Deliverable

Deliverable Number	**D24.6**
Deliverable Title	**White paper on suitability of HNScienceCloud and European Open Science Cloud for synchrotron and FEL applications**
Lead Beneficiary	**PSI**
Authors	**A. Ashton (PSI), R.Dimper (ESRF), A.Götz (ESRF), D.Salvat (ALBA), F.Schlünzen (DESY)**
Type	**Report**
Dissemination Level	**Public**
Due date of delivery	**Month 36**

Introduction

This paper identifies the needs specific to synchrotrons and FELs towards the European Open Science Cloud (EOSC) and how these research infrastructures and their user communities can best take advantage of the EOSC. The CALIPSOplus project [http://www.calipsoplus.eu/] brings together all photon sources in Europe. A significant number of them have participated in the Joint Research Activity 2 (JRA2 also referred to as WP24), to develop a prototype of a remote Data Analysis as a Service (DAAS) platform. The prototype was deployed and tested at the participants site and an online workshop was held at the end of 2020 to present the results [1].

The prototype DAAS service was instrumental in identifying the needs for remote analysis and cloud-like access to IT resources for photon research infrastructures. This experience furthermore allowed the authors to reflect on the promise on "***The European Open Science Cloud (EOSC) is an environment for hosting and processing research data to support EU science***" made by the EOSC and its application to photon sources.

The paper starts by defining the needs of the Photon and Neutron (PaN) community towards cloud computing and then presents the achievements from JRA2. It then goes on to analyse the situation for photon sources according to the services recommended by the EOSC: FAIR data, data storage, data processing, data re-use. It continues with a brief overview of the Helix Nebula Science cloud and then proceeds to an overview of EOSC services, their state and suitability for the PaN RIs.

Needs of the PaN community for Cloud computing

The PaN Research Infrastructures (RIs) are used by a large multidisciplinary scientific user community to carry out experiments for understanding the structure and functioning of matter. Experimental projects are submitted by research teams, peer reviewed and, if successful, scheduled for beamtime. Typical experiments take between hours to several days of beamtime on the selected experimental setup (also called beamline or instrument). During the experiment, the research infrastructure provides the computational means for data acquisition, pre-processing and quality assessment. Full data analysis is usually taking place in the home laboratories of the visiting scientists and takes months to years. The time from experiment to publication typically takes years.

In the last few years, the PaN RIs have experienced a shift to more complex experiments generating more complex and much larger data sets. This has led to the need to extend the access to the IT resources beyond the duration of the experiment and to call on the expertise of the facility staff to help in the data reduction and data analysis process. We also witnessed that some experiments, especially in tomography, generate so much data that carrying the data away from the RI becomes very problematic. These trends put a considerable pressure on the facility staff and the IT infrastructure. Many of our visiting scientists do not have easy access to compute

facilities in their home laboratory or university. All the above is clearly showing in an increased delay between the experiments and the publications and needs to be addressed.

The needs of the PaN community towards cloud computing can thus be seen from four different perspectives:

1. A scale-out solution for the facilities to complement on-site IT resources for peak requirements. This would ideally lead to a hybrid cloud solution where some tasks can be dynamically shifted to a cloud provider. This may initially be limited to applications which do not require to transfer a substantial amount of data, hence more CPU bound than data bound. Alternatively the off-loading of the internal IT resources could be done with applications which require a large amount of CPUs/GPUs and are particularly easy to migrate such as some simulation codes. Larger jobs would also ease the unavoidable monitoring and accounting linked to off-site computing in a cloud environment.
2. An on-demand solution for users of our facilities who do not have large scale compute facilities in their home laboratories or universities. Here the role of the facility would still be to assist users to port and optimize software packages, which are often originating from the RIs, to the selected cloud provider.
3. The PaN community is very diverse and dynamic. Users of our RIs are generally using more than one facility and different techniques to study their samples. Data sets may thus require to be combined and compared, results (publications) must be verifiable, and research outputs spanning from raw data to publications must be made openly available. Since the PaN RIs are “only” one element in this ecosystem, a federating system allowing to interact with research outputs of diverse origins is required. This will be the main thrust for the European Open Science Cloud providing the glue between data repositories and all the tools allowing to interact with the data.
4. Data storage and transfer. The data produced by the PaN RIs is in the order of tens of Petabytes of raw data per year. The data need to be stored and exported to the users home institutes / laptops / cloud infrastructure. The same is true for the processed data. As explained above the processed data can be larger than the raw data in some cases. Most of the PaN RIs have adopted Open Data policies (see below), which means making the data available to the scientific community at large. The PaN RIs need a reliable easy-to-use efficient data transfer and storage solution for long term storage i.e. a decade or more. Ideally the EOSC should provide data storage and transfer services which could be used by the PaN RIs and their user communities. This would be in-line with the objective of the EOSC to provide FAIR data and enable RIs to provide FAIR data without requiring them to install and operate petabyte scale long term data storage. Providing a solution for large data transfer will reduce the financial burden on PaN RIs to purchase expensive commercial data transfer solutions and internet bandwidth. The PaN RIs can concentrate on their core business of producing FAIR data.

