

Proposal name	PaNOSC
Proposal full name	Photon and Neutron Open Science Cloud
H2020 Call	INFRAEOSC-04-2018
Coordinator	Andrew Götz (andy.gotz@esrf.fr)
Coordinating organisation	ESRF

List of participants

Participant No	Participant organisation name	Country
1	European Synchrotron Radiation Facility (ESRF)	France
2	Institut Laue-Langevin (ILL)	France
3	European XFEL (XFEL.EU)	Germany
4	The European Spallation Source (ESS)	Sweden
5	Extreme Light Infrastructure Delivery Consortium (ELI-DC)	Belgium
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	Netherlands

Table of Contents

Executive summary.....	3
Introduction.....	4
Who.....	4
What.....	4
Why.....	5
How.....	5
Vision and Mission	6
1. Excellence.....	7
1.1 Objectives	7
1.2 Relation to the work programme	7
1.3 Concept and methodology	10
1.3 (a) Concept.....	10
1.3 (b) Methodology	12
1.4 Ambition	18
2. Impact	19
2.1 Expected impacts	19
2.2 Measures to maximise impact.....	22
2.2 (a) Dissemination and exploitation of results	22
2.2(b) Communication activities.....	23
2.2.(c) Project internal communication strategy	24
3. Implementation	25
3.1 Work plan -Work packages, deliverables	25
3.1(a) Work Packages	25
3.1(b) Work Package descriptions	28
3.1(c) PaNOSC Deliverables	47
3.2 Management structure, milestones and procedures	50
3.2.1 Project management methodology.....	50
3.2.2 Roles and organisational structure	50
3.2.3 Decision-making process and project control.....	51
3.2.4 Milestones	53
3.2.5 Risk management strategy	54
3.2.6 Risks.....	55
3.3 Consortium as a whole.....	58
3.4 Resources to be committed	61
References.....	66

Executive summary

The Photon and Neutron Open Science Cloud (PaNOSC) is a cluster which proposes to align the efforts of the existing and new photon and neutron sources to link up to the EOSC. The cluster is composed of the ESFRI roadmap research infrastructures: European Synchrotron Radiation Facility (ESRF), Institut Laue-Langevin (ILL), European X-ray Free Electron Laser (XFEL.EU), Extreme Light Infrastructure (ELI), The European Spallation Source ERIC (ESS), and the Central European Research Infrastructure Consortium (CERIC-ERIC), which to a large extent have a shared user community. The consortium also includes the EGI Foundation (EGI.eu) e-infrastructure as partner. PaNOSC is coordinated by the ESRF.

Research Infrastructures (RIs) are an essential part of the European scientific infrastructure. They produce ever increasing quantities of scientific data and belong to the top data producers in the scientific landscape. Analysing huge volumes of data is more and more challenging for scientists using these facilities. The current model of exporting the data to the scientists' home institute and then reducing and analysing the data does not fit any more. The European Commission (EC) has recognised this growing problem and has proposed to combine the resources from Pan-European e-infrastructures (e.g. EGI, GÉANT, EUDAT, etc.) which offer compute, storage and networking services with the data produced by the RIs into a coherent entity called the European Open Science Cloud (EOSC). The EOSC is first of all a concept which now needs to be converted into a working solution. To accelerate its implementation the Commission has called upon the RIs which are on the European Strategy Forum Research Infrastructures (ESFRI) roadmap to engage with the e-infrastructures to link the data they produce to services for treating the data together to form the first implementation of the EOSC. In order to make the data available to everyone the data has to comply to the principles of being Findable, Accessible, Interoperable and Reproducible (FAIR).

The experience of the existing RIs implementing open data policies and curating data will be shared with the new RIs coming online to fasttrack the adoption of data curation and stewardship practices aligned with FAIR data principles. The proposal will enable existing data policies to be updated to be compatible with the latest interpretation of what FAIR data means. New RIs will profit from the updated data policy as well as implement data curation and stewardship across their sites. Data alone is not enough, scientists need programs and services to analyse them. PaNOSC will improve and package new generic data services based on open source tools like Jupyter notebooks, desktop applications and workflow tools for simulation and modelling for all partners and made available for all via the EOSC service catalog provided by EGI. The simulation and modelling services will be enhanced to make them easily accessible for a wide range of experiments. As a result of PaNOSC, huge quantities (Petabytes) of data produced by the RIs in the cluster will be curated correctly and made available as open data in compliance with the FAIR data principles. Existing conventions for the PaN community's NeXus standard for metadata will be enhanced and extended to new fields: lasers and X-ray Free Electron Lasers. The proposal will be innovative and ambitious in creating a new class of users of open data produced at photon and neutron sources. These users will be virtual users who will never come to the RIs but instead they will use the data remotely for creating new insights and doing inter-disciplinary research. PaNOSC will use the e-infrastructures' compute and storage resources to cater for this new class of users thereby contributing to the success of the EOSC. Although PaNOSC represents only the large Pan-European facilities of the larger photon and neutron community in Europe it is not working in isolation. All the results and outcomes will be shared with the national photon and neutron RIs so they can profit from the building of the EOSC. This dissemination is further facilitated by the fact that 24 European countries are members of at least one of the PaNOSC RIs.

Introduction

Who

Who - Photon and Neutron Open Science Cloud (**PaNOSC**) made up of photon and neutron RIs on the ESFRI Roadmap and ERICs (ESRF, ILL, XFEL.EU, ELI, ESS, CERIC-ERIC), Pan-European e-infrastructures (EGI -as a partner-, GÉANT and EUDAT and OpenAIRE) and national RIs and PRACE host members as observers.

European Photon and Neutron sources and their instruments are essential tools for scientists in a plethora of fields (biology, materials, chemistry, technology, nuclear physics, pharmacology, high-energy physics, cultural heritage ...) and their industrial counterparts and as such they are essential components in the European Research Area in addressing at least five out of the seven the Societal Challenges prioritized by the European Commission in H2020¹. Research Infrastructures (RI) are part of the ESFRI roadmap and range from well-established facilities operating for decades up to newer sources opening now or in the near future. We have decided to unite our strengths within PaNOSC in order to build and integrate the EOSC. We not only want to provide open data and a coherent set of services for exploiting them by our existing user community (mainly experimenters), but also reach a wider public and a new level of coherency amongst the whole scientific ecosystem. This includes integrating data scientists into this network. Our RIs, which represent billions of Euros of public investment, strongly believe in the benefits of open science to maximise their scientific and societal impact. Ten years ago, the PaNData-Europe project shaped a common Data Policy framework for analytical facilities, based on FAIR principles. Building on this pioneering work, nearly all members of PaNOSC have today implemented a Data Policy. Our organisations have put in place specialised units with staff dedicated to data management and steering committees to develop our ideas and tools for our users. We are eager to increase our scientific impact by extending data services to a larger audience and integrating them into the bigger vision of the EOSC.

What

What - PaNOSC provides petabytes of curated data from thousands of applied and basic science experiments annually and the analysis software for many fields ranging from materials and life sciences to cultural heritage and palaeontology.

The Photon and Neutron community **provides curated Open Data** from a large variety of experiments. We are curating PetaBytes (PBs) of data, yet few of them are currently openly accessible, but starting from early 2018, the initial embargo period for the first institute (i.e. ILL) implementing open access will end and more and more data is going to be open. Each facility has its own data catalogue to **provide access to experimental data sets, metadata and documentation**. These catalogues all include curated metadata, search indexes and download mechanisms. Any scientist can search for experimental data and request access to the data if still under embargo or download them if not under embargo anymore. Our community is known for their **rich collection of scientific software** encompassing data reduction, analysis and simulations. The PaN-Data software catalogue <https://software.pan-data.eu> collects nearly a hundred applications with information, links, and examples data sets. Our facility users have free access to **scientific support and expertise**. Moving our organisations and users to Open Science is an ongoing and time consuming effort but very rewarding with tangible results. Sharing our experiences with new players and exchanging with other stakeholders will help to create a culture of Open Science in the many domains which are served by the Photon and Neutron communities.

¹ The five challenges addressed by the PaN community are: 1) Health, demographic change and wellbeing; 2) Food security, sustainable agriculture and forestry, marine and maritime and inland water research, and the bioeconomy; 3) Secure, clean and efficient energy; 4) Smart, green and integrated transport; 5) Climate action, environment, resource efficiency and raw materials; (<https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges>)

Why

Why - unify the fragmented research data landscape with a common data policy framework and coherent set of services for data at the RIs and through the EOSC to give scientists better tools to fully exploit their data and facilitate the use of open data.

The EOSC offers a perfect opportunity to overcome the **fragmented research data landscape**. PaNOSC aims to harmonise, optimise and advance the use of data produced at our sites.

As demonstrated by our regular user surveys (2012, 2014, 2016²), scientists are increasingly using several different infrastructures to perform their research. Harmonisation of metadata, interoperability of services and standardisation has become of the utmost importance. Making data from different infrastructures inter-operable, not only simplifies and improves the experience of our users, but also opens up access for new groups of academic and industrial users, with no or little previous experience in X-Ray or Neutron techniques.

While the foundations on data policies were laid in 2012 their implementation has been slow in coming. Consequently, **Open Access** is still in its infancy in the field of photon and neutron research and therefore needs special services and training actions to succeed. PaNOSC will not only offer new data services and tools, but also strongly help to encourage a **new data culture** amongst researchers and supporting parties (IT, librarians, communication, etc.) by communication and training. How do we measure the success of Open Science? Today, we face a clear **lack of useable metrics** that provide indicators and arguments for sustaining efforts. Our role is to integrate current efforts and provide the necessary information to make this happen. Today, sources generate increasingly large **data volumes** and the users are therefore confronted with the challenge of analysing and handling such large data volumes. The European Open Science Cloud and our project PaNOSC will help address this challenge by providing services around the data at the source or in the EOSC cloud thereby lifting the burden of handling large data volumes from the user.

How

How- PaNOSC will build on the experience with Open Data policies from PanData and existing metadata catalogues, extend existing Jupyter notebook services and remote desktop, generalise simulation, link data and services to EOSC.

PaNOSC will build on top of the existing local metadata catalogues and data repositories to provide federated services for making data easily Findable, Accessible, Interoperable, and Re-usable (FAIR). Extracting the scientific value of the experimental data produced in our RIs is not always an easy task. The raw data tends to be larger and larger and quite often requires special skills for being correctly exploited. PaNOSC will develop and provide data analysis services to overcome these difficulties. The services will include notebook (Jupyter based), remote desktop/applications and containers/VMs. These services will be provided locally and will also be made available on the EOSC for general use. These data analysis services will offer a single entry point to the data, the software, the IT capacity and the necessary scientific support. All these services will be fully integrated into the EOSC catalogue, in terms of discovery, accessibility, and user authentication/authorisation, Service Level Agreement (SLA), support and accounting. The PaNOSC cluster will also help introduce a new data culture to the user community – via training and workshops on scientific data management and publishing practices. Best practices in data stewardship will be shared with other laboratories at least within the PaN community. Experiences, trials and results will be shared openly via publications and meetings. The positive experience of implementing an Open Data policy will help convince other PaN institutes, which are still struggling with adopting the FAIR principles, to do the same.

² <http://pan-data.eu/node/105>

Vision and Mission

Vision- PaNOSC will help fastrack new Photon and Neutron facilities, improve data services for existing users, create a new class of virtual users of open data, help build and integrate the PaN RIs to the **EOSC**.

EOSC is a vision of research data Commons based on FAIR principles which “must be pragmatic and technology-neutral, encompassing all four dimensions: findability, accessibility, interoperability and reusability. FAIR principles are neither standards nor practices. The disciplinary sectors must develop their specific notions of FAIR data in a coordinated fashion and determine the desired level of FAIR-ness. FAIR principles should apply not only to research data but also to data-related algorithms, tools, workflows, protocols, services and other kinds of digital research objects”. The mission of PaNOSC is to contribute to the realization of the data Commons for Neutron and Photon science, making it a real and a day-to-day tool for the many scientists from all the many existing and future disciplines using data from Photon and Neutron sources. By collaborating with the EOSC-hub project, PaNOSC will contribute to the EOSC Hub, the Service Integration and Management system (SIAM) accountable for ensuring that all EOSC service and data providers perform to provide a seamless service that is compliant to obligations and align services and data to defined policies and standards.

1. Excellence

1.1 Objectives

Objectives

1. **Participate** in the construction of the EOSC by linking with the e-infrastructures and other ESFRI clusters.
2. **Make** scientific data produced at Europe's major Photon and Neutron sources fully compatible with the FAIR principles.
3. **Generalise** the adoption of open data policies, standard metadata and data stewardship from 15 photon and neutron RIs and physics institutes across Europe
4. **Provide** innovative data services to the users of these facilities locally and the scientific community at large via the European Open Science Cloud (EOSC).
5. **Increase** the impact of RIs by ensuring data from user experiments can be used beyond the initial scope.
6. **Share** the outcomes with the national RIs who are observers in the proposal and the community at large to promote the adoption of FAIR data principles, data stewardship and the EOSC.

The objectives will be measured using Key Performance Indicators (KPIs listed in table 3) including DOIs, citation indexes, numbers of downloads, searches, users of the data services etc.

The large variation in the maturity of the cluster participants is reflected in the varying degrees of implementing FAIR data principles and data services. PaNOSC is a unique opportunity to share the know-how, solutions and experience between those participants who are implementing FAIR data principles and those who are just starting to do so. It is also a unique opportunity to link up to the EOSC hub and e-infrastructures (EGI, EUDAT, PRACE and GÉANT at least) to further develop and provide innovative services for data. The data services provided will allow scientists who already use one or more of the sources in the cluster to get more out of their data. It will also allow new users to find data they did not have access to up until now and use it for cross-disciplinary research. The PaNOSC project comes at the right time for the PaN community to provide FAIR data. A number of the participants have already adopted Data Policies which are FAIR but their uptake is still in its adolescence. PaNOSC will share these principles across the participants including the necessary training for scientists to get maximum advantages out of making and using open data. The landscape of scientific research data is changing and the PaN community needs to evolve its practices to be compatible with this changing landscape and in common with other ESFRI clusters.

The main objectives can be achieved during the duration of the project with help of the EOSC which will emerge at the same time. The EOSC-hub (www.eosc-hub.eu) led by EGI.eu will help providing the underpinning services for many of the infrastructure services required to achieve the objectives of PaNOSC. This is a unique opportunity for PaNOSC and EOSC to succeed.

1.2 Relation to the work programme

The PaNOSC proposal addresses the challenges of the INFRAEOSC-04 call. The PaN community has been following the evolution of the European scientific research landscape and due to its level of maturity, it has already identified and started to address many of the challenges in the call. A good example is the PaNdata Data Policy³ which was a deliverable in 2011 of the FP7 financed project PaNdata-Europe and which has been adopted and is being implemented by ILL, ESRF, European XFEL, and ESS. These facilities' specific data policies are compliant with the PaNdata Data Policy. The PaNdata policy was already very closely aligned with the FAIR principles even before they were published in 2016 ([Wilkinson2016]). Thus, the PaN community already has experience implementing PaNdata-like data policies and will build on this experience during PaNOSC to successfully address

³ <http://pan-data.eu/sites/pan-data.eu/files/PaN-data-D2-1.pdf>

the specific challenges raised in this call to link FAIR data to the EOSC hub and meta catalogues. The following tables demonstrates on a point by point basis that PaNOSC fully addresses the objectives of the INFRAEOSC-04 call.

INFRAEOSC-04 call scope	PaNOSC proposal task / action
<i>This topic will ensure the connection of the research infrastructures identified in the ESFRI Roadmap to the EOSC.</i>	PaNOSC is defined to do exactly that. Interoperable services for 5 facilities on the ESFRI Roadmap + 8 institutes part of an ERIC are developed in WP3-5 adhering to data policies aligned in WP2 while the actual integration with EOSC takes place in WP6. WP3: Users will be able to access their own and publicly available data across all facilities, and be able to search on meta-data across data sets from all involved partners from a single entry point (Task 3.4). WP4: Users will be able to process data using interactive and scripted data analysis services remotely; choosing from a set of prepared analysis recipes and environments. WP5: Users will be able to simulate their experiments before and model the results after the experiment WP8: A shared e-learning service and training material linking the research infrastructures will be developed.
<i>Support to this activity will be provided through cluster projects gathering ESFRI projects and landmarks in each of the following large thematic domains: Biomedical Science, Environment and Earth Sciences, Physics and Analytical Facilities, Social Science and Humanities, Astronomy, Energy. While the ESFRI infrastructures represent the core component of any cluster, other relevant world class research infrastructures with a European dimension, established as ERICs or International Organisations, can also be involved in a cluster.</i>	The project creates a cluster of Physics and Analytical Facilities composed of all existing ESFRIs for materials science (i.e. the ESFRI landmarks European XFEL, ELI, ERSF, ESS, and ILL) and the CERIC-ERIC, which is an umbrella organization of national facilities in materials science in eight countries. Thus all partners involved in the cluster are either ESFRIs and / or ERICs. The cluster includes the Pan-European e-infrastructures EGI (as a partner) and GÉANT as EOSC implementers and providers of IT resources.
<i>Each infrastructure should participate to only one cluster.</i>	None of the 6 involved RIs participates in other clusters. EGI is an e-infrastructure and as such can participate in more than one cluster.
<i>Proposals will address the stewardship of data handled by the involved research infrastructures according to the FAIR principles and in line with the objectives of Open Science.</i>	This is specifically addressed in WPs 2 and 3. Moreover, data generated from experiment simulations (i.e. from virtual experiments, WP5) and associated metadata will be handled, curated, and cataloged in the same way as real experimental data and will thus adhere to data policies developed in WP2 and exploit services developed in WP3. Secondly, WP5 will produce datasets to serve as test data for data handling procedures adhering to the FAIR principles and objectives of Open Science as well as for data analysis services developed in WP4.
<i>This will include the definition of domain specific data policies (e.g. acquisition, deposit, curation, preservation, access, sharing and re-use),</i>	All the data policies currently published by the Photon and Neutron facilities are based on the PaNData experimental data policy framework resulting from the work done in the PaNData-Europe project. We will revisit this work to take into account the new needs (data resulting from analysis, new licence model, ...) and align it with EOSC activities on data policy harmonization.
<i>addressing any legislative or interoperability issues which affect data handling across geographical and discipline borders,</i>	PaNOSC shares the same user community and will address legal issues e.g. conformance with GDPR, in the revised data policy. Ensuring interoperability is addressed in WP3-5 through standards for data formats and APIs. Legal and

	interoperability issues affecting data handling across geographical borders is of particular importance for the distributed infrastructures ELI and CERIC-ERIC.
<i>as well as the development of appropriate tools for depositing, curating and analyzing data.</i>	WP3 addresses the development of tools for data deposition and curation. Analysis services are specifically developed in WP4 and WP5 with WP4 handling traditional data processing for real experiments, whereas WP5 considers services for materials modelling and experiment simulation to be used for experiment planning as well as analysis. Services developed in WPs 4 and 5 are prime use cases for EOSC because they exploit seamless access to data repositories as well as HPC resources for data processing and analysis.
<i>Research infrastructures will have to expose their data and tools under the EOSC catalogue of services and take all the necessary steps to ensure that the used repositories are compliant with the FAIR principles. In doing so proposals should develop synergies and complementarity in data handling between research infrastructures, optimise technological implementation,</i>	PaNOSC will share the data catalogue solutions and will share a common API to expose all data catalogues. Data services for data reduction, analysis, simulation and modelling will be shared to ensure the scientific community can obtain the most value from the experiments performed at RIs.
<i>and ensure integration and interoperability of data and tools within the EOSC.</i>	EGI is a partner in PaNOSC, leading the EOSC-hub project (www.eosc-hub.eu), and works with EUDAT to ensure integration in the EOSC hub. WP5 integrates access to data with modelling and simulation and, via WP4, data analysis tools. The PaNOSC data services will be hosted by the partner sites and the EOSC hub so that scientists can access the services via either system transparently. The PaNOSC data repositories will be linked to the OpenAIRE meta catalogues. GÉANT commits to help us refining our AAI requirements and deploy a sustainable solution to secure the integration of our services in EOSC. Moreover, three PRACE host members are observers (see Letters of Support from CSCS, JSC, and CINECA) facilitating integration with PRACE's federated infrastructure Fenix, currently under development, on a longer term.
<i>Proposals may address the development of domain specific skills for data stewardships and the specific training of research infrastructure staff.</i>	PaNOSC addresses training of RI staff in WP8. This will be made available to the PaN community so that other RIs not part of PaNOSC will also profit from this training and can adopt the same data stewardship practices.
<i>Activities should contribute to a faster adoption of best practices</i>	WP2 will update the community data policy framework. Workshops on best practices in data stewardship will be held as part of WP8 for the PaNOSC partners and PaN community. WP7 will study and propose solutions to sustain Open Data.
<i>and foster the use of open standards and interoperability in data</i>	WP3 specifically addresses the NeXus format which is the common format used by the PaN community. PaNOSC will extend the format for new sources and applications (FELs and lasers). WP5 will adhere to the open community standard with the NeXus data format combined with the openPMD metadata standard implemented and developed in WP4 (SIMEX) of the Horizon 2020 project EUCALL. The NeXus format will also be employed to describe physical parameters (e.g. experimental geometry) and numerical/algorithmic parameters // and publishing the specifics for virtual experiments. This in turn will foster interoperability in data processing and simulation

	workflows. The training activities in WP8 and the dissemination activities in WP9 will promote the open standards to the PaN community at large.
<i>and computing services.</i>	PaNOSC will package common software in WP4 and WP5 so that it can be deployed easily at the participant's sites and on the EOSC infrastructure. PaNOSC has budget reserved for procuring cloud resources for scaling out data services. Moreover, the APIs defined in WP3-5 will ensure the inter-operability of services developed in these WPs.
<i>Consortia should include key participants of the involved infrastructures and/or the infrastructure legal entities as well as other partners needed to address the challenges or develop the required solutions.</i>	Staff in the IT departments, scientific computing groups, scientists and administration of the involved infrastructures have been identified as key participants as described in Section 4. Moreover, several national facilities will follow the project as observers ensuring knowledge transfer to and from these facilities (see Letters of Support).. Finally, PaNOSC will work with several e-infrastructures in order to implement solutions and integrate with the EOSC (see below).
<i>Proposals should build upon the state of the art in ICT and e-infrastructures for data, computing and networking and work in cooperation with e-infrastructure service providers.</i>	PaNOSC is working with EGI and EUDAT for integrating their services in the EOSC hub. Both EGI and EUDAT are Pan-European e-infrastructure service providers and are members of the EOSC-hub project (www.eosc-hub.eu), with EGI leading it. The Pan-European research network provider GÉANT, who is leading the AARC project (https://aarc-project.eu) has committed to work with us to ensure sustainability and compatibility of our AAI community solution inside EOSC. As previously mentioned, three PRACE host members are participating as observers (see Letters of Support from JSC, CINECA, and CSS). Moreover, PaNOSC partners run state of the art ICT infrastructure and are actively engaging with commercial cloud providers directly or via H2020 projects like HNSciCloud (www.hnscicloud.eu).

Table 1: Proposal relation to the work programme

1.3 Concept and methodology

1.3 (a) Concept

Making data from Photon and Neutron facilities available to the EOSC offers significant opportunities as well as challenges – due to the amount of data available and expected to be obtained in the future. Three of the partners (CERIC-ERIC, ELI, XFEL.EU) in the consortium have only recently gone online and need to learn from the experienced partners about what to put in place and how to provide data services for large volumes of data. The concept of PaNOSC is to share the experience and know-how of the experienced partners with the new sites so that a large number of PaN sites can be integrated into the EOSC from an early stage for the benefit of the user community. The same concept can be applicable to the other national RI PaN s who are observers in the PaNOSC project.

To be able to make data available to the EOSC cloud, one needs agreement from facilities and researchers on how data can be shared. These are complex issues involving technological and social tasks, which is further constrained by historical expectations and metrics that do not reward data sharing, but drive the ambition to extract the fullest value out of investments into the research facilities. **WP 2** is focused on Data Policy and Stewardship to address this question.

To make the data findable, some kind of table of contents is required. While this is occasionally available at research facilities, we need to integrate and to some degree unify the different data sets to make such a data catalogue useful across all data contributing facilities and for the shared user community. **WP 3** is on data catalogue services.

