing services are diverse in nature due to already being deployed by the different facilities by separate IT groups. As a consequence the data centres at each facility are completely different, including the software stacks and underlying technologies used.

The diversity of existing services at the infrastructures requires careful consideration as a part of the EOSC integration plan. Services already existing and in use at the facilities are significantly more difficult to modify than new services because they are already in use, creating inevitable inertia to leave them as-is. This inertia is increased because the existing services are relied on by a large existing user community and considerable effort would be required by users to adapt to any new system. The goal of PaN facility users is to obtain scientific results and from the user perspective effort spent on adapting to new IT infrastructure is by and large seen as a cost with no direct benefit to science output. There is

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

thus a considerable practical resistance to imposing changes to existing infrastructure at the facilities.

At the same time, we wish to integrate core services rather than test instances on isolated islands which will disappear at the end of the project in order to ensure sustainability after the project ends. In practice, replacing an existing, operational service with a harmonised one will first require parallel operation of both. Unlike the particle physics community, which has managed to standardise on the Worldwide LHC Computing Grid architecture at each facility, no such standardisation exists between photon science communities and it is unreasonable to expect to achieve complete harmonisation through a horizontal infrastructure such as the WLCG within the scope of ExPaNDS. Rather, the purpose of ExPaNDS is to expose the diverse existing data analysis services through the PaNOSC portal according to the guidelines described in this document. We therefore come to the practical realisation that in order to interface to core infrastructure it will be necessary to deal with this heterogeneity of systems.

3.2. EOSC use case for ExPaNDS analysis services

Recognising that the diversity of infrastructures at PaN facilities makes the deployment of a horizontal PaN infrastructure directly within EOSC challenging, datasets will be exposed within EOSC using the PaN portal described above.

Consider, for example, any one of the sample data sets and associated software resources described in Deliverable 4.2 (DOI: 10.5281/zenodo.4558708). In order to access a dataset and data analysis service from https://eosc-portal.eu/ an EOSC user would:

1. Navigate to the unified PaN portal deployed by EOSC-hub in partnership with PaNOSC;
2. Search for or otherwise find the dataset within the PaN portal. Here, an anonymous user would only see open datasets, while a logged in user may see additional datasets to which their login credentials permit access;
3. After selecting the dataset, resources with access to the data and associated compatible workflows could be selected. Here, it is important to note that not all workflows may be able to be executed at all participating facilities due to the heterogeneity of infrastructure;
4. The user will be directed from the portal to a facility with appropriate access to the data set and associated compute resources;
5. Allocation of compute and/or storage resources in addition to access to any closed data sets may require additional authentication to gain access to resources, ideally through a single-sign-on AAI;
6. Finally, licencing compliance including differentiation of commercial/academic use and compliance with licence terms of the software required to execute the workflow including software re-distribution needs to be ensured.

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

3.3. Implementation

This section describes the common rules and best practices for integrating existing NRI data analysis services into the EOSC. This section concentrates on WP4 related tasks, tasks in other work packages are detailed elsewhere.

Broadly speaking a data analysis service requires:

1. Data to be analysed;
2. Software with which to perform the required analysis; and
3. Infrastructure on which to perform the analysis

The analysis workflow using the analysis service needs to be described. In many cases this could already be well known in the community or already documented elsewhere.

Best practice would require:

1. Data to be analysed should be findable, accessible, and in a common format, ie: abide by FAIR principles;
2. Software with which to perform the analysis should be portable across analysis facilities, ideally in facility-independent containers, notebooks or virtual machines; and
3. Infrastructure on which to perform the analysis should be accessible from outside of the facility and should be able to run the generic analysis software described in point 2 with access to data described in point 1.

In this context the PaN portal directs users to a facility where data and software can be bought together on accessible infrastructure. As already described above, not all workflows may be able to be executed everywhere due to heterogeneity of resource infrastructure. Since datasets are typically large, in many cases this will in effect amount to redirection to compute and software resources at the facility hosting the data. Alternatively, data could be copied to where compute and software resources are available for the particular EOSC user.