For the above scenarios the facility IT experts will be heavily solicited upfront to prepare the software in the cloud environment and to assist users with the particularities of working in a cloud environment. The effort to adapt, optimize, and use the software packages in such environments is not to be underestimated.

The scientific user community of our facilities is not necessarily IT literate. Whatever the data analysis environment, it has to be user friendly to avoid saturating the facility IT staff with support requests.

Current technical implementations of JRA2 and FAIR

FAIR User Facilities

The Photon Science User Facilities are continuously developing their programmes offering scientific opportunities also in particular for remote access, which proved an indispensable asset to maintain research activities and to fight the COVID-19 pandemic.

The user facilities are hence FAIR in a slightly different way by offering findable, accessible, interoperable and reusable data via remote experiments and data services.

The focus of JRA2 was clearly on services supporting interoperable and remote data management and analysis services. The implementations are closely and successfully following the *Blueprint on implementing a DAAS platform* [2].

Interoperability

AAI is a core element of any federated service. The UmbrellaID has been serving as the only common AAI system in the Photon and Neutron community for several years. UmbrellaID is integrated in all User Office systems, and accepted by several other services like data catalogues, wayforlight.eu, gitlab instances and nextcloud based cloud-federations to mention a few.

UmbrellaID is EOSC-ready from a technical point of view, but needs to be developed further to achieve long-term sustainability and full GDPR-compliance. These developments have been brought forward in CALIPSOplus and will be completed shortly by joining GÉANT's eduTEAMS as part of the PaNOSC project (see deliverable D24.7 [3] for the timeline). Once this final step has been taken, the seamless integration of any service on premise or within EOSC will become straightforward, in particular when utilizing central authentication services like Keycloak.

Data analysis services are literally impossible without data catalogues of FAIR data. Most photon RIs have data catalogues in place; implementations, however, differ slightly at each facility. To be usable in an EOSC environment, data catalogues need to be interoperable and support a common set of operations. CALIPSOplus has initiated developments improving interoperability between data catalogues of the user facilities, which were taken up by the PaNOSC [2] and ExPaNDS [4] projects. The resulting search API [5] provides the common interoperability layer supporting meta-data harvesting and discovery by OpenAIRE [6] or B2FIND [7]. The search API is in the process of being implemented and will enhance findability and interoperability in the EOSC ecosystem.

Accessibility and Remote services

Remote data analysis very often requires graphical frontends, which cannot simply be implemented as web-services or Jupyter notebooks. CALIPSOplus JRA2 has taken up this very strong user requirement and developed a prototype of a common data analysis portal. The modularity of the portal components, as designed in the blueprint, allows the composition of tailored, user friendly data analysis services. The portal backend utilizes Django and Django Rest frameworks which connect to underlying compute resources (e.g. Docker) through Apache Guacamole providing VNC or RDP connections through HTTPS. The backend is hence conveniently accessible with any web-browser, naturally also supporting UmbrellaID. The portal frontend supports kubernetes deployment, and integrates with Jupyter services. The portal offers a very convenient and cost-effective way to enable remote data analysis at the user facilities, and at the same time within any cloud environment from a software deployment point of view. In the case of remote cloud, the issue of data transfer for large data volumes needs to be addressed as well.

The JRA2 prototype provided a working demonstration [1] of these principles. The ExPaNDS and PaNOSC projects have recognized the importance of a user-friendly data analysis portal, and have continued the developments with a very similar architecture [8].

The portal is largely ready to be deployed as a generic service within EOSC. There are a few EOSC related services, which are currently not available, but once available would significantly improve the usability of data analysis portals.

A key component is the access to scientific data in the common portal. An on-site deployed portal of course supports direct access to the users' data. Within EOSC the user facilities will become discoverable and accessible through the common search API. Small datasets can simply use URL based file paths in the data analysis applications without requiring any or only very minor modifications. For large datasets, and that is the more interesting use case, the portal would need some data movement service (e.g. a data lake) which serves the purpose with satisfying convenience and performance. The PaN RIs would profit from an EOSC service for data transfer. The ESCAPE project [9] proposes to provide the required services within EOSC by implementing a data lake.

