ENABLING OPEN SCIENCE FOR PHOTON AND NEUTRON SOURCES*

A. Götz, J. Bodera Sempere, A. Campbell, A. de Maria, M. del Rio,
R. Dimper, J. Kieffer, A.Solé, T. Vincent, ESRF, Grenoble, France
S. Caunt, J. Hall, J. F. Perrin, ILL, Grenoble, France
N. Carboni, A.Hafner, R. Pugliese, CERIC-ERIC, Trieste, Italy
M. Bertelsen, T. H. Rod, T. S. Richter, J. Taylor, ESS, Copenhagen, Denmark
J. C. E, H. Fangohr, C. Fortmann-Grote, T. Kluyver, R. Rosca, EuXFEL, Schenefeld, Germany
F. Gliksohn, L. Schrettner, ELI-DC, Brussels, Belgium

Abstract

Photon and Neutron sources are producing more and more petabytes of scientific data each year. At the same time scientific publishing is evolving to make scientific data part of publications. The Photon and Neutron Open Science Cloud (PaNOSC [1]) project is an EU financed project to provide scientific data management for enabling Open Science. Data will be managed according to the FAIR principles. This means data will be curated and made available under an Open Data policy, findable, interoperable and reusable. This paper will describe how the European photon and neutron sources on the ESFRI [2] roadmap envision PaNOSC as part of the European Open Science Cloud [3]. The paper will present the objectives of the project on the issues of data policy, metadata, data curation, long term archiving and data sharing in the context of the latest developments in these areas.

INTRODUCTION

Photon and neutron sources are hitting a data analysis wall with the huge increase in data volumes, new techniques and new user communities. Users are limited by the difficulty in exporting huge data from the source to their home labs and by the lack of easy access to data analysis programs and services. At the same time science is becoming more open by sharing the data, methods and publications - the so-called **Open Science** movement [4]. Making the data used in publications easily available enables others to reproduce the analysis and findings. This has been one of motivations for neutron and photon sources to adopt open data policies. Another motivation has been the need to alleviate the data analysis bottleneck by implementing modern data management so that data analysis services can be built on top of data catalogues. Lack of adequate data management limits the data services which can be be offered to users.

PaNOSC has been financed by the H2020 INFRAEOSC-04 call as part of the EOSC project to bring FAIR data to ESFRI Photon and Neutron sources and to share the outcomes with all national photon and neutron sources. The PaNOSC partners are ESRF (coordinator), ILL (Grenoble, FRANCE), EuXFEL (Schenefeld, GERMANY), ESS (Lund, Sweden), CERIC-ERIC (Trieste, ITALY), ELI-DC (Bruxelle, BELGIUM), EGI (Amsterdam, NETHERLANDS). In addition to the internal partners, the following external partners will assist PaNOSC in its missions - GÉANT (Amsterdam, NETHERLANDS), DESY (Hamburg, GERMANY), CESNET (Prague, CZECH REPUBLIC), STFC (Hartwell, UNITED KINGDOM).

* This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 823852.

DATA POLICIES 2.0

In the past (10 years ago) scientific data was produced without any clear policy on the ownership, life cycle or license of the data. Since 2010 it has become standard practice to define a clear data policy for all scientific data produced in research institutes. This is partly due to the large quantities of data produced but also in order to manage the life cycle of the data better. In order to invest in and manage data for a longer period the ownership and license of the data need to be specified.

All of the PaNOSC partners have adopted or are in the process of adopting (ELI-DC) a data policy. The data policies are all based on the PaNdata-Europe data policy. Adopting a data policy is only the first step. Implementing it is much more work. See [5] for a detailed description of how one of the partners has addressed the challenge of implementing a data policy.

