Road map to open science in nuclear physics

A. Matta & A. Lemasson with the openNP collaboration

European science is moving quickly towards Open Science, accelerating the share of scientific knowledge with society. With the rise of the European Open Science Cloud framework federating many scientific communities, it is clear that any initiative in Open Science should be grounded on the aim to fully operate within the EOSC framework.

Open Science represents an opportunity for the nuclear physics community to improve the management and use of the large data sets produced in nuclear physics facilities around Europe and foster new collaborations. As a goal towards an Open Science nuclear physics, the development of suited data workflow that could be shared among the nuclear physics European facilities would represent a major step forward. The data produced in these facilities should follow the FAIR principles (Findable, Accessible, Interoperable and Reusable) to ensure long term storage and possible future use by, and outside, the community.

Adoption of this new way of conducting research will require a clear and widely adopted road map. Such an endeavour will require work within the community and with the relevant authorities to shape a common, widely adopted, framework.

1 Open Science

Open Science has been made the standard method of conducting research and innovation funded by the European Commission. This approach aims at improving the quality, efficiency and responsiveness of research. Open Science is implemented through various strategic directions such as open access to publications and outreach. In addition sharing knowledge and data early on in the scientific endeavour should allow actors from all walk of society to engage with the progress of science. This document will focus on this latter aspect of open data.

With the latest *Horizon Europe* funding program, come the obligation for all resulting data to be made open access under the *FAIR* (Findable, Accessible, Interoperable, and Reusable) principle by default. To this end the EOSC portal of services provide the central entry point for the services developed by the different communities to reference and offer access to data and associated tools.

The EOSC portal already include many repositories and services from various scientific backgrounds. In particular, two notables examples are ESCAPE (European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures) and PANOSC (Photon and Neutron Open Science Cloud).

ESCAPE, from the High-Energy physics community, provides common solutions for the management, curation and storage of data. It spans over a series of large domains in fundamental research such as astronomy, astrophysics, astroparticle physics, high energy physics, and particle physics.

PANOSC provides a catalogue of data taken at European installations of ESRF, ELI, ESS, XFEL and ILL as well as the software necessary for reprocessing such data.

The benefit of an open data policy for the nuclear structure and nuclear dynamics community is evident. The richness of our experimental installations and detection setup provide a trove of experimental data-set. Unfortunately, most of those data-set are today out of reach for the scientific community at large, and the tools necessary for their exploitation by future generations are either closed source, non-referenced, or non-documented.

The landscape of nuclear physics installations has changed with the arrival of second generation radioactive ions facilities around the world. The specialisation of this installations to specific energy domains, isotopic regions and experimental techniques means no scientific program could be performed at a single facility. This leads to arising of travelling detector used at several installations, and an increased complexity of detection setup assembling various instruments to a purpose build, short lived, high performance multi-detector.

That complexity lead to larger data set and therefore call for optimised means of storage such as data lake. Computing power required to exploit those data set increase and the community is shifting from data exploitation on personal computer to the use of shared computing platforms.

This analysis leads to the natural conclusion that a European wide strategy is needed to tackle the challenges presented to the community. All actors are to be involved, from the ground, with the development and deployment of the tools, to the top with the implementation of practical policies in the matter.

2 openNP

A first step toward open science in the community is the *openNP* initiative. It aims at developing a central service to reference and access all existing data set and associated software. Including, for instance, experimental and simulated data set, theoretical calculations and associated software. It could extend as well to hardware design.

The *openNP* initiative has received its first funding through the EURO-LABS European project HORIZON-INFRA-2021-SERV-01-07. It gathers various actors of the data production and processing chain from several European institutions. It aims at delivering a first functional version of the catalogue within three years and to act as the bedrock of Open Science for the community.

The foreseen implementation for the service is an open catalogue where every user can create new entries and obtain an associated hash, DOI or similar unique identifier. Making them effectively traceable, findable and citable. An entry could describe a data set as well as a data format, associated software, experimental configuration, and of course auxiliary data, such as elog archive.

Rich and machine readable meta data associated with each entry will allow automatic discovery of data-sets and services. This ensures that all ingredient necessary for future exploitation of data are referenced making them interoperable and reusable with any future services.

Entries will be regrouped in collection to build self-consistent data set, them-self associated with an umbrella identifier. A typical collection being an experiment.

Such aggregation would encompass entries created by multiple actors from the data production and exploitation chain. For instance DAQ engineers will register data format and associated readout library as new entries to the catalogue. When an experiment is performed, data officers in charge of the experiment will be able to associate the data format with entries of new data files. This will automatically creates links with data file, format and readout software for future exploitation.

In a similar fashion, collaborations can managed entries of their detector configurations, allowing fast and easy association of a given detector setup with an experiment. This will create links with related material such as configuration files, geometry descriptor or publications. This mechanism will allow interrelation and association of a data set to all part essential to its exploitation.