Tailoring the software stack of a portal instance to specific experiments would benefit from ontologies describing both an experiment type and the most standard software stack, which would be available from trusted container registries.

Finally, Apache Guacamole currently does not support hardware acceleration. For most applications that's not a strict requirement, however quite a bit of standard software used to visualize and interpret 3D/4D data does rely heavily on GPU hardware acceleration. Introducing GPU support in Guacamole, or other frameworks supporting RDP/VNC access, would be highly beneficial.

Jupyter [10] is a great tool to offer data analysis services including visualization in a rather standardized way. An increasing number of analysis pipelines used at the PaN RI facilities are being made available as Jupyter notebooks which makes it easy for scientists as well as citizens to document, follow and view the flow of data analysis or even an on-going experiment. The

CALIPSOplus JRA2 portal prototype supports deployment of Jupyter services on-premise as well as in the cloud, which offers a great deal of interoperability and reuse. Integration of Jupyter services, and possibly also Binder services [11], is a very basic requirement. EOSC offers such services as a prototype, but lacks the possibility to provide such services for non-authenticated, non-registered users, which would be quite important to make tutorials easily available, or for Jupyter based citizen science projects. For Jupyter-based analysis services, there is again the aforementioned problem of transparent access to large datasets.

FAIR data

As stated above, significant progress has been made by Photon RIs during the CALIPSOplus grant, to have interoperable data catalogs in place. However, both the provision of such services as a reliable, scalable solution and ensuring the data being catalogued is sufficiently well described to be considered FAIR, is challenging and not consistently implemented across all the CALIPSOplus RIs.

In 2020, as part of the ExPaNDS project, a survey [12] was published and made available "R*eport on status, gap analysis and roadmap towards harmonised and federated metadata catalogues for EU national Photon and Neutron RIs*". The aim of the survey and associated report was to describe the status, make a gap analysis, and outline a roadmap required to achieve harmonised and federated (meta)data catalogues of the participating national Photon and Neutron (PaN) Research Infrastructures (RIs), aiming for an EOSC-compliant implementation.

The results showed how most facilities had already made significant progress towards ensuring prerequisites such as facility data policies and infrastructures to operate the services all the way through to integrating the chosen catalogues in the facility. CALIPSOplus has shown that this is a key and challenging prerequisite for FAIR user facilities and the implementation work will continue with the follow-on PaN projects PaNOSC and ExPaNDS and beyond.

As part of the process of making FAR data a reality, PaNOSC has updated the PaNdata data policy to be FAIR. The new data policy [13] is in the process of being adopted by all PaNOSC partners. Partners who already have a data policy (ILL, ESRF, EuXFEL and ESS) will update their existing data policies to make them FAIR. CERIC-ERIC and ELI did not have a data policy before PaNOSC and have adopted the PaNOSC data policy adapted to their local requirements. At the end of the PaNOSC project therefore at least 6 PaN RIs will have a FAIR data policy.

Helix Nebula Science Cloud

The Helix Nebula Science Cloud project [https://www.hnscicloud.eu/] was an EU H2020 funded project to explore the use of commercial clouds for scientific use cases. It was set up as a so-called Pre-Commercial-Procurement (PCP) funded project. This means that half of the funding had to go into R&D by commercial companies, to encourage the development of new solutions for scientific research applications by the selected companies. Two photon sources were partners in HNSciCloud - ESRF and DESY. The project was coordinated by CERN. The experience of the project was that PCP is very complicated and a too heavy approach for purchasing commercial services. The HNSciCloud PCP consisted of a tender exercise with 3 selection phases and with

multiple partners. Due to this heavy approach it excluded the main players in commercial cloud services. The partners which were selected did not offer well-adapted services for scientific cloud computing and data transfer. The suppliers for the HNSciCloud project spent the majority of the project developing the missing services e.g. simple commands to setup an HPC cluster. The result was that no meaningful tests could be done until the end of the project and they were not conclusive enough to determine if they had really developed a new service comparable to the main cloud suppliers. A second service which was requested by the photon sources was an efficient and simple to use data transfer service for high volume data. The chosen solution, OneData [14], turned out to not be stable and performant enough to be used in production at the time of the project.

In the end the HNSciCloud project did not enable photon sources to have easier access to new commercial cloud services and suppliers. The main cloud suppliers remain the best suppliers of commercial cloud services for computation. An easy to use and performant solution for moving giga/terabytes of data remains without an obvious solution at the time of writing.