FAIR - PaNOSC will comply with the FAIR principles in the following ways:	
Findable	- all data will have a doi, rich metadata, common api for federated search
Accessible	- api will support open protocol, metadata accessible even without data
Inter-operable	- metadata to follow community standards (NeXus), register metadata
Reusable	- follow community standardise metadata, clear licence (CC-BY)

To make the data usable and re-usable we need to understand that the experiment data recorded (sometimes called 'raw data') needs significant post-processing before knowledge can be extracted. In part, this relates to 'calibrating' the data from complex detectors, and in part to converting recorded and somewhat facility and hardware specific data into data representing the science observed in the experiment. This is referred to as data analysis in the community, and we dedicate **WP 4** to this task.

As part of the experiment planning, execution and post-experiment data analysis, computer simulations of the experiment have become increasingly important: they allow to optimise the experiment in advance (thus allowing to record and store only high-value data) to save precious beamtime, they help obtain unknown parameters from the experiment, and in analysing the data. **WP 5** is focused on this activity.

Our driving ambition is to make data, metadata and the required analysis and simulation services available through the European Open Science Cloud (EOSC). It's worth noting that the size of raw data sets from photon and neutron facilities can exceed hundreds of TB for a single run, and in those cases data are very difficult to move from one storage place to another. This affects multiple work packages. **WP 6** is dedicated to the integration of the above with the EOSC.

One of the big challenges of making data and services available to the community at large is the question of sustainability. How can the PaNOSC and other EOSC activities sustain providing new data services to the existing users and new communities of virtual users? The data services for reducing, simulating and modelling data have not been offered to a wide and potentially unlimited audience in the past, and a sustainable funding and access model needs to be found; also taking into account policy developments at European and national levels. **WP 7** will explore and propose how to sustain these new data services.











Although data producers agree that the principles of FAIR data are a good match and even essential for scientific data the scientists themselves are not always aware of these principles nor do they know how to use them on a daily basis e.g. in citations in articles. Experience from the partners who already have open data policies in place shows there is a strong need to train scientists on how to make their data FAIR (by providing rich metadata for example) and how to use FAIR data. Similarly, the new data services to be provided by PaNOSC like remote Jupyter notebooks will require training and explanation on how to use. **WP 8** is entirely dedicated to setting up an e-learning platform (PaNOSC will reuse e-neutrons.org from the Horizon2020 project SINE2020) and developing and disseminating training material for both staff and users of the RIs

An essential part of PaNOSC is communication of the results and outcomes to the other national RIs in the PaN community and sharing and exchanging with the other clusters in this call on connecting to the EOSC. Regular meetings with the other clusters are planned including a report of their common experience. These tasks will be handled by **WP 9**.

PaNOSC is part of a history of projects in the PaN and European community for addressing the challenges of Open Data, handling of Big Data, data stewardship and in general improving the data management in scientific research. For this reason PaNOSC builds on the outcomes of previous projects especially the **PaNdata-ODI** and **PaNdata-Europe** projects. It will establish close links with other H2020 project which are closely related which are **Calipsoplus**, **SINE2020**, and **Brightness**. PaNOSC will work closely with **EOSC-hub** (led by one of the partners EGI.eu) in order to link to the EOSC. In addition, close links will be established to **OpenAIRE-Advanced** whose mission it is to support the Open Access/Open Data mandates in Europe.

1.3 (b) Methodology

The methodology of the PaNOSC proposal is based on the following steps:

1. **Use** the expertise of the experienced partners in the cluster (ILL, ESRF, XFEL.EU) who have implemented solutions for data management and stewardship at TRL8 and TRL9 levels to bring the other partners (ELI, CERIC-ERIC and ESS) up to the same level

2. **Generalise** data policies and stewardship at all partner sites so that the data and metadata can be curated and archived and be made open (after an embargo period)

3. **Standardise** metadata by following community standards like NeXus. Enhance existing standards where necessary via the community decision making procedure for doing so.

4. **Federate** all metadata catalogs so they can be searched from one portal and harvested by the EOSC portals (OpenAIRE-Advance and EUDAT)

5. **Implement** innovative data services based on Jupyter notebooks and desktop applications as TRL9 services which can be run locally at each partner site (for datasets too big to transport) and on the EOSC

6. **Fully** integrate these services into the EOSC as part of the EOSC applications portal and use the EOSC compute resources reserved for PaNOSC

7. **Develop** with GÉANT a federated Authentication and Authorization Infrastructure (AAI) solution compatible with the EOSC and the PaN community solution UmbrellaId which allows users to access the services and data

8. **Make** the current TRL8 innovative services for doing simulation and modelling available locally and on the EOSC at TRL9.

9. **Share** the results with the photon and neutron community via multimedia material, deliverables and workshops for data stewards

10. **Provide** training and training material for staff and users on how to use the data services and for data stewardship

11. **Provide** a business plan on how to sustain the services in the EOSC and locally

The diagram below illustrates how the services are interlinked and which work package they are dealing with. The labels in the figure 1 indicate in which work package task the respective element will be developed and exposed as a cloud service. Some elements, such as the Data Catalog, standardized data formats and metadata standards will be adopted from WP 3 deliverables and existing open standards

[NeXus: Maddison1997, openPMD: www.openpmd.org, NOMAD: www.nomad-coe.eu]. Development of integrated services for experimental workflows (e.g. as represented in the top-row in figure 1) is addressed in WP 4. Data and simulation services developed will adopt development recommendations and guidelines for scientific software development as defined in the Horizon2020 Project SINE2020 [Markvardsen2017].

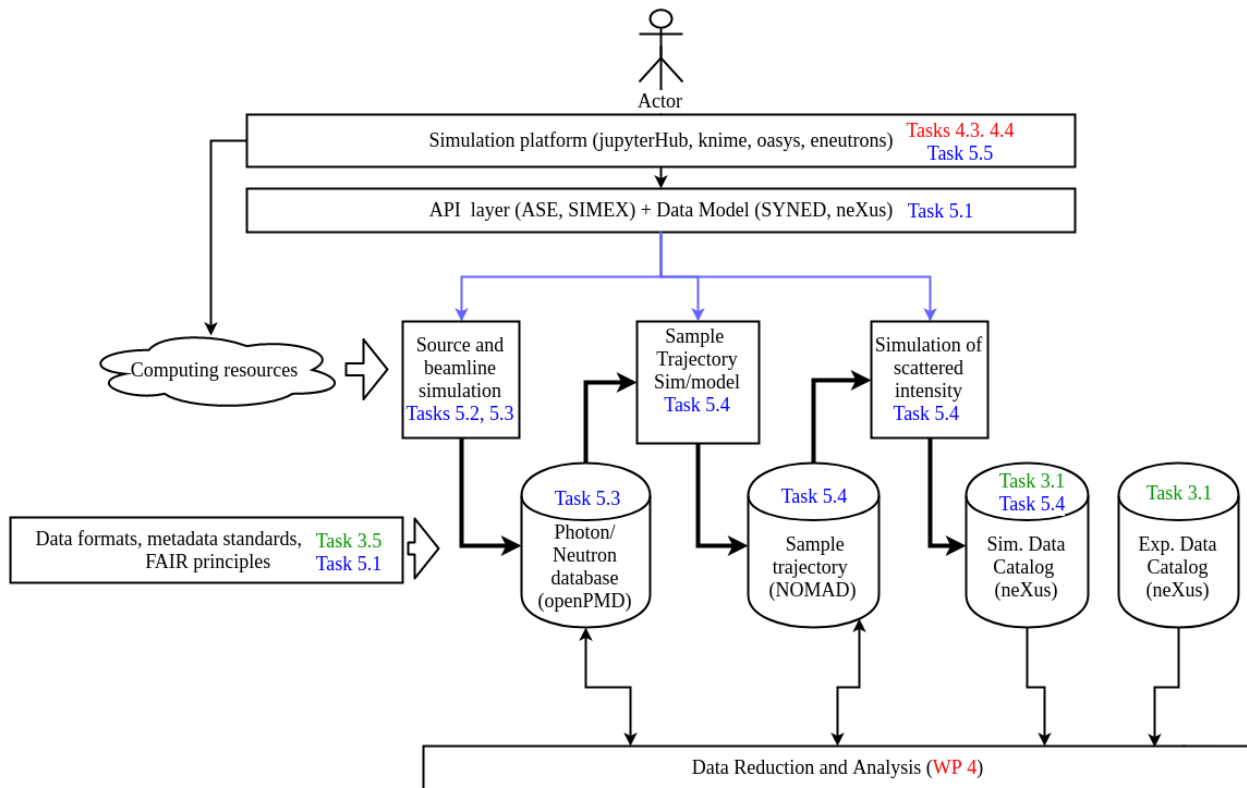


Figure 1: Relation between Data Catalog, Data Analysis, and Data Simulation services developed in PaNOSC.

The following sub-sections describes the data services which will be developed and exposed via the EOSC.

1.3.b.1 Data analysis services in the cloud

Research Infrastructures are tasked with providing data analysis services for their users and provide tools and support to convert raw detector data into reduced data ready for scientific analysis.

Figure 2 shows the data processing pipeline for Single Particle Coherent Diffractive Imaging at XFEL.EU. After calibration and hit finding, 2D megapixel images that typically contain only a few photons are converted into a sparse format storing only the pixel positions with photon counts above 0. The sample structure is then resolved through iterative reconstruction algorithms.

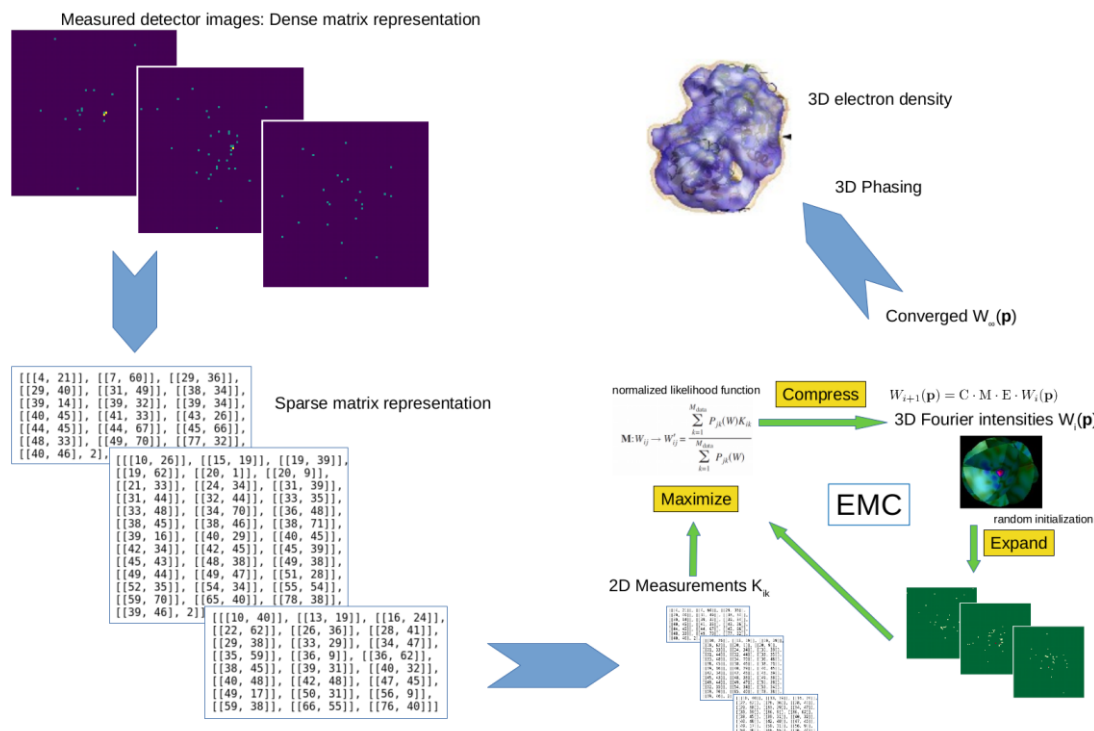


Figure 2: Typical data reduction workflow for Single Particle Imaging Analysis at XFEL.EU

Depending on experimental techniques, scientific instruments and the user community there is a variety of existing and locally used data analysis services across the partner sites and science domains. Data tools used range from command-line based access to the computing machines, to interactive graphic consoles to web notebooks such as the Jupyter notebook. The compute resources used are typically provided by each facility or for smaller computational tasks local desktops and workstations.

Most of the software packages are registered in the Photon and Neutron software catalog - <https://software.pan-data.eu/>. The catalog currently has 98 entries each with a description, documentation (when available), how to use the software and where to get support. The catalog is an essential part of the data services to be provided by PaNOSC. Entries will be enhanced by adding downloadable packages and linking the catalog to the EOSC applications database.

The current remote analysis services used in production facilities fall broadly into four different models:

- 1) **Model 1** - command line based access: Facility users connects using a remote terminal protocol (nowadays secured, SSH) through firewalls to internal facility resources (workstations or HPC installations).
- 2) **Model 2** - remote desktop based access: Many of the software tools used today for analysing data in existing scientific facilities have a Graphical User Interface (GUI) frontend and run on a personal computer or HPC resources. Recent progress in remote display technology has opened the possibility to develop web based access to remote desktops, so that such applications with, for example a Qt interface, can be controlled remotely. These have been used by some facilities (e.g. ILL) to improve security, user friendliness and efficiency in giving direct access to remote data analysis services to scientists.
- 3) **Model 3** - notebook based access: Web based documents (like the Jupyter notebook) give us the possibility to integrate scriptable and command line oriented analysis software (as opposed to graphical software interactively driven by users) into web based services. Combined with containerised execution environments, this model improves the reproducibility of analysis and is growing in popularity.
- 4) **Model 4** - web services: Web services represent a standard way of accessing resources programmatically on the web. They fall into two main classes - REST-compliant web services and arbitrary web services.

Model 1 is already implemented therefore we propose to work on model 2 and 3 in this programme (**WP 4 and WP 5**).

PaNOSC will provide remote desktop (Model 2) and Jupyter notebook (Model 3) simulation and modelling (Model 2 & 3) and metadata catalogue (Model 4) data services as part of the **EOSC**

Model 2 can provide the same interface and functionality that is available locally to remote users, and has no restrictions in applicability: GUI-driven and command line driven applications can be run remotely (including software that can only run on the Windows operating system): the ssh-based access described in model 1 is a subpart of the functionality provided with model 2.

Model 3 shows great potential, in particular for scriptable analysis tools and those for which an interface is available in Python or some other programming language. Combined with containers, this is a modern and relatively lean approach of bundling analysis instructions and compute environment to, for example, accompany raw-data or publications. We see a trend in computational science beyond photon and neutron science that analysis tools start to offer (Python-based) libraries through which they can be controlled, and thus can be natively controlled and integrated into Jupyter Notebooks. While this notebook approach is not suitable for all existing analysis tools, we predict that the momentum of the notebook ecosystem and python-based data libraries and tools will turn the notebook approach into an important and effective technology in the future.

Model 4 represent a big advantage for applications to access resources over ad-hoc scripts. Nonetheless they necessitate a lot of development effort if the application was not designed to access resources on the web via web services. PaNOSC will use web services for exposing and accessing metadata and data catalogues. Data analysis applications on the other hand are mostly desktop based and would require a huge effort to rewrite them to use web services. Rewriting applications is not in the scope of PaNOSC therefore they will be made accessible via Model 2 i.e. remote desktop.

1.3.b.2 Key technology: Jupyter Notebook and ecosystem

A key technology are executable notebooks and the associated ecosystem from Project Jupyter [Perez2015]. Project Jupyter [Jupyter2018] is a set of open source software projects for interactive and exploratory computing and data analysis. These software projects help make data analysis, data science and scientific computing reproducible and multi-language (Python, Julia, R, Haskell, Bash, ...). The main component offered by Jupyter is the Jupyter Notebook: a web-based interactive computing platform that allows users to create data- and code-driven narratives that combine live (re-executable) code, equations, narrative text, interactive dashboards and other rich media.

The figure 3 on the next page shows a Python-based sample session in Jupyter notebook. Through a menu command or a command line interface, notebooks can be converted to widely used read-only formats such as latex, html or pdf: the sharing of notebooks (as a logbook of a simulation or analysis that has been carried out) with collaborators is thus practical and effective.

The Jupyter notebook is being used widely as a data science tool in academia, and industry (including for example University of California, Berkeley, Stanford, MIT, Harvard, Cambridge, Oxford, Google, IBM, Facebook, Oracle, Otto Group, Microsoft, Bloomberg, JP Morgan, WhatsApp) but also at e.g. the Spallation Neutron Source, Oak Ridge National Laboratory, US, where the user feedback has been very positive. Because the architecture and building blocks of Jupyter are open, they are used to build numerous other commercial and non-profit products and services. The Jupyter Notebook has at least 2 million individual users worldwide [Perez2015, page 10].

Jupyter notebook documents provide a complete and executable record of a data analysis that can be shared with others in a way that has not been possible before [Kluyver2017]. Furthermore, Jupyter has been designed with the aim to optimise a data analysis environment for the most expensive resource involved: human time. Finally, it is a tool designed by active users of the tool, ensuring there is no divergence between the tools design and requirements of the science users. These points, among other things, have led to a huge boost in reproducible, interactive research and education documents in recent years. A paradigm that Fernando Perez, creator of the project, has referred to as “literate computing” [Perez2017]. Jupyter notebooks are particular suitable for tutorials on data science related topics and will therefore also be integrated in the e-learning platform as part of the PaNOSC project (WP8).

Code cells show code input and output:

```
In [1]: 1 + 2
```

```
Out[1]: 3
```

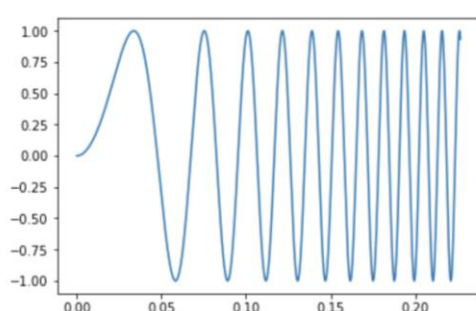
Cells can contain text and latex equations such as $f(x) = \sin(2\pi\omega t^2)$ and $\omega = 220$ Hz. We can use code to define the corresponding functions:

```
In [2]: import numpy as np
def f(t):
    omega = 220
    return np.sin(2 * np.pi * omega * t**2)
```

Let's compute the data and plot the beginning of it:

```
In [3]: t = np.linspace(0, 2, 44100)
y = f(t)
## Show plots inside the notebook
%matplotlib inline
import pylab
pylab.plot(t[0:5000], y[0:5000])
```

```
Out[3]: [<matplotlib.lines.Line2D at 0x1047ada58>]
```



We can integrate media: images, videos, interactive elements and sound:

```
In [4]: from IPython.display import Audio
Audio(y, rate=44100) # plays the data in y as audible signal
```

```
Out[4]:
```

We can connect other languages and tools, for example execution in bash:

```
In [5]: %bash
echo "Some shell command, run at `date`"
```

```
Some shell command, run at Wed 14 Feb 2018 09:56:18 CET
```

Figure 3: A basic session in the Jupyter Notebook: Equations, code, multimedia data from executed code and interpretation in one interactive and executable document.

Multi user notebook environment

The Jupyter Notebook can be run on one machine and accessed over the Internet through the IP address of that machine. However, many organisations feel the need to make notebooks available through a multi-user notebook server, and the JupyterHub project is addressing this need (<https://jupyterhub.readthedocs.io>). A current EC funded project [OpenDreamKit2015] has been working with EGI.eu to install the JupyterHub instance (<https://jupyterhub.fedcloud-tf.fedcloud.eu/hub>) on the EGI.eu-managed federated EU cloud resources early in 2018; demonstrating the feasibility of hosting JupyterHub instances on the EOSC. However, this is a prototype deployment that (i) took quite manual effort to setup, and is only the starting point of making this service useful: (ii) many questions arise about availability of temporary and persistent execution environments and file store, authorisation of usage, allocation of resources, and integration of this into EOSC and potentially compute and data storage resources local to the data providing facilities.

Containerised execution environments

Jupyter Notebooks are flexible and powerful tools for reproducible computational data exploration and analysis: the sequence of cells, each containing one or more commands, stored with the intermediate outputs provides a full

transcript of the analysis process: given the same data set and same compute environment, we can re-execute the notebook and will arrive at the same derived data [Kluyver2017].

It is, however, crucial that both the data and the compute environment (in which the Jupyter notebook commands are executed) are provided initially and preserved to support reproducibility. The compute environment is challenging: research codes often depend on a variety of support libraries and can sometimes only be compiled in particular operating systems or even versions of operating systems. A reproduction and subsequent re-use and extension of an initial data analysis is only possible if exactly the same compute environment is available as for the initial analysis. An elegant way forward is to embed the compute environment in a container (Docker is used widely, but other technologies are feasible), and to execute the Jupyter Notebook inside that container. The approach has multiple advantages, including a very concise definition of the compute environment (through the Dockerfile or equivalent files that build the container), the container image being relatively small (in comparison to virtual machine images), and execution performance being close to native execution.

A combination of a notebook with a compute environment (which could be defined through a Dockerfile) is coupled in the Binder project (<https://mybinder.org>). While the Binder project started in 2015 and is relatively young, there are already some use cases that demonstrate its great scientific value of making data accessible, interoperable and re-usable: for example this review of data analysis for the LIGO experiment that has confirmed existence of gravitational wave [Puget16].

1.3.b.3 Remote desktop based services

From their home laboratories facility users and data scientists are able to access transparently through a standard browser computers for analysing data generated at a facility. They can focus on the scientific aspect of the analysis without worrying about the size of the datasets, software installation and compute resources. This data analysis service is currently in its pilot phase at the ILL, with the support of PaNOSC it will reach production level and be offered into EOSC.

On the technology side, the front end heavily relies on the progress made with HTML5 and especially with the websockets implementation. They allow to open an interactive communication session between the user's browser and a server that proxy the standard remote desktop protocol. Even with a relatively low bandwidth (the limit on asymmetric network being of 5 Mb/s for the downstream link) it allows user to have a good experience and focus on their scientific work.

On the backend side the solution rely on the cloud technology for provisioning virtual machines on demand. These machines get direct access to the data archive and are tailored for the type of analyses that wants to perform the users. For its pilot the ILL is using the OpenStack solution.

The application stack which handles the user authentication, the creation and lifecycle of the user's VMs, manages the sessions and provide the front end application has been developed with Java and incorporate component of the Apache Guacamole project.

The diagram below in Figure 4 presents the architecture of the solution.

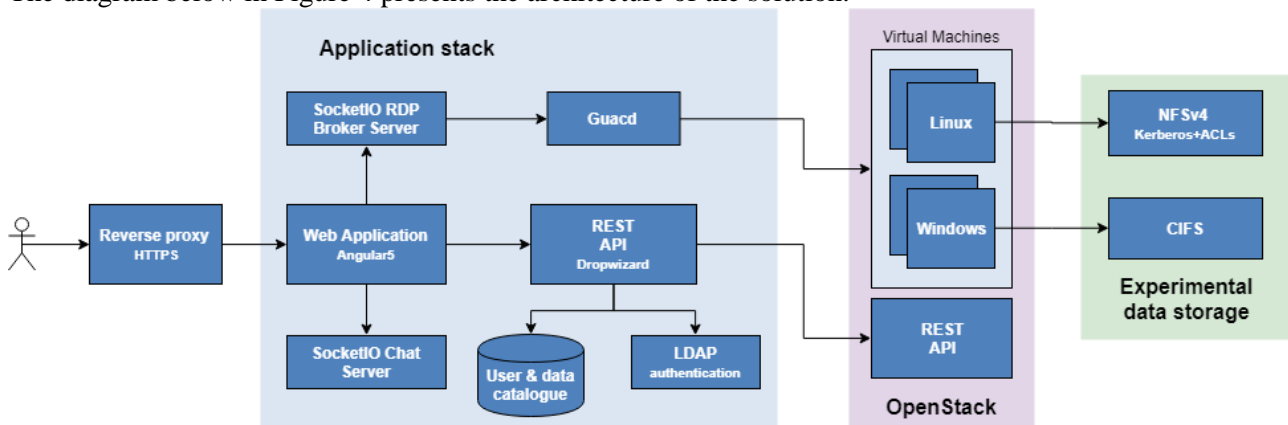


Figure 4: Software architecture of the remote desktop solution

The advantage of a remote desktop solution is that existing desktop applications can export them easily and make them available for researchers without any new developments. This includes all the applications in the PaN software catalogue (<https://software.pan-data.eu/>).

