Grant Proposal

Author-formatted document posted on 05/12/2023

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e116674

PaNOSC/ExPaNDS Science Projects for EOSC Future (WP6.3): Demonstrating EOSC Value through cross-domain Research Science Projects

Miguel Angel Gonzalez, Rudolf Dimper, Patrick Fuhrmann, Fra Schluenzen, Gianluca Santoni, Jordi Bodera, Jayesh Wagh, Irina Safiulina, Paolo Mutti, Arianna D'Angelo, Paul Millar, Krisztian Pozsa, Leonardo Sala, Alun Ashton, Giuseppe La Rocca

ARPHA Preprints Author-formatted document posted on 05/12/2023. DOI: https://doi.org/10.3897/arphapreprints.e116674

Science Projects from PaNOSC and ExPaNDS: Demonstrating EOSC Value through cross-domain Research Science Projects

Authors: Dimper, R., Fuhrmann; R., Schluenzen, F., Santoni,G. Bodera,J., Wagh ,J., Safiulina, I., González,M., Mutti,P., D'Angelo, A, Millar, P., Pozsa, K., Sala,L., Ashton, A., La Rocca, G

Description

The Photon and Neutron (PaN) user facilities serve a large variety of scientific user communities using x-rays and/or neutrons in their research projects in order to understand the structure and function of matter. The most important goal is hence the advancement of science through measures including accelerator and detector developments, evolution of experiment capabilities, or algorithmic and application development.

A number of projects like CALIPSOplus, ExPaNDS or PaNOSC already serve this approach for a small number of use cases in a complementary manner. The activities of the PaN user facilities in EOSC Future will hence focus on key aspects in line with the initial call: (i) integration; (ii) consolidation; (iii) boosting data intense research; (iv) enabling researchers to scalable environment; (v) widening the EOSC user base; and (vi) supporting cross-analysis of data from heterogeneous sources.

In the course of the COVID-19 crisis, the PaN User facilities have impressively demonstrated their ability to react quickly to new scientific and societal challenges. Many of the facilities initiated fast-track access for COVID-19 research, enabling a plethora of structural investigations, resulting in hundreds of thousands of datasets and numerous high-level publications. bioRxiv and medRxiv alone count more than 4300 publications on the topic. Despite the tremendous success, the approach also revealed a few deficits, which could have been avoided if the integration into EOSC had been more advanced. The publication of the raw underlying scientific data does not keep up with the publication of the results, which might need to be taken into consideration by facilities and projects implementing or updating their scientific data policies. Even if all data had been published, the reproducibility and validation of the scientific outcomes would have been nearly impossible. In many cases, in particular in very innovative, rapidly evolving research fields like serial crystallography, bleeding-edge processing pipelines are rarely commonly available, or require an expertise prohibitive for cross-disciplinary assessment. The Science Projects (SPs) will hence concentrate on imaging applications, which can have a strong impact on bio-medical or bio-chemical characterization of biological materials, with the aim to achieve a high level of reproducibility and automated validation by providing easy to handle cutting-edge processing pipelines integrated into EOSC and existing validation pipelines.

The Science projects (SPs) are scientific analysis projects that will produce new scientific results and publications, addressing key interdisciplinary scientific objectives of the participating PaN facilities, fully exploiting the PaNOSC/ExPaNDS services in EOSC. The SPs will demonstrate the innovative impact of the open-science data analysis in the EOSC framework and promote the application of FAIR principles for data stewardship. With this approach, we will be able to address the key expected impacts of the project.

The two SPs selected address different but complementary scientific communities exploring related scientific questions from different perspectives. Despite common objectives and challenges, the cross-disciplinary exploitation of the different approaches is more of an exception than the rule. The proposed SPs will lower the barrier of cross-disciplinary exploitation and advance innovative developments by providing user-friendly, readily available and cutting-edge data processing pipelines, accelerating the take-up of emerging technologies. This approach embraces FAIR-enabling services, many of which are available in the EOSC marketplace, thus increasing researcher's engagement in the EOSC implementation.