Finally, the catalogue must also include features allowing the management of rights associated with entries and a model of rights hierarchy is foreseen,usually represented by an user explicitly added licence. This model allow the rights of a given entries to change over time with a timeline based on the different DMPs (Data Management Plans) associated with it. Those features would facilitate the work of data officers responsible for the data stewardship within collaborations and data producing institution.

All theses features will serve as the basis for interfacing future services of data lake & high performance computing and analysis platforms, but will require the development of a common authentication and authorisation and identification (AAI) service.

3 Authentication and Authorization Infrastructure

The Authentication and Authorization Infrastructures (AAI) are of paramount importance in the Research Data Management and Open Science process given the heterogeneous and distributed nature of the infrastructure. The Identity and Access Management (IAM) services are used to centrally manage the identity and roles of users that will be used for defining access rights to datasets and analysis platforms. While ensuring secured access to the datasets and services, these infrastructure should remain as simple as possible to enable a wide adhesion of the community of these tools and practices

In the last decades, major advances in global authorization and authentication infrastructures have been demonstrated in particular in EGI (European Grid Infrastructure) and WLCG (Worldwide LHC Computing Grid), building upon standard technologies like web tokens. However, these practices have barely penetrated the nuclear physics community and identity and access management is now lacking to make significant progress in integrating the various computing infrastructure used and in effective management of research datasets.

The community should not build new authentication mechanisms but should leverage on existing work to provide the secure composition of data and compute services needed to enable the data lake vision. Naturally, developments made recently within the ESCAPE project could represent a starting points.

The AAI solutions should adopt standards-based AAI solutions that:

- are flexible enough to support heterogeneous authentication mechanisms (federated identities, X.509 certificates, web tokens, social logins).
- provide the abstraction of collaboration/virtual organization, and the tools to manage membership, entitlements and access policies that will regulate access to resources for that organization.
- can support controlled delegation of privileges across the distributed chain of services implementing the data lake vision.
- can be easily integrated in existing data access and computing software leveraging standard, off-the-shelf libraries and components, in particular to map collaboration-level authentication and authorization attributes and capabilities to local access mechanisms.

The adoption of AAI services by the research infrastructures, experimental collaborations and research groups in the nuclear physics community will represent a major step towards the integration of the Nuclear Science in the EOSC ecosystem. This will enable the deployments of the data set within data lakes and their use in future analysis platforms. In addition, this could also allow the creation of virtual organizations across infrastructures that will make possible collaborations to tackle key scientific questions in a multi-datasets approach.

4 Data lakes & computing platforms

The last decade has seen the shift in data intensive industries from isolated *data silos* and processed *data mart* to the more versatile and generic *data lake*. This new kind of data storage infrastructure makes all raw and processed data readily available for new exploitation. This approach aim at fostering new collaborations and exploitation of data in new and unforeseen contexts.

In the scope of our community, a *data lake* would require a federation of existing national research data centres. The service itself would be deployed as a cloud-like service to curate and serve data to the community. The construction of a robust data service is a challenge for the community, however most national data centres already work on the question within the context of ESCAPE, and close synergy with these communities will be essential.

A possible model for computing platforms is the use of containers technology along with orchestration systems. Within this model, the community would provide the software in the form of containers through a continuous deployment workflow. The containers could then run on shared platforms along side data access services. This would partially solve the challenge of interoperability of the different software often needed to perform the full data exploitation chain.

Such an approach requires to work on two aspects of the problem. First the community need to adopt a common workflow for version control, continuous integration and continuous deployment, effectively delivering the base building blocks of the service. At the same time, we need to engage with the relevant institutions to correctly forecast the computing and storage needs of these platforms and help them scale accordingly.

5 Conclusion

We call for the development of a comprehensive and coordinated road map to implement all aspects of Open Science within our community. While openNP will help advance toward open data, it is tighly bound with all aspect of Open Science. Tackling the need for a wider policy in open access, research evaluation, outreach and communication on Open Science should be made a priority. To succeed, these policy need to be implemented in a consistent manner in all European institutes, making NuPECC the ideal place to coordinate future and ongoing national actions through the preparation of an Open Science road map.

6 Endorsement

V. Alcindor IJCLab

M. Al-Turany GSI/FAIR

H. Alvarez Universidad Santiago de Compostela

S. Bianco INFN-INF

S. Bottoni INFN-Milano

E. Clement GANIL

J. J. Gomez Camacho Universidad de Sevilla, CSIC

P. Greenlees JYFL

C. Hornung GSI/FAIR

M. Jouvin IJCLab

Y. Leifels GSI/FAIR

A. Lemasson GANIL

T. Marchi INFN-LNL

A. Matta LPC Caen

D. Menasce INFN

A. K. Mistry GSI/FAIR

P. Morfouace CEA-DAM

R. Nania INFN Bologna

L. Patrizii INFN

C. Scheidenberger GSI/FAIR

O. Stezowski IP2I