Promoting gender equality

The consortium members all fully subscribe to the goals in Horizon 2020 of equal opportunities for both genders. We will promote gender balance by expressing our commitment to equal opportunities in all job advertisements related to PaNOSC. Moreover, for all recruitments to groups where one gender constitutes less than 40% of the practitioners, we will consistently hire a person from the underrepresented group if that person is ranked equal on all other parameters to a person from the majority group. It is the duty of WP 1 management and the executive committee to ensure that this does indeed happen.

One focus to ensuring equal opportunities is ensuring underrepresented groups see that a successful and positive career is attainable. Having positive role models from under represented groups in these areas are essential to this. We will therefore actively expose and promote the women in this project by adopting the “Women in Science” activity from NMI3, where the involved women act as positive role models for other women ensuring that they see a positive and valuable career path, thus attracting female applications to the project. This task will start out with the women in PaNOSC that we are fortunate to already have involved, c.f. Section 4.

1.4 Ambition

The PaNOSC cluster will be part of the coalition of doers building the EOSC. The cluster represents some of the biggest data producers of a large community of photon and neutron RIs - itself a large community serving some ten thousands of users annually. The EOSC is an ambitious project which aims to make a big impact in the lives of researchers. PaNOSC with the EOSC will lower the barriers and address bottlenecks of the current PaN infrastructure e.g. when data are too big to move. By exposing data services to users via the e-infrastructures or local infrastructure, scientists will be able to get more out of the data. Generalising the FAIR data principles will give users better long term prospects for their data e.g. data curation, long term archiving, persistent identifiers, data publishing and citation etc. PaNOSC will accompany these services with an appropriate training program to accelerate adoption by the community. PaNOSC will offer data and services to the world-wide scientific community to increase the outputs from experiments thus benefiting society as a whole.

PaNOSC will have very long reaching consequences. Showing by example, it will strongly influence the adoption of FAIR data principles across the photon and neutron world. It will promote and implement Open Data for publicly funded science thereby ensuring better use of public funded research. It will enable a new group of virtual users to profit from the EOSC and generate new scientific insights. The data from the PaNOSC cluster is from a wide range of domains in life sciences, material sciences, cultural heritage etc. Making these data open and accessible to the scientific community in general will lead to advances in many fields e.g. drug discovery, archeology, new materials, energy saving etc, and will thus be an important asset for addressing the Societal Challenges (see Introduction). PaNOSC will facilitate data mining and discovery leading to a reduction in duplication of experiments. The simulation and modelling service will enable new experiments to be designed.

2. Impact

Summary

The main impact of the PaNOSC will be the adoption and implementation of FAIR data principles for the petabytes of data produced by the partner sites and facilitating the exploitation of these data by linking them to services hosted by the EOSC and the partner sites. This will have a huge impact for the partners like ELI which has 3 sites in the process of commissioning 3 state-of-the-art laser facilities, and ESS which will receive first neutrons during the project. For the remaining partners they form two categories, namely those like ILL, ESRF and European XFEL, who already have a Open Data policy and will enhance the impact of their data by updating their data policy to comply fully to the FAIR principles, and secondly partners like CERIC-ERIC which will adopt a FAIR data policy.

PaNOSC will generalise new remote data services as part of the developments e.g. extending Jupyter notebooks into new areas with new features, generalising the packaging of applications in virtual images or containers, simulation of neutron, synchrotron, free-electron laser FEL and laser experiments to better plan and understand experiments. This will enable new users of the data produced at these sites as well as facilitate the scientific workflow of the current users of the participants. Current users are limited by lack of data services for transferring large datasets, reducing and analysing data. They often do not have the resources in their home institutes or on their laptops. PaNOSC will enable remote access to these services for a wide user community. Specifically PaNOSC will make these services generally available in existing (ESRF, ILL, CERIC-ERIC) and sites who have recently started producing data (XFEL.EU) or will soon produce data during the proposal (ELI, ESS) sites. These sites represent the whole spectrum of photon and neutron sources in Europe and will therefore lead the way to integrating other national photon and neutron RIs to the EOSC and providing data services. PaNOSC will enable more cross-disciplinary research by facilitating the combination of data from different sources into one data set e.g. neutrons and synchrotrons, or synchrotrons and FELs, or cryo electron microscopes and synchrotrons and FELs. A large number of combinations are possible due to the large number of experimental techniques at each source and the variety of sources (neutrons, synchrotrons, free electron lasers, lasers, microscopes). PaNOSC will facilitate combining datasets from photons and neutrons like for example those found in the Small Angle Scattering Biological Data Bank (<https://www.sasbdb.org/> [Valentini2015]).

2.1 Expected impacts

The expected impact of PaNOSC with respect to the work program is detailed in the table below.

Expected impacts from call	Expected impacts from this proposal
<i>In line with the objectives of Open Science, improve access to data and tools enabling new and interdisciplinary research leading to new insights and innovation for the society at large</i>	PaNOSC will generalise FAIR data on 15 sites compared to 4 currently; it will develop data services for reduction and analysis so that users from different scientific domains have easy access and make it easier for data from different PaN sources to be combined as well as with data from other sources (e.g. computational materials science) and produce new insights
<i>Facilitate access of researchers across all scientific disciplines to the broadest possible set of data and to other resources needed for data driven science to flourish.</i>	PaNOSC will federate the Finding and Accessing to open access data from all the partners by federating the data catalogues of all partners thereby making it possible for researchers across all scientific disciplines to search for data from PaN sources; advanced searches will be implemented to speedup finding
<i>Contribute to the creation of a cross-border and multi-disciplinary open innovation environment for research data, knowledge and services with engaged stakeholders and organisations.</i>	PaNOSC is a cross-disciplinary and cross-border cluster covering a wide range of physical and life sciences. By standardising open data policies, federating data catalogues and providing tools to cover the complete scientific data workflow it will foster multidisciplinary research.

<i>Rise the efficiency and productivity of researchers thanks to an easier and seamless access to reliable and open data services and infrastructures for discovering, accessing, and reusing data;</i>	PaNOSC will generalise data services to make them accessible easily and remotely. Remote access will lower the barrier for all users. Data will be curated with metadata in catalogues. A federated search solution will be implemented across all catalogues thereby making it easy to find open data. Data analysis and simulation applications and tools will be packaged and made available from the software catalogue.
<i>Foster the establishment of global standards, ontologies and interoperability for scientific data.</i>	PaNOSC will promote the community standard NeXus for metadata and extend its use to new fields. This will increase interoperability between data produced at the different RIs. To further increase interoperability standard APIs will be developed for associated services.
<i>Develop synergies and complementarity between involved research infrastructures, thus contributing to the development of a consistent European research infrastructures ecosystem.</i>	The PaNOSC cluster will work together to develop a common data policy framework, federated metadata catalogues and generic tools for data analysis and simulation. These will be shared with the other RIs in the PaN community. Links will be established with the other clusters in INFRAEOSC to develop a common understanding of FAIR principles.
<i>Research communities adopt common approaches to the data management lifecycle (data and metadata curation), which leads to economies of scale.</i>	PaNOSC will establish a common approach and best practices to implementing data and metadata management for data providers in the PaN community. Sharing these guidelines with all photon and neutron sources and providing training will accelerate their adoption. A common API will allow a federated catalogue to be implemented for PaNOSC to make all data accessible from a common portal. Partners will share the same metadata catalog where possible.

Table 2: Impacts

Key Performance Indicators (KPI)

The impact of the project will be measured through KPIs. Today we do not have an estimate of how many scientists are going to exploit the Open Data which PaNOSC will provide. In order to follow the evolution of the use of Open Data and the data services offered to scientists via the EOSC and the local infrastructures we will track the following KPIs:

KPI	Description	Collection
DOIs published	Count how many persistent identifiers in the form of DOIs will be published	Statistics on DOIs will be collected each data producer by the metadata catalogues
DOIs cited	Count how often DOIs from the participants are cited	Web of Science, Google scholar , ...
Applications packaged	Count how many applications have been packaged	Statistics will be measured by the PaNdata software catalogue
Downloads of Open Data and data under embargo	Count how many FAIR data sets have been downloaded	Measured by the data portal(s) and FTS service

PaNOSC users on local infrastructure(s)	Count how many users access the data services on the local infrastructures	Monitoring of local infrastructure will measure logins
Jupyter users on EOSC infrastructure	Count how many PaNOSC Jupyter users there are for EOSC	Measured by the EOSC monitoring and accounting
Remote desktop users on EOSC infrastructure	Count how many PaNOSC remote desktop users there are for EOSC	Measured by the EOSC monitoring and accounting
Simulation users	Count how many PaNOSC users use the simulation service on EOSC	Measured by the EOSC monitoring and accounting
Track uptake by PaN RIs	Count how many PaN RIs (Observers and others) adopt the outcomes of PaNOSC	Measured by the Observers and at workshops and general meetings and by contacting PaN RIs
PaNOSC Publications	Publications about PaNOSC, and science publications using PaNOSC services.	Cooperation with partners and early users of PaNOSC services; encourage PaNOSC acknowledgement

Table 3: Key Performance Indicators

Possible impact barrier

The main barrier which could reduce the impact of PaNOSC identified is the lack of awareness of good data citing and data sharing practices amongst the PaN community i.e. lack of a data sharing culture. This is currently a real barrier to the adoption of FAIR data principles amongst the users of the sites. To illustrate this difficulty we take the example of the ILL which mints DOI for all data sets produced since 2012 and request users to cite this DOI in all their publications where the corresponding data set has been used. In 2018, after 4 years of communication efforts only 15% of the publications based on ILL experiments are citing these DOIs (you can visualize the progression of the number of ILL related publications citing their data via the use of DOIs in the graph below). This will be addressed by collectively training staff and users and providing online training material. For partners without a FAIR data policy (CERIC-ERIC and ELI) it will be a challenge to adopt the PaNOSC FAIR data policy framework. PaNOSC will engage with the OpenAIRE-Advance project and the EOSC-hub to see how they can help reduce the barriers in the scientific community.

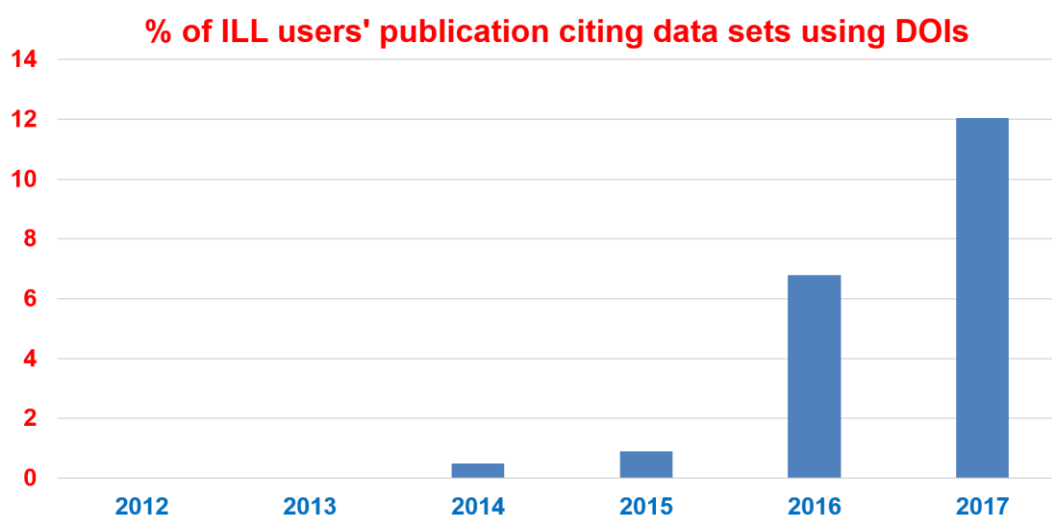


Figure 5: Number of publications citing DOIs for ILL

2.2 Measures to maximise impact

2.2 (a) Dissemination and exploitation of results

The project foresees the deployment of a wide set of dissemination tools (website, social media, workshops, seminars, conferences, meetings, etc.), to maximise the impact of every single WP and of the overall project's results. The main targets addressed are the project's partners and RIs from other clusters, the international scientific and industrial user communities, IT experts and e-infrastructures participating in the EOSC construction, ERICs' and RIs' staff and managers; regional, national and European policy makers in the field of Research, Innovation and Industrial Development. Project's results group into several categories, such as services, reports, business and sustainability plans, methodologies, policies, guidelines and standards, training platform and materials.

The plan for the dissemination and exploitation of PaNOSC's results has the following objectives:

- Transfer the knowledge (also by means of training tools and activities) about the use of the developed EOSC services to the interested photon and neutron RIs, to the RIs from different cluster, and to the users' communities worldwide. To address this objective, and thus increase the knowledge of the PaNOSC services among stakeholders, a promotional campaign is foreseen. All available communications channels will be deployed to increase the awareness about the existence of the e-learning platform developed in WP8 and to stimulate the use of the related training materials.
- Stimulate the Industrial Liaison Offices (ILOs) of the European photon and neutron facilities to adopt the business and funding models developed in the project (D7.3). To foster the achievement of this goal, ILOs will be involved already during the development process.
- Involve the regional, national and European public governing bodies in the development and adoption of the Sustainability plan developed in the project (D7.4). Target events involving these stakeholders will be used for the purpose.
- Work with the different community organisations to get them to help disseminate and adopt the PaNOSC outcomes by the national RIs and community at large. The organisations in questions are PaNdata, LEAPS⁴ for photons, SINE2020⁵ for neutrons, FELs of Europe⁶ for free electron lasers, ELI for lasers.
- Facilitate the work of RIs' managers and staff by providing the tools and guidelines necessary for the adoption of the policies and standards developed throughout the project, while stimulating harmonisation of practices across RIs.

Guidelines with best practices will be made available in the project's website (to appear on www.panosc.eu) and in the e-learning platform (to appear on www.pan-learning.org), which will be active and accessible even after the project ends. Moreover, outputs will be circulated among the European RIs' community from different clusters, both via virtual channels (direct mailing, publication on website, presentations in target events, training materials, etc.) and through meetings and events. They include reports, best practices' guidelines, policies and guidance documents in WP2 (D2.1, D2.2, D2.3, D2.4, D2.5), WP3 (D3.4, D3.5, 3.6), WP4 (D4.1, D4.6) and WP6 (all deliverables), WP7 (all deliverables), WP8 (D8.1 – D8.8), WP9 (D9.1, D9.4) which will be stored in the online repository (D9.3) used by the partners as internal communication tool for saving, sharing and storing project's strategic and operational documents.

The partnership will capitalize on already existing events: users meetings at the different RIs, as well as scientific meetings, such as synchrotrons' and neutrons' communities' conferences, EOSC meetings, ERF-AISBL meetings of the Data working group, RIs' H2020 projects' (ACCELERATE, CALIPSOplus, SINE2020) target meetings and events, as well as the scientific directors' and board meetings regularly planned in the different RIs and in the project. This will allow involving the project's stakeholders, for widely disseminating project's outputs while engaging main target groups, as well as to collect comments and advice for a further fine-tuning of the policies, standards, plans and services developed. To increase the awareness of the project's outputs, articles about all the

⁴ <https://www.leaps-initiative.eu/>

⁵ <http://sine2020.eu/>

⁶ <https://www.fels-of-europe.eu/>

planned activities and the achieved results will also be written and disseminated among relevant contacts and networks.

The project's website will be constantly updated with the publication of all the project's outcomes and results. The news about the main achievements will be also distributed and made available on the project's partners' websites and through other communication tools (e.g. newsletters, social media, leaflets, etc.). The annual project's meetings will be an occasion to show best practices and the results achieved by the project, as well as the further steps to be implemented for their exploitation and sustainability. The discussion in such events will also be the basis for the finalization of the final reports and deliverables in different WPs before their dissemination.

2.2.a.1 Data management

All operational and strategic documents and data generated during the project will be saved and stored in an online repository (D9.3) – in compliance with FAIR principles. Clear instructions and guidelines on how to use such repository will be distributed to all stakeholders involved.

2.2.a.2 Knowledge management and protection

Reports, plans, policies and publications prepared within the PaNOSC project will be public to the extent possible. The aim is to make all publications and the data collected publicly available (gold open access, OA), unless this is in conflict with privacy issues or future commercial activities. The project's website and the e-learning platform developed in T8.5 (D8.2) will further ensure open access of such data, through the publication of the training materials targeting different stakeholders.

2.2.a.3 Open Source Software

All software developed in the course of the project will be under version control in an open source repository (such as GitHub) and licensed under a license compliant with the Open Source Initiative (OSI, www.opensource.org). The software will thus be Findable and Accessible and because it also will be developed with Inter-operability and Reuability in mind, the software will be FAIR in its own right.

All public deliverables will be made available and accessible from the project's website.

2.2(b) Communication activities

In **WP9** (and in coordination with the project's and **WP1** leader), a detailed Communication Strategy (T9.1) – to which the project's communication and dissemination plan (D9.1) will be annexed – will be developed. The project's communication plan will include a section about internal communications, meant to set the ground rules necessary to ensure that all partners are well-informed about upcoming events and deadlines, as well as about steering decisions in a timely manner. All project's partners shall take their role in ensuring a complete and smooth flow of information within the PaNOSC project, while granting to respect confidentiality when required and when referring to matters of strategic importance either to individual partners or to the whole partnership. The project coordinator will take lead of internal communications, whereas the leader of WP9 will have to grant support for putting in place all necessary tools for this purpose. Adherence to these principles is the key to execute the project smoothly.

Main contact persons of partners' institutions will decide on confidentiality level of information at run time, and propagation of information into and out of the internal structure of the participant organization is at the main contact person's discretion and responsibility. The project's steering body has the supreme authority to restrict or loosen access to information, documents and other resources as it deems fit to the aims of the PaNOSC project, always in accordance with the Grant Agreement and Partnership Agreement.

In internal communications, the Partnership will use all modern technologically accessible means of spreading information. Regular physical meetings of the Partnership and task work groups will still be indispensable. However, to reduce costs, preference will be given to conference calls and meetings using both free (gratis) and commercial (paid) software solutions. The partnership will have a shared calendar and use tools for collaborative assembly of documents in working out the policies, white papers, reports, minutes, etc. These will in part use public Cloud services and, where maximum confidentiality is called for, documents will be exchanged via the online

Cloud-based repository dedicated solely to the project (D9.3). All public releases of information will be propagated into the partnership as well, to raise and maintain awareness of the PaNOSC project in the partner organizations.

All communication activities that include other audiences than strictly within the partnership are considered external communication activities, aiming at providing target groups with an easy understanding of the project, raising awareness and increasing the knowledge about its major progress and achievements. All project participants will take part in the external communication activities under the coordination of CERIC-ERIC as leader of WP9.

Main stakeholders of the PaNOSC are the following:

- Community of European Photon and Neutron Research Infrastructures
- Community of Research Infrastructures from other clusters
- Community of current and prospective European Research Infrastructure Consortia
- Directorate-General for Research and Innovation
- National Ministries for research
- IT experts from e-infrastructures involved in the development of the EOSC
- RIs' scientific and industrial communities
- RIs' scientific managers

As for external communications, the communication plan (D9.1) will detail the actions carried out and the tools deployed for properly informing and increasing awareness among the project's stakeholders about the PaNOSC project, its goals, activities and results. In particular, the overall key message will be that "PaNOSC aims at implementing the European Open Science Cloud and the European Data Infrastructure to allow the scientific and industrial users' communities to have access, analyse, process and execute complex data workflows through the newly developed EOSC services and data catalogues". Key sub-messages to each target group will be defined and updated during the lifetime of the project.

As further detailed in WP9, actions taken to pass information to and to interact with stakeholders (i.e. target groups) will have the following forms:

1. Public project updates via the project's partners' website, PaNOSC's website (D9.2), web pages of the partners of the RIs involved in the project, email newsletter and social networks, as well as articles and/or press releases about key PaNOSC's events and achievements.
2. Interaction with multiple stakeholders at the project partners' user meetings, at scientific and industrial workshops and conferences Europe-wide, at the ERIC Forum meetings (attended by ERICs from other clusters) at the ICRI conferences and at the EU presidency's events (which are traditionally attended by DG RTDs).
3. Interaction with contact points of the National Ministries:
 - a. at the partners' governing bodies' meetings (which are attended by representatives of the Ministries from member countries);
 - b. at events organized by the Ministries;
 - c. by standard contact points of government grant agencies in countries where partners are present.
4. Interaction, from an early stage of the project, with other clusters for the coordination and co-development of services. This is needed considering the high level of interdisciplinarity of PaNOSC RIs and the heterogeneity of CERIC-ERIC, comprising facilities beyond the PaN community.
5. Consult and discuss with the Observers and other members in the PaNdata community. Re-activate the pandata.eu website by updating it and posting to it. Participate in the PaNSig group in the RDA (<https://rd-alliance.org/node/926>).

2.2.(c) Project internal communication strategy

In order to ensure that information flows between the different work packages, *Work Package Leaders*, *Project Manager* and the *Executive Committee* there will be regular meetings:

- Monthly video conferences chaired by the *Project Manager* and attended by all *Work Package Leaders* (*Executive Committee* members could attend as observers)
- Annual Workshop attended by the *Executive Committee* members, *Project Manager* and *Work Package Leaders*

Minutes from the meetings and reports will be available to project members via the project management platform.

3. Implementation

3.1 Work plan -Work packages, deliverables

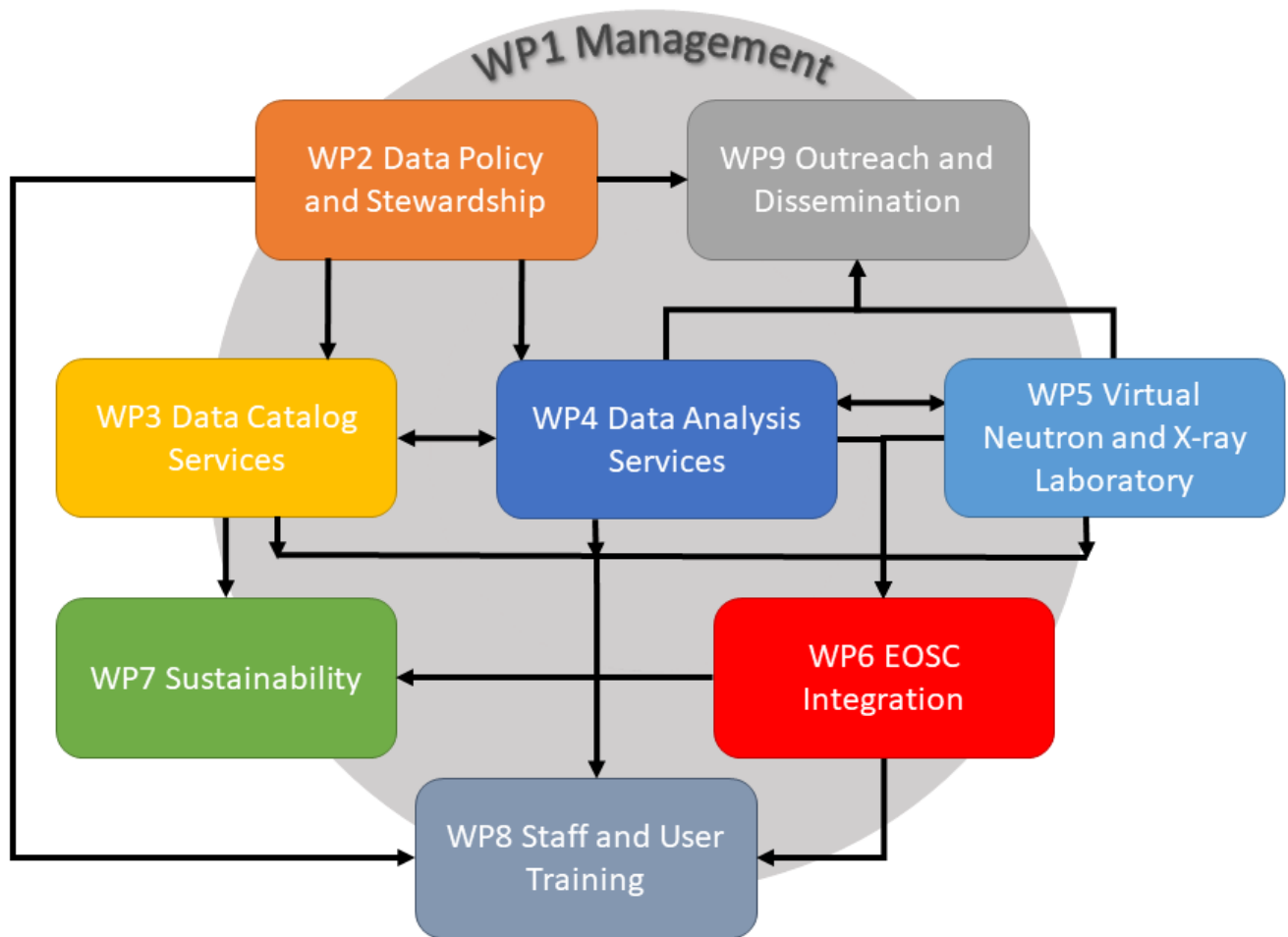


Figure 6: Relationship between work packages

The PaNOSC project is organised into 9 work packages (WP) as listed in Table 7. The structure of the work plan is to have one WP (WP1) for Management of the project, then there are two WPs dedicated to the data policy (WP2) and (meta-)data catalogues (WP3). WP2 will ensure all data policies are updated to align with the FAIR data principles. Actual services are developed in WP3 to WP5. WP3 will ensure the metadata catalogues are implementing the data policy and can be federated and integrated to the EOSC. New services for data reduction and analysis and modelling and simulation will be provided by WP4 (data services) and WP5 (modelling and simulation). These services will be integrated in the EOSC in WP6. For the long term success it is necessary to ensure these new services can be sustained. WP7 will study and propose solutions for sustaining the open data policies and services. WP8 will address the problem of training and educating the data stewards on how to manage data and researchers on how to publish and cite data in the modern digital world and also educate users in using the developed services. Finally, WP9 will be dedicated to disseminating the outcomes of the proposal.