Science Project 7 – “Tracing bio-structures with Serial Crystallography”

Serial (femto-second) X-ray-Crystallography (SFX) is a special variant of macromolecular X-ray crystallography aiming at rapid structural studies at room temperature. This highly innovative technology permits investigation of bio-structures not tractable with conventional X-ray crystallography, and is capable of studying fast in-situ biochemical processes.

The method is still relatively new, but it is already one of the most prominent applications of free-electron lasers (FELs), and increasingly also of very brilliant synchrotron radiation sources. One of the unique characteristics of this type of experiments is the extremely high repetition rate combined with a quite moderate success rate. A crucial task in the rather complex data processing pipeline is the rapid and accurate classification of images: typically, only a few percent of the images contain a diffraction pattern suitable for subsequent integration and structure refinement. AI-supported image classification is hence particularly suited for drastic data reduction, saving precious storage space, compute cycles and processing time.

The experimental techniques and methodologies are rapidly evolving, and the integration of emerging tools into the processing pipeline is an essential task. SFX data sets are big, require substantial storage, and computational power. The main goal of this SP is to establish and develop a data processing platform, which integrates services and developments from PaNOSC/ExPaNDS. The platform should provide integrated processing pipelines for well-established and cutting-edge applications, so that cross-disciplinary users with modest expertise gain rapid and convenient access to tools and documentation of newest developments. On the other hand, it should also provide convenient access to FAIR SFX-data, to foster developments and strengthen collaboration between experimentalists and developers of new algorithms and software implementations. The approach is of very high relevance for all PaN synchrotron and FEL facilities and their users.

Partner institutions on SP7: ESRF , DESY, PSI

Science Project 8 – "Following biological processes with Small Angle Scattering"

Small angle scattering (SAS) techniques are widely used in the scientific communities to determine the shape, distribution and uniformity of particles in solutions. In the case of biological systems (proteins, cell membranes, DNA or RNA, etc.), the measured signal is an average over a representative set of conformations representing the atomic structures that the system can explore. Therefore, combining this experimental information with advanced modelling tools it becomes possible to determine the organization of complex biostructures and their fluctuations, which are essential for their biological activity. Thus, the combination of small angle X-ray scattering (SAXS) and small angle neutron scattering (SANS) can be very effective to study for example the time-dependence of genome release from phages, to investigate entire viral life cycles or the process of assembly of macromolecular complexes, providing deep insights into infection pathways. Neutron and X-rays can be applied in a very complementary mode, as it is the case for the joint SANS-SAXS proposal between the ESRF and the ILL.

This SP aims to advance the field by providing an EOSC based platform, enabling FAIR data and software, unified data processing pipelines featuring robust scaling algorithms for the two different sources, supporting reproducibility and automated validation, and integration with other relevant structural databases (e.g., electron microscopy/tomography or protein structural and ligand databases). Rapid interpretation of the data could be enabled with the inclusion of an AI machinery through multi-modal registration and classification.

Partner institutions involved in SP8: ILL, ESRF

Both SP are quite complementary. SP7 aims more for atomic structural information, whereas SP8 aims more for macromolecular dynamics. SP7 needs to deal with huge data rates and volumes, whereas SP8 has to handle rather moderate data volumes but requires a higher level of integration with available FAIR resources. Both SPs on the other hand are highly cross-disciplinary and share substantial commonalities on the underlying technologies, which will be realized as part of PaN-commons services. The SPs can easily be applied to and benefit from closely related imaging approaches like single particle cryo-electron-microscopy, FEL coherent single particle imaging or the various manifestations of tomography. The PaN-commons developed in INFRAEOSC-03 will serve the whole community by applying the generic services developed in PaNOSC and ExPaNDS to different techniques. The PaN-commons will exploit common software like Jupyter notebooks for a large number of techniques so that users of FAIR data from PaNs will be able to benefit from a catalogue of solutions for multiple techniques. The lack of ready-to-use recipes, i.e., notebooks, is a barrier for new users of PaN data which the EOSC will help to lower. The PaN-commons will extend FAIR policies and services to the PaN community in general.

