TWG9 : Open Science and Data

Convenors: Antoine Lemasson (GANIL)
NuPECC Liaisons: Marek Lewitowicz
WG members:
Hector Alvarez-Pol (USC)
Stefano Bianco (INFN Frascati)
Vivian Dimitriou (IAEA)
Xavier Espinal (CERN)
Michel Jouvin (IJCLab)
Antoine Lemasson (GANIL)
Marin Garcia, Ana Maria (GSI)
Adrien Matta (LPC Caen)
Caterina Michelagnoli (ILL)
Andrew Mistry (GSI/FAIR)
Panu Rahkila (JYFL)
Manuela Rodriguez-Gallardo (Sevilla)
Olivier Stezowski (IP2I)
Enrico Vigezzi (INFN Milano)

Date : 26/01/2024

Notes :

1) Highlight Heterogenous aspects challenge
2) VA- EURO-LABS / STRONG In Box ? Or already in TWG6
3) JENAA / Other Initiatives

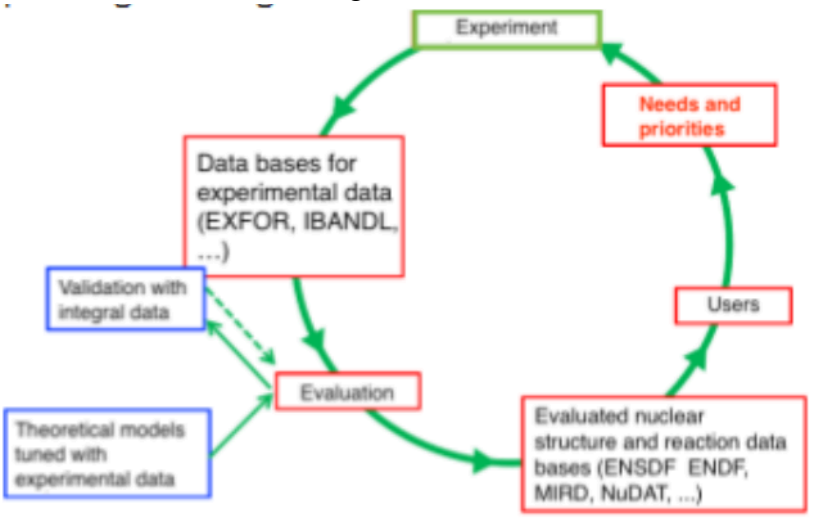
Bilateral Meetings :

1) TWG8-comp:

   Moving Part on ESCAPE to TWG6 / Include some of their text (?)
   Reduce ML/AI paragraph /
   Check Overlap on Chapt4 Software
   -> Discussions Andrew and Johan to finalize

2) TWG 7 :
   Figure XX DLC -> Incorporate in the Section 5 TWG9 (Ask Vivian).

Then refer in TWG7 to the figure



3) TWG 6 :

   Moving Part on ESCAPE to TWG6 / Include some of their text (?)

4) TWG10 :
   Carreer Advancement : Adding sentence/para on “A competitive environment for nuclear talent” => Introduction
   -> Adding Citizen science to ALICE BOX
5) TWG 4 :

6) TWG2 :

   Mention the that the field is at the forefront of open access (Introduction)
   -> Modification on the ALICE Box.

List of changes :

Introduction

Within the nuclear physics community and beyond, it is important to recognise the transformative potential of Open Science, and reap the benefits it can bring to future advancements and growth within the field. Open Science at its core promotes the open exchange of knowledge, enables transparency, breaks down barriers, enhances collaboration, and actively demonstrates sustainability, by making research outputs openly accessible, reproducible and reusable.

The fundamental principles of Open Science are commonly stated as the right of access to the outputs of research with as few barriers as possible. These outputs can include scientific articles, research data, software, and infrastructure. Key to this endeavour are: the implementation of reformed research evaluation criteria with a lower dependence on journal-related bibliometrics; education of researchers aimed at providing the necessary skills to practise Open Science; and the importance of significant contributions from the general public (citizen science). These principles have been adopted in the Horizon Europe funding programme [1] as well as in national Open Science plans so far published. Additional critical issues that must be factored in are the different legislation on copyrights among EU member states and the use of proprietary-owned applications and data repositories.

Embracing Open Science in general can result in several benefits:

- Scientific knowledge advancement through the dissemination of research outputs. Weakening the barriers of access to these outputs accelerates the pace of knowledge transfer and allows researchers to build on each other's work (and their own, in future).
- Encourages collaboration both within the community and cross-disciplinary with other fields. Through the sharing of outputs, and being able to reuse these (e.g. software) can lead to improvements in quality, and offer alternate solutions to problems.
- Enhances transparency of research practices, promoting trust between the community members, on a global research level, and to the general public.
- Strengthens and promotes innovations and technology transfer with industrial partners.
- Attracting future researchers: advertising the research outputs as open demonstrates the field as forward thinking and inclusive. The collaborative nature of Open Science can appeal to the next generation, and attract a diverse talent pool to support the future of the community.
- Sustainability is improved from Open Science: rather than repeating the same work tasks, resource sharing can eliminate waste while preserving scientific competitiveness.
- Research assessment reform: offers an alternative to the traditional, outdated metrics.

The purpose of integrating Open Science within the NuPECC Long-range plan is to outline the importance of embracing Open Science principles within the nuclear physics community, and the methodology for implementation.

This chapter discusses the benefits and application of Open Science within the community, and explores the current and future perspectives for the community. This is divided into the several "pillars" of Open Science, namely: Open Science developments, Open Access publications, Open Data and lifecycle, Open Software and workflows, Infrastructures for Open Science, and Nuclear data evaluation.

1) Open Science Development

Open science policies

Open science policies form the foundation stone of open science practices, playing a vital role to highlight the benefits of open science, raise and increase awareness of ongoing initiatives, and lay out a series of best practices. These are present on an international level (e.g. UNESCO recommendation on Open Science - https://doi.org/10.54677/MNMH8546), national policies (examples given below), or on an institutional level, and may be focussed on one particular aspect of Open Science. Within the nuclear physics community, these policies should be highlighted and promoted at different levels.