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

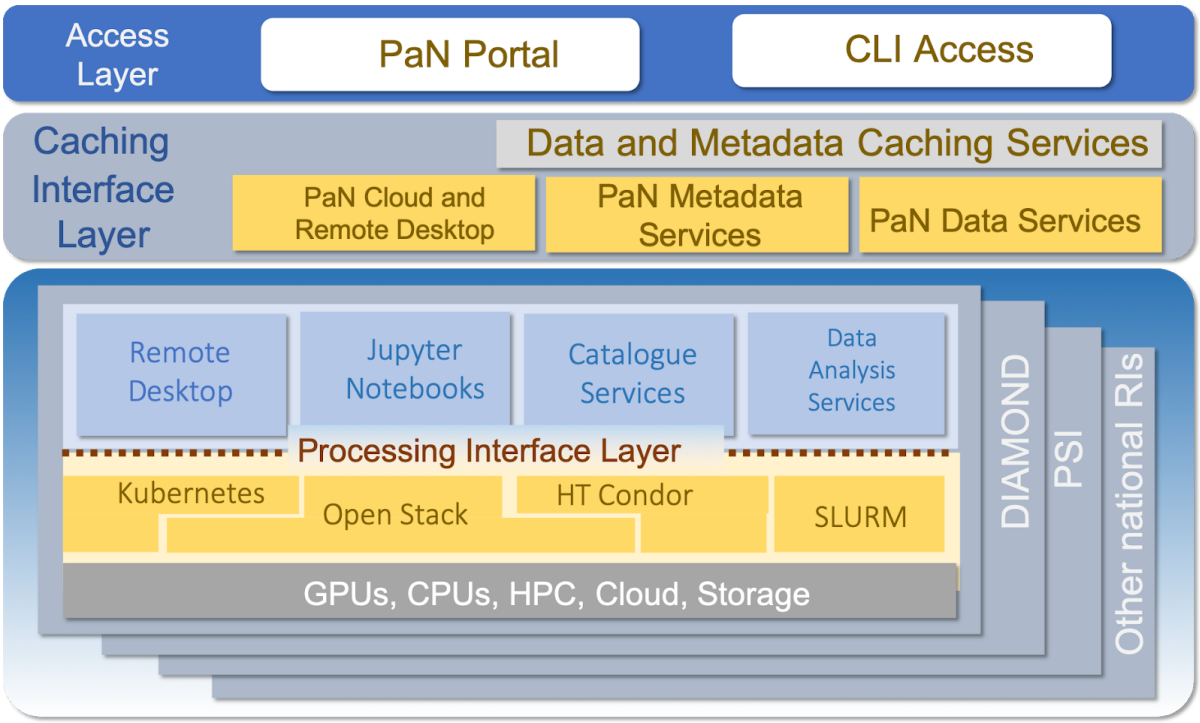


Fig. 1 - ExPaNDS architecture (from 10.5281/zenodo.3697704)

3.3.1. Access to data and FAIR data management

3.3.1.1. Findability

Datasets will be searchable using interfaces developed in WP3, and WP3 intends that each ExPaNDS facility should implement this API against their internal data catalogue. In addition to the usual FAIR metadata requirements, information necessary for selecting analysis services should be included such as data formats, recommended software for raw data sets, and software packages used for derived datasets.

Metadata catalogues should issue persistent identifiers (PIDs) for each data set so that published results can be traced back to raw data. Automatic generation of DOIs for all datasets may not be implemented at all facilities since one requirement for issuing a DOI is that the data be retained permanently and not deleted, which may not be possible for all data sets. Finding the appropriate PID infrastructure for PaN data sets is undertaken as part of WP2. Deployment of search within the PaN portal will be undertaken as a part of WP3[5].

3.3.1.2. Access and authentication to data

Federated AAI such as UmbrellaID will be used as far as practical for access to ExPaNDS services. Open (public) datasets should be findable without local login, and ideally downloadable without further authentication. However, data policies at the institute may require a local login in order to obtain access to open data sets, for example to agree to the facility data policy or to obtain permission to restage data from archive media. In the case of closed data sets a local login may be required for authentication or use of a federated AAI

[5] https://doi.org/10.5281/zenodo.4146819

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

against which access privileges to the dataset can be verified. Integration of a federated AAI infrastructure at all facilities will be done on a best effort basis as access policies at the facilities may require use of a local login for access authentication.