Two new EU H2020 projects, OCRE [15] and ARCHIVER [16], continue the effort started by HNSciCloud for providing easy access to commercial cloud services. OCRE has avoided the complicated PCP process. OCRE provides pre-negotiated prices with cloud suppliers (including the main players). This approach promises to facilitate the access to commercial cloud services for scientific institutes potentially avoiding that the RIs have to tender individually the bespoke cloud services. The OCRE services are being tested by CERIC-ERIC as part of the PaNOSC project. PaNOSC plans to extend the test to other sites and services. The OCRE procurement scheme is also likely to be used for procuring services in the upcoming EOSC-Future project.

The ARCHIVER project is a PCP project intending to deliver end-to-end archival and preservation services covering the full research lifecycle. The project is currently in the first phase of the PCP process during which detailed design reports are evaluated. Like for the HNSciCloud project, it remains to be seen whether the ARCHIVER project will produce results which are useful for the PaN community for preserving and making data accessible to the PaN community and EOSC.

The European Open Science Cloud

EOSC overview

The EOSC is an initiative started by the European Union in 2015 as one of the objectives of its Open Science policy. The implementation of the EOSC followed the recommendations of the High Level Expert group. During the first five years the Commission has financed over 400 M€ in about 50 projects under the H2020 framework programme. There has been a lot of activity over the last five years in many areas which has culminated in the creation of the EOSC association at the end of 2020 [https://eosc.eu]. The CALIPSOplus partners have been involved in building the EOSC, either as actors or as members of various review committees and participants in numerous workshops. They are therefore well informed of where the EOSC implementation stands.

A question which is often asked by researchers or staff of our Research Infrastructures is "what is the EOSC". The EC answer provided on their main information page (see box below) however

needs more background information and interpretation for those not familiar with the EOSC. This paper offers its own interpretation of what the EOSC is in order to report on where the EOSC is in terms of offering services or new possibilities for the PaN RIs.

WHAT THE EUROPEAN OPEN SCIENCE CLOUD IS

THE EUROPEAN OPEN SCIENCE CLOUD (EOSC) IS AN ENVIRONMENT FOR HOSTING AND PROCESSING RESEARCH DATA TO SUPPORT EU SCIENCE.

THE PROCESS TO CREATE THE EOSC WAS INITIATED BY THE COMMISSION IN 2015. IT AIMED TO DEVELOP A TRUSTED, VIRTUAL, FEDERATED ENVIRONMENT THAT CUTS ACROSS BORDERS AND SCIENTIFIC DISCIPLINES TO STORE, SHARE, PROCESS AND RE-USE RESEARCH DIGITAL OBJECTS (LIKE PUBLICATIONS, DATA, AND SOFTWARE) FOLLOWING FAIR PRINCIPLES.

THE EOSC BRINGS TOGETHER INSTITUTIONAL, NATIONAL AND EUROPEAN STAKEHOLDERS, INITIATIVES AND DATA INFRASTRUCTURES TO DEVELOP AN INCLUSIVE OPEN SCIENCE ECOSYSTEM IN EUROPE.

Definition of EOSC according to: *https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en*

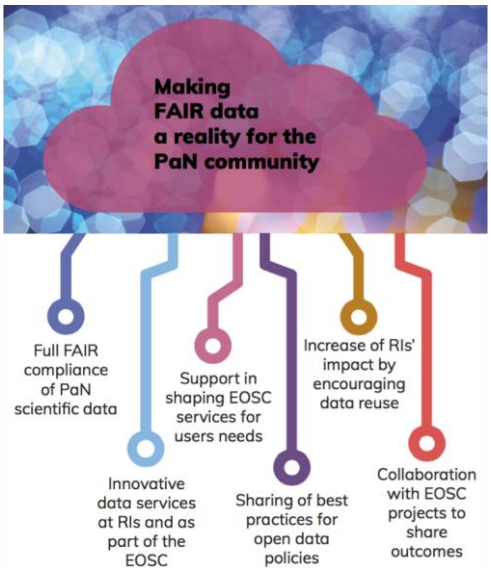
The interpretation of what the EOSC is in this paper is based on how the EOSC has evolved so far and how it will probably evolve taking into account financing from the EC and member states and what on-going and planned EOSC projects plan to deliver. Taking all these aspects into account we think the EOSC should not be seen as a "thing" e.g. a compute cloud, but rather a set of practices and processes for enabling and delivering Open Science and a federation of services (a system of systems). The EOSC does not have the funding model to provide unlimited physical resources to scientists. It is made up of the communities which are part of the EOSC. Communities bring not only data but also their own resources for software and computing power. These are made available via community portals or the EOSC portal. Communities which form part of the EOSC need to provide scientific research data which comply with the FAIR principles - Findable, Accessible, Interoperable, Reusable.