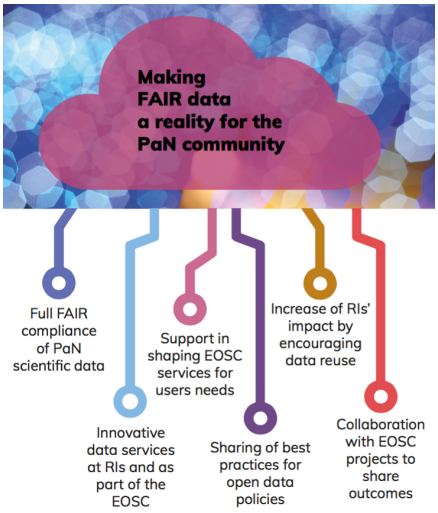

Figure 1: PaNOSC FAIR objectives

One of the main objectives of PaNOSC (see Fig. 1) is to update the PaNdata-Europe data policy framework to include the FAIR (Findable, Accessible, Interoperable, Reusable) principles explicitly. The FAIR principles are implicit in the PaNdata data policy framework but making them explicit involves changes to the existing data policies. This change is part of the push to generalise FAIR data for the EOSC. PaNOSC has done a critical review of the existing data policies for all the partners and is preparing a new Data Policy framework. The new data policy will include changes to address the FAIR principles. The main areas which will be impacted are the metadata which impacts the F, I and R principles.

Data Management Plans (DMPs) are becoming a standard requirement for many funding agencies. Therefore a Data Management Tool is being developed to help facility users to prepare DMPs for their data. The DMP will be a living document which will be updated throughout the lifetime of the data.

COMMON DATA API

Open data catalogues are *the* important first step in order to implement FAIR data. Within PaNOSC partners follow two main tracks to make experimental datasets findable. One is to offer the public contents of their repositories up to EOSC actors like OpenAIRE[6] or B2Find[7], typically after the expiry of an embargo period. These third party aggregation services offer a good exposure to a large scientific community, however their current data model only allows for a fairly generic set of associated metadata and excludes relevant domain specific information, for example about experimental technique, sample properties or beamline configuration.

To overcome these limitations the partners have started to establish a useful set of metadata and to design an extensible query API that enables a search on these terms. All partners are committed to implementing this API through their data catalogues, so that an EOSC actor can easily provide a federated search, spanning all participating facilities, which is not limited to PaNOSC, but can include ExPaNDS[8] members. The common search API provides a user-centric view of the datasets hosted at a facility and does not rely on any specific implementation of a data catalogue, like icat[9] or SciCat[10]. With any returned search results, the user will be given an appropriate list of options of what to do with the data, for example download, transfer to another location or the launch of a data processing or analysis session with the data accessible. This makes the federated data catalogue a major entry point for further services.

An optional authentication step will enable the search for data that still falls under an embargo, i.e. data that the logged in user has been given access to specifically.

SOFTWARE CATALOGUE

The PaNdata software catalogue[11] is a database of software used mainly for data analysis of photon and neutron experiments. It allows scientific institutes to upload information about the software that they have created for others to download although the website does not contain the binaries for the software.

PaNOSC will consolidate the PaNdata software catalogue to make it the main source of generic and bespoke software for data analysis of photon and neutron scientific data. PaNOSC will enhance the software catalogue by providing containerised versions of software and where possible provide web based versions and jupyter notebook recipes for testing the software. This will remove the burden of installing software for users and help make scientific data analysis more reproducible. The most common software entries will be linked to example open data sets.

DATA SERVICES

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.

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

Model 1 - command line based access: Facility user connects using a remote terminal protocol (nowadays secured, SSH) through firewalls to internal facility resources (workstations or High Performance Computing installations).

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) front end and run on a personal computer or High Performance Computing 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 to improve security, user friendliness and efficiency in giving direct access to remote data analysis services to scientists.

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.

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.

In PaNOSC, we consider model 1 to be broadly implemented and therefore will concentrate on model 2 and 3.

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 represents 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.

The vision for PaNOSC is that data sets – for example associated with publications – can be found through a portal, and corresponding data analysis services are made available for this data set; for exampel using model 2 or model 3 as outlined above.

SIMULATION

Simulations are an indispensable part of experimentally driven photon and neutron science at all stages of the data lifecycle: New ideas for experimental schemes can be tested, they can support beamtime application proposals, complement data analysis, and by comparison to experimental findings, validate (or invalidate) the theoretical foundations of the simulation algorithm itself. All these application use cases require that the simulation is capable of synthesizing the experimentally measurable signal with a high degree of accuracy. Even more, the simulation must ideally include all sources of background and fluctuations. In the context of increasingly complex experimental setups employing large scale facilities like synchrotrons, free–electron lasers, high–power optical lasers, or neutron sources, a representation of all components of the experimental apparatus and the instrument as well as the target or sample under investigation must be integrated into the simulation. This requires a framework into which existing simulation codes for e.g. the particle source, the particle transport and beamshaping, the interaction of beams and targets (samples), the signal generation and particle detection, as well as the data processing immediately following detection, can be integrated with relative ease and flexibility.