3.3.1.3. Interoperability and reusability

Common PaN data formats such as NeXus[6], and self-describing container formats such as HDF5, will be used as far as possible for data storage. Datasets should be accompanied by data licencing requirements, for example CC0 (or perhaps CC-BY, but CC0 is recommended for data) otherwise you run into licencing issues. RIs may restrict data, for legitimate reasons (e.g. embargos, sensitivity), and such restrictions need to be respected.

3.3.2. Analysis pipelines (software ecosystem)

3.3.2.1. Software catalogue

Software must be findable and ideally be usable at multiple facilities. Software used as a part of WP4 will be logged into the PaN data software catalogue[7]. Additional information required for ExPaNDS and the PaNOSC portal are guidelines as to recommended compute and storage needs for analysis to help steer selection of compute resources, as well as sufficient information to ensure compliance with software and data licences (see below). Software versions for derived data should be traceable.

3.3.2.2. Source of portable software (containers or notebooks)

Software will be packaged into containers so as to be portable between facilities - as far as practically possible within the constraints of heterogeneous infrastructure and different software stacks. A container repository will be developed to facilitate sharing of containers (subject to licence compliance). Websites at each facility should give instructions as to how to obtain the necessary software.

3.3.2.3. Testing framework

Containerised workflows and notebooks will be tested at each facility to check whether they are compatible (testing framework) and compatibility information posted to the services catalogue of WP3, which in turn will be implemented in the PaN portal. For example, the analysis service could be verified using a CI/CD test running against a test data set for validation at one institution, and CI tests performed at other institutions in the horizontal infrastructure. Development of the testing framework forms a separate deliverable. A minimum requirement is that software should run where the data is stored.

3.3.2.4. Software Licensing requirements

Proper attention to software licencing is required. For example restrictions on commercial use or the need for end users to agree to software licenses. This is the case with MX software such as Phenix and CCP4 for which the end user is required to agree to the software terms, and where academic licences are free but commercial use requires payment

[6] https://www.nexusformat.org/
[7] https://software.pan-data.eu

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

of a fee. Commercial use of an academic licence is a violation of the licence terms and safeguards need to be implemented to ensure software can only be used in accordance with the terms of it's licence.

Redistribution of the containers ('replication', aka 'copying'), and the requirement under some licences to distribute source code with executables (eg: GPLv3) need to be considered. Some software ceases to run after a fixed time unless the licence is renewed (XDS, 6 months). Other software is subject to patents (eg: ptychography) and unable to be openly distributed beyond personally written code without exposing the institution to potential patent violation liability. Alternatively, software may be written and tested against a specific library, for example an Anaconda Python installation, which can not be copied and redistributed in a container without violating software licence terms. There are numerous other examples and many individual variations on this theme.

Within the context of EOSC, and more generally sharing containers between facilities where we enable running code across multiple facilities through containers, we have to be careful not to facilitate licence breaches and thereby inadvertently expose institutions to liability for such licence breaches. In practice, it may be that it is only possible to support certain licences on the EOSC or EOSC-connected open science platforms. This issue needs careful consideration to make sure we don't inadvertently facilitate the violation of licence terms and conditions in some way.

Maintaining and ensuring compliance with software licence terms is essential in order for EOSC services to provide an open science role model to the community. As such ensuring software licence compliance forms an important part of best practices. Here, we need to install methods to ensure:

- Software and resources for academic versus commercial must be clearly separated and that academic licences can not be used for commercial purposes.
- Software in containers, including dependencies, can only be replicated or copied to other facilities in accordance with the software licence terms.
- Software dependencies not able to be packaged in portable containers due to licencing issues must be clearly identified so that horizontal infrastructures providing the necessary software dependencies can be selected.
- Distribution of software in containers must be certified free of patent encumbrances so as to protect other institutions from inadvertently replicating patented software.
- The end user must agree to software licence conditions where this is required by the software licence, and this must be logged.
- Licence compliance requirements may limit the range of analysis services that can be deployed on EOSC to certain classes of software licence.