PaN related EOSC projects

CALIPSOplus JRA2 was a precursor in the EOSC landscape[1]. When JRA2 was proposed in 2016 and funded from 2017 to 2020 the EOSC was in its initial conception phase. The High Level Expert Group (HLEG) had only just proposed to the EC to embark on the EOSC [17]. Encouraged by the report from the second HLEG in 2018 entitled "Prompting an EOSC in practice" the Horizon 2020 program proposed calls for projects to implement the EOSC. The calls were part of the INFRAEOSC series. The most relevant ones for photon sources were INFRAEOSC-04 for projects on the ESFRI roadmap and INFRAEOSC-5b for national projects. Two projects for photon and neutron sources, PaNOSC and ExPaNDS respectively, were funded by these calls. The two projects are briefly described below.

[1] For details on the timeline of the EOSC refer to the report "T*he EOSC strategic implementation plan*" [https://op.europa.eu/s/oIDu] by the EOSC Executive Board.

PaNOSC



PaNOSC is financed by the INFRAEOSC-04 call. This call was aimed at RIs on the ESFRI roadmap. The PaNOSC partners are: ESRF as coordinator, ILL, EuXFEL, ESS, ELI-DC, CERIC-ERIC and EGI as e-infrastructure and GÉANT as close collaborator for the Authentication and Authorisation Infrastructure (AAI). All PaNOSC partners are on the ESFRI roadmap. The main objectives of PaNOSC are listed on the image on the right. Refer to the PaNOSC website [2] for news, deliverables and examples of typical use cases being addressed. The main emphasis of PaNOSC and the INFRAEOSC-04 projects are to "make FAIR data reality" in the partner RIs and “to connect the RIs to the EOSC”. To do this PaNOSC has developed a FAIR data policy framework together with the national PaN RIs. PaNOSC is developing a common search API to search for FAIR data across the PaN RIs. Based on the experience gained from the CALIPSOplus JRA2 prototype a common portal for innovative data services is being developed. Special emphasis is being put on promoting the use of Jupyter notebooks as a solution for reproducible data analysis. Simulation services for experiments and optical elements are being extended and made available as a PaNOSC service. A training portal for PaN users [18] has been deployed and is being extended within the project. All these services will be made available via the EOSC services portal. PaNOSC will provide a solution for transferring data to user home institutes and in and out of the EOSC. PaNOSC services support an EOSC ready AAI solution based on eduTEAMS from GÉANT. Last but not least these services need to be sustained. PanOSC has a dedicated work package for analysing and proposing a sustainability roadmap for the future. All the services and activities of PaNOSC are shared with its sister project ExPaNDS comprising all the national PaN RIs in Europe.

ExPaNDS

ExPaNDS is in many respects the small sibling of PaNOSC. The project has largely overlapping goals, and hence both projects are mutually adopting the majority of developments and enabling all PaN RIs to make their data open following the FAIR principles. The core elements are an UmbrellaID based AAI, interoperability of data catalogues through a common search APIs and enhanced remote access and reproducibility through virtual desktops and Jupyter-based services. The CALIPSOplus blueprint and prototypes are collaboratively advanced by Geant, PaNOSC and ExPaNDS, which guarantees the sustainable continuation of the main achievements of the CALIPSOplus project. Most components are technically ready to be deployed in an EOSC ecosystem, and largely interoperable with OpenAIRE, B2FIND, EOSC AAI are other core components of the EOSC platform.

Figure: Infographic on how the ExPaNDS project works ([https://expands.eu/infographics/])

Science Clusters

The INFRAEOSC-04 projects are referred collectively to as the Science Clusters. They are ENVRI-FAIR, EOSC-Life, ESCAPE, SSHOC and PaNOSC together with ExPaNDS. Together they represent a large variety of scientific domains and services. A number of the scientific domains overlap with the scientific domains of the experiments carried out at the PaN RIs for example life sciences, environmental sciences and space sciences to mention a few. The holy grail of Open Science and EOSC in particular is to combine data from different domains to produce new insights thereby enabling and encouraging cross domain scientific studies. The 5 cluster projects have worked together on common topics and assisted in EOSC meetings and discussions with the EC. A common paper on the services offered and needs wrt EOSC of the science cluster was published in 2020 [19]. Common workshops on knowledge exchange or common solutions have been organised. During the remaining two years of the 5 projects the frequency of these common meetings are planned to increase culminating in common solutions for workflows, notebooks, data transfer etc. The EOSC-Life project is custodian of the EU COVID-19 repository [20] which by the end of these projects should link to all COVID-19 related data from the various clusters. The clusters are working closely together participating in EOSC-Future representing the scientific users communities. The PaN RIs should continue working with the other science clusters in the future after PaNOSC and ExPaNDS to maintain the synergy between the different scientific domains. This work could be done under the umbrella of the LEAPS [21] and LENS [22] projects or by following the examples of three of the other clusters (i.e. ELIXIR, ENVRI and CESSDA) as an ERIC dedicated to managing data from the PaN RIs.