A comprehensive overview of Open Science policies in seven European countries can be found in [4]. These policies typically vary in scope and implementation, but have the same underlying message of strong support for open science practices. In particular emphasising open access to research results, data, software, and workflows, while highlighting the benefits to researchers and beyond. Additional insights can be gained from recent statements of intent and plans across Europe : Second French plan for Open Science (2021-2024)" [5] ; The Italian "Piano nazionale scienza aperta (2022)" [6]; The Spanish “National Strategy for Open Science (Estrategia Nacional de Ciencia Abierta, 2023-2027)” [7]; The German research foundation (DFG) [8]; and the Finnish Declaration for Open Science and Research [9]. These examples highlight the increasing recognition and significance that Open Science has gained within the community.

While Inter/national policies form the basis for promoting and highlighting open science, policies on the institutional level play a crucial role in defining the exact open science procedures and techniques that researchers should use in their projects. These can include both policies that state mandatory aspects, as well as broader representative guidelines which include specific examples. In addition, institutions may choose to publish several complementary policies focussing on a particular aspect of Open Science. For example, a set of policies and guidelines on research data management which couple data management aspects with open science initiatives.

Evaluation in the open science and publication era

The rise of open science has led to emerging ways of evaluating research and publications, which are more transparent and inclusive. Traditionally, the impact factor of a published article was used as an indicator for the “quality” of research. However, this is deemed to be an insufficient metric, as it does not assess the impact on society nor the quality of the underlying research. Open Science enables alternative methods of evaluation, such as open peer review and external collaborative evaluation. As an example, assessing the level of ‘FAIRness’ of a published dataset using metrics such as F-UJI score and using this to optimise the strategies and infrastructure for publishing open data and software code. These new methods will aim to provide an enhanced evaluation of research, while promoting open science values.

Promotion and recognition of Open Science and data activities

To develop a widespread culture of openness in research, it is important to establish a framework that promotes the benefits of Open Science, rewards researchers that follow its principles and widely communicates these benefits.
Researchers and communities should be encouraged to adopt Open Science practices. As examples, the following strategies could be employed within the NuPECC community:

- Generating award schemes and providing visibility to researchers and institutions that make contributions towards Open Science endeavours.
- Offering training and resources to researchers to develop practices within the community, on an institutional, national and international level (in line with the European Open Science Cloud (EOSC))
- Developing collaborative platforms and tools to enable researchers to work together and provide contact points and networks.
- Implementing Open Science policies on an institutional and national level as a guide for best practices and provide a marker for support.
- Offering public outreach and incorporating Open Science into education to raise awareness and promote understanding of Open Science practices.
- Supporting the Open Science clusters such as ESCAPE and PANOSC and future activities to ensure that the goals of Open Science are met and obtain the necessary resources for implementation within NuPECC.

Coordinating effort within the community and across the domains

A wide variety of national and international Open Science projects and initiatives are ongoing aiming at supporting Open Science practices by offering infrastructure, resources and support to the research community. Presently, the nuclear physics community has a limited implication in these initiatives. Another important aspect resides in utilising existing infrastructure such as Open Science platforms and repositories, ensuring sustainability and identifying common standards and best practices among the projects. The EOSC plays a key role towards this, through coordination and alignment, and will continue to develop these actions. The current scenario in Europe in the quest for sustainable and efficient Scientific Computing, Open Science and Open Data are boosting collaborations to investigate and test approaches towards common models and tools. The strong implication of the stakeholder on the nuclear physics community in the present and future projects represents a major asset to accelerate adoption and implementation of open science practices. An example of this, the ESCAPE project is described in box 1.

Box 1 : The ESCAPE Open Collaboration

The European Science Clusters approach is a good example of an institutional effort to connect scientific communities by providing services for interdisciplinary research. In particular the ESCAPE project (https://projectescape.eu/) with the recently announced long-term ESCAPE Open Collaboration Agreement. This collaboration has been signed by the Directors of Landmarks Research Infrastructures, such as the European Organization for Nuclear Research (CERN), the Cherenkov Telescope Array Observatory (CTAO), the KM3NeT Neutrino Telescope Research Infrastructure (KM3NeT), the European Gravitational-Wave Observatory (EGO-Virgo), the European Southern Observatory (ESO), the European Solar Telescope (EST), the Facility for Antiproton and Ion Research (FAIR), the Joint Institute for VLBI-ERIC (JIV-ERIC) and the Square Kilometer Array Observatory (SKAO).

This agreement reflects the strong support of diverse scientific communities to pursue joint R&D programs in computing with strong focus on commonalities in computing models and tools to promote economies of scale and provide data FAIRness to experiments and laboratories.

Encouraging Open Access Publication

Open Access (OA) publications are crucial for the active, efficient, fair and sustainable dissemination of academic research. Several models of OA publication are shown in Fig.1 : diamond, green, gold or hybrid gold. In diamond OA, the journal publishes the article and makes it available with no fees involved for neither author nor reader, being the publications costs covered previously by institutional support. The "Action Plan for Diamond OA " launched by Science Europe is one of the drivers of this, proposing to align and develop common resources to enable diamond OA publishing.

- Diamond OA should be recognised as the pinnacle of good practice in publishing, and be set as a long-term target across the nuclear physics community. Diamond OA should be highly supported, and resources for community based diamond access journals is essential. At this time, only a few diamond journals exist, and the incentives for researchers to publish in such journals remains limited due to e.g. low impact factors.
- As a short term plan, researchers should be encouraged to publish in green OA when possible.

The scientific community should be aware of the importance of reforming intellectual property and copyright law, maintain communication with institutions and organisations active in the reform projects at both national and European level, and possibly participate.

In summary, open access journals should continue to be supported and resources dedicated to support diamond open access journals. Researchers who publish open access should be incentivised to do so.

