PaNOSC – Making FAIR open data a reality for Photon and Neutron Science

Adobe Stock © Ar_TH

Photon and neutron (PaN) facilities are research infrastructures (RIs) using lasers, X-rays or neutrons on samples to obtain data and high-resolution images up to the nano-scale that can help us better understand how matter and biological processes work. They are essential for the understanding of matter and its properties. Together these facilities produce petabytes of raw data, which with the appropriate data treatment and interpretation, can give us a more complete picture of the world around us.

To meet the challenges given by the ever-increasing amount and volumes of data and by data fragmentation, in 2015, the European Commission (EC) proposed a new concept: the European Open Science Cloud (EOSC), aiming to provide scientists with access to data, software and services from as many scientific data sources in Europe as possible, by making them FAIR (findable, accessible, interoperable, reusable) across research facilities and scientific domains. To achieve this, the EC financed a large number of projects to kickstart the building of the EOSC in many domains.

The Photon and Neutron Open Science Cloud, PaNOSC project—which started in December 2018 and is close to completion—is one such project. Together with its sister project, ExPaNDS, it gathers the majority of PaN facilities in Europe. The project's main goal is to address the FAIR principles in the PaN RIs by equipping them with all the necessary software, policies and the required legal and administrative frameworks. Throughout its implementation, PaNOSC has provided common policies, strategies and solutions for enabling open science through the adoption of FAIR principles across three European PaN RIs and three European Research Infrastructure Consortia (ERICs), helping to make their data open and available to the EOSC.

By closely working together for the PaN research community, PaNOSC and ExPaNDS have paved the way to make data produced at PaN facilities across Europe easily accessible: data has started to be curated and made available under an open FAIR data policy. Even the two project partners (ESS ERIC and ELI ERIC), which are still in the process of construction, are now ready to publish open data from the beginning of operation.

Overall, all facilities have adopted, or are in the process of adopting, the PaNOSC FAIR research data policy framework, which serves as guidance for FAIR data stewardship. It does this by defining the curation of (meta)data from the generation of raw data from each experiment, to analysis, through to publication and re-use.

There are many reasons and benefits for PaN facilities to adopt a FAIR data policy. These range from the need to make science reproducible and replicable by adopting an open science approach to improving the quality of scientific data. This implies following the recommendations of international bodies (e.g. OECD, ISC, IUCr, G7), implementing FAIR principles to enable data re-use, providing scientists with new data services and archiving important datasets.

Data stewardship also implies the implementation of effective **data management plans (DMPs)** to ensure that RIs' users and support teams are aware of the data volumes that will be produced and how to process them throughout the whole data lifecycle. Such information is also beneficial to forecast the IT infrastructure required to support an experimental programme based on a more detailed understanding of users' needs. This is why a solution for generating and managing DMPs for each experiment has been proposed and implemented across PaN facilities.

How to find and access the data?

Prior to the start of the project, PaN facilities had their own data catalogues and search tools for users to retrieve their datasets and related metadata at single facilities. However, to make PaN data easily findable and accessible across a multitude of PaN facilities in Europe, domain-specific searches across the PaNOSC data repositories needed to be enabled. This has happened by developing and adopting a **federated search** API (application programming interface) for PaN data catalogues and a common protocol for harvesting data and metadata to make public datasets available to third-party EOSC cross-discipline repositories. This service provides a unified way across facilities for PaN scientists to find, filter and score/rank datasets and publications from any number of configured sites based on relevant domain-specific metadata using a variety of parameters (source characteristics, sample information, detector details, etc.), and can be used by third parties to find data released from any facility after the embargo period.

The work towards building up a custom graphical search user interface has been carried out as part of the **Open Data portal** (https://data.panosc.eu/) development task, which consisted of the implementation of a web portal to use the federated search API across metadata catalogues and data repositories to search, find and download open data across all PaN RIs in Europe which deploy the federated search and provide open data. Some facilities are still in the process of populating site-specific data catalogues with metadata for open-access investigations. At newer facilities, the work has focused on designing and building the necessary background infrastructure to capture, secure and store raw experiment data, together with its associated auxiliary data and metadata.