3.1(a) Work Packages

Work Package	Leader	Effort	Start	End
WP1: Management	ESRF	64 PMs	M1	M48
WP2: Data policy and stewardship	ESRF	76 PMs	M1	M36
WP3: Data catalogue services	ESS	291 PMs	M1	M48

WP4: Data analysis services	XFEL.EU	309 PMs	M1	M48
WP5: Virtual Neutron and X-ray Laboratory	XFEL.EU	219 PMs	M1	M48
WP6: EOSC Integration	ILL	192 PMs	M1	M48
WP7: Sustainability	CERIC-ERIC	56 PMs	M1	M48
WP8: Staff and User Training	ESS	108 PMs	M1	M48
WP9: Outreach and dissemination	CERIC-ERIC	70 PMs	M1	M48
TOTAL		1385 PMs		

Table 4: List of WP and their leaders

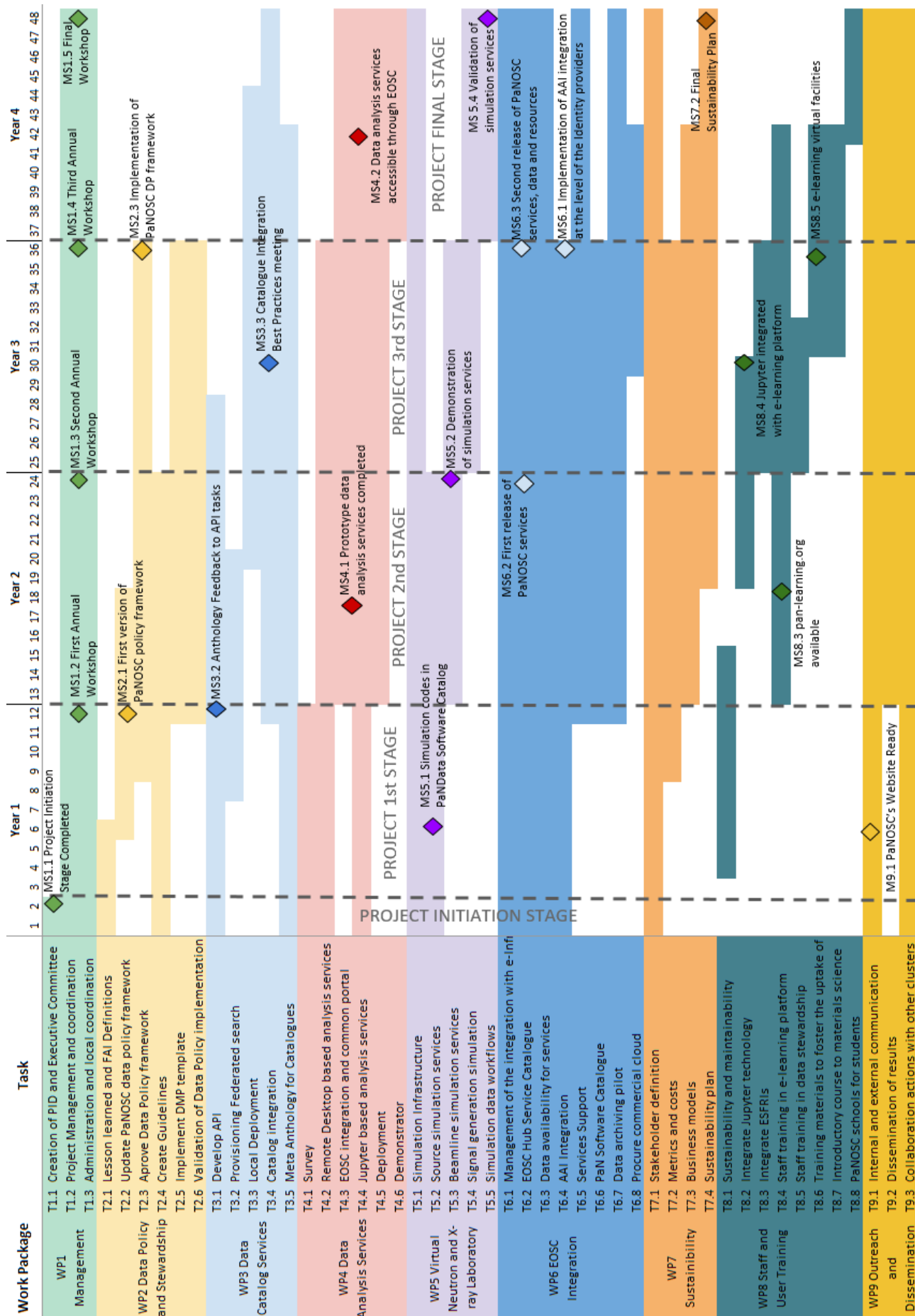


Figure 7: PaNOSC tasks and major milestones

3.1(b) Work Package descriptions

Work package number	1	Lead beneficiary				ESRF
Work package title	Management					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC
Person months per participant	49	3	3	3	3	3
Start month	1		End month		48	

Objectives

Manage and coordinate the project to ensure that the objectives are delivered on time. Organise regular follow-up meetings and annual workshops to ensure progress and results are communicated between participants, observers and the community at large. Interact with and follow-up all other work packages while managing change and risk.

Description of work

Task 1.1 Creation of Project Initiation Documentation (M1-M2), appointment of Executive Committee and selection of the tools to be used for project management.

Leader: ESRF Contributors: ILL, XFEL.EU, ESS, ELI and CERIC-ERIC

Task 1.2. Project management and coordination (M1-M48): kick-off meeting, monthly video conference and an annual conference will take place. Submission of regular reports, financial statements and deliverables as defined by the contract, including all liaison with the EC. Change, progress, scheduling, communication and risk will be managed as part of this task as well as monitoring of job advertisements across project to improve gender balance

Leader: ESRF Contributors: ILL, XFEL.EU, ESS, ELI and CERIC-ERIC

Task 1.3. Administration (M1-M48): capturing and reviewing all actual information about actual costs (financial and human resources efforts) and comparing it with the planned forecasts.

Leader: ESRF Contributors: ILL, XFEL.EU, ESS, ELI and CERIC-ERIC

Deliverables

Deliverable 1.1 Project Initiation Documentation (M2, R, PU, ESRF)

This set of documentation will include an executive summary for the project, its governance (including rules to appoint Executive Committee members), scope, organisation, risk management strategy, communication strategy, list of stakeholders, the list of executive committee members, executive committee appointment rules and initial set of risks and issues identified.

Deliverable 1.2 Mid-year summaries of regular video conferences (M6, M18, M30, M42, R, PU, ESRF)

This will consist of a summary of the regular video conferences that will take place as part of *Task 1.2 Project management and coordination* and it will include news from the partners about their progress and a snapshot of the current health of the project.

Deliverable 1.3 Report of annual workshop (M12, M24, M36, M48, R, PU, ESRF)

These reports will focus on the activities of the annual workshop, status of the project, summary of progress achieved during the year, residual risks, main changes to the project and a report from the Executive Committee.

Deliverable 1.4 Data Management Plan reviewed and agreed to by partners (M6, R, PU, ESRF)

Data Management Plan following H2020 guidelines, reviewed and agreed to by all partners.

Work package number	2	Lead beneficiary				ESRF
Work package title	Data Policy and Stewardship					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC
Person months per participant:	17	10	3	14	20	12
Start month	1		End month		36	

Objectives

The overall goal of the work package is to enable facilities to ensure their data policies honor the FAIR principles in the way they curate data. Currently some of the participating facilities have defined data policies (ILL, ESRF, XFEL.EU and ESS) others still have to define and apply a data policy (ELI and CERIC-ERIC). All members have committed to adopting an Open Data data policy as part of the PaNOSC proposal. The existing policies are based on the PaNdata policy 10 years ago, and have provisions for Open Data. Nowadays, the FAIR data principles⁷ more clearly define the concepts of Open Data. The goal is therefore to update the current policies while respecting the specific needs of the PaN community, to better align with current understanding of FAIR principles. PaNOSC will foster links to experts in applying FAIR principles to research data like OpenAIRE and Force11 to validate the data policies. Based on the current barriers of adopting FAIR principles, the following objectives have been identified:

1. Definition and harmonisation of PaN specific data policies and management of Intellectual Property Rights (IPRs) and ethical issues; addressing legislative and interoperability issues which affect data handling across geographical and discipline borders specific to the PaN community.
2. Definition and adoption of common open standards for interoperability. Registering with and citing of these standards by standards bodies and publishers.
3. Stewardship of data handled by the involved research infrastructures according to the FAIR principles. Citing of PaN data repositories and data descriptors by publishers e.g.
<https://www.nature.com/sdata/policies/repositories>
4. Produce guidelines for best practices based on experience of those PaN partners who already have Open Data policies since a few years now to help partners adopting Open Data data policies correctly from the start. The guidelines will be shared with the rest of the PaN community. Guidelines for dealing with typical PaN issues like huge data sets will be dealt with by exploring data reduction and compression schemes which reduce the burden on the data infrastructure.

Description of work

Task 2.1: Lesson learned and FAIR Definitions (M1-M6) *Leader: CERIC-ERIC. Contributors: ESRF, ILL, XFEL.EU, ESS, ELI*

Compile lessons learned from previous policies of the partners and other members of the PaN community who have experience putting data policies in place. To identify existing barriers for making the community accept FAIR principles, we will collect the existing experience and give recommendation for how to overcome such barriers. Consult with expert organisations like Force11 and OpenAIRE to understand latest research and best practices in applying FAIR principles to data.

Task 2.2: Updated PaNOSC Data Policy framework (M6-M18) *Leader: ESS. Contributors: ESRF, ILL, XFEL.EU, ESS, ELI*

Based on the existing PaNdata policy⁸ create a new PaN data policy framework that all facility specific data policies should adhere to. Ensure that the data policy framework is aligned with EOSC activities on data policy harmonization. The aim of the policy is to ensure that FAIR principles are applied as broadly as possible.

Task 2.3: Approve Data Policy framework (M9-M36) *Leader: CERIC-ERIC. Contributors: ESRF, ILL, XFEL.EU, ESS, ELI*

Participating facilities will publish and apply or align their current data policies with the new policy based on the common key principles defined in task 2.2 for their facility or amend their existing policy to be consistent with the FAIR principles.

⁷ <https://www.force11.org/fairprinciples>

⁸ <http://wiki.pan-data.eu/images/GHD/0/08/PaN-data-D2-1.pdf>

Task 2.4: Create Guidelines (M1-M24) Leader: ESRF. Contributors: ESS, ELI, XFEL.EU, CERIC-ERIC

Creating Guidelines for DOIs, long term archiving, and FAIR data (legal aspects) Support best practice by e.g. formulating a click through license agreement to be accepted when submitting proposal, creating or downloading data covered by policy. Create Guidelines for handling GDPR and other legal aspects of federating data. To ensure legal aspects will not be a barrier for adoption of FAIR data, clear guidelines are needed. Part of this will be to establish a common set of definitions for policies, using and publishing data.

Task 2.5: Implement DMP template (M12-M36) Leader: ESS. Contributors: ILL, CERIC-ERIC

Define and implement a template for DMPs for experiments performed at the PaNOSC research infrastructures. Have support for automatic filling out of the template based on existing information about experiment (e.g. proposal text).

Task 2.6: Validation of Data Policy implementation (M12-M36) Leader: CERIC-ERIC. Contributors: ELI

We will follow up with multi facility partners (CERIC-ERIC and ELI) and other partners who already have a data policy to track and document the progress on adopting or adapting their existing policies to the PaNOSC data policy framework.

Deliverables

Deliverable 2.1 PaNOSC data policy framework updated (M18, R, PU, ESRF)

Deliverable 2.2 DMP Template for facility users published (M36, R, PU, ESS)

Deliverable 2.3 Guidelines on best practices implementing the PaNOSC data policy framework published. (M24, R, PU, ESRF)

Deliverable 2.4 Integration of the policy in the User Access and facility information systems (M36, R, DEC, CERIC)

Work package number	3	Lead beneficiary				ESS
Work package title	Data Catalog Services					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC
Person months per participant:	25	21	36	43	78	88
Start month	1		End month	48		

Objectives

The overall goal of the work package is to provide an EOSC service that allow for users to seamlessly and easily access data from the diverse set of catalogs at the existing facilities. The situation today is that there is a plethora of different catalog services that each allow for access to data in slightly different ways. Some solutions are used in more than one place (e.g. ICAT) but quite commonly with local adaptations, and not allowing for federation of other similar catalogues. The work package will not supplant the existing services, but rather define a unified API and enable the existing and future services to be used by EOSC through the API. The API, test harness, demonstrator implementation as well as lessons learned from deploying it will be made publicly available, making it easy for facilities outside the PaNOSC to adapt and have their data exposed in the EOSC. The work package objectives in more detail are:

- Provide a federated data catalog service across the Photon and Neutron community, compatible with OpenAIRE.
- Definition of standard metadata for scientific domains at the partner facilities to access to data beyond the generic search features of OpenAIRE, enabling new and interdisciplinary research leading to new insights and innovation for the society at large.
- Facilitate access of researchers across all scientific disciplines to the broadest possible set of data and to other resources needed for data driven science to flourish.
- Integrate the data catalog with the existing data sources (e.g. experimental stations). This includes the integration of the data production facilities with the catalogue service.

Datasets in the public domain (after an embargo period) shall be findable to the interested public and wider scientific community at large, possibly after registration to enforce policy compliance (expected input from WP2). During the embargo period for a dataset access may be restricted to the original proposer or facility.

Description of work

The work package will create an API that existing catalog solutions can adopt to allow for seamless integration into EOSC via OpenAIRE. The API description will be accompanied by a test suite that can test a given implementation for compliance. To further illustrate the intended behavior, a demonstrator implementation will be developed. Furthermore a web service will be deployed that will allow for search across facilities exposing the API. Finally, all the participating facilities will expose their data through the API according to their data policies and the joint data policy framework developed in WP2.

Task 3.1: Develop API (M1-M28) *Leader: ESS. Contributors: ESRF, ILL, XFEL.EU, ELI, CERIC-ERIC*

Define an API to be used in the Photon and Neutron community that will allow for FAIR exposure of the data at the individual institutions through a catalogue service. The API will allow federation, and exposure of metadata relevant for the area, in a way that will enable search and facilitate access of researchers across scientific disciplines. Existing APIs (e.g OAI-PMH) and communities (e.g. openarchives.org, Dublin Core Metadata Initiative (DCMI), OpenAIRE) will be taken into account. The API will enable domain specific search extensions aware of the metadata definitions and usage at photon and neutron facilities.

In order to test any implementation at facilities for compliance, a set of API tests will be developed. The test harness will be executable against a given site catalogue service and result in a report stating the status towards compliance.

An implementation, based on an existing solution is developed and deployed at a facility to show the feasibility of the approach. The implementation is to be fully compliant with the API, but not necessary performant for very large catalogues.

Task 3.2: Provisioning Federated Search (M8-M20) *Leader: ELI. Contributors: ESS, CERIC-ERIC*

This task will link the PaNOSC beneficiaries' data catalogs to the EOSC hub. The EOSC hub will provide the API needed to share and search metadata. In the absence of a definition following the OpenAIRE DOI equivalent scheme should yield sufficiently wide exposure. A web service demonstrator will be provided that allows searching all PaNOSC partner sites for available datasets using the common metadata API. The demonstrator will showcase how access to the catalogue will provide identifiers that will allow the found data to be accessed and used for analysis. Once the demonstrator is working the next step will be to work with EOSC hub to provide a production ready service to be provided as part of EOSC.

Task 3.3: Local Deployment (M20-M44) *Leader: ESS. Contributors: ESRF, ILL, XFEL.EU, ELI, CERIC*

The participants will implement the mandatory part of the API for their local data repository. This includes effort required to ensure that all facilities metadata is compliant with the standards from WP2. The implementation is expected to happen through extending the existing catalogue services at the partner sites with the API or by an equivalent solution.

Task 3.4: Catalog integration (M12-M48) *Leader: ELI. Contributors: CERIC-ERIC, ILL, XFEL.EU, ESRF, ELI*

Integration of data production facilities with the data catalog which is especially important for heterogeneous and distributed facilities. This task will document best practices and support heterogeneous and distributed facilities getting their workflow to support cataloging data.

Task 3.5: Meta Anthology for Catalogues (M1-M42) *Leader: ESS. Contributors: ESRF, ILL, XFEL.EU, ELI, CERIC-ERIC*

Extend NeXus metadata standards to enhance interoperability. In order to operate on their own data across facilities or explore relevant foreign datasets in the public domain, searches on the scientific metadata need to yield the correct results. For large parts of the communities NeXus is the most commonly used file format. It is the only one with an ambition to extend into all relevant scientific fields, for both raw and derived data. Building search terms and keywords from the NeXus dictionary, would make use of the community buy in and expertise

that went into this standard. However, NeXus is based on a hierarchical backend, tree like storage, making use of parent-child relationships, that do not straightforwardly map into usually flat search terms. In addition developing a standard mapping for existing NeXus definitions, in this task we can add missing definitions for raw data, as well as for processed derived data. With the results to be proposed to the NeXus committee for community adoption.

Deliverables

Deliverable 3.1 API definition (M18, R, PU, ESS)

Deliverable 3.2 Demonstrator implementation (M28, Other, PU, ESS)

Deliverable 3.3 Catalog service (M40, DEC, PU, ESS)

Deliverable 3.4 Implementation Report from Facilities (M44, R, PU, ESS)

Deliverable 3.5 NeXus Metadata Mapping Schema and Proposed New Definitions (M42, R, PU, ESS)

Work package number	4	Lead beneficiary				XFEL.EU
Work package title	Data Analysis Services					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC
Person months per participant:	36	71	60	32	50	60
Start month	1		End month		48	

Objectives

Data analysis is the process of extracting meaning from recorded data, and thus enables the transition from measurements to insight and new science. In the context of photon and neutron science, this is an essential and non-trivial process and can extend over weeks and months for a single experiment. To make (so-called) raw data usable and re-usable for research, it is critical that we provide such data analysis services together with the data. The size of the activity planned here is commensurate with the importance of it to enable new science from existing data. We expect some of the outputs will benefit science beyond photon and neutron facilities and users.

The objective of this work package is to make such data analysis services available through cloud hosted services and on the EOSC. In particular, this means that it must be possible to choose, control and execute analysis services remotely. Ideally the user interface for local and remote execution is identical or at least similar. In the context of the FAIR principles, a complementary objective is to support traceability, persistent identification, and reproducibility of the data analysis process from raw data to publication data.

Description of work

In this work package, we make analysis tools and services available for remote execution in the cloud.

We have identified two technologies that we plan to explore as priorities (see Section 1.3.b.1) but will in addition carry out a review of other possibilities at the beginning of the project, and at later points to stay connected to leading edge developments and opportunities. The first technology is based on browser driven remote desktop execution (Section 1.3.b.3), i.e. the provision of virtual machines that are tailored for particular data analysis tasks and which can be controlled remotely through a web browser which hosts a graphical desktop interface. This is the most generic approach as any application currently used can be hosted in the virtual machine and the ‘usual’ interface (be it graphical or command line) can be displayed remotely in the browser (T4.2). We will deploy this to project partners (T4.3) and the EOSC in WP6.

The second technology is to combine the Jupyter Notebook⁹ with bundled execution environments through containers (T4.4) as described in Section 1.3.b.2. This has advantages of being designed for remote access, allowing users to access the service with the webbrowser of their choice on their machine, and allowing much better support of the FAIR principles through intrinsic reproducibility. While we predict the importance of this approach to grow, the Notebook hosted data analysis service is not usable for all data analysis requirements. This will be deployed to all partners (T4.5) and to the EOSC in WP6.

⁹ <http://jupyter.org/>

These services have also different maturity levels whether they are in pilot or production phase, ranging from TRL 7 to 9.

Currently most of the analysis services are run on the IT infrastructure of each facility, and the computing hardware is thus typically physically close to the data storage hardware. Some of the (derived) data sets are small (~GB) and can be moved through the Internet to other locations, but for some data sets the effort of moving the data is substantial (~TB), and can be prohibitive on short timescales (hours). In principle, we have the options to 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 and the EOSC, technology for moving computation and data, and willingness of users to wait for data migration. This is an important issue affecting most tasks in this work package.

The issues with reproducible and remote data analysis, that are advanced in this work package, are not unique to photon and neutron science but shared by many areas of data-driven science and computational science. Furthermore, the technologies we will bring to the EOSC – such as the tools from the Jupyter ecosystem – have much wider applicability. As we will contribute to making this tools more robust and flexible, we will create value for many more researches and domains beyond photon and neutron science.

Task 4.1 Survey data analysis requirements and solutions at the partner sites, and horizon scan other emerging tools and technologies (months 1 - 12), *Leader: ILL. Contributors: ESRF, XFEL.EU, ESS, ELI, CERIC-ERIC*
We will create a survey for each partner aiming to collect the different solutions, workflow and technology already offered for data analysis services. This includes raw data access, analysis software availability and preservation of results.

We will analyse the survey results and identify best practices and tools that can be used more widely. Feasibility analysis will take into account technology, cost and security for the partners and the EOSC integration. We will identify particular pilot data analysis services that will be made available first through EOSC, and which will be realised in Tasks 4.2 to 4.5.

We will review developments that affect remote provision of analysis services (such as workflow and compute environment management tools, tool sets for reproducible analysis etc). If deemed appropriate, we may amend later tasks in this work package accordingly to benefit from these emerging technologies.

Task 4.2 Remote desktop based analysis services (1-36 month). *Leader: ILL. Contributors: ESS, CERIC*
Using remote desktop and cloud technology we can offer a remote data analysis experience that appears as if the display of a local computer is available remotely. This is achieved by providing a graphical desktop of data analysis computing machines accessible via the users web browser. In the background, we use virtual machines, hosted at the facility and located close to the data archive or in the cloud (providing that the data have been transferred to the same location), to provide the analysis software, storage and computing capacity that is tailored to each use case. These services are currently in their pilot phase at the ILL.

Feedback from the pilot

We will improve the usability, features and support tools in preparation of a larger user audience based on the feedback received from the pilot. This initially concerns the integration of screen sharing/broadcasting capabilities, screencasts and communication tools (video conference, chat, discussion boards). Other users' requests received during the remaining course of the pilot may also be taken into depending on the resources still available.

Software distribution repository

The photon and neutron facilities cover very diverse scientific fields and therefore the ecosystem of analysis software is extremely large (more than 100+ referenced software).

Offering virtual machines with all of the analysis software locally installed has proved to be unmanageable in terms of maintenance, software update and consequently security.

