General Architecture description in relation to the EOSC services

Document Control Information

Settings	Value
Document Identifier:	D1.5
Project Title:	ExPaNDS
Work Package:	WP1
Document Authors:	Diego Scardaci (EGI), Daniel Salvat (ALBA), Anton Barty (DESY), Alun William Ashton (PSI), Patrick Fuhrmann (DESY), Sophie Servan (DESY)
Responsible Partner:	EGI
Doc. Issue:	1
Dissemination level:	Public
Date:	05/03/2020

Abstract

The ExPaNDS project aims at deploying into EOSC data catalogues and data analysis services. This document describes the proposed architecture for these services and how they are planned to integrate into EOSC.

Licence

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 deliverable is describing the envisioned architecture of the ExPaNDS project, focusing on the technical aspects only. It is composed of contributions by the technical work packages on catalogues and analysis as well as from EGI, our European infrastructure partner.

The document enumerates a selection of core services required to maintain the envisioned PaN overall system, essentially AAI, accounting and monitoring as well as scientific services provided by the particular ExPaNDS PaN facilities, like data catalogues and analysis pipelines. Similarly, software catalogues are an important prerequisite of our analysis pipelines. They are already provided by the scientific communities[1] and therefore not further described in this document. The first part of the document elaborates on our plan to abstract the data and analysis services with common APIs in order to make them available to portals and processing pipelines without knowing where the actual facility data, analysis stacks and ICT infrastructure resides. Agreed interfaces, e.g. APIs not only allow to abstract and federate underlying services, but they enable to run a variety of portals for different target users, like the shared PaN portal and the EOSC portal as a minimum. Moreover, as this is one of ExPaNDS high level objectives, the document describes the relationship between services provided by ExPaNDS RIs and the EOSC as well as their interdependencies. Furthermore we are elaborating on synergies between our project and the PaNOSC ESFRI cluster project[2] with which we naturally share a large number of objectives.

1. Introduction

Besides improving the establishment of FAIR principles in data, meta data and analysis pipelines of national PaN facilities, the major technical aspect of ExPaNDS is to unify the access to core and scientific ICT resources for PaN communities and ideally for associated or interested non PaN scientists.

During the first two quarters of the project we have identified what we will do to move forward and in particular where synergies with other projects and initiatives and the knowledge of already available, well established services can get us ahead within the scope of our funding.

One example, which will be further elaborated in this document, is the inevitable Authentication and Authorisation Infrastructure. As this is of common interest for almost all distributed scientific endeavours, large efforts have been made in the past to provide tools as well as legal and technical guidelines to setup an AAI infrastructure compatible with

[1] https://software.pan-data.eu/

[2] PaNOSC project website: https://www.panosc.eu/

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

European and national laws and which can be connected to cross scientific identity and services providers. Similarly important, if not essential, is the alignment of ExPaNDS activities with the PaNOSC project as outlined in our Grant Agreement (GA). Putting some deviations aside; PaNOSC is targeting almost identical objectives for those facilities, being members of the ESFRI group, in the PaN domain. Specifically, and further detailed in the remain of this document, ExPaNDS will work on the unification and standardisation of the three pillars of distributed scientific computing:

- Access to data, covering the provisioning of common standards in accessing the data itself as well as its meta data catalogues.
- Access to analysis and processing pipelines, easily portable between facilities as well as support for traditional access to monolithic pipelines through remote desktops to batch facilities operated on standard or HPC hardware.
- Access to the actual computing infrastructures which could be traditional batch systems, including HPC clusters but increasingly modern Cloud Management Frameworks, like OpenStack, container based systems, like Kubernetes or even server-less systems, like OpenWhisk.

For all three areas, ExPANDS envisions standardised or at least commonly agreed APIs which will facilitate and foster the use of PaN ICT infrastructures and will allow scientists to focus on their science and with that will make access to computing and data more efficient and FAIR.

2. Description of the ExPaNDS architecture

The ExPaNDS architecture is designed to allow seamless connection to catalogue and analysis services hosted either at participant facilities or elsewhere through a common front-end portal, thereby allowing transparent access to services for users. Essential elements of this are:

- A catalogue of available data sets and their location;
- A catalogue of data services available at host institutions;
- Correlation of which datasets and analysis services are available together.

In designing the architecture, the following properties are desirable but may not be universally available at all host institutions depending on local configuration and policies:

- Single sign-on authentication;
- Matching of datasets to compatible or suggested analysis services (depends on this metadata being available and searchable);
- Display only data and/or services available to a particular user (not all services or data are available everywhere so display only those which the user can use).

The following schema represents the proposed architecture to achieve the above goals for the ExPaNDS project.