Contribution to EOSC and from EOSC, e.g., cross-domain and composability features, contributing to EOSC; the feasibility of integrating the proposed services into the EOSC ecosystem; the type of services the proposal asks from EOSC (e.g., security, monitoring, AAI);

PaNOSC/ExPaNDS will provide the following services for integration to EOSC:

- HDF5 data viewer available in EOSC (M6)
- Umbrella AAI federated with EOSC AAIs (M12)
- Decision on data transfer solution (PaNOSC/ExPaNDS vs EOSC) (M12)
- Data search API for domain specific searching of PaN data, in collaboration with OpenAIRE (M12)
- Jupyter notebook prototype, PaN data analysis portal (M15)
- Jupyter notebook and PaN data analysis portal integrated with EOSC AAI, data transfer, and data search (M20)
- PaN learning platform with interactive content (M20)
- Data compression service (M20) (eventually from LEAPS-INNOV)

PaNOSC/ExPaNDS need from EOSC the following services:

- Umbrella AAI federated with EOSC AAIs (M12)
- Data search enabled over all connected data repositories based on metadata ontologies and digital identifiers (M18)
- EOSC HelpDesk integration (M18)
- Access to storage and compute with clearly defined access mechanisms (M18)
- Monitoring services for storage and compute (M18)
- Long-term data archival beyond the RI data policies (M30)

Engagement

- Target groups: Users of photons and neutrons facilities, structural biology community in Europe, students and teachers for experimental data analysis and method developers.
- Dissemination plan:
    - Scientific publications:
        - Publish the data analysis platform.
        - Describe the small angle scattering analysis.
        - Possible scientific results from first users.
    - Conferences:
        - Facility users meetings.
        - Crystallography and structural biology.
    - Training:
        - Users training programs at PaN facilities.
        - Possibility to organize on-line training sessions for a specific technique.

Contact Persons:

Overall coordination for the two PaNOSC/ExPaNDS SPs: Gianluca Santoni, (Patrick Fuhrmann) and since January 1st, 2023, Jordi Bodera (ESRF) replacing Gianluca Santoni.

ESRF -	Gianluca Santoni, (Jordi Bodera)
DESY -	Patrick Fuhrmann, Frank Schluenzen
ILL -	Miguel Gonzalez, (Paolo Mutti)
PSI -	Leonardo Sala, (Alun Ashton)

Work Plan:

PaNOSC/ExPaNDS will provide portable (EOSC marketplace) Jupyter notebooks for both SPs, including Continuous Integration, MetaData search, storage events, data transfers and data management in general.

This is an unprioritized list of tasks we will work on:

Task A: EOSC Onboarding, AAI integration, Marketplace integration, Help Desk
Task B: Algorithm software preparation for notebooks and containers, integration to SLURM
Task C: Data management, storage events, data transfer and metadata
Task D: Demonstrator preparation and execution

M1 - M12: (Task B) Preparation, algorithms,

- Designing Jupyter notebooks for both use cases, e.g. CrystFEL software
- Containerization, continuous integration.
- Simulating trigger for Serial Crystallography with prefiltering with ML
- Testing different algorithms

M12: (Task D) First demonstrator (**Milestone**)
M13: (Task C) First agreements on Data Management solutions.
M13: (Task A) Agreement with EOSC-FUTURE on the support structure
M13: (Task A) Starting onboarding processes of the SP and Services into the EOSC
M15: (Task A) Enabling access via EOSC AAI (Umbrella ID, eduTEAMS)
M16: (Task A) Launching Jupyter notebooks in the EOSC (EOSC marketplace)
M17: (Task B) Continuing software adoption to Jupyter notebooks and containers
M18: (Task C) Start work on signaling creating action in the EOSC
M18: (Task C) Using ExPaNDS/PaNOSC search API
M21: (Task D) Second demonstrator (**Milestone**)
M21: (Task A) Help Desk support structure and operation (to be discussed with EOSC-Future)
M22: (Task A) Finishing onboarding of services and users
M22: (Task C) Starting to implement data management tasks

- Access to data via MetaData (B2find, OpenAire,)
- Data Transfer Orchestration (ESCAPE Data Lake)
- Long Term Archiving

M30: (Task D) Last demonstrator (**Milestone**)
M30: Paperwork finalisation