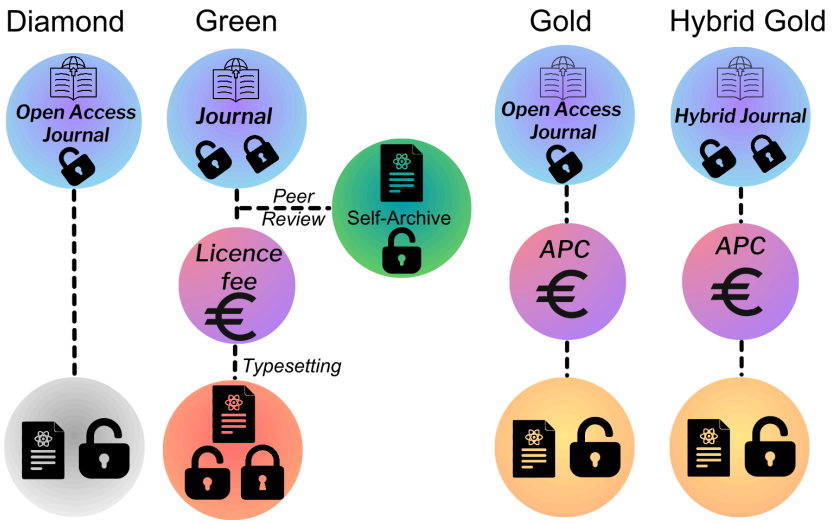


Fig. 1 Depiction of the different categories of open access publication. See text for further details. Diamond OA: content is freely available with no charges to authors or readers. Green OA: allows a version of the work to be deposited in an open access repository (self archiving), the publisher's version may be behind a paywall. Gold OA: content is made available freely, an article processing charge (APC) must be paid by authors or institutions. Hybrid Gold: a traditional subscription based journal offers the option for open access subject to an APC.

Recommendations :

- Advocate for the creation and adoption of open science policies and guidelines within individual institutes.
- Allocate sufficient resources to ensure successful implementation of these policies.
- Encourage joint initiative across scientific domains.
- The community should promote and commit to reforming research assessment through collaborative efforts and shared ethical principles.
- Encourage the development and support of Diamond OA journals of the field.

Box 2: Open Science in practice at ALICE

CERN committed to an Open Science policy as a key to maximize the global impact of research conducted at the facility (https://openscience.cern). Among the different actions, CERN established an Open Data Policy, to be followed within a consistent approach by all LHC experiments. This was endorsed by the ALICE Collaboration in November 2020. The policy commits to publicly releasing level three scientific data: namely, the input to most physics studies together with the corresponding Monte Carlo, as well as the software and documentation needed to use the data. The aim is to start data releases within 5 years of the conclusion of the run period, while the full datasets would be made available at the end of the collaboration. CERN has setup a common Open Data Portal (https://opendata.cern.ch) where collaborations can upload their data samples. For example, ALICE uploaded 5%(7%) of the 2010 Pb-Pb (pp) event summary data totalling 6.5 TB. An analysis demonstrator is also available to execute analysis processes directly through the CERN Open Data Portal.

ALICE aims to make data from the first two runs publicly Open Data from 2024+ (105 TB/year), using a new data format for resource optimization. Data corresponding to the third LHC Run will be made Open from 2030+ (2 Pb/year). Dedicated staff and resources have been allocated to realise this.

The international masterclass program is a success example of the Open Data policy. Each year high school students are invited to participate into the international masterclass program, where for one day students become particle physicists by analysing real data from an experiment at the CERN's LHC.

2) Towards Open Data Life Cycle in Nuclear Physics

The nuclear data evaluation community has developed over several decades a very efficient and robust evaluation pipeline as highlighted in section 5 of this document. This pipeline allows for the production and maintenance of a variety of high quality curated nuclear data repositories used in many societal applications. Inspired by this approach and its success, the nuclear physics community aspires to reach a similar situation when it comes to research dataset repositories.

The concrete implementation of FAIR (Findability, Accessibility, Interoperability and Reusability principles [3] to the large variety of research datasets produced by the diverse nuclear physics community is in itself a challenge. To achieve this goal, it is important to perceive the research data life cycle (DLC) as a production process that demands a commitment to ensuring its quality (See Fig. 2).

The DLC should commence upon submission of the experimental proposal and conclude upon the publication of the final physical observables, with a proper data storage for reuse and revision and a proper policy for unprofitable data removal. At every step along the cycle, collecting comprehensive metadata is crucial to ensure that FAIR principles are implemented. Efforts in standardising metadata is required to ensure reproducibility and interoperability of datasets.

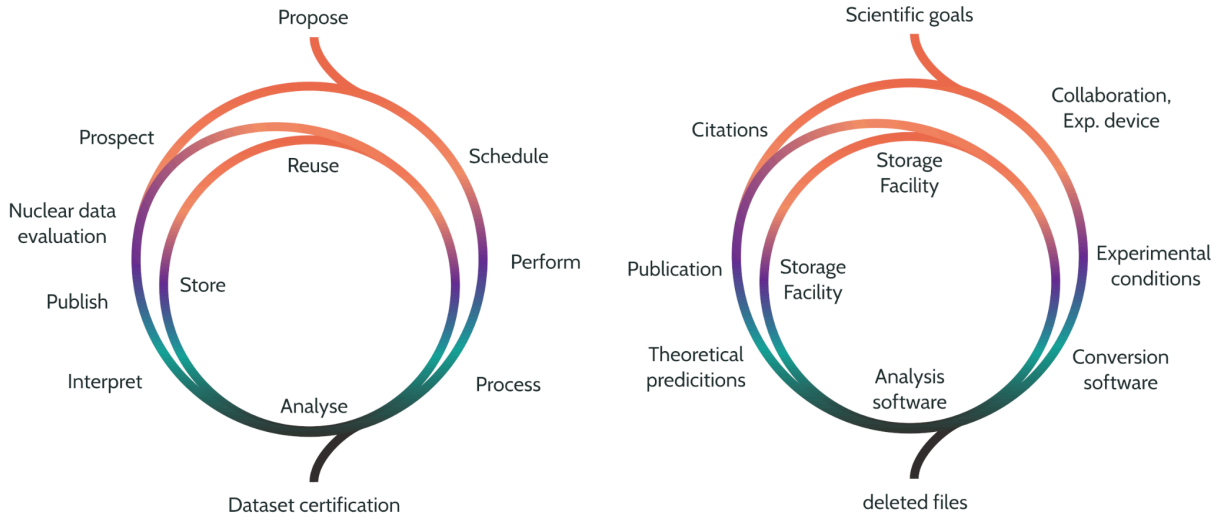


Fig. 2 : (left) Schematic view of the Data life cycle in nuclear physics and (right) example of associated metadata that could be accumulated during the lifespan of data.