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

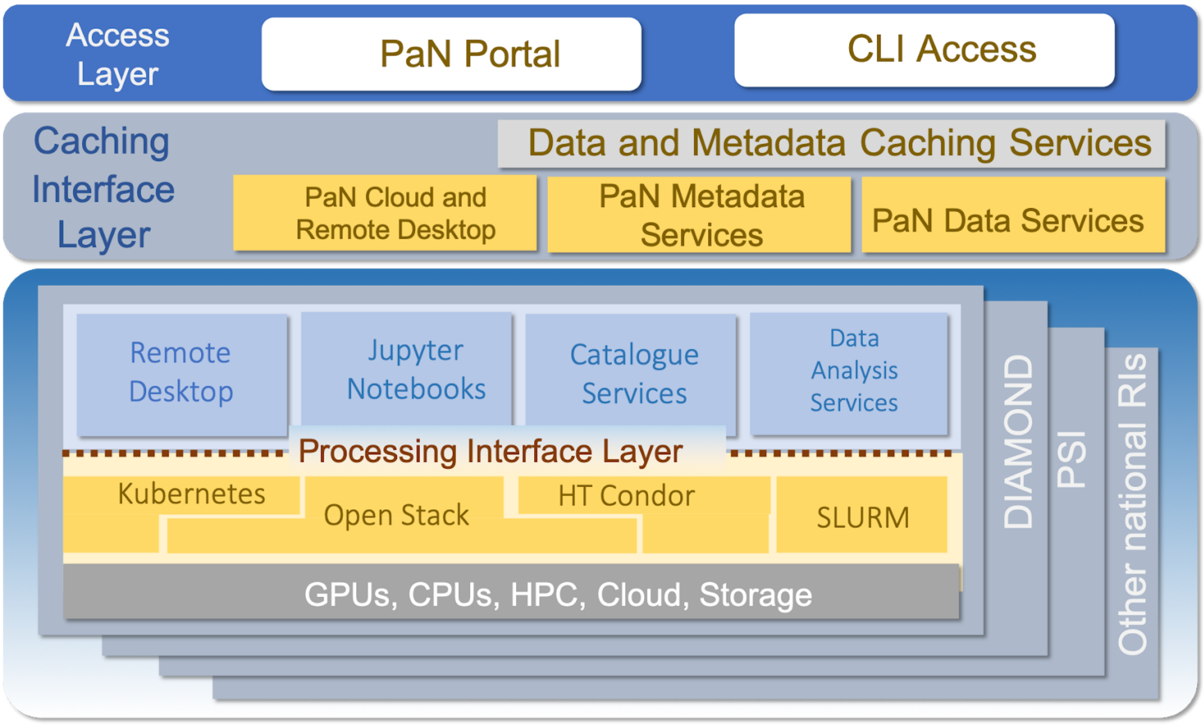

Fig. 1: ExPaNDS architecture - overview

This architecture proposes integrating services and APIs to the existing infrastructures, taking advantage of the commonalities detected on one side, and pursuing synergies and economy of scale on the other.

The architecture is composed of independent modules, commonly referred to as microservices, linked by APIs to enable maximum flexibility:

- **PaN Portal:** where the PaN users will connect to search data and request for analysis services;
- **PaN Data Catalogue Services:** Metadata search services for locating data sets;
- **PaN Compute Services:** Data analysis services;
- **AAI:** Authentication and Authorization Services.

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

2.1. PaN Portal

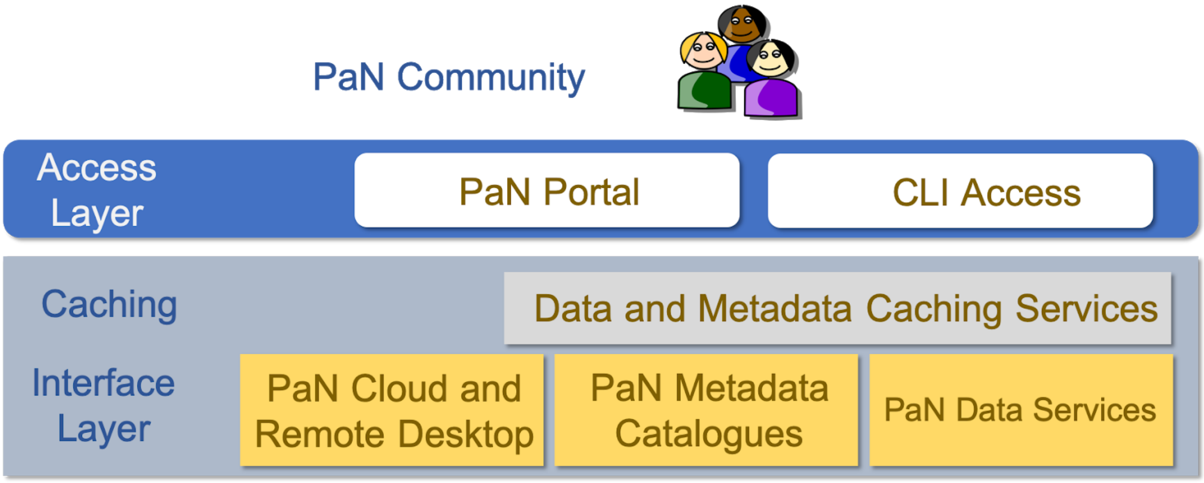

Fig. 2: ExPaNDS architecture - PaN Portal

The main portal is the entry point for all PaN users to connect to either search services or data analysis services. The portal will be common between PaNOSC and ExPaNDS so the community can reach all PaN resources through the same interface. Once authenticated, users can search for specific datasets according to metadata, for example by samples or a set of common and previously accepted list of metadata parameter options. This list of metadata parameters structure will be commonly agreed by all participants of the ExPaNDS and PaNOSC project and, on a first iteration, there will only be a selected group of ontologies having their metadata parameters defined, within the *Work Package 3 - EOSC data catalogue services for EU Photon and Neutron national RIs*.