To solve this issue, the analysis software can be offered on demand instead of being installed locally on a virtual machine. This will be achieved by offering a software distribution repository for the PaN community based on

the CERNVMFS technology. In this task we are going to setup a repository server, install a pre-selection of analysis software for distribution and then integrate the repository client inside the virtual machine. We also need to organise the workflow and responsibility of distributing the selection of software to the virtual machines.

Reference implementation documentation

Provide technical documentation to allow all partners and others to implement the Remote Desktop service for their facility. This will include all the required components and interfaces to allow each partner to deploy an installation tailored to specific requirements of their facility. This will take into account the fact that each facility has different user registration, proposal management and data archive systems.

Task 4.3 EOSC integration and common portal for remote data analysis services (M13-48), *Leader: ILL. Contributors: all*

We will provide a service that will allow a user to remotely analyse their data from any facility via a common portal. Federated authentication and cross facility data transfer will be required for the common portal to reach its full potential. These requirements will be addressed in WP6 Task 6.3 and 6.4. In this task we will implement a facility connector for the remote desktop services and adapt these services for the EOSC AAI and data transfer mechanisms.

EOSC Authentication and Authorization Infrastructure

In this task we will implement the technical solutions chosen in WP6 to authenticate the users in the different layers of the portal architecture (the portal itself, the data access and the machines). This authentication should allow the different providers to grant the proper authorization and to implement service usage accounting. We will review the security of the system to identify risks and necessary security measures.

Data sharing

We will modify the compute services to benefit of the work in WP6 concerning the movement of data. By integrating the data transfer solution we will allow users to transparently work on data, irrespective of its location, and perform data analysis.

Common platform

We will extend the single site remote desktop portal for the selection of the compute and data providers. Each participating facility will implement a connector that allows the common portal access to their compute infrastructure and manage the transfer of data. A user will be able to select any facility and start remotely analysing their data via a single interface. This common platform, based on existing solutions, should also provide the possibility of directly archiving and sharing results after a user has completed their data analysis using the services provided.

Task 4.4 Jupyter ecosystem based data analysis services (M1-48) *Leader: XFEL.EU. Contributors: ESRF, ESS, ELI, CERIC-ERIC, ILL*

The Jupyter Notebook (Section 1.3.b.2) is an executable document, hosted in a web browser. The notebook is composed of a sequence of input cells, each of which can contain text or code. Code input cells are numbered and can produce textual or multimedia output. These outputs are displayed inside a web browser that is connected to a computational backend. This backend can be the same machine on which the notebook is displayed in the browser, or a remote server, making this technology immediately usable for remote and cloud access while providing exactly the same interface.

Throughout this task we will work with scientists at our facilities to ensure practical value and ease of use from the users' perspective. User groups across the partner facilities have already started to use Notebooks where currently possible. We will also work with the Jupyter team and contribute required modifications to the code base back to the community in order to avoid duplication of code and effort, to make the code base of our services as sustainable as possible, and to contribute towards better data analysis beyond neutron and photon facilities and users.

Local use of Jupyter Notebook based data analysis services

Following from Task 4.1 we will prioritise existing data analysis services and make them available through the notebook, deploy this locally at our facilities, share experiences across partners, and gather feedback from users. The complexity varies: data analysis tools and libraries that can be scripted (through any of the Notebook

supported languages, including Python and bash) are relatively easy to integrate, while for some others we may have to provide such interfaces. Amongst other packages, we will port some of the functionality of the Scientific Library for eXperimentalists (silx - <http://www.silx.org>) data reduction and graphical library to the Jupyter Notebook, focussing on high priority items that are not already available through plotting tools and Jupyter Widgets. Silx is the base platform used by the ESRF to develop data reduction and analysis applications for data from synchrotron sources. It is used by at least 5 flagship applications for spectroscopy, diffraction, and ray-tracing developed by the ESRF, ELETTRA and MAXIV, and these applications are widely used in the synchrotron community.

JupyterHub for multi-user remote data analysis

In this task, we will adapt and use the JupyterHub project for data analysis service provision, that can be used remotely or locally. The JupyterHub software provides a multi-user server to host multiple instances of single-user Jupyter notebook servers, to offer data analysis services remotely. The notebook servers run in separate containers, and we need to allow users to select appropriate containers for the desired analysis chain, and will have to create the containers. We will have to make the existing deployment more robust and flexible and provide a near-automatic deployment of the system on a variety of servers (at least for involved facilities and EOSC), linking the installation to appropriate authentication and authorisation and corresponding visibility of data to analyse and user specific persistent storage. The JupyterHub service has to be integrated into the orchestration of available computing resources, and needs monitoring of resources used and required. We need to extend mechanisms that allow to upload and download data through the Jupyter interface, including making use of the data catalog and its API developed in WP 3.

Binder for data analysis and FAIR principles

In this task, we explore and exploit solutions to bundle the execution environment with notebooks.

We will make use of the binder project (<https://mybinder.org>) to support reproducible flexible definition and selection of compute environments, in which data analysis procedures - that are coordinated through Jupyter Notebooks - can be carried out. This supports realisation of the FAIR principle as for any derived data computed, we have recorded all computation steps (in the Notebook) and we have recorded the compute environment (through the container). There are at least two canonical places for notebooks and analysis environments to be stored: (i) with the raw data as part of the metadata, and (ii) with results obtained from some data set. We need to develop guidelines and solutions where resulting notebooks are stored, considering social and technical aspects, with links to WPs 2 and 3.

Because of the lightweight nature of the containers, we expect that we can use this model to take the computation environment to the data if required. For example if a user needs to analyse a data set that is hosted by the ESRF synchrotron facilities and the data set is so large that it cannot be easily moved, we can instead relocate the computation to take place at the ESRF: we need to move the relevant container and the notebook that contains the analysis recipe to the ESRF. Another use case for moving notebook and container – which is suitable for analysis on small data sets – is to run the execution of the notebook in the container on the user's Desktop.

Towards reproducible publications

As Jupyter notebooks can be saved, re-loaded and given the data and required compute environment (Task 4.4.2) also be re-executed, they help significantly in moving towards reproducible data analysis. The data science community in academia and enterprise has embraced the notebook as an executable document describing and documenting reproducible data analysis and as a high productivity tool. Beyond reproducibility, researchers can also easily extend a given reproducible analysis and thus build on existing research outputs without having to re-create the already published analysis as the first step.

In an ideal world, researchers would publish a scientific publication together with a (computer executable) script, a reference to the original data, and the computation environment in which the original data has been processed into the numbers and figures used in the publication. This would provide full reproducibility of the published results, and encourage other researchers to further exploit the data (currently researchers spent significant amounts of time to reproduce the earlier findings before embarking on new work). In this task, we extend the notebook to provide important functionality: if the code segments of the notebook could be hidden where desired, it would be possible to create the publication inside the notebook as the primary document and to export one version for publication (as LaTeX for example). The primary document should be published as an electronic supplement that

allows, for example, to re-create the data and figures and to modify the analysis for further research. Here, we will explore this option for data analysis in neutron and photon science. In particular, this requires the use and enhancement of the “nbconvert” and “bookbook” projects of Jupyter, and investigation of requirements from authors and publishers.

Exploitation of emerging technology and methods

We will exploit new emerging trends and ideas from other Jupyter users to benefit the vision of the EOSC. JupyterLab, for example, can be equipped with an in-built viewer for hdf5 files and with state-preserving widgets. It could further be used to conveniently log experiments by inserting code handles to retrieve the relevant data already during the experiment, and thus make it easier to provide more complete and explicit metadata. The new-ish projects NoteBook VALidate (NBVAL) and NoteBook DIff and MERge (NBDIME), funded by the OpenDreamKit project, can be integrated into our EOSC services to help users to understand where a notebook has changed (NBDIME), or to validate that all displayed data and derived entities are current and valid (NBVAL).

Task 4.5 Deployment of remote analysis services at PaNOSC facilities (M13-48), *Leader: XFEL.EU. Contributors: ESRF, ILL, ESS, ELI, CERIC-ERIC*

Following development of the technology at selected sites, we need to commission servers to offer the Jupyter based services (as described in Task 4.2 and 4.4) at the sites of project partners. We also need to work with project partners to make additional data analysis services available at their sites, for example through provision of suitable containers and enabling of selected libraries and packages for use through Jupyter. We will invite users as soon as possible to benefit from these services, even if initially only local data with a subset of available services can be analysed. The provision of all remote analysis services to the EOSC is covered in WP6.

Task 4.6 Publicly accessible demonstrator (M36-48), *Leader: CERIC-ERIC. Contributors: all*

To demonstrate the value of the EOSC and the science carried out at the research infrastructures, we will make available some data and services for which no authentication is required and which allow to appreciate the value of cloud hosted and accessible data and data analysis services (with links to WP8 and WP9).

Deliverables

Deliverable 4.1 Report on the current technical elements of data analysis at each partner site (M12, R, PU, ILL)

Deliverable 4.2 Prototype remote desktop and Jupyter service (M18, DEM, CO, ILL)

Deliverable 4.3 Remote desktop and Jupyter analysis service deployed at EOSC (M42, DEM, CO, XFEL.EU)

Deliverable 4.4 Publicly accessible demonstrator (M48, DEM, PU, CERIC-ERIC)

Work package number	5	Lead beneficiary				XFEL.EU
Work package title	Virtual Neutron and x-raY Laboratory (VINYL)					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC
Person months per participant:	40	36	48	36	24	40
Start month	1		End month	48		

Objectives

1. Expose existing instrument and experiment simulations capabilities as a virtual facility service in the EOSC to promote the access and integration of simulated data in complex analysis workflows.
2. Build state-of-the-art e-infrastructures, providing a flexible simulation framework that enables users to rapidly implement simulation and analysis workflows specific to their facilities, instruments, and experiments.
3. Make simulation data services inter-operable among themselves and with data analysis services and data catalogs through development of appropriate APIs and adoption of open data standards.
4. Enable RIs to seamlessly link the EOSC experiment simulation services to their in-house data reduction, analysis, and visualization infrastructures. Enable computational scientists to use the EOSC services and data catalogues (WP3) and analysis services (WP4) for validating their own bespoke structure or dynamics predictive modelling algorithms and to embed their tools in the EOSC.

5. Foster the acceptance and adoption of open standards for data formats and APIs related to simulation services in the photon and neutron science community by developing simulation applications suitable for education and outreach and by simulating data sets for testing purposes of data tools and services.

Description of work

Simulations of the various parts and processes involved in complex experiments play an increasingly important role in the entire lifecycle of scientific data generated at RIs: Starting with the idea for an experiment (often triggered by results from numerical and theoretical work), via design and optimization of experimental setups, estimation of experimental artifacts, generation of supporting material for beamtime proposals, assisting in decision making during an ongoing experiment to interpretation of experimental data, data analysis, and finally extrapolation from the obtained results which then leads to new experiments. In particular, the analysis of experimental data often involves simulations e.g. to refine molecular or crystalline structures measured in diffraction experiments [[White2012](#)]

Comparison between simulated and experimental data leads to insight not conceivable from experimental or simulated data alone:

1. For the acquired raw data (on-line comparison with previously run simulations to assess and monitor data quality, enabling (automated or guided) optimization of source, beamline, and instrumental configurations.
2. For the reduced data, as a sanity check for both experimental and simulated data.
3. For the results from the analysis step which had included all instrumental and experimental data and ended with a refined set of data for the particular sample.

Comparing reduced and analysed data is readily possible today, whereas it is not readily possible to compare raw data acquired from a real and a virtual experiment. Thus, algorithms to infer the likelihood that two sets of observations, one experimental and one theoretical, originate from the same underlying probability distribution need to be developed before it becomes possible to compare data without reducing them. Such algorithms would enable a direct validation of structural and dynamical sample (target) models and simulation techniques as they eliminate the dependency on the data reduction and analysis steps.

An objective of this work package is to facilitate the rapid prototyping and execution of data workflows that combine experimental data and simulations inside user friendly application frameworks as an EOSC service. Ultimately, this will be achieved by creating a cloud based virtual research facility that represents all major components of real photon and neutron RIs and thereby allows the exchange and coupling of data and services between the real facility and its virtual simulated counterpart with the overarching objective to boost the extraction of meaning and information from raw experimental and simulation data.

The elements of the virtual facility are schematically shown in the block diagram figure 1. A virtual photon or neutron facility experiment consists of a sequence of simulations describing the physical and conceptual entities of the experiment. Starting from a simulation or model of the photon or neutron source followed by propagation of photons or neutrons through beamline and instrument optics to yield a precise characterization of the beam (temporal, spectral, and spatial structure, degree of coherence, divergence, and polarization) and the very complex process of interaction of the beam with the sample (radiation damage) including the scattering of radiation and eventually the signal generation including a model or simulation of conversion of scattered intensity into a digital detector signal. Every process is simulated in a specific way. Often more than one implementation of a simulation algorithm and models of different levels of physical detail (e.g. ray tracing vs. wavefront propagation or atomistic first principle simulation vs. continuum models for radiation damage) exist. Establishing (where required) and maintaining (where already present) interoperability and consistency of simulated data between these codes and modules as well as harmonization of APIs is a central task of this work package and key to the realization of the virtual facility. Careful design and documentation of our APIs will enable partner and non-partner RIs to plug-in their specific simulation softwares into the simulation chain and to create customized workflows for the specific virtual experiments needed in their facility. Finally, our APIs enable the integration of simulation capabilities and workflows including configuration and execution of simulation jobs, as well as retrieval, processing, and visualization of resulting simulation data in high level user interface frameworks such as jupyter-notebook [[Kluyver2016](#)], Oasys [[Rebuffi2017](#)], and others. This will be readily used in WP8 for expanding the usage of an existing e-learning platform.

APIs developed in the SIMEX workpackage in EUCALL [Fortmann-Grote2017], the Atomic Simulation Environment (ASE [Larsen2017]), and MDANSE [https://mdanse.org/] will serve as prototypes and seeds for our developments that focus on the creation of simulation and analysis data services in the EOSC.

Task 5.1 Simulation Infrastructure (M1-M48) *Lead: XFEL.EU, Contributors: ESRF, ESS, CERIC-ERIC*

Harmonization of simulation code APIs (SIMEX, ASE, WOFRY) and data formats to enable and to support interoperable simulations as a cloud service.

- Harmonize APIs for beamline, sample trajectory, and signal generation simulation codes
- Adopt simulation data formats: openPMD for particle and mesh data, NeXus for detector data (see WP3)

Task 5.2 Photon and Neutron Source simulation data services (M1-M24) *Lead: ESRF Contributors: XFEL.EU, ELI, CERIC-ERIC*

Using the APIs from T5.1, expose photon source simulations as a cloud service for synchrotron and free-electron laser sources. Description, stockage and access to the parameters describing the source (storage ring Twiss parameters, insertion devices). Computation of the radiation. Decomposition of coherent modes with COMSYL [Glass2017] and remote storage. Simulation of photons and neutrons from ultraintense laser-plasma interaction. Population of a radiation source database with precomputed beams containing intensity distributions and wavefronts.

- Jupyter-notebook and execution environment (see WP4)
- local OASYS workflows to access remote instrument description and data
- remote desktop session and execution environment (see WP4)

Task 5.3 Photon and Neutron beamline simulation data services (M13-M36) *Lead: ILL, CERIC-ERIC Contributors: ESRF, XFEL.EU, ELI*

Expose photon and neutron beamline optics simulation services for photon and neutron facilities. Description of the beamline elements, deployment of scattering models for interaction of photon beams or wavefronts with optical elements (mirrors, crystals, lenses) and simulation data deposition in a database. Reuse existing libraries such as as SYNED, WOFRY, and support workflow-based high level user interface OASYS. Populate an instrument simulation database.

- Jupyter-notebook and execution environment (see WP4)
- local OASYS workflows to access remote instrument description and data
- remote desktop session and execution environment (see WP4)

Task 5.4 Simulation of signal generation including radiation-matter interaction (M25-M48) *Lead: ESS, Contributors: ILL, XFEL.EU, ELI, CERIC-ERIC*

Enable simulation of scattering signals from given sample structural dataset and beamline propagation data as a cloud service.

Simulate interaction of radiation (from T5.2) with sample structural data stored e.g. in NOMAD as well as scattering and absorption signals.

- Jupyter-notebook and execution environment
- remote desktop session and execution environment
- Protocol for comparison of raw simulated data vs. raw experimental data

Task 5.5 Integrated simulation data workflows (M37-M48) *Lead: XFEL.EU Contributors: ESRF, ILL, ESS, CERIC-ERIC*

- Expose simulation data services in data analysis frameworks accessed via Jupyter notebooks or remote desktop solutions.
- Iterative data analysis workflows including experiment simulations

Deliverables:

Deliverable 5.1 Prototype simulation data formats as openPMD domain specific extensions including example datasets (M12, R, PU, XFEL.EU, ESS)

Deliverable 5.2 Release of documented simulation APIs (M24, O (Software), PU, XFEL.EU+ILL)

Deliverable 5.3 Repository of documented jupyter notebooks and Oasys canvases showcasing simulation tasks executable via JupyterHub or remote desktop. (M42, O (Software), PU, XFEL.EU+ESRF)

Deliverable 5.4 VINYL software tested, documented, and released, including integration into interactive data analysis workflow with feedback loop. (M48, R+Software, PU, All)

Work package number	6		Lead beneficiary				ILL
Work package title	EOSC integration						
Participant number	1	2	3	4	5	6	7
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC	EGL.eu
Person months per participant:	21	38	13	13	12	13	82
Start month	1		End month		48		

Objectives Integrate the Photon and Neutron data catalogues and services in the EOSC.

This work aims to integrate the PaNOSC cluster with EOSC through a strong collaboration the EOSC-Hub project and more generally with the e-Infrastructures and other Research Infrastructures jointly contributing to the realization of the EOSC Hub. In order to achieve a smooth experience for users, this effort will take place at various levels.

- Strategic: by engaging with other EOSC stakeholders in order to contribute to the definition of the EOSC implementation roadmap.
- Executive: by contributing PaNOSC data, resources and services to the EOSC service catalogue.
- Technical: by making use of relevant EOSC services (e.g. AAI, marketplace, Cloud Compute, Data Archiving and Data Management services) contributed by other providers and initiatives to ensure economies of scales, and a more integrated service offering to the end-user.

As a cluster of Research Infrastructures wanting to fully integrate within EOSC, with data at the core of our production, PaNOSC will contribute the following resources and services:

- Curated Open Data and metadata of the highest quality
- Reliable services dedicated to understanding and to further exploiting these data
- Technical and scientific support on these data and data services
- Our experience on FAIR data policies and FAIR implementation guidelines for Photon and Neutron science
- Our knowledge and understanding of our scientific community
- Our ability to promote FAIR culture amongst our community

In order to integrate these assets within EOSC, we have engaged to work with the European e-Infrastructures and more particularly those participating in the shaping of the EOSC. We have identified core activities that are vital to its success:

- Active participation in governance
- Active Participation in open policies activities (WP2 - Task 2.2)
- Integration of our data catalogues into the EOSC data catalogue (WP3 -Task 3.4)
- Use of E-Infra IT services to deploy more specific services targeted at Photon and Neutron data type and users.
- Provisioning of models and solutions to bring small datasets to the compute resources and vice versa for very large datasets.
- Commonly defined service quality levels (Service Level Agreements) and if necessary upgrade the services to reach and maintain reliably this level of quality.
- Commonly defined usage metrics and the adoption of the necessary tools to collect and publish them
- Harmonization of solutions for federated identity provisioning, authentication and authorization.
- Set up a technical and scientific support structure for handling data scientist (not necessarily facility users) requests
- Promoting FAIR data culture.

Some of these activities will take place in dedicated work packages and tasks (specified in brackets), the others that support more than one work package, will be handled by this WP.

Description of work

Task 6.1 Management of the Interaction with the e-Infrastructures. (M1-M48) *Lead: ILL, Contributors: ESRF*

To ensure the success of the EOSC platform we should build strong links and get a mutual understanding with the different e-Infrastructures (EGI, OpenAIRE, EUDAT, GÉANT) and other relevant e-projects participating in the construction of the EOSC. This task aims at collecting needs and requirements from the PaNOSC partners and provides a collective response to the other projects. This is absolutely necessary at this initial stage of the EOSC construction process. Foreseen activities in this task will be:

- Participation in conferences, workshops and meetings to share the needs of the community and provide information and feedback to the PaNOSC partners on the EOSC progress and more specifically on how we can efficiently integrate.
- During the course of the project new questions will arise, to address them collectively we will build surveys and ensure comprehensive responses and analyses.
- Participation in the EOSC governance. Even if the global governance is not yet well defined, we anticipate to contribute at least through participation into strategic boards of projects like the EOSC-Hub.

Task 6.2 EOSC Hub Service Catalogue (M1-M48). *Lead: ESRF, Contributors: all*

Following the work done on the cataloguing of e-Infrastructure services by the eInfraCentral project, EOSC Hub is going to provide the same type of catalog for the EOSC services including RI services.

We will participate actively in the discussion for the definition of the services harmonisation. These activities include elaborating standards for service description, classification, selection of metrics/KPIs, tools for collecting these metrics and definition of appropriate service level agreements (SLAs) and monitor the respect of these SLAs. Once the standards are defined, we will prepare accordingly the PaNOSC services for the integration into the EOSC service catalogue. Which includes:

- Provide for each service information and description in standardised form
- Implement tools into the services allowing to collect and report standard metrics
- Ensure for each services, through testing, that SLAs are met, if not, we will provide and implement action plans.

Task 6.3 Data availability for the services (M1-M48). *Lead: ELI, Contributors: all*

One of the ultimate goals of this project is to support data scientists, who are not necessarily experienced users of the facilities, with a combined offer of distributed open data repositories, co-located with cloud compute IaaS and high level applications for data analysis. Beside the difficulty of federating distributed open data, the technical challenge will be to make these federated data transparently accessible by computing resources running on different cloud environments (e-infrastructure, research infrastructures, ...).

- For services, where data has to be moved to computers, implement the integration of the EGI data-hub technology into the facility repositories. Test movement of data and understand the limit of such model.
- For services where data are too big to be moved, we would like to test the integration of local resources into the EOSC compute cloud. We also need to evaluate security constraints and necessary measures.

In this task we will first pilot the technical solutions with one facility before rolling out to the other partners.

Task 6.4 Authentication Authorisation Infrastructure (AAI) integration (M1- M36). *Lead: ILL, Contributors: all*

A common identification of the users by the different service providers is key to the construction of the EOSC. This will be ensured by the different identity providers either through unification or through the means of technical solutions ensuring complete interoperability. Authorisation, level of Assurance (LoA), Security Incident Response and compliance to personal data regulations also need to be addressed globally. The Photon and Neutron community is currently operating its own AAI infrastructure: umbrellaID.org. This AAI solution has benefited from a long history of fruitful collaborations with GÉANT and we are pleased that they have accepted to work with us on this project. GÉANT, via a letter of support (see annex after Section 5) has committed to work together with the PaNOSC project partners to help scope their AAI requirements, design and deploy a sustainable AAI solution that meets those requirements and ensures the secure integration of the PaNOSC services in the EOSC. In this task, in collaboration with GÉANT we will:

- Study the feasibility, potential impacts and sustainability of the possible models for integrating the Photon and Neutron AAI with EOSC.

- Present, discuss and reach agreement inside the Photon and Neutron facility community at large (the PaNOSC partners and other members of the AAI consortium) on the integration of the PaNOSC AAI infrastructure, delivered with GÉANT, into EOSC.
- Implement this integration at the level of the Identity providers (IdP).
- Provide solution and documentation for the integration into the different services that PaNOSC is providing.

Task 6.5 Services Support (M12-M48). *Lead: ELI, Contributors: all*

Organise an integrated technical and scientific Helpdesk that will give support to data scientists (i.e scientist that would like to use the PaN open data but are not necessarily users of the facilities). This organisation should ensure that all requests are addressed following the Service Level Agreements published in the EOSC service catalogue, and that activities are aligned with the customer relationship process to be jointly defined with the support of the EOSC-hub project

Task 6.6 PaN Software catalogue (M12-M36). *Lead: ILL, Contributors: all*

The Photon and Neutron community is using a software catalogue that not only references the analysis and simulation software in use and supported at the facilities but also provides complete examples with data sets and practical information for scientific instruments where they are used. This is an important documentation tool for facility users that we need to integrate into the EOSC, to make it more accessible to the whole scientific community interested into our open data. We have identified two main tasks:

- Implement in the current catalogue the missing features, for instance docker/image registry, to meet the level of EOSC standard.
- In collaboration with the EOSC-hub, define and implement APIs that could allow its integration into the EOSC database catalogue. Selected software presented in the PaNData software catalogue and officially supported by at least one of the partners will benefit of this integration by being also referenced in the EOSC database catalogue.