The work package 5 of PaNOSC develops such a simulation environment under the title “Virtual Neutron and X–ray Laboratory” (ViNyL).

ViNyL requires three components to be put into existence:

1. A library of harmonized “high–level” application–programming interfaces to the “low–level” simulation codes
2. Definition of metadata standards for simulation data.
3. Documented usage examples showcasing the applications of simulation workflows in web based and in desktop based user interfaces.

ViNyL builds on top of existing simulation frameworks: The photon science simulation (python) module SimEx (Simulation of Experiments) [12, 13], the neutron ray–tracing environment McStas [14], and the x–ray optics simulation and design framework Oasys [15] to name the most prominent ones.

SimEx: SimEx is a python library. It defines abstract classes representing the generic stages of a photon facility experiment: the photon source, the photon beamline, through which the photons propagate from the source to the point of interaction with the sample, the interaction of photon beam and sample, the signal generation, photon detection, and signal processing.

McStas: The aim of the McStas package is to simulate neutron beam lines from the source to the detector. This is accomplished mainly by compiling a meta language instrument file into a c code describing the experiment. The user writes the instrument file by instantiating components that describe physically separated beam line components that are placed in the simulated experiment. The components are smaller c codes which are frequently contributed by the community at large, resulting in more than 200 components in the latest McStas release. With the large component library, the McStas package is capable of simulating the entire beamline, including guide, choppers, sample, sample environment and detectors. The resulting detector counts can be saved in a hierarchical file format.

A python API for McStas called McStasScript have been developed under PaNOSC workpackage 5, enabling the user to write the instrument file in python. The API also covers executing the simulation, provides easy access to the data and convenient plotting functionality.

Oasys (*OrAnge SYnchrotron Suite* [15]) is a modular environment, built for simulations of X-ray optics in an easy-to-use graphical user interface. Several add-ons for it have been developed, each providing a different kernel for calculation of the electric field propagation along the beamline. Only a few are listed here for brevity: *ShadowOui* for ray-tracing calculations [15], *SRW* for wavefront propagation using fast Fourier transforms and *WISEr* for wavefront propagation using numerical integration usable for calculation of grazing incidence optics. As OASYS is highly modular, the calculations can be performed using one method and then

converted to another. This gives the user an opportunity to do fast prototyping and then resort to a more computationally expensive technique for more accurate results. The main effort in PaNOSC is directed at two improvements: computational optimization of the code, including scalability; user accessibility of the environment from a remote machine, either through SSH or remote desktop.

The second pillar of ViNyL is the harmonization of metadata standards and data formats for simulation data. This will enable the construction of complex simulation workflows, where data flows between the various simulation codes without the burden of error prone reformatting of output data from one code to become input data for the next simulation step.

For each generic experimental stage, ViNyL will support a common metadata standard. We will not (re)-define these standards but rather build on community driven open standards. Examples are the openPMD metadata standard for particle and mesh data [16] or the Nexus standard [17] based on HDF5 for experimental data, which we will employ for simulated signal data, in cooperation with workpackage 3 of PaNOSC.

Our approach, to define a simulation software infrastructure will allow users to access our simulation capabilities from various types of user interfaces including classical graphical or command–line user interfaces, but also more modern frameworks such as the jupyter notebook/jupyter lab ecosystem [18], or workflow environments such as knime [19] Oasys.

A major advance of ViNyL compared to earlier simulation environments is that it will be exposed as a service to the user community, delivered either as a web (browser) service or as a remote desktop application. Simulations as a web service will be based on jupyter notebooks. We will provide well documented and preconfigured notebooks, that a user can modify to his needs and execute in an encapsulated computing environment based on the binder service [20] hosted on a high–performance computing system. This aspect will be developed in close collaboration with work package 4 of PaNOSC.

We also foresee to expose our simulation services through remote desktop solutions. This is particularly interesting for the Oasys platform which is developed as a pure desktop application. Novel additions to Oasys, such as coherent wavefront propagation, require substantial compute resources. Here, a user could login via a remote visualization service on the compute server that hosts the Oasys application.