There is in the EOSC community an increased demand for technical solutions to address these legal aspects intrinsic to FAIR and open source. We will benefit here from the work carried out in the other INFRA-EOSC-5 call projects, e.g. NI4OS's Licence Clearance Tool[8].

[8] https://wiki.ni4os.eu/index.php/License_Clearance_Tool_-_Description_and_Documentation

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

3.3.2.5. Attribution and traceability

In order to properly cite the software chain in publications, methods for generating the 'methods statement' for papers from the repository will be instantiated. This will, where possible, provide references to software publications. We will also investigate providing persistent identifier references and/or repository commit information for software, including version histories.

3.3.3. Access to infrastructure

3.3.3.1. Locating suitable infrastructure

The first assumption is that analysis pipelines should run at the facility where the data is located. Although the goal is to have software portable across infrastructures using containers, it is possible under some circumstances this may not be possible eg: due to need for specific hardware (eg: certain CPU or GPU types) or facility-specific frameworks that are not inherently portable. Thus, we require the minimum requirement software should run where the data is stored with the goal of having portable software if possible.

In order to identify appropriate third party resources for running analysis pipelines, additional information should be available as to recommended compute and storage needs for analysis to help steer selection of compute resources. Additional information will be required to ensure licence compliance and that appropriate software dependencies are available (see Licencing section above). This is necessary to run on horizontal infrastructures or other generic compute infrastructure such as AWS.

3.3.3.2. Access and authentication to compute infrastructure

Federated AAI such as UmbrellaID will be used as far as practical for access to ExPaNDS services. In cases where access to core infrastructure is required, the security policies of various facilities may require use of a local login for access to data and/or computing resources. Further details on AAI implementation can be found in the appendix and will be implemented where practical.

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

3.3.4. Integration to PaN portal

The main focus of ExPaNDS WP4 is the coordination, adaption, and alignment of existing data analysis services at national RIs with the new EOSC services available via a common PaN services portal. PaNOSC has been developing the Common Portal for Data Analysis Services to facilitate starting a data analysis session after a dataset of interest has been collected. The portal aims to provide access to both remote desktop environments and Jupyter Notebooks, enabling users to remotely analyse data from PaN facilities. Within WP4, we do not intend to repeat this effort. NRI data analysis services will be integrated into the EOSC using the PaN portal developed by PaNOSC. Services should also be accessible without use of the PaN portal such as directly through the user facility.

As already mentioned above, alignment with the common PaN portal under development by PaNOSC maintains alignment of the unified PaN community with EOSC, addressing the main focus of this work package as the coordination, adaption, and alignment of existing data analysis services at national RIs with the new EOSC services available via the EOSC Service Catalogue and EOSC Portal and integration into the EOSC using PaNOSC services.

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

Appendix: A longer term perspective on integration of NRIs into EOSC

This appendix contains a lengthier description of the operation of horizontal infrastructures within EOSC. This section describes how the PaN integration with EOSC may evolve in the future and goes beyond the rules and best practices described in section 3. We therefore keep the appendix separate from the guidelines outlined above as it refers to tasks that, while important, are beyond the scope of deliverables in ExPaNDS WP4.

Harmonise access to PaN facilities

As already mentioned in the ExPaNDS architecture[9] and earlier in the document, one of the key points to implement national RI's analysis services is to have in place an Authentication and Authorization Infrastructure (AAI) which enables single sign-on access for users and granting access to services. To be compliant with EOSC this AAI has to follow the AARC Blueprint Architecture (BPA) 2019[10] and the AARC guidelines[11] to guarantee interoperability with other EOSC AAIs. From an historical point of view, the Photon and Neutron (PaN) community has a long-running collaboration with GEANT that is currently operating the PaN AAI community proxy based on UmbrellaID.

Following the AARC-BPA-2019 architecture recommendations, in order to support the interoperability of the PaN Community AAI with the R/e-Infrastructure proxies in EOSC the high-level architecture described in Fig. 2 was proposed. From a technical perspective, Fig. 2 shows the initial setup where PaN users, using the Umbrella AAI provided by GÉANT eduTEAMS, should be able to access the data analytics service offered by EOSC (e.g. the EGI Notebook service), which is federated through the EGI Check-In Infrastructure Proxy using OpenID Connect.