Datasets and samples will be "exposed" for searching under the recommendations and best practices provided by the *Work Package 2 - Enabling FAIR data for EU Photon and Neutron nation RIs*.

The data analysis services to be provided will be selected by the *Work Package 4 - EOSC data analysis services for EU Photon and Neutron national RIs* upon their potential impact for the PaN community and their level of readiness. They will then be adapted to comply with the ontologies, APIs and standards developed by WP3. Redirection from the ExPaNDS EOSC hub to local services is handled by WP3 and WP4 as it concerns catalogue as well as analysis services.

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

2.2. PaN Data Catalogue Services (Search)

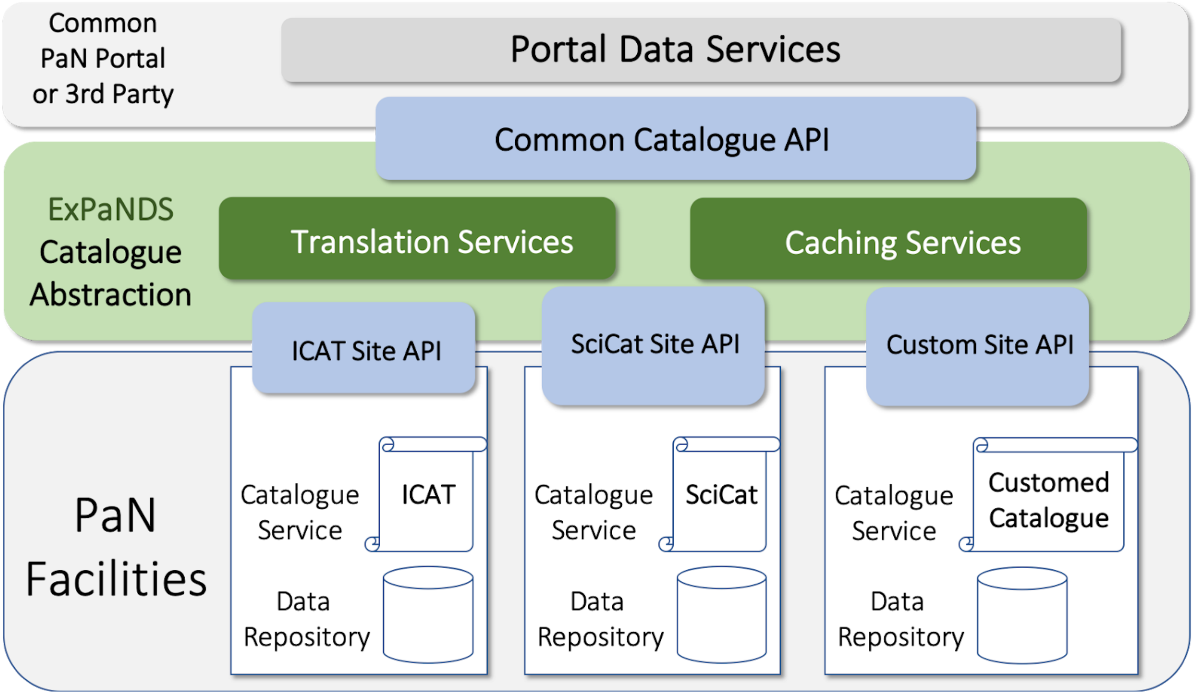

Fig. 3: ExPaNDS architecture - Catalogues services

The search action will trigger a request to search the PaN portal database, containing a cached copy of relevant metadata across all facilities. The result will be a list of aggregated information on available analysis services for each contributing facility. Optionally, if cached information is not available, out of date or if the portal user decides so, the search engine of the PaN portal can search databases of online service providers directly, through a well defined API. The PaN metadata services APIs will be developed to make existing catalogues compatible with the common portal or with other, third party portals, like the EOSC portal.

Once services, like ICAT, SciCat or others are implemented, they will be integrated by the contributing facilities, making the endpoints of the search available.

Assuming services comply with the suggested standards, the number of those endpoints is only limited by the scalability of the selected framework.
By sharing documents and meetings between ExPaNDS and PaNOSC we are spending significant efforts into making APIs between both projects compatible and coordinating API integration work.