Task 6.7 Data archiving pilot (M12-M48). *Lead: ELI, Contributors: ESS , EGI.eu*

Some of the facilities do not have a long term archiving capability and need to explore solutions for archiving their data directly into EOSC. Through this activity we will get insight on the real technical requirements necessary for archiving high throughput production data over the Internet and on the suitable economical and organisational model. This task will provide a global and practical feasibility study limiting the data to be archived to 4 PB in total over 4 years.

Task 6.8 Procurement of commercial cloud services (M30-M42). *Lead: ESRF, Contributors: ESRF, ILL, XFEL.EU, ESS, ELI*

Commercial cloud services may constitute an important alternative “scale-out” solution for peak demands of data analysis needs in the RIs. However, as of today the RIs do not have adequate mechanisms in place to procure and use commercial cloud services in a flexible and secure manner. This task will allow to tender commercial cloud services for all partners in a mutualised manner and acquire practical experience of how to allocate resources to individual scientists. Although initially targeted at “in-house” scientists it may at a later stage allow to enlarge the service offering of the RIs. The procurement activity will profit from experience with the e-Infrastructures, but also from the PCP procurement project HNSciCloud led by CERN. This task will require to work in close relationship with the purchasing departments of the partner RIs.

Deliverables

Deliverable 6.1 EGI data-hub integration with the facilities’ data repositories (M18, R, PU, ELI)

Deliverable 6.2 Integration of local compute resources into the EOSC cloud (M12, R, PU, ELI)

Deliverable 6.3 Integration of the PaN AAI into the EOSC (M36, R,DEM, PU, ILL)

Deliverable 6.4 Demonstration of the PaN software catalogue integration into EOSC (M24, DEM, PU, ILL)

Deliverable 6.5 Report on EOSC integration (M48, R, PU, ILL)

The deliverable will report on PaNOSC organizational, technical and strategic activities contributed to establish a PaNOSC data, application and services Commons. In particular, the deliverable will report on EOSC service management processes and the related performance, including incident management, data archiving and distributed computing in EOSC, and experience in cloud procurement and adoption.

Work package number	7		Lead beneficiary			CERIC-ERIC
Work package title	Sustainability					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC
Person months per participant:	3	3	3	3	12	32
Start month	1		End month		48	

Objectives

Propose a business plan on how to sustain the data catalogs and services in the Photon and Neutron community and as part of the EOSC. In particular:

- Coordination with national or international related initiatives and support to the deployment of global and sustainable approaches in the field including coordination with EGI and the other EOSC stakeholders like RDA, the PaNdata community and LEAPS (League of European Accelerator-based Photon Sources) initiative (<https://www.leaps-initiative.eu>) .
- Study - even by using advanced methodologies - of the cost per partner for maintaining the infrastructure required for providing FAIR data (archiving, data services etc.) and explore different scenarios for financing the long term costs.

Description of work

The work package is organised in 4 tasks most of which are connected to iterative processes that are best organized following the principles underlying the Deming cycle. The connection with the stakeholders and the collection of their feedback is a critical aspect of the work package.

Task 7.1 Stakeholders for the Photon and Neutron community EOSC (M1-M48)

(Lead: CERIC-ERIC; Participants: ESRF, ILL, XFEL.EU, ESS, ELI)

Definition of a database of stakeholders. Creation of links with the main players of the EOSC-hub, to the RDA and PaNdata community, ERF Data working group, and relevant industries in order to be able to collect input and feedback from them. The stakeholders will be involved in surveys during project execution in order to collect their important feedback. Meetings with stakeholders will be organised to facilitate interactions with the community and possibly other cluster projects in conjunction with other meetings and as part of events related to WP8 and WP9.

Task 7.2 Metrics and cost for the Photon and Neutron community EOSC (M9-M36)

(Lead: CERIC-ERIC; Participants: ESRF, ILL, XFEL.EU, ESS, ELI)

Analysis and development of metrics for the evaluation of costs and added value of the services provided to the community. This clearly depends and connects to the developed data policies and on the overall architectural choices for the Photon and Neutron community EOSC.

Task 7.3 Business models for Photon and Neutron EOSC (M13-M42)

(Lead: CERIC-ERIC; Participants: ESRF, ILL, XFEL.EU, ESS, ELI)

Development of advanced business and funding models in connection with Industrial Liaison Offices of each facility, the user communities and all the relevant industrial and research community EOSC stakeholders.

Task 7.4 Sustainability plan for the Photon and Neutron EOSC (M19-M48)

(Lead: CERIC-ERIC; Participants: ESRF, ILL, XFEL.EU, ESS, ELI)

Development of a formal long-term mission and vision for the sustainability of the PaNOSC infrastructure and software developed which will balance the viewpoints of the different stakeholder and the developed business models.

Deliverables

Deliverable 7.1 Photon and Neutron EOSC Stakeholder Feedbacks (M18, R, PU, CERIC-ERIC)

Deliverable 7.2 Photon and Neutron EOSC metrics and costs model (M36, R, PU, CERIC-ERIC)

Work package number	8	Lead beneficiary				ESS
Work package title	Staff and User Training					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC
Person months per participant:	6	9	4	30	48	6
Start month	1		End month		48	

Objectives

The objectives in this work package are:

1. *Provide infrastructure and service for e-learning* (Task 8.1 to 8.3). Provide an e-learning platform and service that can be used by all facilities to provide training to staff and users. The e-learning platform will be based on e-neutrons.org that integrates three teaching components;
 - a. Moodle,
 - b. MediaWiki,
 - c. virtual facility that enables students to perform virtual experiment simulations.
2. *Staff training in data stewardship and e-learning platform* (Task 8.4 and 8.5). Develop training material for data stewardship to foster faster adoption of best practices and for how to use the e-learning platform for developing courses, and train staff at relevant RIs at specific workshops
3. *User training in PaNOSC services and facilities* (Task 8.6 to 8.8). Develop training material for the PaN user community to promote the FAIR principles and best practices as well as for introducing users to the PaNOSC services and capabilities of PaNOSC facilities.

The work package will be led by ESS and co-led by ELI. Thus, all tasks are led by one of those two facilities. Where needed, other facilities are also involved.

This work package is reliant on WP3-5.

Description of work

The WP objectives will be fulfilled by the following tasks:

Task 8.1 Sustainability and maintainability of e-learning platform (M4-M15) *Leader: ESS. Contributors: ELI*
 The purpose of this task is to ensure the sustainability and maintainability of the e-neutrons.org service where +800 users currently have an account. Sustainability will be pursued by migrating e-neutrons.org to ESS, which has the resources to sustain and maintain it long term as a part of their user programme, and integrate the service with the EOSC in collaboration with WP6. Effort will also be dedicated to setting up software development infrastructure (test suites and build servers) to make the service maintainable beyond the current project. An analysis of different solutions for ensuring that sufficient CPU resources are available during peak-loads, e.g. during courses, will be performed followed by the implementation of an actual solution. A new domain name pan-learning.org, will be employed in order to cater for the PaN community as a whole (MS8.3)

Task 8.2 Integrating Jupyter technology (M19-M30) *Leader: ELI. Contributors: ESS*
 In collaboration with WP4 Jupyter technology will be integrated into the e-learning platform so that Jupyter notebooks can be launched from the platform and used to provide teaching material. The integration of Jupyter technology will be particular beneficial for developing training material in Tasks 8.4 and 8.5 for the services developed in WP3-5. The dependency on WP4 is identified as a risk and will be mitigated by creating a joint plan for WP4 and WP8 early in the project (MS8.1) that will be updated on a monthly basis in video conferences between the two involved WP leaders. Milestone MS8.4 indicates that integration is completed.

Task 8.3 Integrate ESRFs in the e-learning virtual facility (M25-M36) *Leader: ESS. Contributors: ESRF, ILL, XFEL.EU, ELI, CERIC-ERIC*

Only the two neutron sources ILL and ESS have adapted a community standard for providing e-learning materials in the form of e-neutrons.org. The aim of this task is to extend e-neutrons.org with the APIs provided in WP5 with Deliverable D5.2 thus enabling the x-ray / light sources to create their own training modules that can benefit from facility specific virtual instruments. This task will make it possible to make training modules that highlight how the different facilities complement each other. The dependency on WP5 is identified as a risk and will be mitigated by creating a joint plan for WP5 and WP8 early in the project (MS8.2) that will be updated on a monthly basis in video conferences between the two involved WP leaders. Another milestone (MS8.5) indicates that at least one virtual instrument per facility is integrated into the platform

Task 8.4 Staff training in e-learning platform (M13-M42) *Leader: ESS. Contributors: ESRF, ILL, XFEL.EU, ELI, CERIC-ERIC*

A workshop will be prepared and held during the project to foster uptake of the e-learning platform by the PaNOSC beneficiaries. In the workshop, the participants will be trained in using the platform for developing both passive and interactive training material for their own teaching. Mentoring to developers of training material will be provided subsequently to the workshop. During the mentoring phase new functionality from Task 8.2 and Task 8.3 will be introduced. Likewise, the mentoring phase will be used to retrieve feedback on usability. Besides providing a service to the PaNOSC facilities, this task is also seen as an important step to ensure the sustainability of the e-learning platform. A report covering lessons learned from the workshop and mentoring phase and with an outlook to the future will be delivered (D8.2).

Task 8.5 Staff training in data stewardship (M25-M32) *Leader: ELI. Contributors: ESRF, ILL, XFEL.EU, ESS, CERIC-ERIC*

The purpose of this task is to upgrade staff's skills in data stewardship. At workshops targeting staff at the PaNOSC beneficiaries and other interested RIs (e.g. the national facilities), the FAIR principles will be promoted and the toolset of a modern data culture (concept of PID's, Orcid, DataCite, etc.) will be introduced. Moreover, the developed PaNOSC policies and services (WP2-WP5) will be introduced in the context of employing proper data stewardship procedures. Course material in the form of videos, how-to's, webinars, and testimonials from fellow scientists already following such practices, will be available from the e-learning platform, which also will be used for developing interactive course material (e.g. quizzes) based on existing functionalities and functionalities developed in Tasks 8.2 and 8.3. In this way the material can also be used by others and for self-learning. A report covering lessons learned from the workshop and mentoring phase and with an outlook to the future will be delivered (D8.1).

Task 8.6 Training materials to foster the uptake of PaNOSC services (M31-M42) *Leader: ELI. Contributors: ESRF, ILL, XFEL.EU, ESS, CERIC-ERIC*

The aim of this task is to ensure that adequate self-training material is available to enable users to use the services developed in this project (WP3-5 and Task 8.1) for their own needs. Thus a set of training material, which enables trainees to retrieve stored open access data using services from WP3, analyse them using services from WP4, and also perform a virtual experiment using services from WP5 and analyse these data using services from WP4. This task will thus be done in close collaboration with WP3-5. The training material will be based on the e-learning framework and made available through the e-learning platform. Videos and Jupyter tutorials linked to the virtual facility are anticipated to be the preferred didactical tools. The services together with this training material will be presented at relevant user meetings of the participants and the observers from the PaN community.

Task 8.7 Introductory course to materials science LSFs in the European Research Area (M31-M42) *Leader: ESS, Contributors: ESRF, ILL, XFEL.EU, ELI, CERIC-ERIC*

An e-learning course will be developed for students with the purpose of introducing them to the specific strengths of each of the PaNOSC beneficiaries. This will be based on the existing technology in the e-neutrons.org as well as on new functionality and training material developed in this WP. Most noticeably the students will be able to compare the outcome of virtual experiments for the same sample at different instruments at multiple facilities and analyse the data. The tutorials will also show and discuss discrepancies between virtual and real experiments. The e-learning course will specifically contain functional Jupyter tutorials that guide the students through making a virtual experiment at multiple facilities, analyse and compare the results including with results from real experiments. A quiz will also be available that demonstrates that students understand basic concepts of light- and neutron scattering techniques and that they got the expected outcome from the virtual experiments and analysis exercises.

Task 8.8 PaNOSC schools for students (M42-M48) *Leader: ELI, Contributors: ESRF, ILL, XFEL.EU, ESS, CERIC-ERIC*

At least one summer school for graduate students will be organised towards the end of the project in collaboration with the Hercules school for neutron and synchrotron radiation (see Letter of Support). The purpose of the course is to 1) enable scientists to better leverage the European Research Area by guiding scientists towards the facility where the cost-benefit is the highest, 2) promote the FAIR principles, and 3) introduce students to the services developed in PaNOSC. The course will be based on the materials developed in this WP. In collaboration with WP9, other channels for courses will be investigated and exploited, such as the facilities' user meetings.

Deliverables

Deliverable 8.1 Report on lessons learned and future prospects for adopting best practises on data stewardship at the PaNOSC facilities, task 8.5 (M32, R, PU, ELI)

Deliverable 8.2 Report on lessons learned and future prospects for adopting the e-learning platform at the PaNOSC facilities, task 8.4 (M42, R, PU, ELI)

Deliverable 8.3 Teaching material for users of PaNOSC services, FAIR principles, and the PaNOSC facilities accessible in the e-learning platform at pan-learning.org, task 8.5-7 (M42, DEC, PU, ESS)

Deliverable 8.4 Closing report including report from summer school, task 8.8 (M48, R, PU, ESS)

Work package number	9	Lead beneficiary				CERIC-ERIC
Work package title	WP9 Outreach/Communication and Dissemination/Impact					
Participant number	1	2	3	4	5	6
Short name of participant	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC
Person months per participant:	6	6	6	6	6	40
Start month	1		End month		48	

Objectives

This WP aims at engaging all project stakeholders in further developing and integrating services into the EOSC (external communications), and at ensuring a smooth flow of information among project partners (internal communications). Another goal is to ensure the dissemination and foster the exploitation of the project outputs and results by all available means.

Therefore, WP9's objectives include:

- Keeping all project partners and stakeholders informed and up to date about the PaNOSC progress.
- Support the implementation of outreach activities and events, to provide information about the EOSC functionalities and operation, and to disseminate the project's outputs (policies, standards, methodologies, technical and operational information, etc.) to different stakeholders.
- Assist the exploitation of outputs, in particular adoption of common standards, sharing of policies and the use of the developed services by research institutions, the industry and the research community as a whole.

External communication activities will target the following stakeholders:

- Ministries responsible for science in the Member States and the European Commission with the goal of facilitating and fostering the coherent development and adoption of the EOSC locally, and of broadly promoting the developed services in all EU countries.
- Research institutions and infrastructures, to facilitate the integration and harmonisation of Photon and Neutron catalogue and services into the EOSC, and to support the further development of EOSC services according to specific scientific needs and requirements.
- Project managers of similar EOSC projects from other clusters to exploit synergies.
- The PaNdata (<http://pan-data.eu>) community of IT professionals from different research infrastructures, with the goal of stimulating interactions, exchange of knowledge and best practices.
- Other clusters involved in the implementation of the EOSC, to foster coordination, interoperability between disciplines and a more efficient use of resources.

- National research communities and science and technology professionals, to inform them about the functionalities and possibilities offered by the EOSC to storage, manage, analyse and re-use data linked to their research activities, across borders and scientific disciplines.
- Representatives from the industrial sector, with the aim of getting them acquainted with the EOSC's catalogue and services.

Description of work:

This WP includes all activities related to the project's Communication, as well as to the Dissemination and Exploitation of the project's results.

The WP Leader will be responsible for setting-up and managing all the tools necessary for ensuring a smooth communication both within the partnership, and between the project's partners and its stakeholders (T9.1).

All project partners will be involved in the dissemination of the project's results, as well as of the standards, policies and procedures developed throughout the project. They will present the developed policies, methodologies and tools to the national authorities and, in a coordinated way, to the European Commission, and they will participate in dissemination events involving their scientific and industrial networks (T9.2).

The WP leader and the partners will support the promotion of the training platform and actions developed in WP8 – Users Training.

Finally, this WP will promote the transfer of best practices (e.g., policies, strategies, tools and technologies) to other INFRAEOSC-04 clusters (T9.3).

Task 9.1. PaNOSC's internal and external communications (M1-M48)

(Lead: CERIC-ERIC; Participants: ESRF, ILL, XFEL.EU, ESS, ELI)

This task includes all activities meant to ensure a proper information flow within the partnership, and between the partnership and its main stakeholders, so that project objectives are clearly communicated to all target groups.

The project coordinator and manager will take lead of internal communications, whereas the leader of WP9 will have to grant support for putting in place all necessary tools for this purpose. Moreover, the WP9 leader will coordinate external communications and provide support in the promotion of the public activities foreseen in the different WPs.

All communications tools meant for this purpose will be set-up and implemented in the first months of the project, and will be regularly managed and updated throughout the whole period of implementation. Such tools will be defined in the project communication plan (D9.1) and will include the project website (D9.2), an online repository (based on cloud technology, e.g. D4SCIENCE, Basecamp, Slack, Asana) to share internal documents and information within the partnership (D9.3), the creation and management of the PaNOSC Twitter account, the regular publication of news and updates on partners' social media channels incl. a Women in Science section on the project website, the preparation and delivery of the project communication material (.ppt templates, brochures, leaflets, rollups, conference folders, gadgets) and, when appropriate, the release of news and articles among relevant contacts and networks. Key actions and events of the community will be also part of the yearly plans, in order to better deal with the big variety of partners and clusters, and to maximise communications efforts.

The communication strategy (M9.1) will provide useful details about the target groups to be addressed and strategic actions for their outreach, in order to ensure that all key audiences are reached by relevant communication for the whole duration of the project.

The specific objectives of this task are the following:

- Ensure that all partners and stakeholders are up to date about the project's goals and progress, by putting in place all the necessary tools.
- Foster the involvement of the photon and neutron community in the development process of the data catalogue (WP3), the data analyses services (WP4), and virtual facility services (WP5) to be harmonised and included in the EOSC;

- Stimulate the participation of national ministries in providing the information about the catalogue and data analysis services available in different countries, and inform them about further developments.
- Accompany the activities in the WPs to ensure that the right information reaches the target audience.

Task 9.2. Dissemination of PaNOSC's results (M13-M48)

(Lead: CERIC-ERIC; Participants: ESRF, ILL, XFEL.EU, ESS, ELI)

The project's dissemination plan (D9.1) will include the actions foreseen to inform and increase the awareness among PaNOSC stakeholders about the main outputs of the project. The deployed tools will often coincide with the ones used in communications (e.g. website, social media, press articles, etc.), and the message will be tailored for the target groups who will use the results, with the final aim of stimulating the use of the technologies and tools developed throughout the project. Specific meetings/conferences with a dissemination goal (D9.4) are also foreseen, and will involve main stakeholders: national science ministries, EC representatives, IT experts, RIs' scientific directors/personnel and industrial representatives. In particular, to address ministries and the EC, the partnership will capitalise on already existing events (ICRI Conference, target events of the EU presidency, meeting of RIs' governing boards, RDA etc.); to target the users' communities, the project partners will attend users meetings at the different RIs, as well as target events, such as the conference on Synchrotron Radiation Instrumentation (SRI), the European Research Facilities – ERF-AISBL meetings, ACCELERATE, LEAPS, EIRO-Forum CALIPSOplus and SINE2020 project events, as well as the scientific directors' meetings regularly planned in the different RIs. To ensure outreach to the experts involved in the development of the research e-infrastructures, scheduled EOSC events will be attended.

Finally, the annual PaNOSC's meetings will be used to invite target actors to boost the dissemination of main project outputs and achievements, and to foster the exploitation of results in the long term.

In order to ensure that results are properly communicated, the partnership will make efforts to keep regular contacts with other clusters throughout the whole project.

Task 9.3. Collaboration actions with other clusters (M1-M48)

(Lead: CERIC-ERIC; Participants: ESRF, ILL, XFEL.EU, ESS, ELI)

FAIR principles, their interpretation, associated policies and necessary implementation methods, as well as EOSC integration options, are developed in all projects funded in the INFRAEOSC-04 call. This task concentrates on communications, information flow and collaboration between the cluster projects that are funded as part of the INFRAEOSC-04. It includes maintaining frequent updates on the Project progress and developed policies, strategies, tools and technologies to organise the collaboration between the clusters via physical and virtual meetings, workshops and working groups. The task also includes planning and execution (supporting travel and workshop organisation or publication costs in the case of reports) of these collaboration activities within the Consortium as such opportunities are identified during project time. More possibilities of cooperation will be explored in the course of the project.

Deliverables

Deliverable 9.1 PaNOSC's Communication and Dissemination Plan (M7, R, CO, CERIC-ERIC), which will define communication and dissemination tools and actions according to the project's specific objectives.

Deliverable 9.2 PaNOSC's Website (M6, DEC, PU, CERIC-ERIC) - Set-up, content creation and update and online publication

Deliverable 9.3 PaNOSC's repository for internal communications (M3, DEC, CO, CERIC-ERIC) based on Cloud technology, e.g. D4SCIENCE, Basecamp, Slack, Asana

Deliverable 9.4 Dissemination and Outreach activities (M48, DEC, PU, CERIC-ERIC) - Report

3.1(c) PaNOSC Deliverables

The initial list of deliverables is:

WP	Deliverable number	Deliverable name	Short Name of lead participant	Type	Dissemination level	Delivery date (in months)

1	D1.1	Project Initiation Documentation	ESRF	R	PU	M2
1	D1.2	Mid-year summaries	ESRF	R	PU	M6, M18, M30, M42
1	D1.3	Annual Workshop report	ESRF	R	PU	M12,M24,M36, M48
1	D1.4	Data Management Plan	ESRF	R	PU	M6
2	D2.1	PaNOSC data policy framework updated	ESRF	R	PU	M18
2	D2.2	DMP Template published	ESS	R	PU	M36
2	D2.3	Guidelines published.	ESRF	R	PU	M24
2	D2.4	Integration of the policy in the User Access and facility information systems	CERIC-ERIC	DEC	PU	M36
3	D3.1	API definition	ESS	R	PU	M18
3	D3.2	Demonstrator Implementation	ESS	Other	PU	M28
3	D3.3	Catalog service	ESS	DEC	PU	M40
3	D3.4	Implementation Report from Facilities	ESS	R	PU	M44
3	D3.5	NeXus Metadata Mapping Schema and Proposed New Definitions	ESS	R	PU	M42
4	D4.1	Report data analysis capture	ILL	R	PU	M12
4	D4.2	Prototype remote desktop and Jupyter service	ILL	DEM	CO	M18
4	D4.3	Remote desktop and Jupyter analysis service deployed at EOSC	XFEL.EU	DEM	CO	M42
4	D4.4	Publicly accessible Demonstrator	CERIC-ERIC	DEM	PU	M48
5	D5.1	Prototype simulation data formats	ESS	R	PU	M12
5	D5.2	Documented simulation APIs	XFEL.EU	O	PU	M24
5	D5.3	Documented simulation tasks executable	ILL, ESS	O	PU	M42
5	D5.4	Software tested and released including interactive simulation and analysis workflow	ESRF	R,O	PU	M48

6	D6.1	Data-hub	ELI	R	PU	M18
6	D6.2	Compute cloud	ELI	R	PU	M12
6	D6.3	AAI	ILL	R, DEM	PU	M36
6	D6.4	Software catalogue	ILL	DEM	PU	M24
6	D6.5	Report on EOSC integration	ILL	R	PU	M48
7	D7.1	Photon and Neutron EOSC Stakeholder Feedback	CERIC-ERIC	R	PU	M18
7	D7.2	Photon and Neutron EOSC metrics and costs model	CERIC-ERIC	R	PU	M36
7	D7.3	Photon and Neutron EOSC Business model reference document	CERIC-ERIC	R	PU	M42
7	D7.4	Photon and Neutron EOSC Sustainability plan	CERIC-ERIC	R	PU	M48
8	D8.1	Report on lessons learned and future prospects for adopting best practises data stewardship at the PaNOSC facilities	ELI	R	PU	M32
8	D8.2	Report on lessons learned for adopting the e-learning platform at the PaNOSC facilities, task 8.4	ELI	R	PU	M42
8	D8.3	Teaching material for users of PaNOSC services, FAIR principles, and the PaNOSC facilities accessible in the e-learning platform at pan-learning.org, task 8.5-7	ESS	DEC	PU	M42
8	D8.4	Closing report including report from summer school, task 8.8	ESS	R	PU	M48
9	D9.1	PaNOSC's Communication and Dissemination Plan	CERIC-ERIC	R	CO	M7
9	D9.2	PaNOSC's Website	CERIC-ERIC	DEC	PU	M6
9	D9.3	PaNOSC's repository for internal communications	CERIC-ERIC	DEC	CO	M3
9	D9.4	Dissemination and Outreach activities	CERIC-ERIC	DEC	PU	M48

Table 5: Deliverables

3.2 Management structure, milestones and procedures

3.2.1 Project management methodology

The project management methodology used for PaNOSC will be based on same principles as PRINCE2, a well-known structured project management methodology. The project will be managed primarily by the coordinator, ESRF, in WP1. ESRF has strong experience in managing projects with a success, the most recent being the EBS (100 million euros project) and its associated beamlines projects (15 new end stations). The ESRF project manager is a member of the PaNOSC proposal and will be leading the management WP.