The nuclear physics community faces a few specific challenges which make solutions developed by connected fields not always adapted. In particular, the heterogeneous sizes of the collaboration, ranging from individuals to hundreds of collaborators (e.g. AGATA), the size of the detection setup, from single electronic channel up to complex multidetector system (e.g. R3B-GLAD-NEULAND), and the various lifetime of an experimental setup from single experiment up to several decades of operation (e.g. INDRA). As a result it is both unrealistic and impractical to expect uniform technical implementations such as a common Data Management Plan, common file formats or a common repository across the diverse nuclear physics community. However, the definition of standard requirements, such as the existence of a Data Management Plan, documented file formats or the usage of a repository represents an achievable goal for the community. Such standards should be easily scalable depending on the collaboration size, meaning little to no effort for simple and small experiments. The standard is bound to evolve with practices and the incorporation of new technologies, and a revised standard should be produced on a planned schedule.

The formalisation of such a production process will require coordination within all actors among the chain. Physicists, IT departments, facilities, data evaluators and funding agencies should be involved in the determination of DLC requirements demonstrating the quality of the conducted research. The development of a common open standardised ontologies that is frequently used by researchers working in the field for the data and metadata description is essential to allow access and sharing of information across the fields.

To solve the complex problem of standardisation other communities have developed a consortium approach. A good example is the ISO C++ committee overseeing improvement of the language through a three-stage pipeline involving different working groups. A similar approach could be developed for the research data community wherein domain-specific groups (e.g., beam production, detectors, analysis, etc.) could propose new standardised metadata fields to be included in the collection pipeline in a first stage. In a second stage, the metadata collection implementation would be reviewed, to guarantee the proper process and tools exist in the real world. Finally, a wording and consistency stage would take place to enrich the common open vocabulary in a consistent way. The creation of such a metadata collection pipeline would set, and update, standard practices.

Moving towards a functional DLC, a new community of research data specialists should emerge and form the backbone of the thematic working groups of the consortium.

Research data stewardship will be overseen by data officers. The data officer tasks will be handled by members of the collaborations or research groups (e.g. physicist). His/Her role will consist in defining the collaboration data policy (e.g. DMP) and ensure that the DMP is properly implemented at every step of the DLC, making sure the appropriate tool and infrastructure exist, and that research datasets are readily available.

A second role to emerge is data curators, with the task to ensure that the metadata schemes are interoperable between data sets and the technical solutions follow the community standards. Typically working at data producing or processing institutions, these curators will also play an essential role in proposing new data fields to the consortium, as well as educating the community to renewed standards.

During the “propose and schedule” phase of an experiment, an initial metadata pool is built, including scientific motivation, partial authorship, and experimental details. This information will help facilities and collaboration to further optimise experimental conditions. In addition, clear policies, using DMP, are integrated into the metadata scheme. Ideally, more advanced information related to the experimental designs (mechanical, acquisition and designs, etc …) should be deposited. These metadata should be published on identified catalogues supported by the community to ensure the findability of the datasets.

When the experiment is running, auxiliary data, describing the experimental conditions, will be accumulated. Those auxiliary data take various forms today, from physical notebooks to their electronic counterparts. An essential program needs to be run within the community to create new tools adapted to the reality of this difficult data collection. Interoperability between the machine and the detector is for instance a friction point in schemes deployed today. Auxiliary data taken in an archivable form, with the ability to mix automated and human input, is therefore extremely important to be able to analyse the data in new contexts.

The processing step consists in converting the raw-data in a more efficient format. Most data acquisition systems produce binary files in native format, with a premium on software efficiency and simplicity. However, this type of format is usually poorly performing size and reading speed wise. A simple conversion from a native DAQ format to a readout-optimised, higher level, file format (e.g. root tree, hdf5) should be provided. Such software should not alter the data in any form (calibration, selection,...) but limit re-organisation and optimisation. The software developed for this process phase will be integrated into the metadata pool of the experiment.

Data analysis requires both analysis and simulation software. An essential change in paradigm is the acknowledgment that software is an integral part of the dataset and should be treated as such. An overlooked issue today is the preservation of analysed and simulated files during the analysis process. Most current DMP in the field focus only on the raw data with the idea that any other file is in principle reproducible. However, a significant part of the manpower effort of an experiment is assigned to the data analysis process between the experiment and the publication of physical results. It is therefore essential that these datasets are included in the DMPs. All software used for analysis should be open source, documented, preserved, and referenced in the metadata. Ideal cases would include the production of running images of the software and its environment (e.g. container, virtual machine), for long term use. The full panel of technical possibility will be discussed and in the subsequent section 3 on software development and section 4 on shared computing infrastructures.

When the different files produced during this process are stored and curated to form a consistent dataset, the metadata scheme should be enriched with information on storing facilities, file format, and embargo information. This will allow proper access, preservation and curation of the data. At this stage a first selection of the file relevant for future use could be made. This process of data file decimation will contribute to sustainability of the storage

infrastructure with careful and time sensitive suppression of files no longer relevant to the data set while keeping trace of their past experience. A concrete example could be the choice to store untreated data only in a space efficient file format and unstage the original DAQ native data file at the end of the initial analysis. In a similar fashion, untreated data files could be removed after a defined period post publication to retain only refined data. Each of these options should be described in the DMPs.

Long term data preservation represents a challenge for the scientific communities with quickly evolving environments and large size datasets. Accumulating all data produced over a long period is neither a sustainable goal, nor a pertinent use of the community and society resources. DMP and catalogues should be the tools of choice for the community to forecast and monitor storage use for research datasets and help design a sustainable and ecology conscious storage infrastructure in line with our open science objectives.