[9] https://doi.org/10.5281/zenodo.3697704
[10] https://zenodo.org/record/3672785#.XlkiT2hKg2w
[11] https://aarc-project.eu/guidelines/

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

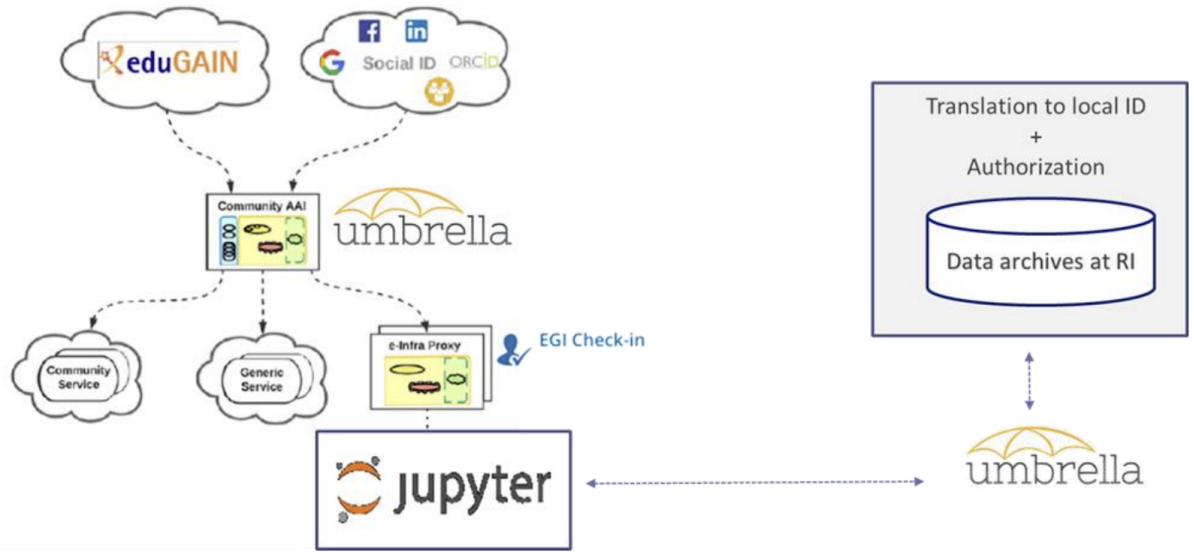


Fig. 2 - Access to compute service from different communities to analyse data.

In order to achieve the above, the following steps have to be taken:

- The UmbrellaID AAI, the proxy, should be added as a community AAI in EGI Check-In.
- Attribute release from Umbrella AAI to EGI Check-In with all the required attributes, so that the user does not have enter person information manually on the EGI Check-In.
- Alignment of the Acceptable Usage Policies (AUP) of UmbrellaID AAI and EGI Check-In with the WISE[12] Baseline[13] so that the user can accept the Umbrella AAI AUP once they register.
- Configuration of EGI Check-In to map Umbrella AAI entitlements to the access rights used internally in EGI.
- Enable the computational flows running on the EGI Compute Infrastructure to access data in the Data Archive service of the RIs.

While implementation across all ExPaNDS facilities is beyond the scope of ExPaNDS WP4, it does form a part of PaNOSC and will be further pursued through LEAPS and LENS working groups.

Enable remote data transfer and sharing between RI archives and EOSC remote sites

Data Transfer services are required in order to easily move data between the various Research Infrastructures and to external computing resources. Although data transfer services are technically outside the scope of ExPaNDS, data transfer services are needed to facilitate the transfer of data to user's home organizations after running experiments at the facilities and to allow data archival at remote locations. Those three different Data Transfers

[12] https://wise-community.org/
[13] https://wiki.geant.org/display/WISE/Baseline+Acceptable+Use+Policy+and+Conditions+of+Use

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

requirements have been analysed in PaNOSC WP6 and different piloting activities have started trying to address them, by leveraging solutions which are also available in EOSC.