This will result in the project being divided into different stages with deliverables and milestones used as control points. Clear roles and responsibilities will be assigned to all persons involved in the project.

3.2.2 Roles and organisational structure

All the partners will agree on the creation of an *Executive Committee* and the rules for the appointment of its members by the end of the second month after the approval of the project as per section 3.2.2.1. This committee will be the decision making body for the project and the ultimate responsible for its success.

The *Project Coordinator* has been leading the creation and coordination of this proposal and during the project phase will be in charge of overseeing the project manager and ensuring that the data management plan and project initiation documentation are implemented (risk management strategy, internal communication strategy, etc.).

The *Executive Committee* will describe the level of delegated authority for the Project Manager who will be appointed by the PaNOSC coordinator (ESRF).

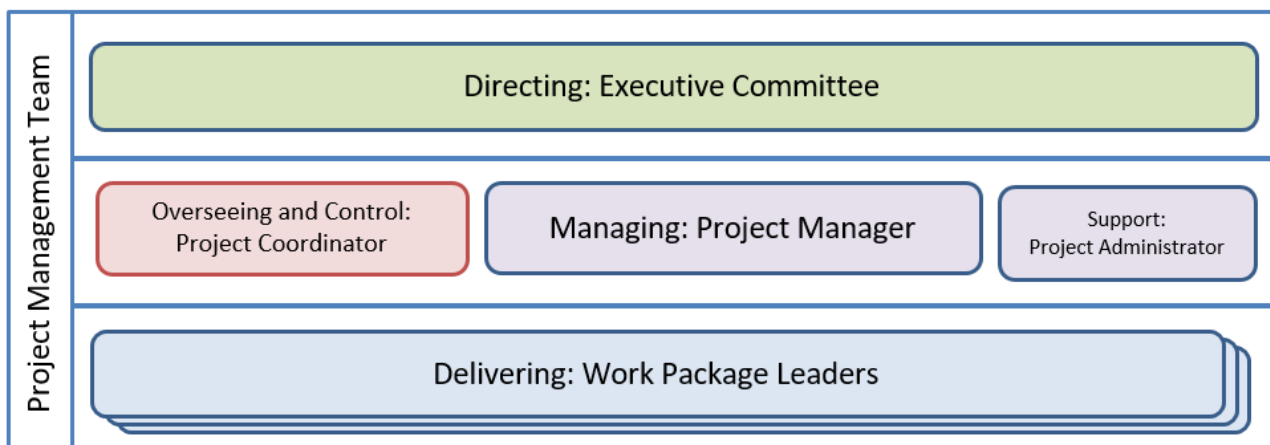


Figure 8: PaNOSC Project Management Team

The *Project Manager* will be notified from each leading beneficiary of the appointment of a *Work Package Leader* in charge of delivering each respective work package and inform the *Executive Committee* and all other stakeholders of these appointments. The *Project Administrator* will provide administration support to the *Project Manager*.

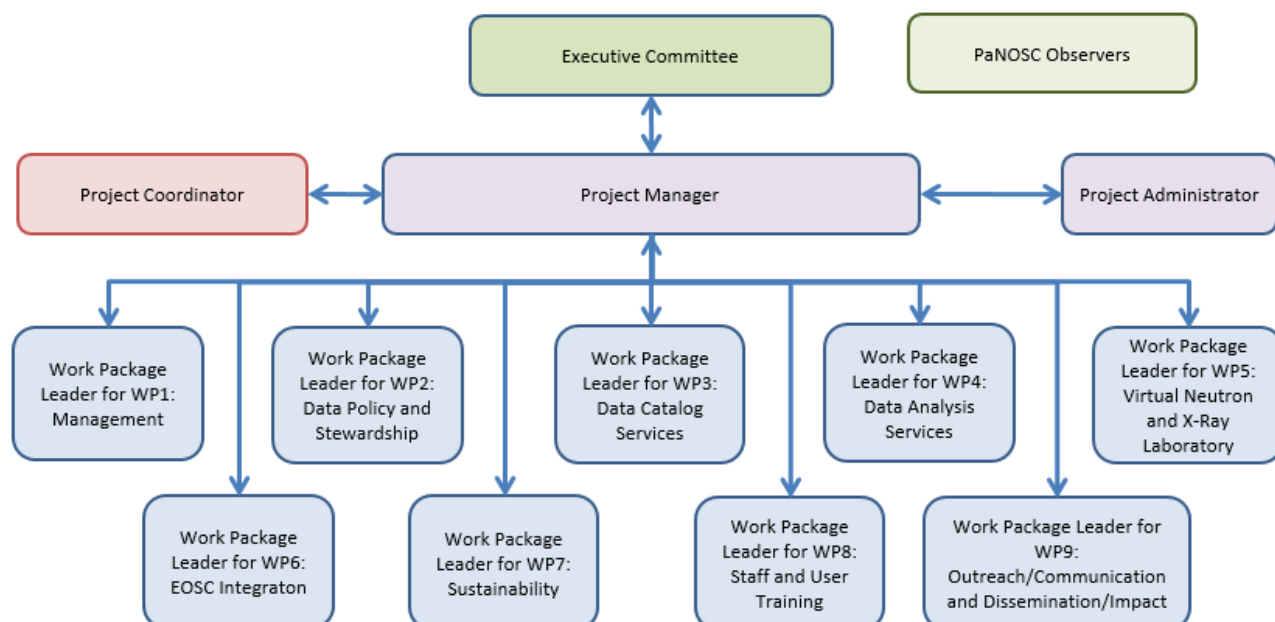


Figure 9: PaNOSC Project Management team structure

Work Package Leaders will be managing the day-to-day execution of their respective work packages (within their level of delegated authority) and informing the *Project Manager* during the monthly video conferences of plans, risk, change and progress. The *Project Manager* will in turn also inform the *Executive Committee* at regular scheduled points in time of the status of the project.

A list of all the roles, their responsibilities and the person(s) assigned to each role will be documented and maintained during the project, including major stakeholders.

3.2.2.1 Executive Committee appointment rules

The Executive Committee is the decision-making body for PaNOSC and will be made up of seven members, each one appointed on behalf of one partner.

The Executive Committee will require more votes for than against in order to make a decision.

Among the seven members one of them will be appointed Chair (in case of a draw between several candidates the leading partner's vote will have double value).

The Chair of the Executive Committee will

1. Have the tie-breaking power when no majority can be obtained for any decision among the Executive Committee members.
2. Request information or give ad hoc guidance to the Project Manager at any given time
3. Request an extraordinary meeting of the Executive Committee

An extraordinary meeting of the Executive Committee will be called:

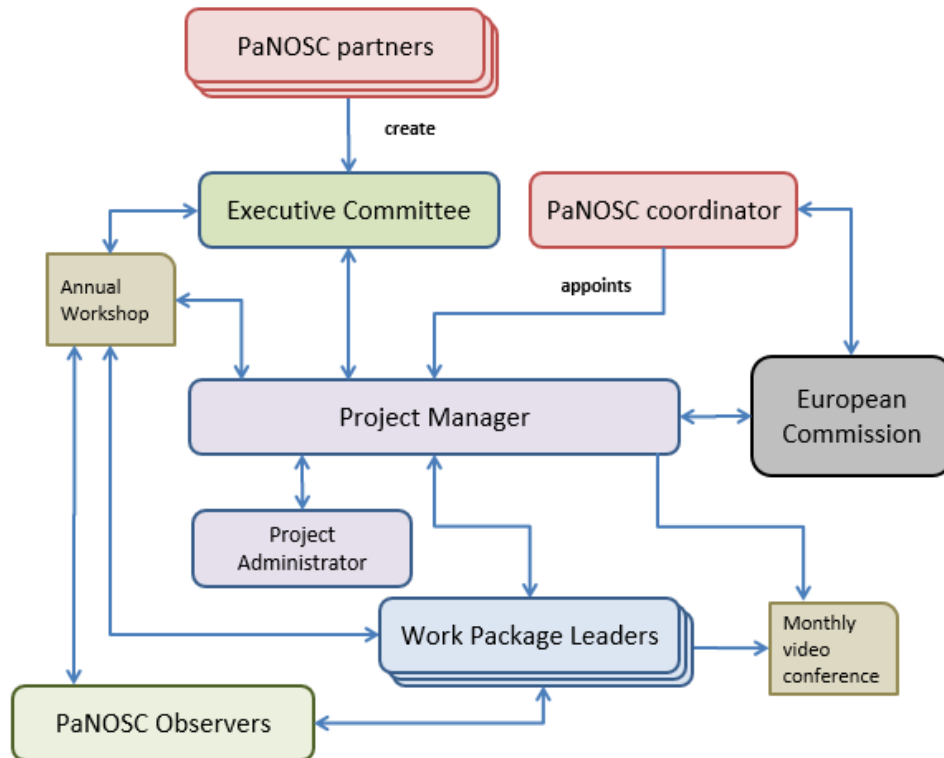
1. In case of resignation of the Executive Committee Chair
2. For any other reason by:
 - a. The Chair of the Executive Committee
 - b. Half of the Executive Committee members

3.2.3 Decision-making process and project control

In order to be able to quickly respond to the day-to-day demands of project management the *Executive Committee* will delegate an agreed level of authority to the *Project Manager*, which in turn will delegate to the *Work Package Leaders*.

The *Executive Committee* will meet once a year during the PaNOSC annual conference and at least a further meeting arranged via teleconference between each annual face-to-face meeting, however further meetings will be arranged if the Executive Committee decides that are necessary.

As long as the agreed levels of authority are not breached, *Work Package Leaders* and the *Project Manager* will manage the execution of PaNOSC and report regularly to their higher level. An escalating procedure will be used to handle the cases that would otherwise deviate from the agreed levels of authority (management by exception).



The flexibility given to *Work Package Leaders* by the *Project Manager* will ensure that they are able to adapt to each work package and the culture and structure of each participant.

The *Observers* will have a special role in the project. They represent the national RIs in the PaN community, PRACE host members, and other stakeholders (see list of Supporting Letters in annex). They will be consulted in the beginning of the project at a dedicated meeting to ensure their requirements for applying the outcomes of PaNOSC are met. The Observers will be consulted all along the project to give feedback and input.

Each work package will have a set of milestones, which will be used as control points to assess progress.

3.2.4 Milestones

The list of milestones for PaNOSC is:

WP	Milestone Number	Milestone name	Related WPs	Due date (in month)	Means of verification
1	MS1.1	Project Initiation Stage completed	1	M2	D1.1 PID available and executive committee members appointed
1	MS1.2	First Annual report	All WPs	M12	First approved annual report from first annual workshop
1	MS1.3	Second Annual report	All WPs	M24	Second approved annual report as per D1.4
1	MS1.4	Third Annual report	All WPs	M36	Third approved annual report as per D1.4
1	MS1.5	Final Annual report	All WPs	M48	Fourth approved annual report as per D1.4
2	MS2.1	First version of PaNOSC DP framework	2	M12	The availability of a first version of the text that can be used for applying it a participating facilities
2	MS2.2	Adoption of PaNOSC DP framework	2	M24	Data policies are endorsed officially and are published on official pages
2	MS2.3	Implementation of PaNOSC DP framework	2	M36	Main features of DP are implemented in the workflow of User access and data production up until publication of DOIs
3	MS3.1	Survey of Catalogue APIS and Roadmap to EOSC Integration	3	M12	Proposed API and tasklist with distribution between partners
3	MS3.2	Anthology Feedback to API Tasks	3	M12	Written report with suggestions for an extended API with domain specific features
3	MS3.3	Catalogue Integration Best Practices Meeting	3	M30	Minutes from meeting with all partners covering lessons learned and suggestions for future adopters
4	MS4.1	Prototype data analysis services completed	4	M18	Prototypes accessible and usable via Internet
4	MS4.2	Data analysis services accessible through EOSC	4	M42	Services accessible via Internet and EOSC portal
5	MS5.1	Simulation codes in PaNData Software Catalog	5	M6	PaNData software catalog website
5	MS5.2	Demonstration of simulation services	5	M24	Written report

5	MS5.3	VINYL software release	5	M42	Software released via open source repository
5	MS5.4	Validation of simulation services	4,5	M48	Written report
6	MS6.1	Implementation of AAI integration at the level of the Identity providers.	6	M36	usage
6	MS6.2	First release of PaNOSC services	6	M18	First group of PaNOSC services added to the EOSC catalogue
6	MS6.3	Second release of PaNOSC services, data and resources	6	M36	Second group of PaNOSC services, data and resources added to the EOSC catalogue
7	MS7.1	Stakeholder database ready	7	M6	Document to be included in D7.1
7	MS7.2	Final Sustainability Plan	7	M48	D 7.4 Document
8	MS8.1	Joint WP4 & 8 plan	4,8	M6	Gantt chart with milestones cross-linking WP4 & 8
8	MS8.2	Joint WP5 & 8 plan	5,8	M6	Gantt chart with milestones cross-linking WP5 & 8
8	MS8.3	pan-learning.org up running	8	M15	Existing functionality at e-neutrons.org accessible from pan-learning.org. Continuous integration testing up running at ESS. Implemented solution can handle 50 simultaneous users without lagging.
8	MS8.4	Jupyter integrated with e-learning platform	4,8	M30	Jupyter can be launched from the e-learning platform and relevant Python modules can be imported
8	MS8.5	e-learning virtual facilities	5,8	M36	At least one virtual instrument per PaNOSC facility can be used in the e-learning platform.
9	MS9.1	PaNOSC's Website Ready	WP9	M6	Deliverable

Table 6: Milestones

3.2.5 Risk management strategy

All risks initially identified will be documented in the risk register together with the owner of the risk, mitigation strategy, actions and an assessment based on the impact and probability. This initial assessment and the approval of a risk management strategy will take place during the initiation stage and be part of the project initiation documentation deliverable within work package 1.

Risk Assessment Matrix			
Impact Probability	Minor	Moderate	Major
Very Likely	Medium	High	Very High
Moderate	Medium	Medium	High
Unlikely	Low	Medium	Medium

Figure 11: Risk Assessment matrix

The seriousness of the risks will be part of the delegated authority agreements between the *Executive Committee*, *Project Manager* and *Work Package Leaders*, therefore any risk that increases its seriousness beyond the agreed levels will require an exception and review by the higher level.

All risks will be continuously monitored by their owner, who will be in charge of ensuring that the risk does not negatively affect the project.

Extreme and high risks will be reviewed at least during each monthly video conference.

All risks will be reviewed as part of the annual conference and report.

3.2.6 Risks

The initial list of risks is:

WP	Risk Description	Probability	WP involved	Proposed risk-mitigation measure
1	Participants become less engaged in project	Medium	All WPs	Monthly video conferences tracking progress, actions and risks and frequent communication to keep all participants engaged. Interact with other EOSC projects to ensure motivation stays high.
1	Executive Committee deadlock prevents decision making	Low	All WPs	The Executive Committee and its guidelines will be created in a way to prevent blockages.
1	Key staff members become unavailable (staff leaving, illness, etc.)	Medium	All WPs	Minimise single points of failure among staff in the participants, ensure good communication and document work being done.
2	Local and/or legal conditions prevent a common data policy.	Low	2	All parties in the proposal have committed to an open data policy. Make a policy with minimal core and optional parts
2	Data policy formally but not effectively accepted. The users may be forced to sign it but not	Medium	2	Adequate involvement of the user communities of each facility especially

	fully comply with it (e.g. minimal metadata)			for heterogeneous and distributed facilities.
3	Some of the facilities will not be able to meet the mandatory requirements for catalogue API	Low	3	All parties have the source code of their metadata catalogues so they can adapt them. In case changes cannot be made allow exceptions to be granted.
3	Data Catalog not integrated with data sources	High	3	Adequate effort to integrate data sources (e.g. experimental stations) with the catalog has to be foreseen at least for highly distributed and heterogeneous facilities.
4	Duplication of efforts/branching of projects/lack of sustainability	Medium	4	Where we use open source and need to modify it to adopt to our needs, we aim to feed those changes back to the open source project.
4	Lack of skilled staff	Low	4	We have relevant expertise at the facilities and can train new staff if required in most areas. Where this is not the case, we have risk-mitigation such as the co-operation with the Jupyter team.
5	Failure of partner to meet task or deliverable date	Low	5	Weekly teleconference meetings to monitor progress and update intermediate tasks.
5	Data formats not published or not compatible with simulation data	Medium	3,5	Fall back on existing data formats (NeXus, CXIDB) and metadata standards (openPMD)
5	Developed APIs not compatible with data analysis framework	Medium	4,5	Monthly meeting with WP 4 contributors to address compatibility issues
5	Compute resources needed for testing and demonstrations not available or insufficient	High	5	Apply for HPC resources, e.g. PRACE preliminary access
6	Recruitment of skilled staff	Medium	6	Start recruitment as early as possible.
6	Difficulties due to the EOSC core services implementation delays	Medium	6	Close collaboration with the EOSC involved partners will allow issues early identification and response.
7	Not clear definition of the EOSC structure	Low	7	Direct collaboration with the EOSC-hub to participate in the definition of the Hub (service management processes, activities, policies and its governance) together with main EOSC stakeholders

				(e.g. EGI, EUDAT, GÉANT, and other major research infrastructures from other disciplines)
7	Not clear definition of the EOSC stakeholders	Low	7	Stakeholder Feedback
7	Not clear understanding and harmonisation of the stakeholders' viewpoints	Medium	7	Stakeholder Feedback
7	Development of unsustainable business models	High	7	Preparation of different versions of the Sustainability plan to match with stakeholder feedbacks .
8	Delay in migrating e-neutrons.org to ESS	Low	8	Migrate and setup software development infrastructure before starting to remove technical debt
8	WP4 deliverables to WP8 are not in due time	Low	4,8	Make detailed plan (MS8.1) at start of project and subsequently update regularly (monthly) to ensure that WP4 deliverables to WP8 are delivered in due time
8	WP5 deliverables to WP8 are not in due time	Low	5,8	Make detailed plan at start of project (MS8.2) that ensures that WP5 deliverables required for WP8 are delivered in due time and update the plan on a regular (monthly) basis. Focus effort on easily doable PaNOSC beneficiaries to ensure success for at least some of the PaNOSC beneficiaries.
9	Difficulty to ensure outreach to key audiences	Low	9	The project's communications strategy will detail the target groups to be addressed, as well as the plan to make a mapping of key audiences at the beginning of and throughout the project
9	Risk not to successfully communicating results	Medium	9	The WP leader and the partnership will put efforts in keeping regular contacts with other clusters to ensure there is a complete and homogeneous communications of results
9	Difficulty managing communications due to the big variety of partners and clusters	Medium	9	The yearly communication plans annexed to the strategy will include key actions and events of the community, to avoid double efforts, and to ensure that common objectives are met and main outcomes are communicated

Table 7: Risks

Innovation management

PaNOSC will create innovative data services as part of the EOSC which will allow researchers and the general public to access their data and/or Open Data to generate new insights. The PaNOSC community knows the research landscape well so it can ensure the services are adapted to the data and needs of researchers. This is handled in WP4, WP5 and WP6. PaNOSC does not plan to develop any commercial products itself and therefore does not plan to do any market studies. This does not exclude a user of the PaNOSC services from developing a new product based on the insights they have gathered from the data using the analysis services. This applies particularly to the industrial users of the photon and neutron sources and data.

3.3 Consortium as a whole

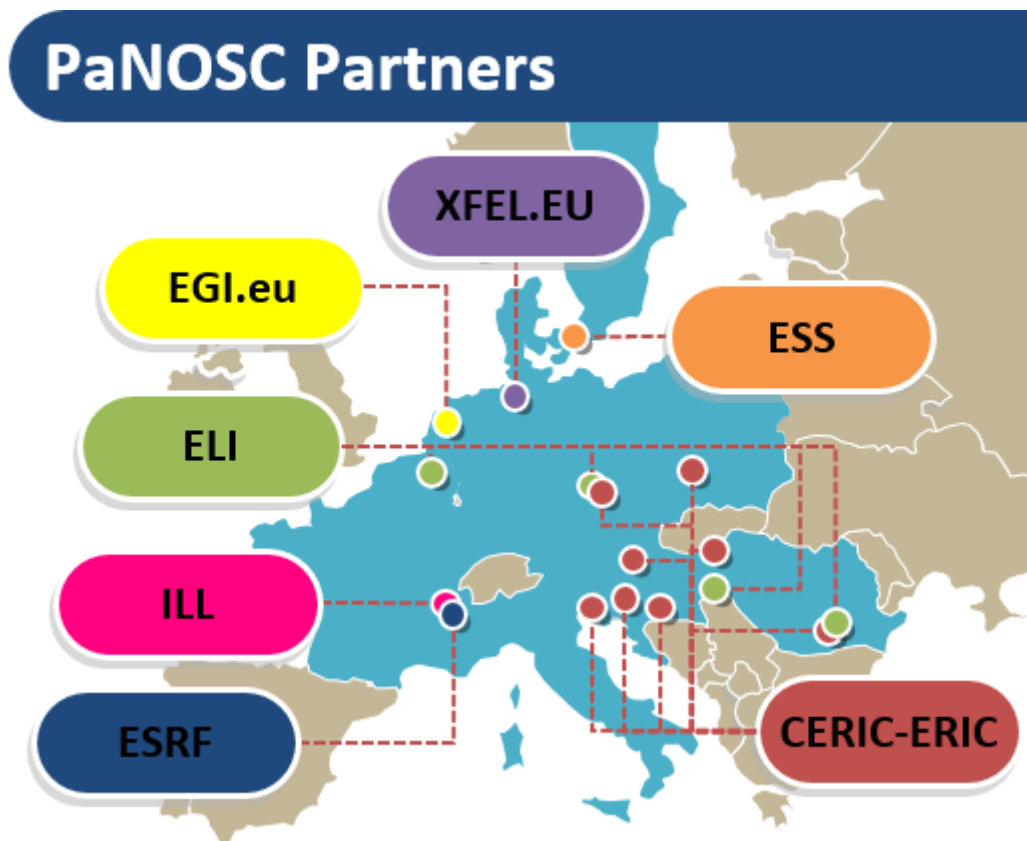


Figure 12: PaNOSC partners

Members

The PaNOSC cluster is made up of a mix of old and new RIs viz. ESRF, ILL, XFEL.EU, ELI, ESS, CERIC-ERIC and the e-infrastructure EGI. The RIs are all photon or neutron sources and have a strong overlap in the scientific experiments they offer. The mix of old and new RIs in the cluster has been carefully selected to ensure that PaNOSC covers as large a spectrum of RIs in the photon and neutron community across Europe while still staying compatible with the requirements of the call i.e. restrict partners to RIs on the ESFRI roadmap and/or ERICs. The cluster shows good geographic coverage across Europe, at least 24 European countries are members of at least one of the PaNOSC facilities. PaNOSC has representatives of all the type of sources found in the PaN world. These are neutron reactors (2), 4th generation storage ring (1), 3rd generation storage rings (2), spallation source (1), free electron laser (1), laser facilities (3), and academic institutes (5). With such a large coverage PaNOSC will be able to address all issues which national RIs typically encounter thereby strengthening the outcomes and making them easy to be adopted and adapted by national RIs. The cluster will be able to extend metadata standards in all fields. The small size will make it more efficient and able to connect to the EOSC. This is in line with the idea to start implementing the EOSC with a small group of doers and then extend it to include the whole scientific community. All of the PaNOSC partners firmly support the EOSC, EGI.eu is coordinating the first incarnation of the EOSC (<http://eosc-hub.eu>). Three of the RIs (ESRF, ILL and XFEL.EU) have signed the EOSC declaration and the others are firmly committed to linking up to the EOSC.