Clearly, one major use case of our simulation platform is in the educational domain: Students, new employees at research facilities and their users could perform virtual experiments to learn about the various components of the experiment, and to gain experience in handling research data. Within PaNOSC, we will contribute to the teaching material for the photon and neutron science e–learning platform developed in workpackage 8.

DATA PORTAL

This project aims to provide a common portal which will offer access to the remote data analysis services at each participating facility. These services will allow a user to analyse experimental data through the provisioning of remote desktops and Jupyter notebooks.

Initially work will be carried out to design a unified architecture which will be implemented at each individual facility. The objective here is to provide these data analysis services locally. The portal will be developed collaboratively across all partners and the code will be made available to the wider open-source community.

It is assumed that each facility will have to develop certain site-specific elements to enable integration into their local systems and infrastructure. The architecture should take this into account in order to ease deployment.

Once this primary phase has been validated, the next step will be to develop a single access point that will allow transparent access to the data analysis services provided by each facility.

Consequently, a scientist will not only be able to search for data across all facilities but will able to seamlessly access online data analysis services.

EOSC INTEGRATION

In order to integrate our services in EOSC, we have decided to partner with other stakeholders already involved in the construction of the EOSC infrastructure such as GEANT and EGI. With their active support, we have started to prepare our AAI service following the AARC Blueprint Architectecture (`https://aarc-project.eu/architecture/`), as the AARC BPA is de facto the standard AAI model for EOSC. The Photon and Neutron AAI service, UmbrellaID (`https://umbrellaid.org`) is currently being integrated with the eduTEAMS service operated by GEANT. This integration, planned to be completed by the end of 2019, is expected to provide full compatibility of our community AAI with EOSC and the other community services.

We are also actively working on data transfer where we have 3 pilots running to identify the best solutions for addressing simple use cases where a scientist wants to access the PaN RIs data through EOSC services, typically an analysis service not necessarily provided by the community. In this scenario the data are archived on the RI premises and the user service is remote. The 3 solutions currently being explored are OneData and dChache both providing some sort of cache on the service side and two more basic solutions based on Webdav and GlobusOnline.

We will also have to consider user support activities, service monitoring and accounting in the EOSC federated model. These activities are currently being specified. Work is planned in 2020 with the other EOSC builders to define and implement common mechanisms that should allow smooth service operation.

EOSC VISION

PaNOSC sees EOSC as an opportunity to generalise the adoption of FAIR data policies at all photon and neutron sources. Adopting FAIR data will enable data sharing across a wider community and the provisioning of services for remote data analysis. In order for these objectives to be realised the EOSC must provide the following services:

1. a common way of identifying, authenticating, and authorising users (AAI) across Europe;
2. a free service for downloading data efficiently (distributed and high bandwidth);
3. a (commercial or free) solution for long term archiving of large quantities of open data (petabytes) coupled to (commercial or free) high-performance storage and compute resources for the (re)analysis of this open data;
4. a search machine for searching and finding scientific data in a wide variety of domains;
5. a catalog of (free and commercial) services for analysing data ranging from generic services like Jupyter notebooks to specific applications per scientific domain.

The above needs are considered the Minimum Viable Ecosystem for the EOSC from the PaNOSC point of view. It would be desirable for the EOSC to become more than just a data warehouse. The EOSC should become the *GitHub of Open Science in Europe*! To achieve this it must provide scientists with a personal space where they can create content (data analysis notebooks, workflows, publications), store data and share their work with collaborators via a versioning system like GitHub.

The PaNOSC RI's would in return offer (1) petabytes of raw and processed data in a wide variety of scientific domains; (2) tools for generic and specific data simulation and data analysis ; (3) notebook recipes and expertise for reducing and analysing data; (4) training material for understanding photon and neutron science. Figure 2 depicts PaNOSC's vision of the EOSC as multiple layers of clouds.