The selection of the best technology to be used also depends on the volume of data sets to be processed. Relative small datasets can be easily transferred from one facility to a computing infrastructure and vice-versa over the network. For large data volumes it is more convenient either carry out the computation where the data is, or move the data to the compute resource. Which model is feasible will depend on a complex set of factors including available compute and data resources at facilities, bandwidth within and between facilities, compatibility of technologies for moving computation to data, and willingness of users to wait for data migration versus waiting for compute time. Resources are not provided for free and availability of compute and disk space to the specific user relative to cost and speed will be a major consideration in which model is selected.

One of the pilot documented in PaNOSC[14] describes the case where a scientist wants to access a data analysis service offered by another organisation to perform analysis on data sets obtained from one experiment at one of the PaN facilities. The computer resources, where the service lives, are distant from the RI archive holding the data.
First of all the typical storage archive available at a PaN facility stores data produced by the detectors of the scientific instruments. As data is the main output of the RI, this is a highly critical role when the instruments are running that quite often necessitate a protected network bandwidth to ensure the minimal performance required by the experiment. The central archive collects the data directly from the instruments, but also serves the access for user transfers and analysis through gateways to control that these accesses do not disturb the acquisition processes. The gateways also offer standard access protocol (NFS, SMB, FTP, rsync, etc). The users can access the services directly exposed by the gateways using their own RI credentials, these are the local IDs that are used by the authorisations mechanisms (ACLs).
For the data transfer pilot developed in PaNOSC, EGI DataHub[15] was used (based on the Onedata[16] technology), which offers a way to expose the RIs data using a locally deployed component (OneProvider) and implement custom mapping between local user accounts or credential on storage resources (e.g. POSIX user ID/group ID, LDAP DN, Ceph username, GlusterFS UID/GID, etc.) to AAI credentials. In the case the adoption of the UmbrellaID as Community Proxy for the project contributed to implementing the authentication and authorization mechanism and making sure that only authorized users from the PaNOSC (and ExPaNDS) facilities can access datasets under embargo. Fig. 3 shows the details of the pilot, which allows transparent data transfer, access and analysis from remote EOSC facilities. The EGI notebooks service has been used for this particular pilot as a data analysis service, but the same can be done when using Cloud VMs or HTC behind an instance of the PaNOSC portal as previously described.

[14] https://www.panosc.eu/wp-content/uploads/2020/12/D6.1_DataHub.pdf
[15] https://www.egi.eu/services/datahub/
[16] https://onedata.org/