Box 3: At the forefront of FAIR and Open data at the Institute Laue Langevin

The Institute Laue Langevin was at the forefront of data management by publishing a "Scientific Data Policy" in November 2011. The text came into force in October 2012. Following the publication of the policy, the ILL created an interdisciplinary working group, the DPP (Data Protection and Processing) to drive the development of the software tools needed to put the policy into practice, with a focus on usability (especially during experiments) and security. The data portal (https://data.ill.eu) was launched in 2014. DOI persistent identifier allow readers to obtain more information about the referenced experiments, access the ILL Data Portal and even request access to the experimental team if the data are not yet publicly available. The DPP continuously upgrades the data policy so as to reflect the evolution of the Data Management tools available at the ILL. Further, under the PANOSC (Photon and Neutron Open Science Cloud) scientific cluster, the VISA platform was developed to provide an integrated environment for data analysis (remote desktops, jupyter notebook) to the scientific community. It has a strong impact on capacity and productivity of scientists, by facilitating the access to data and analysis programs as well as exchanges.

Publication or citation of the data should be included in the metadata scheme, providing clear metrics for the usefulness of the data. Through dataset citation in publication using DOI or other machine readable identifier, the metadata scheme will be enriched with this information.

All the produced metadata should be machine readable, allowing interaction between different catalogues and services, whether at the community-wide or facility-specific level. The richness of the accumulated metadata will provide unprecedented understanding of the dataset, its scope and limitations, and will help with "open review" process to increase the scientific quality, nuclear data curation for societal use, and re-use of datasets in new and unforeseen contexts.

At the final step of the scientific production, published results are reviewed and curated by the nuclear data community, thus disseminating the incremental and reliable knowledge for societal benefit. The long tradition of a worldwide open collaboration offering the net output of our scientific community to society has a profound impact on the world we live in. From nuclear energy to medical use and going through space exploration, nuclear data play a major role in modern society. Our rich and fruitful community produces a vast number of results, and the curation process of the nuclear data community faces new challenges, as discussed in section 5 and the report from TWG7. By producing a catalogue of research datasets to the highest standard, the nuclear physics community will grant access to all necessary information for this review process to take place in the best and most efficient way, increasing our field impact on society.

The availability of open research dataset and associated analysis and simulation tools will also play a major role in training the next generation of professionals on modern data analysis techniques and simulations. With lectures and schools based on FAIR experimental data, a wide range of students will benefit from the community effort, helping boost the field and its reach within society.

To reach this ambitious goal the community needs to establish new standard practices. This includes a new collaborative framework, allowing scientists to take stewardship of their data's future. A set of new tools, particularly a comprehensive community-wide data catalogue, and a methodical approach to aggregating auxiliary data, must be developed.

To this end, we recommend to coordinate this effort through the creation of a consortium of the different actors, committed to designing, maintaining and updating, a “real-world proven” metadata scheme for the community at large.

Recommendations:

- Strongly support the application of the FAIR Principles: encourage training and investment in human resources for data management (data officers, data curators, …) at the various levels (institutions, labs, collaborations) to effectively advance the open data practices.
- Support the creation of coordination bodies to pursue standardisation of the DLC to ensure interoperability. Work on tools and guidelines development towards researchers and collaboration to help real-world implementations.
- Engage active collaboration with other communities (e.g. ESCAPE for HEP/COSMO/ASTRO and PANOSC for photons and neutrons).

3) Open software and analysis workflows

In the last two decades, the community has undergone a drastic change in research software development practices. The progressive percolation of industry standards within the research community (in particular, the use of tools improving collaborative development and code quality) led to the emergence of software specialists along with dedicated infrastructure such as code repositories. The overall quality of developed software increased while manpower dedicated to software development is now used much more efficiently.

The community has undergone a transformation, progressing from the development of a constellation of small, usually short-lived software packages, to the establishment of larger collaborative software packages, maintained over several years by an active group. This led to a redistribution of development manpower towards the enhancement of new features and the improvement of the scientific quality of the software.

This shift allowed the community to maintain, and sometimes reduce, the time gap between data taking and publication, while the complexity and data volume of experiment increased with the widespread use of triggerless and/or digital electronics, and multidetector systems.

This rich background and quickly evolving environment push the community to lead the front in modular software packages and innovate in the domain. It has then become possible to run heterogeneous acquisition systems allowing a smooth hardware evolution in the community. On the analysis and simulation side, nptool package allows for fast and efficient mixing of simulation and analysis of multidetector systems. It brought modern tools to a large variety of smaller experimental collaborations such as MUST2, SHARC, FATIMA, ISS at a low developing cost. FairRoot on the opposite side of the spectrum provided a solid framework for large scale, fixed setup experiments such as CBM and R3B, and offered a variety of ready to use tools for high performance computing that lead to the O2 collaboration for the ALICE experiment at LHC.

As software packages become more widely adopted, associated collaborations need to be formally built. Those collaborations are today very informal in most cases, with little recognition from their host institution. Two situations typically arise: either the code is limited to a single collaboration, and managed as a sub-task of the formal collaboration or the code

is not tied to any collaboration and no formal collaboration is recognised, and therefore no resources are allocated. Institutions should take ownership of the software production on par with detector manufacturing, and manage them as formal and recognised collaborations, effectively promoting the effort made by developers to adopt good practices in software development.

With such formal collaboration, the challenging question of the end of life of software could be managed properly. This involves the careful consideration of options such as maintenance, evolution, or replacement by a new project based on more efficient technologies or even complete retirement in case the application is not needed anymore. For the latter, the collaboration should archive and document the code appropriately to guarantee the community is not losing knowledge in the process.

Software, understood as an interface between developers and hardware, is produced in a given language and programming paradigm, may depend on running environments and third-party libraries and renders some overall performance. All these elements constitute an essential context which should be documented in order to ensure usability, maintenance and preservation.

In the case of hardware specific code whenever using FPGA, GPU and co-processor or quantum computers, a tension between the performance and the long term support is also a challenge. All these promising tools require maintenance of legacy hardware or virtual emulator environments to keep the data reproducible. FPGA are critical components of the data acquisition chain, dealing with both data readout, treatment and decision. In most cases today, firmware used in embedded electronics are fully closed and unavailable to the community. Finally, quantum computing presents a great potential for our field, particularly for theoretical prediction. Early works on the subject have already started, however the field is still new and no standard practices exist yet.