Some of PaNOSC RIs databases of data collected over the last 3 to 4 decades are currently under-exploited e.g. paleontology data in the `https://paleo.esrf.fr` is an example of processed data which are not widely known or exploited yet. These data are ideal for cross disciplinary applications and linking up with data from museums and/or other scientific disciplines

SUSTAINABILITY

The sustainability of the PaNOSC research data infrastructure that will be integrated in the federating core of the EOSC depends not only on sound business models that create added-value for the end-users but also on the incentive and rewards for researchers that encourage them to participate in a culture of sharing the results of their research. Without

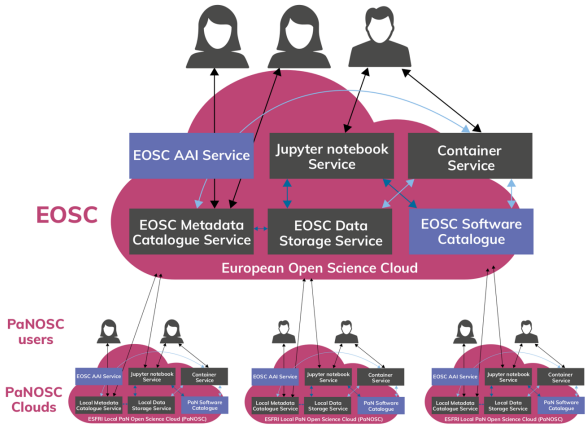

Figure 2: PaNOSC vision of EOSC cloud services

the engagement from researchers, we could be technically successful but the impact of PaNOSC would drastically be reduced.

The sustainability work package has the purpose to reduce that risk by involving all the PaNOSC stakeholders, by calculating the cost of the PaNOSC infrastructure in an "auditable" way. A tool for measuring impact of data and publications will be adapted to the proposal system of the user offices to track citations of data and experiments based on proposals and DOIs. By defining suitable metrics to evaluate the impact of a FAIR data infrastructureand exploring suitable business models will enable a sustainability plan to be defined and implemented.

Considering the complexity of the task involved an iterative approach based on the Deming PDCA (Plan-do-check-act/Plan–do–check–adjust) cycle will be adopted.

A database of stakeholders will be used to involve stakeholders and get feedback via targeted questionnaires and interviews. As engaging with the stakeholders only via questionnaires may be not enough, a presence of PaNOSC representatives to the Research Infrastructure User Meetings and direct engagement with individual scientists who can be considered early is foreseen. The feedback from stakeholders and early adopters will allow us to address the other tasks of the work package. An audit-able cost model will be developed and metrics will be defined to evaluate the added value of services.

The business models suitable for the sustainability of the PaNOSC infrastructure and the sustainability plan will follow. An important aspect that will be considered in this work package is the analysis of the scope of the PaNOSC infrastructure as part of the federating core of the EOSC, its dependence on the EOSC sustainability and the associated risks.

TRAINING

Training is relevant for PaNOSC, and in a wider remit, EOSC, in two regards:

1. Training is a central element for any scientific community and thus digital platforms facilitating such training should be part of the EOSC service catalogue particularly because such platforms will serve users of multiple facilities. At the same time, such platforms can benefit greatly from leveraging other EOSC services, such as access to data and computing power.

2. A fully comprehensive description of how data are produced can rarely be automated but depends on the commitment of the involved scientists for providing adequate information and hence making the data truly FAIR. Thus, for open science to become a reality it is not sufficient that the underpinning technologies and policies are in place. Equally important is the behavior of the individual facility staff member and user and that they see the importance and benefit of making their data FAIR.

To address the two points above, PaNOSC has dedicated a work package to staff and user training that contains three high-level activities, namely e-learning platform, staff training, and user training.

E-LEARNING PLATFORM

PaNOSC will convert the e-learning platform `https://e-neutrons.org` to a platform suitable for both photon and neutron sources and adaptable by EOSC. The e-learning platform e-neutrons.org originates from two EU funded projects NMI3 (`https://nmi3.eu/`)and SINE2020 (`https://sine2020.eu`) and consists of a wiki, moodle, and a simulation platform where virtual experiments with (virtual) neutron instruments can be performed. As part of PaNOSC it will be migrated to ESS for sustainability and the platform will be extended with Jupyter technology, which will enable interactive python based tutorials to be developed as part of the platform (e.g. for data analysis software). Moreover, the activities in simulations will be leveraged to extend the simulation platform with instruments from involved photon sources. In PaNOSC the e-learning platform will be used to support the activities in staff and user training.

STAFF TRAINING

Staff at the involved facilities will be trained in data stewardship in order to facilitate the uptake of open science and FAIR data at the facilities and to have them serve as evangelists of open science. Moreover, staff will be trained in how to use the e-learning platform for developing and providing their own courses.

USER TRAINING

Training materials will be developed for users to foster the uptake of services developed in PaNOSC, but the e-learning extensions will also be utilized to develop courses relevant for all facilities, for instance to demonstrate for students how the involved facilities complement each other. The training will be provided through the e-learning platform and summer schools.