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

2.3. PaN Compute Services (Data Analysis)

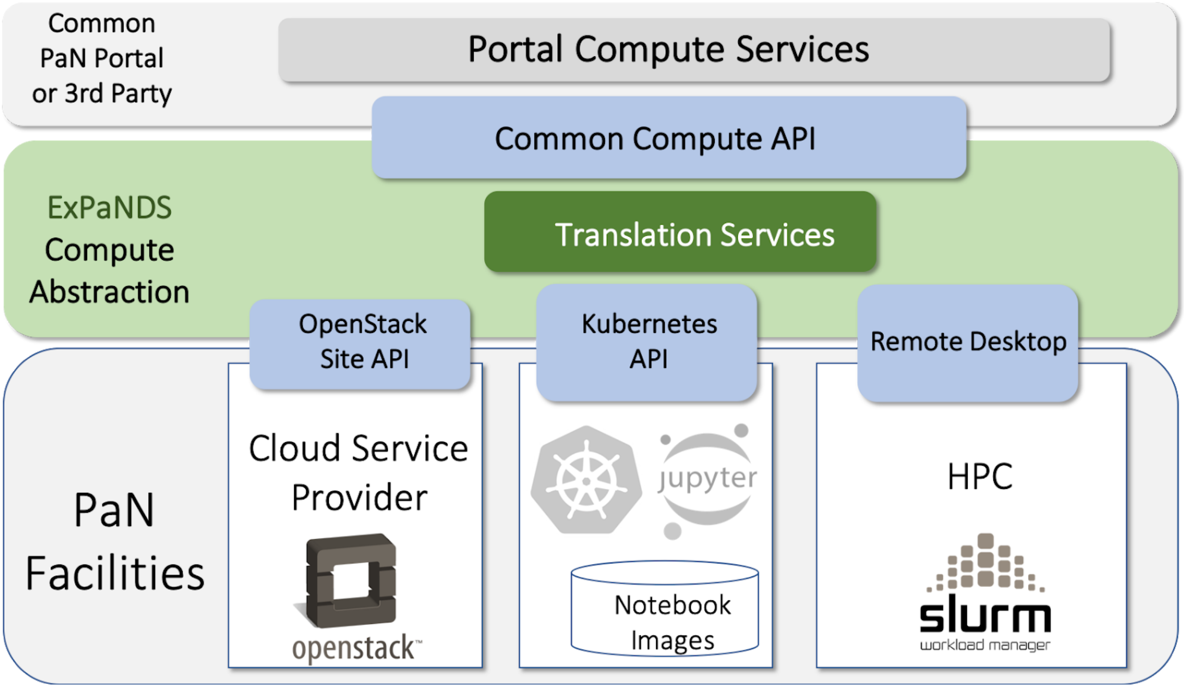

Fig. 4: ExPaNDS architecture - Data analysis services

The proposed Cloud Services architecture covers two different technologies: Virtual Machines and Containers.

Users select datasets to work with and open a new connection to the cloud services of the facility hosting those datasets. Depending on the type of service, interfaces provided by the PaN portal to each of the facilities will offer the best strategy to deliver the underlying service. At the time being, this can either be a remote desktop, giving access to complex applications, running in VMs within Cloud Management Frameworks (OpenStack) or, using more modern approaches, offering “Jupyter Notebooks” through e.g. containers, orchestrated by Kubernetes or similar technologies.

We also foresee to interface HPC resources required for large-scale data analysis or if particular hardware setups are needed, like access to GP-GPUs or low latency networks. Although attempts are made to operate HPC resources with Cloud technologies, we will make those resources primarily available through remote desktop access as already mentioned in the section above. As HPC clusters are mainly operating through job control systems such as the “Slurm workload manager” or the IBM “Load Platform Sharing Facility (LSF)”, ExPaNDS will offer interfaces to those technologies too. Providing this option is required as significant effort is needed to port commonly used analysis software in the PaN community to cloud like infrastructures. The provision of an HPC friendly interface will enable the availability of more analysis options albeit with a less sophisticated cloud services model.

Similar to search APIs, service APIs are coordinated between ExPaNDS and PaNOSC.

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

2.4. Authentication and Authorization infrastructure

The Authentication and Authorisation infrastructure (AAI) is one of the core components of a technical architecture composed of distributed services and users. Prerequisites imposed by European and national laws, policies of our project partners, and the fact that we intend to closely collaborate with the PaNOSC ESFRI cluster as well as with other players in the EOSC, lead us to believe that the only way to move forward is to comply with commonly accepted standards as described in the AARC blueprint document. Another advantage of following the AARC recommendations, besides the benefit of technical standardisation, is the fact that the blueprint provides templates for the necessary legal policies to operate identity provides, services and proxies. The legal aspects of imposing cross-facility authentication are not to be underestimated, further motivating the use of an existing standard.

Although currently dominated by the use of a full mesh SAML federation (UmbrellaID[3]), we will investigate connecting our services to Identity proxies, providing advantages, as token translation between OpenID Connect and SAML, authorization based on identities as well as on group memberships and merging of identities, as often scientists have assignments at different RIs. Moreover, the use of identity proxies would allow us to connect to an EOSC infrastructure in the future. Options, besides others, are the GEANT eduTEAMS system[4], the EGI Check-in service or the INDIGO IAM, provided by INFN CNAF.

Independently of the AAI we are going to use, users must have at least one trusted identity provider issuing OpenID Connect or SAML credentials. Furthermore, it may require accounts to be created at the facilities, hosting data and data analysis services as a separate step. **This guarantees that the service-hosting facility keeps complete control over the scientists using their services.** In general authorisation can be based on individual credentials or group membership.

[3] www.umbrellaid.org/

[4] GEANT eduTEAMS https://www.geant.org/Services/Trust_identity_and_security/Pages/eduTEAMS.aspx

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

3. Relationship of the ExPaNDS architecture with EOSC

One of the main ExPaNDS objectives is enabling the EOSC to the whole Photon and Neutron Community helping the very diverse scientific user groups to benefit from the EOSC to boost scientific outputs.

ExPaNDS will enhance services from the Photon and Neutron Community adopting EOSC best practices and standards facilitating the interoperability with the horizontal services offered via the EOSC (AAI, Helpdesk, computing and storage resources etc.) and integrating them when their adoption is considered useful and convenient (e.g. data transfer services, computing or storage facilities to scale-up, new analytics, etc.). ExPaNDS will also contribute to the selection and the definition of EOSC standard interfaces leveraging the experiences of its community on dealing with large distributed IT infrastructures and data sources.