The Nuclear Physics landscape is intrinsically extremely rich encompassing a wide variety of physics cases, numerous theoretical models spanning large energy scales, many laboratories and collaborations of various sizes, many detectors assembled in dedicated and often in short lived configurations. As a consequence, over several decades, countless software projects have been realised for different purposes. While the outcome of developed software has been largely published in papers and many results are available in well organised, evaluated, databases for experimental results (see section 5), the software themselves were not systematically published and referenced. As a result, many software and their related data are not usable any more, if not outright lost. The consequences of such status can be different depending on the goal of the application and on the availability of similar operational codes. With the trend towards open data, software and science, a particular attention should be dedicated to that part of the Nuclear Physics heritage. Currently operational applications should operate their migration if not already started.

While still quite fragmented, the Nuclear Physics landscape has evolved during the last decades, notably thanks to projects requiring important resources (large research infrastructures developing high intensity or radioactive beams, complex travelling arrays, highly computing demanding applications). Following the example of large projects such as Geant4 or ROOT, some practices have been developed towards collaborative, distributed open software: version control systems (such as git for the most recent one) are now pillars of many collaborations (for instance ALICE, FAIRROOT, …) Release versions and publications (including in dedicated repositories such as zenodo, licences) of well documented codes tend to be more and more general. As in the case of data, software can be subject to embargo periods. This might be particularly true for complex theoretical codes since the software itself is the source of scientific progress, contrary to for instance data processing code for which the potential of discovery is expected to be inside the data stream. While embargo periods are understandable, there are clear benefits to fully open

any codes since this way can also be associated with clear evaluation processes and measurable merits. Thanks to modern tools such as github or gitlab, some collaborations have included in their software production pipelines, systematic continuous integration and deployment approaches. This however is probably not as spread as the use of version control systems. It should be noticed of course that testing reproducibility of codes is a common practice even if it is not based on modern systematic ways.

With open software, any researcher has the potential to check any result: however, this is true only if one has access to the necessary data and meta-data (see section 2). The processing pipeline may in many cases not be as simple as running a single application. As a matter of fact, this requires not only that the source code should be open but also a solid description of the workflow that has been set to obtain a result. Only very few collaborations have started to investigate this issue in Nuclear Physics. The systematic adoption of machine readable workflow description should be made a priority to insure the reproducibility of research results for exemple by the use of containerized applications. Such usages seem marginal in the NP community and more effort should be dedicated to this kind of technology. Ultimately, this path should lead to full environments to run reproducible data analysis on shared computing platforms (See section 4).

Reproducibility of research results is a pillar in open science and this requires first to have open software to fully understand how a particular outcome (experimental or theoretical) has been obtained. Versioning tools provide a detailed tracking of changes made in a given software package and automatically generate unique revision identifiers that can be tied to a software output as a metadata. As more and more scientific results require complex processing pipelines, descriptions of the processing workflow should also be documented.

Containers or virtual machines offering portability and relative long term preservations are likely to be heavily deployed. This allows efficient systematic software quality checks, continuous integration (CI) and deployment (CD). With the wide adoption of institutional software repositories, CI and CD are already becoming mainstream practices within the community. To accelerate the adoption of these practices in all collaborations, the definition of standards practises at the institutional level is the best course of action.

These practices are also mandatory with the rise of AI and Machine Learning (AI/ML) approaches. For instance, Neural Networks do require training phases which are based on input datasets and rely on hyperparameters to be tuned. With this new, additional, link to data, the processing workflow becomes even more complex and should be managed properly with dedicated tools. AI/ML is still in its early age and the field is rapidly evolving. The coming decade will be decisive in shaping common practices and tools and the community needs to invest in training to keep up with those evolutions. (See TWG8-section XX)

Recommendations :

- Encourage formal software collaborations to improve structuration, oversight, and enable institutional recognition through awards and career advancement.
- Invest in software development methodology training: versioning, collaborative development, CI/CD, AI/ML, workflow management
- Formalise the primordial role of software in the reproducibility of the scientific results through systematic software publication and software session in workshops and conferences.

4) Infrastructures for and effective open science

In the nuclear physics community, the research activities often lead to the generation and the treatment of large volumes of data. These are typically impractical to simply download and process on single machines. Therefore, it is necessary to develop an ecosystem of infrastructure that allows the end user access to data, software, and computing resources; follow the analysis workflow; and visualise the outputs in a meaningful way. Strong synergies with other domains of physical science like particle physics and astrophysics, where significant steps have already been achieved, will strongly benefit the nuclear physics community.

Authorization, Authentication and Identity Management (AAI)

The goal is to provide a recognisable infrastructure with a framework of common tools to enable the Nuclear Physics community, experimentalists and theoreticians, to exploit distributed computing resources and enable data management capabilities. A first fundamental step to achieve this is to agree on a common layer of trust. Allowing to seamlessly integrate identities of the users and experiment experts with the computing resources and the experiment workflows. The AAI frameworks should provide the users with a unique entry point and view to perform their work: data analysis, software development, experiment data management. This can be achieved by adopting technology standards promoted in other disciplines and resource providers. In addition, the chosen federated identity infrastructure should ensure the connectivity to other identity providers (IdP) in the community, to enable access to computing resources (cpu and storage), data/metadata catalogues or infrastructure providers. The chosen federated identity infrastructure should allow the experiments, laboratories, projects and the collaborations to manage user memberships.

Data Storage Infrastructure and Interconnectivity

The community would largely benefit in agreeing and adopting a common data model, an agreed set of tools to establish a common and coherent data storage and data management infrastructure for the Research Infrastructures (RI) in the NP community. This would largely facilitate the implementation, operation and manipulation of the experimental data by the RIs, and potentially have a positive impact by improving efficiency and costs for the experiments computing responsibles and for the sites providing and running resources for one (or many) of the RIs.