EOSC-Future

EOSC-Future is an INFRAEOSC-03 project under preparation which will drive the next level of integration and development of EOSC core components and the EOSC interoperability framework. The project will most likely start in Spring 2021, has a total budget of 40MEuros for a duration of 30 months, and brings together the European e-infrastructure organisations EGI, EUDAT, GEANT, OpenAIRE, and RDA and the five INFRAEOSC-04 ESFRI science cluster projects ENVRI-FAIR, EOSC-Life, ESCAPE, PaNOSC, and SSHOC. The science clusters and the e-infrastructure organisations will closely work together to ensure that services developed in the science cluster projects are made available generically in EOSC and that the EOSC overall is suitable for research, and is promoted and used by the scientific user communities. The EOSC-Future project will also closely link to the recently approved INFRAEOSC-07 projects and the EOSC association and ensure that new requirements where appropriate are taken onboard.

EOSC-Future aims to facilitate interoperability between EOSC deployed services, and therefore also facilitate cross-disciplinary research and reuse of FAIR data. This is particularly challenging for the rather heterogeneous scientific communities utilizing the PaN user facilities, and it bears the risk that a too strong harmonization will make the adoption too expensive for smaller scientific communities, or for newly established scientific fields like serial crystallography or single particle imaging which have yet to develop standards and well-defined workflows. The emerging scientific fields are however driving forces of new developments and innovation, and EOSC should try to support adoption of EOSC core services and frameworks at an early stage. For this, the strictness of compliance with EOSC would need to be balanced with the occasionally fast changing requirements of scientific communities. We are therefore glad that the science clusters are deeply involved in the project and can link to scientific communities and user facilities to contribute to the success and usefulness of the EOSC.

Bringing rather generic PaN services to EOSC will also require certain changes in the way the user facilities support their users, or in this case the consumers of the service. Extending the support beyond individual facilities' user base might require some conceptual changes and adoption of more collaborative workflows, but most likely also additional resources for sustainable support and further developments.

Sustainability of Data Analysis Services

Beyond the technical scope, in which the participants studied the technological feasibility, facilities have also raised the need to look at this model's sustainability and operating cost. A survey was conducted on the situation concerning data analysis service policies at all partner sites, the outcomes of which are summarised as follows.

Today the users who send their scientific proposals and get them approved, have permission to use the different instruments the facilities offer. This process sometimes includes various activities around the measurement or experiment, depending on the scientific case, and ends with the scientists leaving the facility with, either the raw data or an initially processed data, which should facilitate the analysis at their home institutions culminating in a publication in a peer reviewed journal. Facility resources are identified, clearly visible, and can only be consumed

by the accepted experimental team. In economic terms, and with respect to science production, one may identify the scarce resource as the instrument which collects the data in this particular case. No parallel operation is possible, so the users must wait in the queue for their turn to utilize it.

This new model adds a new item, not considered before, to the existing service catalogue: data analysis. This implies transforming the current operational model by adding a new type of scarce resource, computing resources, and additional personnel to operate it and provide support, for which users compete. In this case, a parallel operation is possible and will depend on how the computing resources, and support staff, are distributed.

Facilities adopting this new model would have to initiate the whole service definition, including the committed capacity (e.g. the number of nodes in an HPC Cluster), the service level to be provided (e.g. 24/7 or 8/5), and workforce (e.g. the number of support engineers). Facilities should internally consider any eventual underpinning process (outsourcing computing resources to commercial cloud or hyperscalers to manage peak loads). It is also highly recommended to include the cost structure. Each facility will then have to define the funding model to cover this new requirement.

Since this new service is directly provided to external users, some aspects of its definition (e.g. service level) would have to be publicly available, as an annexe to the existing data policies, or as a new data analysis policy. Currently no RI has a data analysis policy. It will be necessary for RIs who aim to provide data analysis as a service to define the rules of usage.

Current state of EOSC for PaN RIs

It is timely to reflect on the added value of the EOSC for the PaN RIs and progress after 5 years that the EOSC was officially proposed and 2 years since the official launch and Vienna declaration [23]. The progress can be evaluated from two different perspectives - (1) the progress of the EOSC infrastructure and standards by the e-infrastructure led EOSC projects, and (2) the progress by the PaN led projects like CALIPSOplus, PaNOSC and ExPaNDS.