Furthermore, ExPaNDS is working to enrich the EOSC offer on data management services and fostering the “Open Data” paradigm for the benefit of the wider EOSC community. A similarly important objective of the project is to increase the number of scientists which can rely on an efficient and interoperable European Photon and Neutron ICT infrastructure. Scientists working on several different scientific disciplines (biology, materials, chemistry, technology, nuclear physics, pharmacology, high-energy physics, cultural heritage …) could take advantage of these facilities.

ExPaNDS will also work on harmonising the services offered by the national RIs. National RI FAIR data will be made accessible and shareable across broadening user communities by ensuring RI’s data catalogues are conform to common EOSC standards and APIs and made available within the PaN portal and, if need be, the core EOSC Portal[5].

3.1. ExPaNDS services for the EOSC

The ExPaNDS services for the European Photon and Neutron Community will leverage the distributed architecture depicted in section 2 and will be offered to the wider EOSC community through the PaN portal.

Although the PaN portal will be the main point of access for these services, the project will evaluate the publication of its services in the EOSC Portal to facilitate and promote their uptake in EOSC.

ExPaNDS will closely follow the activities related to the design and evolution of the EOSC Portal to assess the best way to publish its services in the central EOSC Portal. ExPaNDS would beneficiate of a programmatic interface to automatically publish ExPaNDS services in

[5] The EOSC Portal: https://www.eosc-portal.eu/

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

the EOSC portals once validated against the EOSC rules of participation and will discuss this requirement with the projects currently in charge of the EOSC Portal development (EOSC-hub[6], EOSC-Enhance and the project that will be funded under the call INFRAEOSC-03-2020[7] expected to start in 2021).

The following picture shows the possible different access channels to the ExPaNDS services.

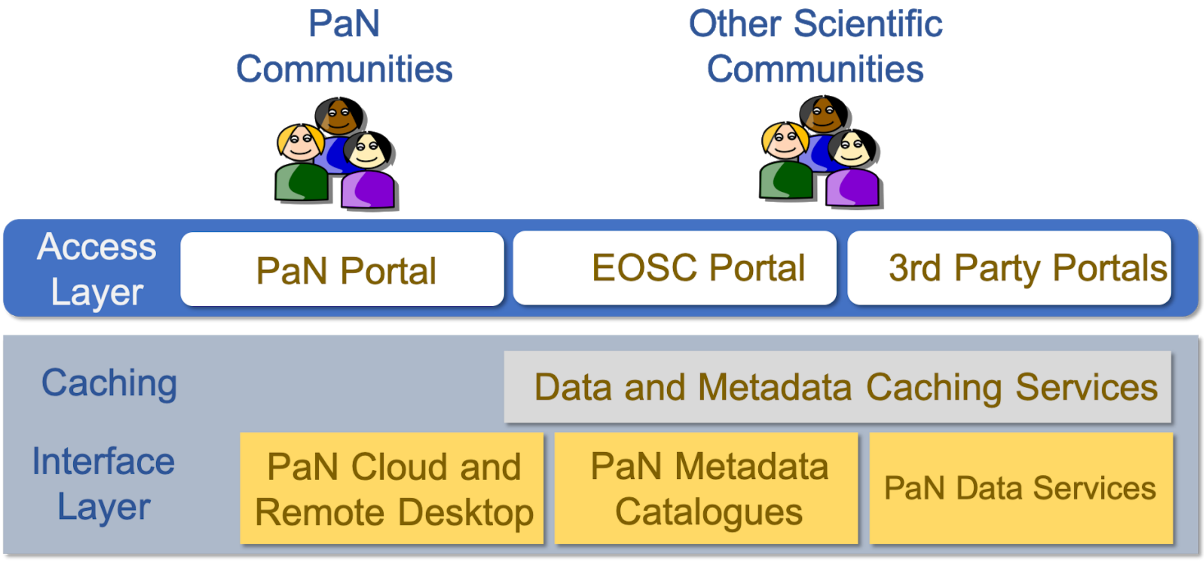

Fig. 5: Accessing ExPaNDS Services in EOSC. Photon and Neutron communities and other EOSC user communities will be able to access the ExPaNDS services through the PaN portal and, possibly, via the EOSC portal.

3.2. Service integration and composability

EOSC has been envisaged as a key instrument to facilitate access to scientific services, lower barriers to integrate and compose services and promote the usage of services between adjacent communities. To reach this aim, an effort is being done to suggest EOSC standards and define interoperability guidelines that would allow to identify EOSC 'compliant' services. These services would offer well-established and documented interfaces for usage and integration, based on well-known standard or APIs, facilitating their exploitation and the combined usage of more EOSC services.

This activity is currently led by the EOSC-hub project that proposed a reference EOSC Technical Architecture[8] that includes functions, interfaces, APIs and standards as technical

[6] EOSC-hub web site: https://www.eosc-hub.eu/

[7] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/infraeosc-03-2020

[8] EOSC-hub D10.4 EOSC Hub Technical Architecture and standards roadmap v2: https://www.eosc-hub.eu/deliverable/d104-eosc-hub-technical-architecture-and-standards-roadmap-v2-under-ec-review

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

concepts, with the final aim of fostering interoperability and, ultimately, service composability.
The adoption of standard interfaces is particularly relevant for horizontal services that can be adopted in different contexts like operational services (AAI, helpdesk, accounting, monitoring, etc), named **federation services** in the EOSC-hub work (they allow to federate services offered by multiple providers), and IT services (cloud/computing facilities, data transfer services, data management services, data sources, etc.), named **common services** in EOSC-hub (they can be shared by many research facing services).