TRAINING SCHOOLS

One of the goals of PaNOSC as well as the other EINFRAEOSC-04 cluster projects is to share scientific software tools training and best practices. A very efficient way of doing this which is common in the Open Source community is by sponsoring events where users can meet, exchange and be trained. PaNOSC has started doing this by co-organising two events on software tools:

1. **OASYS school** [21] - the first OASYS school was organised in May 2019 at the ESRF for training new scientists how to use OASYS to design the optics for XRay beamlines

2. **h5py code camp** [22] - a code camp to work on issues in the HDF5 Python library (h5py) was organised in September 2019 at the ESRF in conjunction with a workshop on HDF5 the data format used by most photon and neutron sources.

Both events were well attended and highly appreciated by newcomers and existing members of the community. They help share the knowledge and increase the number of developers who understand the tools. This is vital for the long term sustainability of Open Source software used by the community. More events are planned together with the other EINFRAEOSC-04 clusters to share best practices of common software tools. The EOSC should play an important role fostering such events.

COMMUNICATION

PaNOSC has been implementing a combined communications-dissemination strategy, with the aim of raising awareness among target groups about project results and best practices, highlighting the benefits that the project brings to the community. The final aim is to drive a change in culture by embracing Open Science among the scientific communities that the project will target. This translates into the contribution to the definition of a common FAIR vocabulary and narrative shaping open data policies and their implementation, and underlying the training activities of the users and staff who will be using the services and tools developed in the project.

Effective communications also mean an increased coordination and networking among the members of PaNOSC, and between them and the other EOSC clusters and stakeholders, in order to foster the exchange of best practices towards a coordinated and harmonized construction of the EOSC and its adoption by target user communities. To this aim, regular meetings take place among the partners and actors involved in the development of the Open Data policy, and the services for data catalogue and storage, analysis and simulation, to plan future developments and update about the progress of the project.

To ensure the community is on track and up to date, the information about all advancements is made publicly available through different web tools and platforms,

i.e. the project's website (https://panosc.eu) and the project's open repositories on GitHub (https://github.com/panosc-eu). The latter is an open source tool, which has been used since PaNOSC's start for the documentation and management of the project and its team, allowing and facilitating communications and interactions between the partners on the different project's topics and WPs. PanOSC is has adopted a FAIR strategy for its internal communication!

Actions taken to pass information to and interact with stakeholders, as well as to assist exploitation of the outputs and results also include:

1. Public project updates via the project's and its partners' websites, newsletters and social media posts, as well as articles and/or press releases about key PaNOSC's events and achievements.

2. Interaction with multiple stakeholders at the project annual meetings and at partners' user meetings, at scientific and industrial workshops and conferences Europe-wide, at the ICRI conferences, EU presidency's events, EOSC events, ERIC Forum meetings, etc.

3. Interaction with contact points of the National Ministries at the partners' governing bodies' meetings, at events organized by the Ministries, as well as by standard contact points of government grant agencies in countries where partners are present.

4. Interaction with other clusters for the coordination and co-development of services.

5. Consultation and discussion with the Observers and other members in the PaNdata community.

MANAGEMENT

The project management methodology used for PaNOSC will be based on PRINCE2[23], a well-known structured project management methodology and managed by the ESRF in Work Package 1. This has resulted in the project being divided into different stages with deliverables and milestones used as control points.

The organisational structure for the project will comprise the following bodies:

1. **Executive Board** as the ultimate decision-making body for the project

2. **Project Management Committee** as the supervisory body for the project execution that is accountable to the Executive Board.

Each Work Package has a leader in charge of coordination and delivering the work agreed in the proposal and feeding back status information to the Project Management Committee. The Project Coordinator will lead the coordination with the European Commission and supervise the Project Manager who will coordinate the day-to-day execution of the project. Clear roles and responsibilities are be assigned to all persons involved in the project.

To enable effective project management, regular internal communication channels are used (bi-weekly Project Management Committee videoconferences, mailing lists, GitHub repository, etc.), the proposal funded by the European Commission is used as the baseline project plan and a escalation procedure is in place for any kind of exception and/or issue.