As mentioned previously the EOSC is not a "thing" which can be easily identified, however, it strives to provide a set of common services to which the RIs will add by publishing their services. The main project in charge of delivering these services since the start of the EOSC is EOSC-Hub [24]. EOSC-Hub will terminate in Spring 2021 and should be continued by EOSC-Future. The main visible outcome of EOSC-Hub is the EOSC portal [21] where the PaN RIs should register their EOSC compliant services. At the time of writing this document only very few PaN services have been registered. PaNOSC is in the process of registering two but there are issues in registering community services when the community does not have a legal entity to represent it (neither PaNOSC nor LEAPS nor LENS are legal entities). The first services to be registered for PaNOSC will be the software catalogue and the training platform. In passing we notice that the situation is similar for the other science clusters i.e. they only have a few if any EOSC services registered in the EOSC portal.

What about the other services in the EOSC catalogue? There are 274 services registered at the time of writing this document (January 2021). How many of them are suitable for the PaN community? First of all it is difficult to know from the catalogue entries alone if the service offered is available for the PaN user community. Each service has its own Terms of Use conditions which can stipulate that the resource is reserved for a specific scientific user community. Secondly most of the resources require IT experts to use them. This means we cannot point PaN users to the EOSC catalogue to find a solution. The PaN RIs IT experts would have to identify the resources which are useful for the PaN community and act as an interface between the community and the provider. So far the PaN RIs have not identified any obvious resources in the EOSC catalogue which could be suited to the PaN community however an in depth search of which resources would be appropriate should be conducted. This could be done by the PaNOSC and ExPaNDS projects. A further difficulty of the EOSC catalogue is the "marketplace" type approach which makes the service look like commercial services with a shopping basket and a complex workflow to access them. This is not compatible with the open approach being encouraged for data by the EOSC.

The EOSC catalogue is a database of resources each of which point to the web page of the provider. This means the providers have to provide the necessary data, software, infrastructure etc. required by the resource. In the case of the PaNs this will be done by each of the PaN RIs by providing resources locally for their user community. As mentioned above the PaNOSC project is in the process of registering two pilot services for the PaN community: software catalogue and pan e-learning platform. These are of general interest and can be accessed by all. The next step is to provide data analysis services for the PaN user community. These services are being worked on inside PaNOSC and ExPaNDS with the implementation of a generic data analysis and search portal - a continuation of the DAAS prototype of JRA2. The PaN data portal will be deployed by all partners and give access to the PaN community to resources for doing data analysis. This brings up the question of who can have access and for how long and will require each PaN provider to define the Terms of Use.

The EOSC is far from finished and the next step will hopefully see the outcomes of the working groups on AAI, PID, FAIR, Training and RoP [25] being adopted and implemented by EOSC-Future. The PaN RIs will be able to profit from these efforts once they are ready assuming the outcomes are adapted to the needs of the PaNs and are not too complicated to use. PaNOSC and ExPaNDS projects are preparing to integrate their solutions into the EOSC by adapting the PaN AAI (UmbrellaID) to use the EOSC compliant eduTEAMs solution and continue to develop the PaN data analysis portal, data search API, simulation services, training and sustainability to be EOSC ready and available to the PaN user community.

Conclusions

The EOSC is a golden opportunity for the PaN RIs to embrace the principles and practices of Open Science and pave the way for them to become standard practices in the Photon and Neutron RIs and user communities. The two projects, PaNOSC and ExPaNDS, which have been funded as part of the INFRAEOSC-04 and INFRAEOSC-5b calls, are a unique opportunity to continue the work started by JRA2 of CALIPSOplus. A lot remains to be done however the two projects are only financed for 2 more years. This brings up the issue of sustainability of the work started in JRA2

and continued by PaNOSC and ExPaNDS. Clearly the best approach is for the PaNs to work closer together on implementing the necessary data management and data services for the PaN user community and to share the cost for sustaining them. The two initiatives, LEAPS and LENS, could be the platform for collaborating closer together and reaching the necessary agreements required for sharing the costs. Another solution is an independent body like a PaNdata ERIC which could be financed by LEAPS and LENS and which has the necessary resources to implement common solutions for data from the PaN community. This is the solution adopted by three of the science clusters mentioned above (EOSC-Life, SSHOC and ENVRI-FAIR).

What is clear from the current development and future financing of the EOSC is that the PaNs will have to develop, maintain and sustain their own solutions for a PaN Open Science Cloud (OSC) be it distributed or centralised. The EOSC will offer best practices and define standards but does not have the financing to develop solutions or provide significant infrastructure for the PaN RIs. Although the EOSC is unfinished the contours are becoming clearer and the PaN RIs are filling in the missing pieces for the PaN community with the help of PaNOSC and ExPaNDS.