ExPaNDS is intended to contribute to the definition of the EOSC standard interfaces with the expertise of its consortium and to evaluate the adoptions of such interfaces in its architecture to maximise the interoperability with other EOSC services and harmonise the services offered by the national RIs. Furthermore, ExPaNDS is interested in assessing EOSC horizontal services (federation and common) for adoption on its infrastructure with a focus on operational/federation services. Usage of these services by ExPaNDS would allow, from one side, to reduce the cost for implementing the distributed infrastructure (re-using already existing basic services instead of implementing new ones) and, from the other side, to better "plug" its infrastructure in EOSC and align its architecture with the EOSC one. Indeed, for example, adopting an EOSC compliant AAI in ExPaNDS would facilitate the access of ExPaNDS services to other EOSC communities or implementing an accounting system integrated with the EOSC central one would allow to have a clear measurement of the impact and uptake of ExPaNDS services in the whole EOSC. ExPaNDS will also make its infrastructure interoperable with the data management services developed by the PaNOSC project and evaluate EOSC data transfer services, advanced analytics tools (e.g. Notebook) and computing and storage facilities to scale-up the amount of IT resources available to its end users.

3.2.1. Map ExPaNDS Architecture to the EOSC Technical Architecture

When feasible, ExPaNDS will adopt best practice and standards suggested by EOSC to make its services interoperable with other EOSC services fully integrating its architecture in the EOSC one.

This section describes how the ExPaNDS architecture, previously introduced, can be mapped to and become fully aligned with the EOSC Technical Architecture proposed by EOSC-hub.

The functional view of the EOSC technical architecture is shown in the following picture. It highlights interactions between different service categories, federation and common, introduced in the previous section, and thematic (the research facing services, e.g. a data analytics tool for a certain scientific discipline).

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

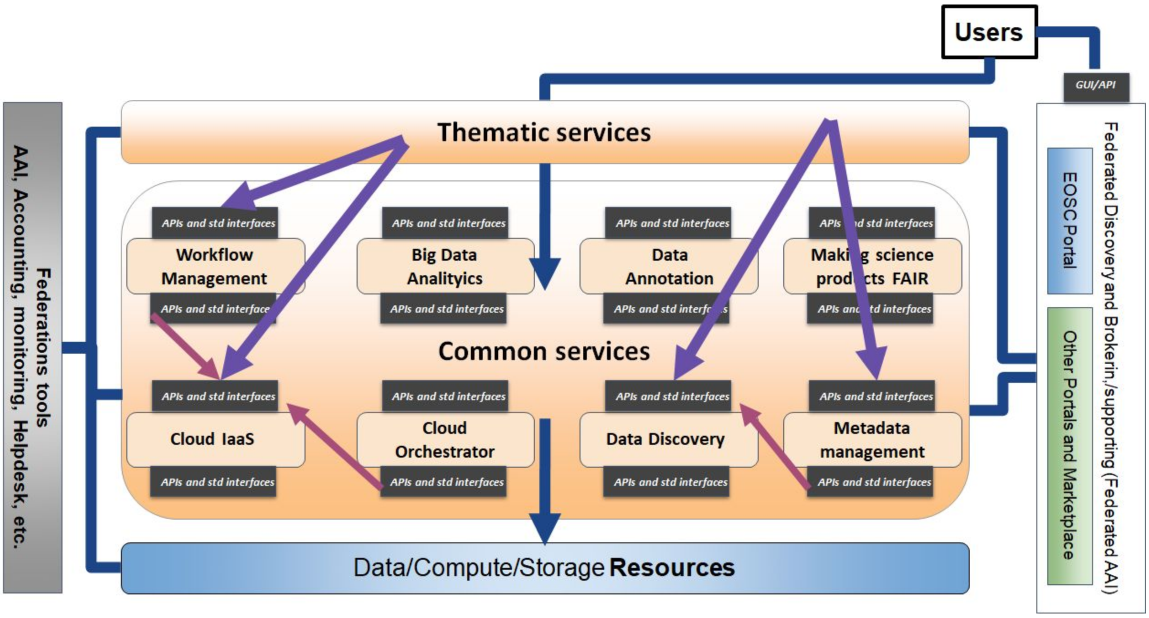

Fig. 6: EOSC Technical Architecture. Functional view.

Each component within the Common and Federation services exposes API and standard interfaces to facilitate its interactions with other components. User facing services can be accessed through the EOSC Portal, through any other portal publishing them or directly.

The next picture shows how the ExPaNDS architecture can be mapped into the EOSC Technical Architecture. PaN analytics can be directly accessed through the PaN portal, in addition their publication in the EOSC Portal will be considered (in gold in the picture). The PaN analytics leverage the underlying computing services (cloud IaaS and Kubernetes) and data catalogues (ICAT, SciCat and custom catalogues) to perform their tasks. Both compute services and data catalogues will expose well defined API and standard interfaces to facilitate their exploitation and integration with other services/components. ExPaNDS will integrate the EOSC AAI adopting an AARC blueprint compliant AAI system and will evaluate the adoption of other EOSC federation services (helpdesk, accounting, monitoring, etc. in gold in the diagram).