Role of members

- **The European Synchrotron (ESRF)** is spearheading the development of a 4th generation storage ring with the Extremely Bright Source (EBS) project (<http://www.esrf.eu/about/upgrade>) in the synchrotron community. Data management is a cornerstone of the upgrade. ESRF is the first major synchrotron to implement an Open Data policy (<http://www.esrf.fr/datapolicy>) including archiving data for 10 years. The experience of the ESRF implementing an Open Data policy since 2015 will be used in WP2 to update the Open Data policy framework for PaNOSC. The EBS source will provide new beamlines with unprecedented flux which in turn will generate huge quantities of data. Dealing with the future data deluge will be one of the challenges PaNOSC will help the ESRF address. Work packages WP4 will provide services on site and as part of EOSC to help users deal with the huge data volumes on site or remotely. ESRF will need to extend the NeXus standard to incorporate the new experimental setups. This will be done as part of WP3. The ESRF is leader in a number of open source software packages and therefore has a long experience in developing software and collaborating in Open Source projects. ESRF has been involved in the EOSC via the stakeholders meetings, EOSCPilot and is well aware and aligned with the goals of the EOSC. ESRF has committed itself in a number of actions directly linked to the EOSC (open data policy, data management, data services etc.). PaNOSC is the latest in the line of actions to link up with the goals of the EOSC and provide users with better data services.

- **European XFEL (XFEL.EU)** has just entered its operation phase and has produced about 500 TeraBytes of data in only a few weeks of early user experiments in the last months of 2017. With these data volumes and the high repetition rate of 27000 pulses per second to be used in the future, concepts of data stewardship, data policies, online and offline data reduction and analysis, as well as a metadata catalog are essential for successful operation. In this proposal, European XFEL will lead the work packages on data analysis (WP4) and on “Virtual Neutron and X-ray Laboratory” (WP5), making in particular use of its experience with the Jupyter ecosystem and experience in integrating simulations and modeling of experimental observables into the data lifecycle. Both areas of expertise are currently supported by EC-funded research projects (OpenDreamKit and EUCALL, respectively). European XFEL works closely with DESY, in particular in the topics of data management and data analysis that are relevant to this call. Development of experiment simulation capabilities and associated open metadata standards (openPMD) for simulation data is performed in close partnership with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR).

- **Central European Research Infrastructure Consortium ERIC (CERIC-ERIC)**, is a distributed Research Infrastructure integrating and providing open access to some of the best facilities in 9 European Countries (Austria, Croatia, Czech Republic, Hungary, Italy, Romania, Serbia and Slovenia). Thanks to the long experience of some of the partner facilities (e.g. Elettra) and to the complexity of the task at hand, CERIC-ERIC will be involved in most of the work packages and will lead in particular WP7 and WP9 with important participations in WP2 and WP3 that will co-lead. Due to its distribution across several European countries with different approaches to EOSC and due to the high level of multidisciplinary a considerable effort will have to be invested in order to implement PaNOSC. However, the expected impact of including CERIC-ERIC in PaNOSC will be high, as it will contribute to synchronisation of policies across a number of countries and scientific disciplines. It is of particular importance since for some of the CERIC-ERIC member countries PaNOSC will be the only project submitted in the field of physical sciences (Slovenia, Croatia, Poland, Italy), while for some others it will strengthen the coherence of the approach. This is the case of Czech Republic, Romania and Hungary, which participate in PaNOSC through ELI. It is also important to mention that some of the partner facilities like Elettra have long term experience in the ideas behind the FAIR principles as they participated from the very beginning to the PaNdata (<http://pan-data.eu>) efforts, and will therefore significantly contribute to the delivery of the high quality solutions developed by PaNOSC.

- **Institut Laue Langevin (ILL)**, is the world’s flagship centre for neutron scattering science, providing scientists with a very high flux of neutrons feeding some 40 state-of-the-art instruments, which are constantly being developed and upgraded. In 2011, the ILL was amongst the first analytical facilities to adopt a Data Policy based on FAIR principles (resulting from the work done in the PaNdata-Europe project). By the start of the project the ILL will be the only facility able to provide massive open data for piloting the services. With the PaNOSC project, ILL will be part of the ‘coalition of doers’ for building the EOSC. In line with the endorsement of the EOSC declaration, we will be involved in all work packages, co-lead WP4 and lead WP6.

- **European Spallation Source (ESS)**, currently under construction in Lund, SE, will be the brightest source of cold neutrons in the world, by an order of magnitude. Better sources and better instruments means orders of magnitude more data to collect and analyze. That means the scientists doing research at ESS must be positioned to

take full advantage of the last decade of exponential growth in data management capacity and computational power. As a greenfield project, built from the ground up not only physically, but organisationally and philosophically, ESS is positioned as the vanguard of the next century of experimental science and the world's next great Big Science facility, and has the unique opportunity of having the European Open Science Cloud being built in from the very start. Without the legacy of old data sets collected under old policies, ESS will be able to share all their data under FAIR principles.

- ❖ The **Data Management Center (DMSC)** in Copenhagen, is a division in ESS and an integral part of the value proposition of ESS. DMSC have the focus of being a true enabler of science using neutrons. It has attracted a large number of experts in the fields of neutron science and IT, and is a key player in many of the software project that will allow the next generation of neutron based science, including NeXus, Mantid, McStas and a number of analysis packages. It collaborates closely with several European neutron source in-kind partners on delivering the software for data management and treatment. ESS is representing the PaN community in the EOSCPilot, both through shepherding the Photon and Neutron Science demonstrator, involvement in the work on Service Portfolio Management, and a firm commitment to be part of the “coalition of doers” from DG John Womersley. Moreover, ESS leads a work package on data treatment software in SINE2020. In PaNOSC, ESS will lead WP3 Data Catalog Services, leveraging its expertise in NeXus data formats and development of a new data catalog system in collaboration with Swedish synchrotron MAX IV and the Swiss Light Source (SLS) at PSI. ESS will also lead WP8 on Staff and User Training made possible due to its strong collaboration with University of Copenhagen. Moreover, ESS has significant experience with virtual experiment simulations of use in WP5 and 8.

• **Extreme Light Infrastructure Delivery Consortium (ELI-DC)** is an international association under Belgian law in charge of supporting the coordinated transition of ELI from implementation to operations. ELI is a research infrastructure based on a new generation of laser technologies that will deliver sources of ultra-intense high-energy particle beams and ultra-bright radiations up to the femtosecond and attosecond timescales. As the first truly international laser research infrastructure, Based initially on 3 complementary research facilities located in Szeged (Hungary), Dolní Břežany (Czech Republic) and Măgurele (Romania), ELI plans to start giving access to international users in 2018. It will operate as an integrated organisation under an ELI-ERIC in the process of being established (ELI-ERIC is expected to take over ELI-DC as a beneficiary once established).

When fully operational, the three pillars will serve thousands of users generating an estimated 5-10 PBytes of scientific data every year. ELI is committed to provide its users with state-of-the-art tools, methods and services for the acquisition, analysis, curation, and preservation of experimental data. Participation in PaNOSC is an essential step for ELI to make this objective position and ensure that ELI develops in an integrated way with EOSC. ELI is willing to provide significant contribution to every work-package of PaNOSC and co-lead WP8 (Staff and User Training), which is very closely related to the capacity building requirements of ELI at this stage. ELI's participation will greatly help the adoption and implementation of a data policy and data management practices that comply with FAIR principles within the future ELI-ERIC. Last but not least, ELI will ensure the representation of the laser community within the consortium.

• **EGI.eu** is leading the implementation of the EOSC-hub and is a partner in PaNOSC. The EOSC-hub technical support and service provisioning activities will be provided in order to ensure an effective integration of PaNOSC scientific applications and open data in the European Open Science Cloud, These include:

- ❖ **EOSC service catalogue integration.** EGI.eu with linked third parties will provide technical support and tools to include PaNOSC service and data providers into the Hub, the service integration and access management framework, ensuring that services and resources in EOSC can be seamlessly managed and operated.
- ❖ **Federated Cloud provisioning.** PaNOSC will rely on a cloud compute infrastructure computing leveraging the federation services offered by EGI with the support of EOSC-hub. The platform includes: cloud IaaS, storage virtualisation, VM provisioning and management through the Application Database, scaling and cloud bursting. The activity includes service enabling, operations and management of the infrastructure and technical support the services. The operations support will perform problem-solving steps and manage proper involvement of second-level support for issue resolution. Cloud Compute will be provided by CESNET and DESY.
- ❖ **Notebook and Applications on Demand.** EGI will be responsible for the deployment and operation of Jupyter notebook service for PaNOSC community integrated with federated authentication, persistent storage for user notebooks, and access to PaNOSC datasets. Technical support to application porting will be ensured.

- ❖ **Data transfer.** A central data transfer service for data transfer scheduling will be provided and supported by STFC. The support activity includes requirements analysis, piloting, porting of existing applications to the service, and FTS3 operations.
- ❖ **Data archiving.** A distributed data archiving facilities will be federated in order to experimentally host PaNOSC open data and co-locate that with computing facilities. A 1.5 PB facility will be collectively provided by CESNET, DESY and STFC.
- ❖ **Metadata harvesting.** The PaNOSC community already makes use of standardised metadata formats in which the data is described. To extend the accessibility of PaNOSC data, all members will make their data available via data repositories, via the standardised metadata formats and made persistent via DOI's. To increase the findable of PaNOSC data PaNOSC will develop a central metadata catalogue in which all metadata from PaNOSC members is collected. To provide added value to the PaNOSC researchers central metadata catalogue will provide domain specific search and browse functionalities. The PaNOSC central metadata catalogue will be made harvestable through other search engines to support cross disciplinary research and to increase the scientific impact. PaNOSC will collaborate with the EOSC-hub to leverage the expertise on metadata and to make the metadata harvestable. To development the PaNOSC central metadata catalogue, EOSC-hub technologies will be explored and especially a feasibility study and pilot will be conducted on the EUDAT B2FIND service.

• **Observers** - The six PaN participants in PaNOSC represent the ESFRIs as the *avant garde* of the PaN community connecting to the EOSC. It is essential that the national PaN RIs profit similarly from the outcomes of PaNOSC so they can stay aligned with the evolution towards Open Data and Open Science and connect to the EOSC in a second phase. For this reason PaNOSC has identified a role called Observer. The majority of PaNs have expressed their support for PaNOSC (see letters of support in the annex after section 5 from ALBA, DESY, HZB, JCNS, PSI, SOLEIL) and requested to be observers. Moreover, three PRACE host members will also be observers (see letters of support from CSCS, CINECA, and JSC). This is important for ensuring future linking to PRACE and associated supercomputing centers, which in turn is important for being able to link molecular and materials modelling and simulations to PaN experimental data. Those who missed the deadline to supply a support letter will be invited to be observers again once the project starts. The Observers will be consulted from the start of the project to find out what specific requirements they might have compared to the PaNOSC partners. Their input will be included in the design of the solutions. They will be invited to attend the annual meetings and the technical workshops (some budget is foreseen for paying for travel for some of them). They will have the role of testing and providing feedback on the data services deployed on the EOSC. The Observers will play a critical role in ensuring the results of PaNOSC can be rolled out for the whole PaN community.

3.4 Resources to be committed

The financial request to the European Commission for PaNOSC totals 11.96 M€ and is a result of careful cost containment and optimization by each partner. The budget calculation follows the Horizon 2020 financial rules, including personnel costs, travel, equipment, and other direct costs, and 25% overheads.

Person months of effort per partner and work package								
Work Package	ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC	EGI.eu	Total
WP1: Management	49	3	3	3	3	3	0	64
WP2: Data policy and stewardship	17	10	3	14	20	12	0	76
WP3: Data catalogue services	25	21	36	43	78	88	0	291
WP4: Data analysis services	36	71	60	32	50	60	0	309
WP5: Virtual Neutron and X-ray Laboratory	40	36	48	31	24	40	0	219

WP6: EOSC Integration	21	38	13	13	12	13	82	192
WP7: Sustainability	3	3	3	3	12	32	0	56
WP8: User Training	6	9	4	35	48	6	0	108
WP9: Outreach	6	6	6	6	6	40	0	70
Total person months	203	197	176	180	253	294	82	1385
Average Annual Cost (k€)	85	87	96	99	55.5	54.5	71.2	
Total Personnel Cost (k€)	1438	1428	1408	1485	1170	1335	486	8751
Total Personnel Cost with overheads (k€)	1797	1785	1760	1856	1463	1669	608	10938

Table 8: Personnel partner's costs

The resources requested within PaNOSC will be primarily used for covering additional staff resources and for collaboration activities, such as workshops, expert meetings and development and integration of services in the EOSC. The existing experienced staff will contribute to the activities through in-kind contributions as part of their working time, both for initiation of the tasks and activities and subsequently for guidance of the new staff. The request for personnel resources is more than 85% of the requested funding and is absolutely vital to engage in the tasks as described in the proposal. Each of the partners operates a data center of a few thousand cores and Petabytes of storage which will be made available in kind for users of the data services for analysing data from experiments.

The majority of the remaining request for funds is related to EGI as provider of compute resources, the organization of meetings and workshops and for travel. All of these costs relate to the fact that, within PaNOSC, an intense exchange among experts shall be established and maintained. We will invite external experts, attend conference and spend a reasonable amount of effort on user training, communication and outreach to ensure that users can access data and services in the EOSC and contribute to their success. Given the huge amounts of data volumes produced in the partner's RIs every year and the computational intense data services around 5% of the PaNOSC budget will be spent with EGI.eu and its linked third parties in implementing a federated cloud infrastructure and enabling the PaNOSC scientific applications to use it for the analysis of reference data sets in EOSC. The AAI for PaNOSC will be managed by GÉANT. PaNOSC will sign a collaboration agreement with GÉANT at the beginning of the project to do the development of the AAI solution taking into account the developments which have already taken place in the PaN community with UmbrellaID (<https://www.umbrellaid.org/>).

ESRF Costs within PaNOSC		
Concept	Cost (in €)	Justification
Personnel Cost	1,437,917	Effort required in order to deliver PaNOSC tasks
Travel	58,905	Travel and subsistence to meet other partners and to conferences
Other direct costs	124,800	Purchase to explore procuring and integrating commercial cloud in PaNOSC, annual conferences organisation, conference fees, workshops, open access publications, etc.
Total	1,621,622	2,027,027 including overheads

Table 9: ESRF's costs

ILL Costs within PaNOSC		
Concept	Cost (in €)	Justification
Personnel Cost	1,428,250	Effort required in order to deliver PaNOSC tasks
Travel	58,905	Travel and subsistence to meet other partners and to conferences
Other direct costs	15,000	Conference fees, workshops, open access publications, etc.
Total	1,502,155	1,877,694 including overheads

Table 10: ILL's costs

XFEL.EU Costs within PaNOSC		
Concept	Cost (in €)	Justification
Personnel Cost	1,408,000	Effort required in order to deliver PaNOSC tasks
Travel	80,905	Travel to meet other partners, to project meetings, to conferences and other events as well as visiting/inviting the Jupyter team
Other direct costs	22,500	Annual conference organisation, conference fees, workshops, open access publications, etc.
Total	1,511,405	1,889,256 including overheads

Table 11: XFEL.EU's costs

ESS Costs within PaNOSC		
Concept	Cost (in €)	Justification
Personnel Cost	1,485,000	Effort required in order to deliver PaNOSC tasks
Travel	58,905	Travel to meet other partners and to conferences
Other direct costs	23,500	Annual conference organisation, conference fees, workshops, open access publications, etc.
Total	1,567,405	1,959,256 including overheads

Table 12: ESS' costs

ELI Costs within PaNOSC		
Concept	Cost (in €)	Justification
Personnel Cost	1,170,125	Effort required in order to deliver PaNOSC tasks
Travel	94,105	Travel to meet ELI colleagues in different sites, other partners and to conferences
Other direct costs	68,500	Training materials, conference fees, workshops, open access publications, etc.
Total	1,332,730	1,665,913 including overheads

Table 13: ELI's costs

CERIC-ERIC Costs within PaNOSC		
Concept	Cost (in €)	Justification
Personnel Cost	1,335,250	Effort required in order to deliver PaNOSC tasks
Travel	108,185	Travel to meet CERIC-ERIC colleagues in different sites, other partners, to conferences and for communication and dissemination of PaNOSC
Other direct costs	72,500	Public website, communication and promotional materials, conference fees, workshops, open access publications, etc.
Total	1,515,935	1,889,919 including overheads

Table 14: CERIC-ERIC's costs

EGL.eu Costs within PaNOSC		
Concept	Cost (in €)	Justification
Personnel Cost	486,498	Effort required in order to deliver PaNOSC tasks
Travel	25,000	Travel and subsistence to meet 3rd linked parties and partners
Other direct costs	3,000	Certificate of Financial Statement
Total	514,498	643,123 including overheads

Table 15: EGL.eu's costs

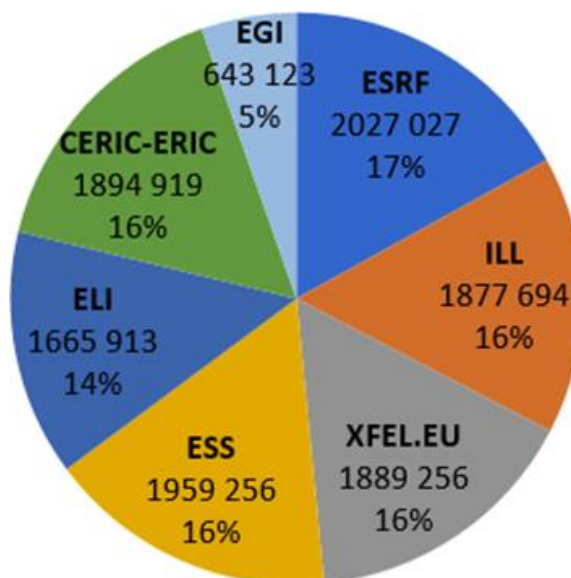


Figure 13: Cost breakdown per partner

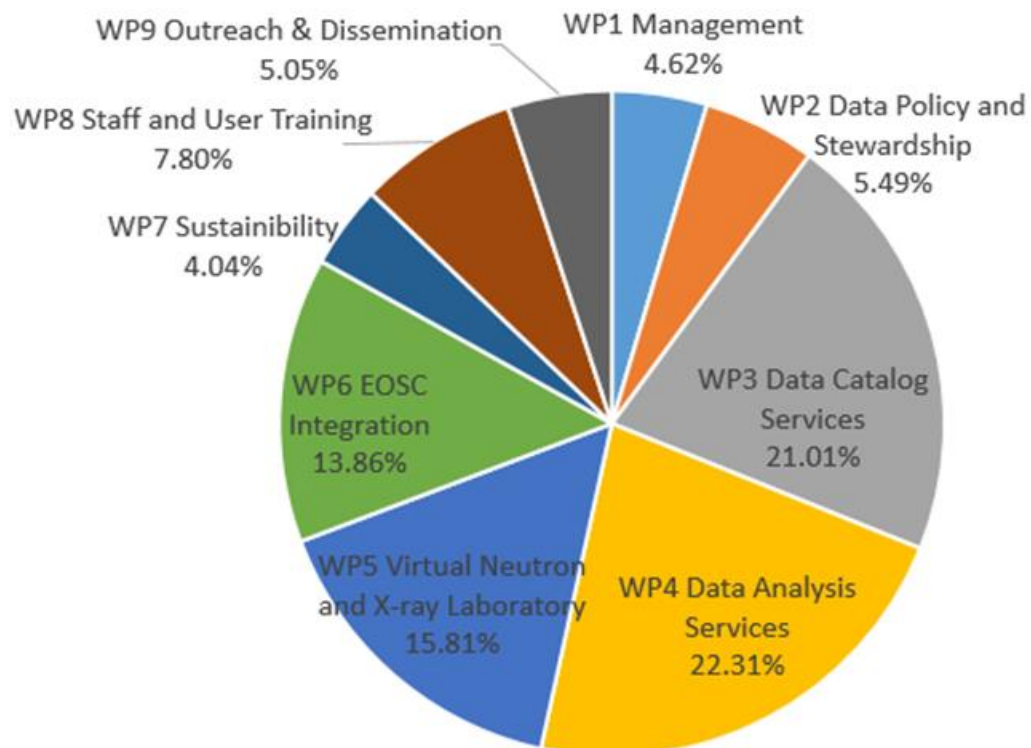


Figure 14: Effort breakdown per work package

Aggregated total cost of PaNOSC in €							
ESRF	ILL	XFEL.EU	ESS	ELI	CERIC-ERIC	EGI.eu	Total
2,027,027	1,877,694	1,889,256	1,959,256	1,665,913	1,894,919	643,123	11,957,187

Table 16: PaNOSC's aggregated costs

References

- [Fortmann-Grote2017] C. Fortmann-Grote et al. Simulations of ultrafast X-ray laser experiments Proc. SPIE, Advances in X-ray Free-Electron Lasers Instrumentation IV, International Society for Optics and Photonics, 2017, 10237, 102370S (2017). [10.1117/12.2270552](https://doi.org/10.1117/12.2270552)
- [Glass2017] M. Glass, M. Sanchez del Rio “Coherent modes of X-ray beams emitted by undulators in new storage rings”, Europhysics Letters, 119 3 (2017). [10.1209/0295-5075/119/34004](https://doi.org/10.1209/0295-5075/119/34004)
- [Jupyter2018], Project Jupyter, <http://jupyter.org>; formerly IPython. Accessed 15/3/2018
- [Kluyver2016] T. Kluyver. et al. Jupyter Notebooks – a publishing format for reproducible computational workflows IOS Press, 2016, 87 - 90 (2016). [10.3233/978-1-61499-649-1-87](https://doi.org/10.3233/978-1-61499-649-1-87)
- [Larsen2017] Ask Hjorth Larsen et al. The Atomic Simulation Environment—A Python library for working with atoms. J. Phys.: Condens. Matter Vol. 29 273002 (2017). [10.1088/1361-648x/aa680e](https://doi.org/10.1088/1361-648x/aa680e)
- [Maddison1997] Maddison, D. R.; Swofford, D. L. & Maddison, W. P. Cannatella, D. (Ed.) NeXus: An Extensible File Format for Systematic Information Systematic Biology, Oxford University Press (OUP), 1997, 46, 590-621 (1997) [10.1093/sysbio/46.4.590](https://doi.org/10.1093/sysbio/46.4.590)
- [Markvardsen2017] A. Markvardsen, “Report on Guidelines and Standards for Data Treatment software,” SINE2020 Deliverable 10.1 (2017) <http://sine2020.eu/files/d10.1-guidelines-and-standards-1.pdf>.
- [OpenDreamKit2015], N. Thiery, H. Fangohr, et al: OpenDreamKit project, <http://opendreamkit.org>
- [Perez2015] F. Perez, B. Granger. Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science (2015) <http://archive.ipython.org/JupyterGrantNarrative-2015.pdf>
- [Perez2017] The state of Jupyter [<https://www.oreilly.com/ideas/the-state-of-jupyter>]
- [Puget2016] Puget: Use Python Notebook to Discover Gravitational Waves (2016), <http://ibm.co/2tPpLLB>
- [Rebuffi2017] Rebuffi, L and Sanchez-del Rio, M., “OASYS (OrAnge SYnchrotron Suite): an open-source graphical environment for x-ray virtual experiments”, Proc. SPIE 10388, Advances in Computational Methods for X-Ray Optics IV, 103880S (2017). [10.1117/12.2274263](https://doi.org/10.1117/12.2274263)
- [Valentini2015] Valentini E, Kikheny A.G., Previtali G., Jeffries C.M., Svergun I., *SASBDB, a repository for biological small-angle scattering data*, Nucleic Acids Research, Vol. 43, D357–D363 (2015). [10.1093/nar/gku1047](https://doi.org/10.1093/nar/gku1047)
- [White2012] T. White et al. CrystFEL: a software suite for snapshot serial crystallography, Journal of Applied Crystallography, International Union of Crystallography (IUCr), 45, 335-341 (2012) [10.1107/S0021889812002312](https://doi.org/10.1107/S0021889812002312)
- [Wilkinson2016] Wilkinson, M. D. et al. “The FAIR Guiding Principles for scientific data management and stewardship”. Sci. Data3:160018 doi: 10.1038/sdata.2016.18 (2016)