Through PaNOSC and ExPaNDS, the PaN community has an active role in the building of the EOSC. A number of results are being developed and extended e.g. adoption of open data policies, search API, data analysis portal, data analyses services, jupyter notebooks and simulation services, data transfer services and sharing best practices. The production of FAIR data is well accepted by the RIs but a lot of work remains to be done to fully implement FAIR for all techniques. The user community is still need help adopting FAIR data best practices and will need training and documented advantages on why to adopt FAIR data practices. New types of metrics and acknowledgements are required to incite scientists to adopt FAIR faster. Some journals are helping by insisting on making data available with the publications. For data that are not published more effort is required to ensure they are reusable by scientists who were not part of the initial team.

The full potential of the collaboration amongst the science clusters still needs to be achieved. Common workshops and sharing of common tools will hopefully take place over the remainder of the projects and beyond. Data visualization, workflows and jupyter notebooks are emerging as common tools which could be shared.

FAIR data management and data services require additional work and higher standards compared to “un-FAIR” data but result in higher quality data and enable data services to be built and offered to users. CALIPSOplus JRA2 was a small but significant step towards FAIR data services. The next steps are being led by PaNOSC and ExPaNDS for the PaN community as part of the EOSC process. LEAPS and LENS will hopefully provide means for sustaining FAIR data and data services when PaNOSC and ExPaNDS end in 2023.

Author contributions

The document was the collective work of all the authors with specific sections being written by each of them. A.Ashton wrote the section on FAIR data. F.Schlünzen wrote the sections on FAIR facilities and ExPaNDS. D.Salvat wrote the section on Sustainability of Data Analysis Services. R.Dimper wrote the sections on Needs of the PaN community and EOSC-Future. A.Götz wrote the sections on Helix Nebula, EOSC, PaNOSC, and a first draft of the Conclusion. S.Egli re-organised and corrected the document. All editors participated in the reviewing and correcting of the document.

The authors acknowledge the feedback from the reviewer J-F.Perrin.

References

[1] “ExPaNDS/PaNOSC and CALIPSOplus Technical Coordination Workshop I - The Portal Architecture test experience”, http://www.calipsoplus.eu/wp-content/uploads/2021/01/D24-7_reported.pdf

[2] “Blueprint on implementing a DAAS platform”, Aidan Campbell (ESRF) et al, http://www.calipsoplus.eu/wp-content/uploads/2019/03/D24.2.pdf

[3] “The Photon and Neutron Open Science Cloud”, https://www.panosc.eu/

[4] “European Open Science Cloud (EOSC) Photon and Neutron Data Service.”, https://expands.eu/

[5] “Photon and Neutron Search Api”, Henrik Johansson et al, https://github.com/panosc-eu/search-api

[6] “Connecting open science”, https://www.openaire.eu/

[7] “Discovery service based on metadata steadily harvested from research data collections from EUDAT data centres and other repositories”, https://www.eudat.eu/services/b2find

[8] “Code related to the development of the PaNOSC Data Analysis as a Service portal”, https://github.com/panosc-portal

[9] “The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures“, https://projectescape.eu/

[10] "Jupyter project home page", https://jupyter.org/

[11] “Turn a Git repo into a collection of interactive notebooks”, https://mybinder.org/

[12] “R*eport on status, gap analysis and roadmap towards harmonised and federated metadata catalogues for EU national Photon and Neutron RIs*” , https://doi.org/10.5281/zenodo.4146818

[13] "*PaNOSC FAIR Research Data Policy framework*" , https://doi.org/10.5281/zenodo.3826039

[14] "OneData project home page", "https://onedata.org"

[15] “Open Clouds for Research Environments”, https://www.ocre-project.eu/

[16] “ARCHIVER: Archiving and preservation for research environments”, https://www.archiver-project.eu/

[17] “*Realising the European open science cloud”*, Directorate-General for Research and Innovation, https://op.europa.eu/s/oIDt

[18] “PaN training portal”, https://pan-learning.org/

[19] "*ESFRI cluster projects - Position papers on expectations and planned contributions to the EOSC*", https://doi.org/10.5281/zenodo.3675080

[20] "COVID-19 Data Portal", https://www.covid19dataportal.org/

[21] “LEAPS home page” , https://leaps-initiative.eu/

[22] “LENS home page”, https://www.lens-initiative.org/

[23] "*The Vienna declaration on the European Open Science Cloud*", https://eosc-launch.eu/declaration/

[24] "EOSC-Hub home page", https://eosc-hub.eu/

[25] “EOSC Working Groups”, https://www.eoscsecretariat.eu/eosc-working-groups