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

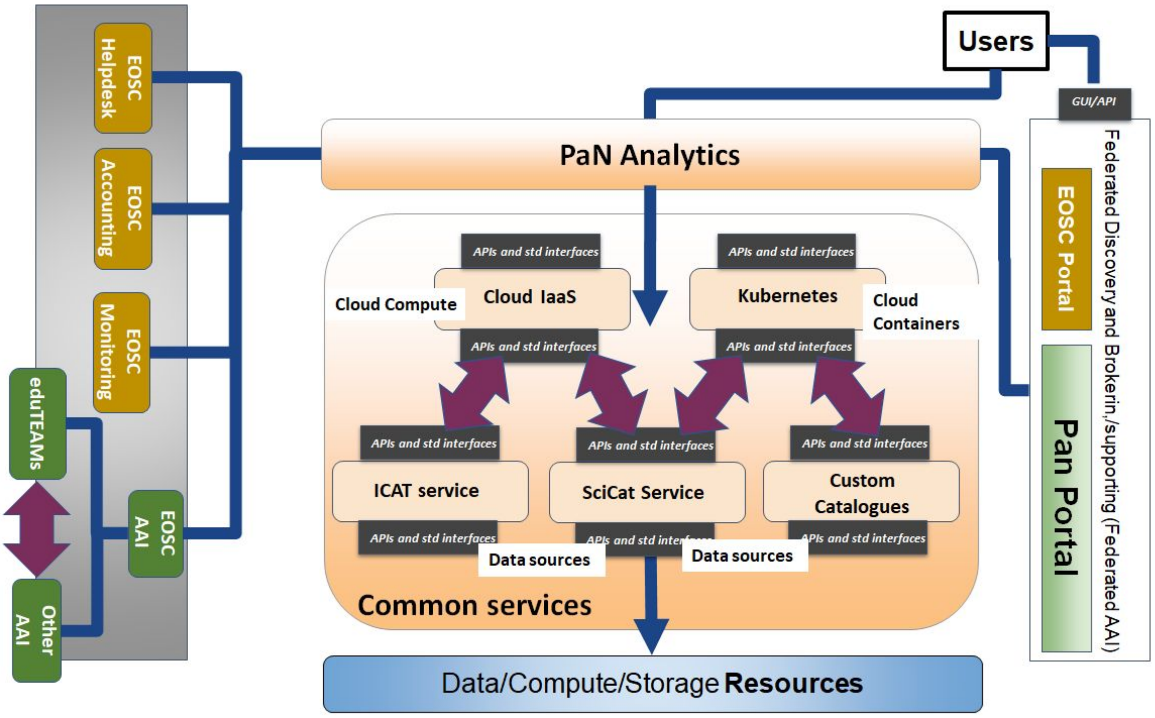

Fig. 7: ExPaNDS architecture mapped into the EOSC technical architecture. EOSC services in gold will be evaluated for integration by ExPaNDS during the project lifetime.

It is worth mentioning that the adoption of EOSC well-known API and standard interfaces in the “Common services” layer would facilitate the inclusion in the architecture of new EOSC services delivered by providers outside the Photon and Neutron communities. This would allow (1) to easily scale compute resources when needed and (2) to connect the PaN analytics with other EOSC data catalogues.

3.2.2. Integration with EOSC federation services

This section presents the most relevant cases of integration of the ExPaNDS architecture with the EOSC federation services that have been considered until now.

3.2.2.1. AAI

As described in section 2.4, ExPaNDS will implement an Authentication and Authorisation Infrastructure (AAI) compliant with the AARC Blueprint Architecture 2019[9] and the AARC guidelines[10] to guarantee interoperability with other EOSC AAIs enabling single sign-on access for ExPaNDS users in EOSC and granting access to ExPaNDS services to other EOSC users.

The following diagram shows a researcher’s perspective following the AARC Blueprint Architecture. According to this architecture, ExPaNDS will implement a community AAI for the Photon and Neutron Community and deploy an infrastructure proxy on top of each of its services/facilities to enable the access from any AARC compliant community AAI.