CONCLUSION

Approaching the end of its first year of execution PaNOSC has so far brought together its partner institutes to work on the challenges of implementing FAIR and handle the huge increase in data volumes. The agreed milestones and deliverables are being respected, however plenty of work is still required over the next three years in order to make the most of the grant awarded and help push forward the data, policies and services required to make the photon and neutron sources FAIR-compliant.

PaNOSC as part of the EOSC will depend on the EOSC for a number of services as outlined in the section on the EOSC vision. The EOSC is still in the process of being defined and developed which therefore increases the uncertainty around its roadmap. PaNOSC has taken this risk into account by targeting services which can be developed and consumed locally first to ensure that existing user communities benefit first in adopting Open Science practices. The larger community of users of Open Data will depend on the EOSC for services.

ACKNOWLEDGEMENTS

The authors acknowledges fruitful discussions with the other EINFRAEOSC-04 clusters and the ExPaNDS project.

REFERENCES

[1] *PaNOSC*. https://panosc.eu/. [Online; accessed 07-September-2019].

[2] *ESFRI*. https://www.esfri.eu//. [Online; accessed 07-September-2019].

[3] *EOSC*. https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud. [Online; accessed 4-September-2019].

[4] *Foster Open Science*. https://www.fosteropenscience.eu/. [Online; accessed 07-September-2019].

[5] R. Dimper et al. "ESRF Data Policy, Storage, and Services." In: *Synchrotron Radiation News* 32.3 (2019), pp. 7–12.

[6] *OpenAIRE*. https://www.openaire.eu/. [Online; accessed 30-September-2019].

[7] *B2FIND*. http://b2find.eudat.eu/. [Online; accessed 30-September-2019].

[8] *ExPaNDS*. https://expands.eu. [Online; accessed 30-September-2019].

[9] *ICAT*. Collaboration, T I 2014 The ICAT Project. DOI: https://doi.org/10.5286/software/icat.

[10] *SCICAT*. https://scicatproject.github.io. [Online; accessed 30-September-2019].

[11] *PaNdata Software Catalogue*. https://software.pan-data.eu/. [Online; accessed 22-August-2019].

[12] Carsten Fortmann-Grote et al. "Start-to-end simulation of single-particle imaging using ultra-short pulses at the European X-ray Free-Electron Laser." In: *IUCrJ* 4.5 (2017), pp. 560–568. https://doi.org/10.1107/S2052252517009496

[13] *The github repository for simex_platform*. https://github.com/eucall-software/simex_platform

[14] Kalliopi Kanaki et al. "Simulation Tools for Detector and Instrument Design." In: *Physica* B551 (2018), p. 386. DOI: 10.1016/j.physb.2018.03.025. arXiv: 1708.02135 [physics.comp-ph].

[15] Luca Rebuffi and Manuel Sanchez del Rio. "OASYS (OrAnge SYnchrotron Suite): an open-source graphical environment for x-ray virtual experiments." In: *Advances in Computational Methods for X-Ray Optics IV*. Ed. by Kawal Sawhney and Oleg Chubar. Vol. 10388. SPIE, 2017, 103880S–103880S. DOI: 10.1117/12.2274263.

[16] Axel Huebl et al. *OpenPMD 1.0.0: A meta data standard for particle and mesh based data*. en. 2017. https://dx.doi.org/10.5281/zenodo.33624

[17] Mark Könnecke et al. "The NeXus data format." In: *Journal of Applied Crystallography* 48.1 (2015), pp. 301–305. DOI: 10.1107/s1600576714027575.

[18] Thomas Kluyver et al. *Jupyter Notebooks – a publishing format for reproducible computational workflows*. IOS Press, 2016, pp. 87–90.

[19] https://www.knime.com

[20] Project Jupyter et al. "Binder 2.0 - Reproducible, interactive, sharable environments for science at scale." In: *Proceedings of the 17th Python in Science Conference*. Ed. by Fatih Akici et al. 2018, pp. 113–120. DOI: 10.25080/Majora-4af1f417-011.

[21] *First OASYS Schoole*. https://indico.esrf.fr/indico/event/26/overview. [Online; accessed 29-September-2019].

[22] *h5py Code Camp*. https://indico.esrf.fr/indico/event/33/overview. [Online; accessed 29-September-2019].

[23] *PRINCE2*. https://www.axelos.com/best-practice-solutions/prince2/what-is-prince2. [Online; accessed 25-September-2019].