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

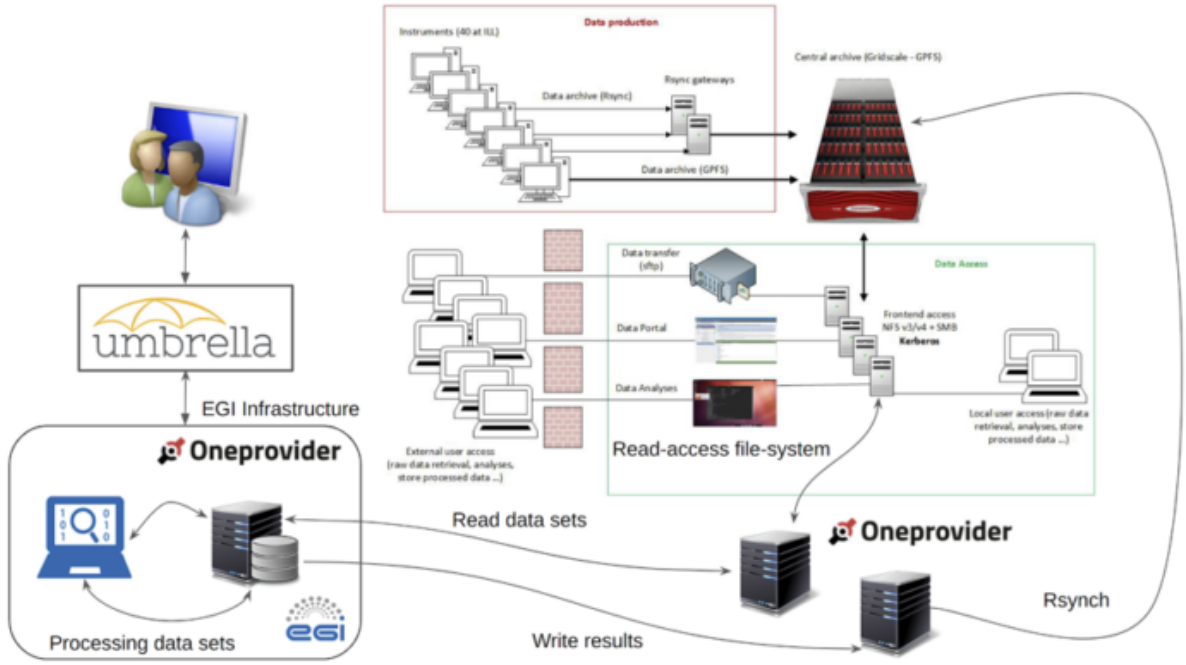


Fig. 3 - The DataHub PaNOSC Data transfer pilot

Possible other solutions for data transfer available in the context of EOSC could be also piloted for this use case, for instance the usage of the FTS[17] service, in junction with Rucio[18] to offer policy based data orchestration. This is the data management solution that has been adopted by the ESCAPE project[19]. 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). An example of the solution developed in ESCAPE (the ESCAPE Data Lake) is depicted in Fig. 4.

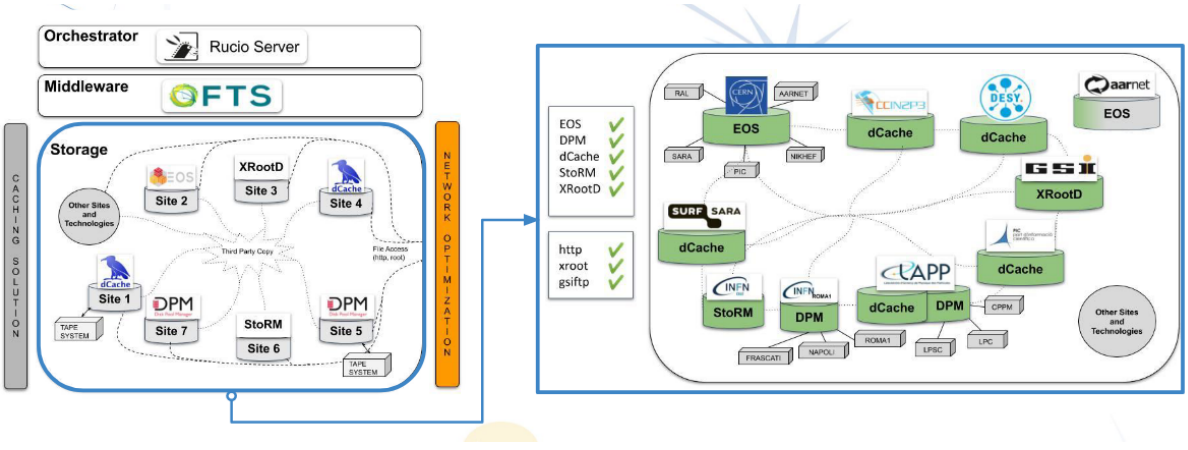


Fig. 4 - The ESCAPE Data Lake

[17] https://fts.web.cern.ch/fts/
[18] https://rucio.cern.ch/
[19] https://projectescape.eu/

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

Publishing services in the EOSC portal

The EOSC Portal is envisaged to become a key component of the European Open Science Cloud (EOSC) by providing an access point for services and resources for Europe's research sector. For this reason the EOSC Enhance project[20] is responsible to further support the development of new functionalities in order to boost the performance and the usage of this portal, in collaboration with the EOSC-hub project.

To drive the evolution and the development of new features of the EOSC portal and address the needs of the different stakeholders, EOSC Enhance started to collect technical requirements with a dedicated survey. The goal of the requirements gathering activities was to collect, analyse and prioritise functional and non-functional requirements from users, providers and European Open Science Cloud implementation projects to continuously improve the EOSC Portal functionalities and value proposition. The following stakeholders were invited to contribute and submit requirements to drive the evolution of the EOSC Portal:

- EOSC Portal users;
- EOSC Portal Resource providers;
- EOSC-related projects;
- EOSC Governance & Policy Makers.