This data storage infrastructure is commonly referred to as the Data Lake and includes a common set of tools for: Data Management, Data Transfer, Data Access and a common AAI framework. The Data Lake exposes the dataset of a scientific community as a single repository, and eases the implementation of data life-cycles, policies and access rules to cater with data FAIR-ness. By providing a common interface and standard protocols to expand data processing capabilities by facilitating access to the data: from user laptops to lab’s batch systems or HPC/clouds. A conceptual scheme of the infrastructure is shown in Fig.3.

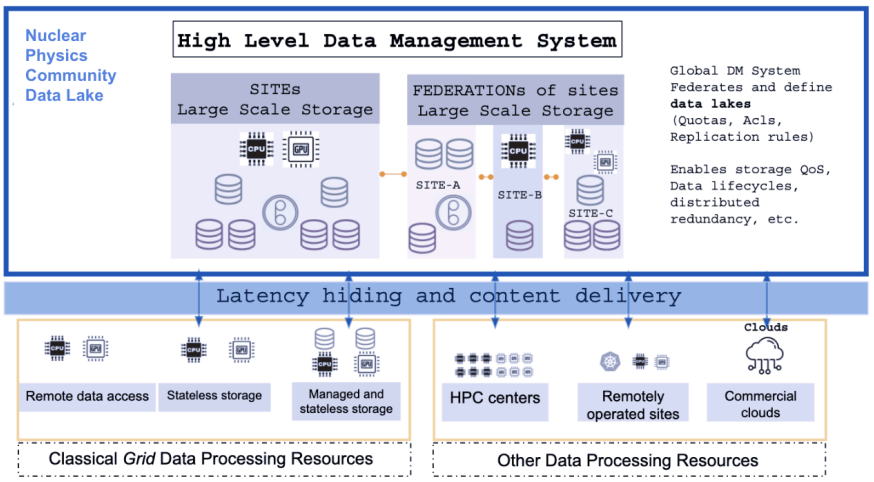


Fig 3. Data Infrastructure (Data Lake) conceptual view, adapted from ESCAPE Data Infrastructure for Open Science report https://arxiv.org/abs/2202.01828v1. A distributed data repository is perceived as a single entity, and is able to deliver the data products to an heterogeneous set of facilities for data processing and data analysis purposes. The ecosystem is formed by labs and sites providing resources to the Nuclear Physics experiments but is also opened to embrace punctual contributions from external providers (eg. clouds and HPCs).

There are several fundamental aspects to consider inherent to the data infrastructure :

- The data infrastructure should ensure the accommodation of full life-cycles of the data: from raw data recording to data distribution for user access (See Section2). Also covering long term needs in terms of Data Management Plans, Data Preservation and Analysis Preservation (ie. Rucio is the data management system, widely used in LHC and adopted by ESCAPE).
- A Data Distribution and File Transfer service should be selected to guarantee a high-level transport layer with the required protocols and interfaces with the storage systems in the data infrastructure. This service acts as the “middleware” to provide third party copy transfers without proxy-ing the data (ie. as FTS has been used on ESCAPE).
- The data infrastructure should be flexible and able to be used in a very simple manner for end users hiding all the complexity. It should also cater for high-level data management experts needs: access control, data replication, and pre-placement, implement life-cycle.
- The data infrastructure should be able to deliver data and provide access from a varied type of processing facilities: large batch systems at the experiment’s facilities and labs, computing clouds (private and commercial) and potentially to HPC centres. The mechanism to optimise remote data processing is to leverage content delivery networks and latency hiding mechanisms (data caching services) interconnected with the Data Lake.

Analysis Platforms

Several nomenclatures are given to the new trends in performing (visual or interactive) data analysis: Analysis Facilities, Analysis Platforms or Virtual Research Environments. The scope of these frameworks is to facilitate the development of end-to-end physics workflows, providing researchers with access to an infrastructure and to the digital content necessary to perform a scientific analysis and preserve the result in compliance with FAIR principles. This is a rapidly evolving field as the scientists' methods to perform their analysis are changing. These (usually) containerised environments are able to hide the infrastructure’s complexity

from the user and provide: 1) Seamless access to the experiment data by having browsable catalogues and easy (clickable) download and upload options to/from the Data Lake, 2) Access to the latest experiment software (mounted in the framework), 3) access to external repositories for user codes.

The community should follow up trends of new analysis tools and methods used by the new generations of scientists, and actively participate in the activities and the projects developing prototypes for the new generation frameworks for user data analysis.

Analysis Preservation and reproducibility

The awareness of “Analysis Preservation" and the need to move towards consistent reproducibility should be enforced. The required tools should ideally be embedded and facilitated by the Analysis framework itself, providing researchers an easy way to package and store the analysis for the future and for sharing purposes. It is important to highlight that this does not apply only to experimental data, but also to the experiment software, user code and the framework to run the analysis in a stand-alone way (OS, software packages, etc.). Such tools exist and some Funding Agencies, experiments and communities start to enforce that publications comply with the Analysis Preservation in view of scientific computing FAIR-ness and open data and open science.

The reproducibility has to be extended to any theoretical calculation that is published. So, the suggestion must be that all publications should include the detailed information on the inputs (potentials, integration parameters, etc) necessary to reproduce the calculations.

Recommendations :

- Investigate the feasibility and interest in the community to have a common distributed data infrastructure (Data Lake) for the Research Infrastructures (RI) in the NP community. To facilitate Data Management, Data Processing and the implementation of data FAIR-ness policies.
- Strong suggestion to adopt technology standards being promoted also in other disciplines and resource providers for: Identity Management, Data Management, Data Analysis and Reproducibility/Preservation. To search for commonalities with the potential to bring economies of scale.
- The various stakeholders (researchers, collaborations, research infrastructures and institutes) should actively contribute to joint initiatives and technical developments in coordination with scientific communities for the deployment of scalable infrastructures adapted to its specific needs.

5) Nuclear Data (Databases) and Evaluation

Nuclear data refers to curated data describing the properties of atomic nuclei, nuclear decay, cross sections for nuclear reactions, and other quantities relevant to nuclear science and engineering. These data are typically obtained through experimental measurements performed in small- and large-scale nuclear physics facilities. Apart from their direct use in a host of applications including nuclear energy, nuclear waste management, nuclear safety, nonproliferation, nuclear medicine, environmental control, and cultural heritage, that have been discussed in TWG 7, nuclear data are also indispensable for basic nuclear physics research. As scientists plan future experimental activities that may lead to new discoveries, they seek to improve their knowledge by interrogating the databases. Theoretical developments also rely heavily on the availability of reliable and up-to-date nuclear data.