A community metadata standard for PaN sources (NeXus/HDF5) has been widely adopted to make data interoperable and reusable. Electronic logbooks have been developed to capture what happens during experiments and keep track of the various steps and settings of the experiments for future usage. Also, facilities have dedicated resources to generating **DOIs (digital object identifiers)** for each experiment and for one or more specific datasets to be cited in publications. DOIs ensure data are findable and accessible and enable tracking of data re-use by other research

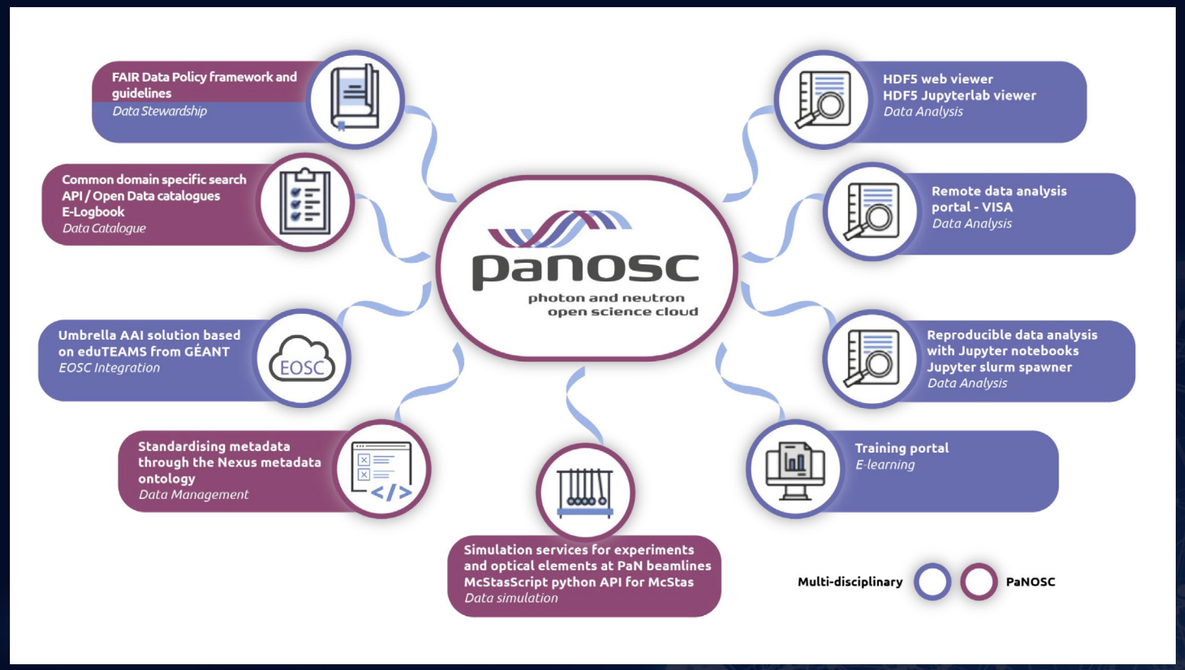

Image: PaNOSC overview infographic.

teams in the same or different domains. In addition to data generation, collection, processing, storage and management, data need to be visualised, analysed and interpreted. Given the increased data amounts and volumes, this requires an increasing level of computer power and storage space and makes a download and local computation for data analysis by single users partly unfeasible. To further lower the barriers to re-use of open data, users should be enabled to explore data through their web browser after identifying a dataset of interest. To this aim, PaNOSC has provided a remote access infrastructure to enable and contain FAIR data services for users of the PaN community and scientists from across domain borders, through the EOSC. This has been achieved by making available and developing two types of data analysis services: **remote desktops** for graphical software use and **Jupyter Notebook**[1] for programmatic data analysis. These run in virtual machines and can be accessed remotely via the open-source data analysis online portal, **VISA—Virtual Infrastructure for Scientific Analysis** (watch the video). Initially designed at the Institut Laue-Langevin and further developed in the frame of PaNOSC, VISA offers remote control, simulation services of experiments and experimental set-ups and data analysis (of user data or open data from the data portal). So, basically, VISA is a new way for academic and industrial researchers to access data and advanced analysis tools from anywhere. It offers support for real-time collaboration through data sharing and can have a great impact on the work of researchers in the PaN community, as it increases their scientific capacity and productivity. Users can select their experiment and resource options (memory, CPU, display) for the virtual machine to be used and the type of analysis service to operate within the virtual machine: a Jupyter Notebook or a remote desktop with access to the software stack contained in the virtual machine image.