The ExPaNDS project contributed to this process submitting requirements that were also reported in D1.5 - General Architecture description in relation to the EOSC services_v1[21].

The first release of the new EOSC Portal was rolled out in production in Sept. 2020. This important milestone was followed by a webinar series organized by the EOSC Enhance project to promote the new Portal release focussing on the features to target new providers and end users. Training materials and video recording to promote the new features of the EOSC portal are also published in the EOSC Portal web site[22].

For the service providers perspective, and in particular for the ExPaNDS providers, a new registration flow is now in place. Rather than on-boarding a service by completing a spread-sheet, a dedicated wizard is now available in the EOSC Portal to guide the service provider through the on-boarding steps. From a technical perspective, the registration of new service providers in the EOSC Portal is done in two different steps: we start with the registration of a service provider, providing a set of information to describe the service provider/organization offering the service and acting as main contact, website, level of maturity, physical location, etc, and we continue by adding resources to the provider. A quality check of the information submitted will be verified by the EOSC core team before to proceed through the next steps. More specifically, at this stage will be checked the appropriateness of the information provided and the legal entity of the organization that represents the provider. If the registration of the provider is approved it will be possible to describe resources. This step can be done either using APIs or web portal. The resources will be further checked by the core team before to allow the provider to self-publish the resources and make them available in the Marketplace and enable service ordering.

[20] https://www.eosc-portal.eu/enhance
[21] 10.5281/zenodo.3697704
[22] https://eosc-portal.eu/using-the-portal/tutorial-new-providers

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

Integration with Federation Services

As previously said, the EOSC-hub project, coordinated by EGI, plays a central role in the EOSC landscape as it established and runs the central services for EOSC, like the Portal, the AAI or the service onboarding team.
The project has delivered a "EOSC-hub project Integration Handbook"[23], where details for the integration with the current EOSC Federation services are given. Besides the already mentioned AAI and Portal, it's important to highlight the integration with:

- Availability and Reliability Monitoring
- Usage Accounting
- HelpDesk

PaNOSC WP6 is integrating these requirements in the development of the services published to EOSC: the PaN portal, PaN software catalogue and PaN learning platform. The contribution of ExPaNDS partners in maintaining these services during and after the projects is discussed in the PaNOSC WP6 meetings where ExPaNDS is represented and within the sustainability WPs of both projects.

Availability and Reliability Monitoring

Monitoring is the key service needed to gain insights into an infrastructure. It needs to be continuous and on-demand to quickly detect, correlate, and analyze data for a fast reaction to anomalous behavior. The challenge of this type of monitoring is how to quickly identify and correlate problems before they affect end-users and ultimately the productivity of their organizations. EOSC-hub provides a service monitoring service based on the ARGO system. This ARGO Service collects status results from one or more monitoring engine(es) and delivers status results and/or monthly availability (A) and reliability (R) results of distributed services.

Usage Accounting

The EOSC Accounting service can collect, store, aggregate, and display usage information about the following types of services:

- High Throughput Compute
- Infrastructure-as-a-Service cloud virtual machines
- Storage space providers
- Data set providers

Accounting information is gathered from the service by probes and sensors according to certain data formats. Probes and sensors are deployed locally at the service providers. Data is forwarded from the sensors into a central Accounting Repository where those data are processed to generate various summaries and views for display in the Accounting Portal.

[23] https://zenodo.org/record/3826907#.YCzQQWMo--q

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

HelpDesk

The EOSC-hub Helpdesk is the entry point and ticketing system/request tracker for issues concerning EOSC services. New service providers of EOSC can integrate into the Helpdesk and this results in:

- a corresponding support topic listed on the Helpdesk user interface, for users to ask questions or raise issues directly to the provider
- the provider support team to receive notifications about tickets that are assigned to this topic by the users, or by the ticket handler team of EOSC-hub.

The Helpdesk therefore serves two groups, offering features to end users and to the provider support team.

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