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

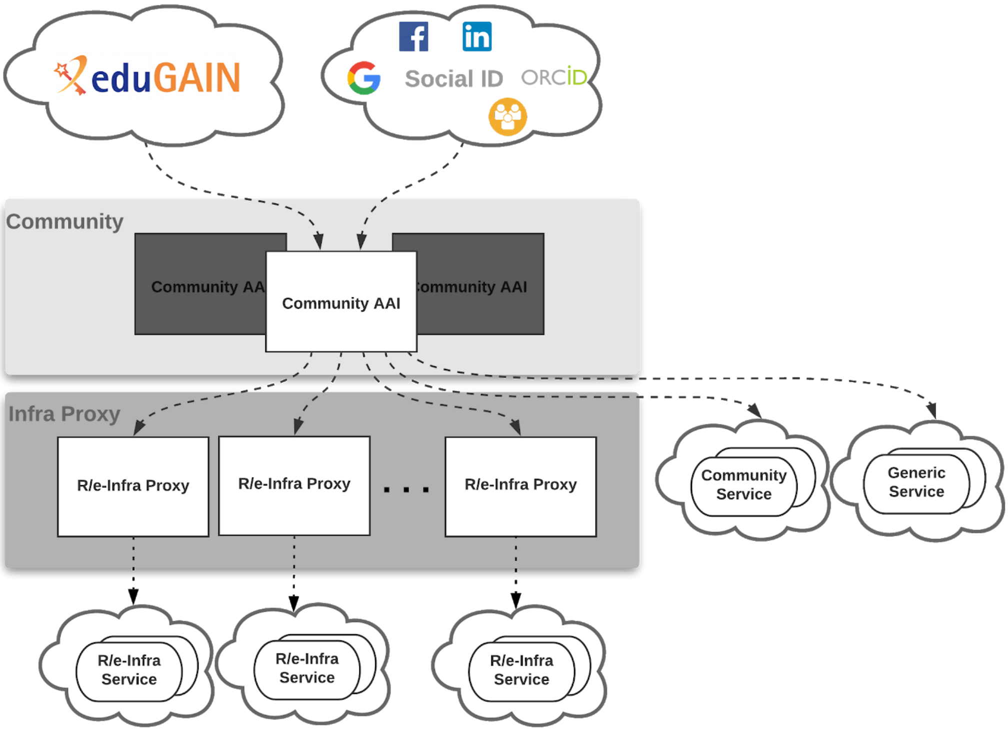

Fig. 8: High-level view of AAI architecture for access to EOSC resources: A researcher's perspective following the AARC Blueprint Architecture.

3.2.2.2. EOSC helpdesk, accounting and monitoring

ExPaNDS will evaluate the adoption of EOSC helpdesk, accounting and monitoring tools in its architecture. These services take on great importance to connect ExPaNDS with EOSC. For example, if ExPaNDS services will be published in the EOSC Portal, it will be necessary offering a way to users accessing the services via this channel to interact with the ExPaNDS support team, this can be achieved with the EOSC helpdesk. EOSC accounting would allow to have a measurement of the uptake of the ExPaNDS services in EOSC allowing the project to demonstrate its impact to the EC. Finally the monitoring would allow to demonstrate the reliability of the ExPaNDS services in EOSC.

EOSC helpdesk[11], accounting[12] and monitoring have been conceived as distributed systems with a central catch-all instance and interfaces towards federated instances. Each provider in EOSC can choose different ways to adopt and integrate with these services (using central instance, interconnect its own instance with interface A or interface B, etc).

The project will assess the worth of these tools for its infrastructure and, in case of adoption, will consider both to use the EOSC central instances or to interconnect its own instances

[11] EOSC Technical Specification - Helpdesk: https://wiki.eosc-hub.eu/display/EOSCDOC/Helpdesk
[12] EOSC Technical Specification - Accounting: https://wiki.eosc-hub.eu/display/EOSCDOC/Accounting

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

with the EOSC central ones. ExPaNDS will adopt as much as possible interfaces compliant with the technical specifications of the EOSC federation tools[13].

3.2.3. Integration with EOSC common services

There are a series of EOSC common services that are interesting for ExPaNDS and that will be assessed for integration during the project lifetime. These include the PaNOSC data management services, data transfer services, notebooks, e-infrastructure computing and storage resources,etc.

3.3. ExPaNDS requirements for EOSC

Although the project is still in an analysis phase, a certain number of requirements for EOSC have been already identified as it can be in part deduced reading the previous sections of this document. These are:

- An easy to use, possible automatic, interface to publish ExPaNDS services in the EOSC Portal;
- The definition of EOSC APIs and standards to facilitate the integration between services and components;
- The adoption of AAI systems compliant with the AARC Blueprint Architecture and its guidelines;
- A central helpdesk to collect user requests from the EOSC Portal with interfaces to connect external helpdesks;
- A central tool for accounting the usage of resources with interfaces to connect external accounting systems;
- A central tool for monitoring services with interfaces to connect external monitoring systems;
- Availability of a data transfer services;
- EOSC computing and storage services to scale-up when needed.

This is an initial list that will be enriched and refined during the project lifetime.

[13] https://wiki.eosc-hub.eu/display/EOSCDOC/Federation+services

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

4. Conclusion & Outlook

This deliverable describes the technical architecture of the ExPaNDS project on how to interface and federate current and possibly future services, locally provided by national PaN infrastructures, to a high level scientific service, accessible via state of the art cloud interfaces. In consequence, this enables ExPaNDS to enrich the EOSC marketplace with specific, high profile PaN offerings.

As a first step, we will identify commonalities among tools provided by project partners in their own facilities and define and implement common interfaces to abstract those services. This involves, but is not limited to, catalogues like ICAT and SciCat, access to cloud resources via Jupyter Notebooks or to HPC resources by means of remote desktops.
Defining standards and best practices on handling of data and on the composition of metadata is the first step to make ExPaNDS data complying with FAIR principles. Similarly important for FAIRness is to work towards interoperability of Data Analysis platforms amongst project partners in ExPaNDS and PaNOSC, to guarantee future reproducibility of scientific results.

A topic, not sufficiently covered in our architecture, is the actual access to data and the popular question of either to move data to the compute infrastructures or to process analysis pipelines at the facilities holding the data. As this work is not part of our DoW, we rely on findings provided by PaNOSC WP6. Additionally we are in contact with WP2 of the H2020 ESCAPE project on 'data lakes'. The 'data lake' concept abstracts access to data, including prestaging of data and on demand caching.

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