Free, open-source software and services for simulation and modelling of PaN sources, beamlines and experimental instruments, and start-to-end simulations to describe entire experiments, are also accessible via VISA, as part of the **Virtual Neutron and X-ray Laboratory (ViNYL)** developed within the project. ViNYL enables PaN users to rapidly implement simulation and analysis workflows specific to their facilities, instruments and experiments. This is important, as 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. Among the software packages available are:

- **McStasScript**
  McStas simulation code, which is world-leading in the simulation of instrumentation for (virtual) neutron scattering experiments
- **OASYS**
  Open-source graphical environment for x-ray virtual experiments
- **SIMEX**
  Photon experiment simulation environment.

Providing training modules in PaN science is also a project goal. In this domain, PaNOSC, in collaboration with ExPaNDS, has developed an e-learning platform hosting free education and training for scientists and students, with online interactive courses on both the theory of PaN science and how to use python code or software for data reduction and modelling. Moreover, a training catalogue for PaN science allows browsing instructional material and resources from institutes around Europe.

The platform has also been added to the service catalogue in the EOSC Portal[2]. In addition to it, the following services are now provided and accessible via the EOSC, using a single AAI—authorisation and authentication infrastructure service (Umbrella ID)—which enables users to log in to multiple applications and websites with one single set of credentials:

- **PaNOSC Software Catalogue**, with over a hundred standard software tools used for analysing data from PaN RIs
- **Human Organ Atlas**, an open data portal of 3D scans of human organs with micron resolution for different pathologies, including COVID-19
- **PaNOSC open data portal**
- **Search API service**

How to make the PaN EOSC sustainable in the long run?

Sustainability has been a core issue since the project's start. A number of services developed by PaNOSC are being integrated into the core business activities of the facilities and will need to be sustained. Certain fundamental objectives like FAIR data and open science will change the way the facilities work and will be considered the new norm. Where additional costs are incurred by EOSC users, these have been collected and are being used to produce a business model and a formal long-term mission and vision for the sustainability of the PaNOSC infrastructure and software developed. As also stated in the **LEAPS roadmap** presented to the EC in June 2022, investments in sustaining and strengthening open science activities are needed to develop the federated service **PaN Open Data Commons** as an essential component of EOSC for showcasing and accessing data.

However, PaN facilities alone cannot achieve the goal of making data open and FAIR. The contribution of researchers is key to making EOSC and FAIR data a reality in the long run. Authors need to embrace FAIR research practices as well. They are strongly encouraged to link the software used to obtain the results of their analyses with the raw (meta) data and to make software and results openly accessible, together with the analysis procedure description, scripts, software and software environments that completely describe the process of data analysis from the raw and metadata to the published results, to allow others to reproduce that analysis.

1 Web application that allows users to create and share documents containing live code, equations, visualisations and descriptive text.

2 The EOSC Portal is a gateway to information and resources in EOSC, providing updates on its governance and players, the projects contributing to its realisation, funding opportunities for EOSC stakeholders, relevant European and national policies, important documents, and recent developments. The EOSC Portal Catalogue & Marketplace acts as an entry point to the multitude of services and resources for researchers.

PROJECT NAME

Photon and Neutron Open Science Cloud (PaNOSC)

PROJECT SUMMARY

The Photon and Neutron Open Science Cloud (PaNOSC) is a European project enabling open science through the adoption of FAIR principles across photon and neutron (PaN) facilities in Europe. To this aim, PaNOSC has developed and provided common policies and software and services connected to and made accessible via the European Open Science Cloud (EOSC).

PROJECT PARTNERS

European Synchrotron Radiation Facility (ESRF)
Central European Research Infrastructure Consortium (CERIC-ERIC)
European XFEL
Institut Laue-Langevin (ILL)
Extreme Light Infrastructure ERIC (ELI ERIC)
European Spallation Source ERIC (ESS ERIC)
EGI Foundation

PROJECT LEAD PROFILE

The ESRF (European Synchrotron Radiation Facility) is the world's most intense x-ray source and a centre of excellence for fundamental and innovation-driven research in condensed and living matter science.

The intense source of synchrotron-generated light produces x-rays 100 billion times brighter than the x-rays used in hospitals.

Located in Grenoble, France, the ESRF owes its success to the international cooperation of 22 partner nations.

PROJECT CONTACTS

Andrew Götz, Software Group Leader at ESRF, PaNOSC Project Coordinator

Jordi Bodera Sempere, Project Coordinator at ESRF, PaNOSC Project Manager

Nicoletta Carboni, Senior Communication Officer at CERIC-ERIC, PaNOSC Communication Officer

andy.gotz@esrf.fr
www.panosc.eu/
/Panosc_eu (@Panosc_eu)

FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852.