The research data life cycle can only be complete if the measured data are curated and incorporated into the nuclear data databases and libraries. Only then are the data readily available for use by the broader user community, be it in basic sciences or applications. The main advantages of nuclear data are that they are evaluated, verified, and validated before being disseminated as data files, data libraries or databases. The evaluation procedure involves assessment of the experimental methods and uncertainty budgets provided by the experimentalists, treating discrepant data or outliers, and recommending best values for a nuclear parameter using a statistical method or a fitted model with associated uncertainties. The evaluated data are prepared in formatted data files with well-defined formats that form part of a well-organised and documented database or data library. These libraries are put to various tests to verify the correctness of the formatted data files but also to validate the evaluated data against benchmark integral experiments, before finally being released online. Nuclear data or curated data are essentially the harbingers of **FAIR** research data as they meet all these criteria through their very definition and construction.

The collection/compilation, evaluation, verification, and validation of nuclear data are arduous tasks that rely heavily on contributions from experts in both the basic and applied research communities. International organisations, such as the International Atomic Energy Agency (IAEA) and the Nuclear Energy Agency (OECD/NEA), facilitate the curation of nuclear data by bringing together knowledge, expertise and infrastructures from all over the world, thus enhancing progress while reducing the financial burden on national authorities and funding agencies. The major general purpose nuclear databases that form the foundations for all other derivative databases, interfaces, and simulation codes are maintained by international networks and collaborations as follows:

- Exchange Format (EXFOR): compilation of all published measured reaction cross sections, angular distributions, excitation functions, thick-target yields, and fission-related data (cross sections, fission yields, etc.) hosted at the IAEA (https://www-nds.iaea.org/exfor). EXFOR is maintained by the international network of Nuclear Reaction Data Centers (NRDC) which comprises 13 Data Centers from 8 countries and 2 international organisations (IAEA, NEA Data Bank) under the auspices of the IAEA.
- Evaluated Nuclear Structure Data File (ENSDF): a unique compilation and evaluation of measured nuclear structure and decay properties across the nuclear chart that is hosted at the National Nuclear Data Center, BNL (https://www.nndc.bnl.gov/ensdf). ENSDF is maintained by the international network of Nuclear Structure and Decay Data evaluators (NSDD) under the auspices of the IAEA. The NSDD network comprises 17 Data Centers from the USA, Europe, Russia, India, China, Japan and Australia including one international organisation (IAEA).
- Evaluated Nuclear Data File (ENDF): evaluated nuclear reaction data originally developed for neutrons and later extended to charged-particle/photon transport codes. Different evaluated files are produced and maintained at a national or broader collaborative level (CENDL, China; ENDF/B, USA; JEFF, Europe; JENDL, Japan; INDEN, IAEA; ROSFOND, Russia). The Joint Evaluated Fission and Fusion File (JEFF) (https://www.oecd-nea.org/jcms/pl_20182/jeff) is produced by a collaborative effort of about 100 scientists from 21 countries including 3 international organisations (EU, IAEA, NEA Data Bank) and is coordinated by OECD/NEA Data Bank.

Other horizontal evaluation efforts with predominantly European involvement include the Decay Data Evaluation Project -coordinated by LNHB-CEA- which is responsible for providing recommended decay data for about 230 radionuclides for metrology applications (https://www.lnhb.fr/ddep_wg/).

While there has been significant investment in the development of small- and large-scale experimental facilities and in measurements by European member states and EU projects, nuclear data curation activities in Europe have been largely underfunded in the past decades. The main source of funding of nuclear data has been the EURATOM projects such

as CHANDA, SANDA, ANDES, ARIEL, and ERINDA. These initiatives not only offer financial support but also enable European nuclear data groups to access facilities, share infrastructures and expertise, and play a pivotal role in educating and training the next generation of nuclear data experts. However, this form of funding falls short of what is required for a sustainable European nuclear data curation effort commensurate with the region's rate of production of experimental data.

The absence of a coordinated nuclear data effort in Europe in conjunction with limited funding has resulted in a shortage of expert data evaluators with serious repercussions for the international databases. Over the past two decades, Europe's contribution to ENSDF has dwindled to less than 20% of the overall. This figure is notably low considering the substantial volume of experimental data generated by the European nuclear physics facilities. Combined with a global decline in the ENSDF effort worldwide, the cumulative ENSDF evaluation effort is no longer adequate to keep the database up to date within a 10-year cycle. To maintain a regular 10-year update cycle, a minimum of 12 full-time-evaluators is required, whereas the current global effort amounts to just 5. What this situation implies is that the new, accurate, and precise experimental data generated by state-of-the-art European nuclear physics facilities will not be promptly incorporated into the databases, thus delaying their utilization in various applications.

Another consequence of the underfunding of nuclear data evaluation in Europe is the loss of expertise in the nuclear data sector. The absence of career paths and promising prospects in the field of nuclear data is discouraging young nuclear physicists and engineers from entering the field, while simultaneously forcing early-career nuclear data scientists to seek opportunities elsewhere. When coupled with the retirement of senior nuclear data experts who may not have immediate replacement, it becomes clear that Europe is at a risk of losing its expertise in nuclear data evaluation, and hence its capacity to maintain its databases and data libraries up-to-date.

Following a dedicated Consultants' Meeting organised by IAEA and NuPECC in 2023, ann IAEA report on the needs for a "Comprehensive European Plan to acquire and curate nuclear data" (IAEA Technical Report INDC(NDS)-0875) concluded that: **The main challenge facing the European nuclear research and applications community is establishing a sustainable source of funding of measurements and nuclear data evaluation, and ensuring well-defined career paths in nuclear data to maintain and enhance the available expertise.**

Concomitantly, the principles of **OPEN** data, as well as **FAIR** data, are paving the way for new opportunities in nuclear data curation. Both **OPEN** and **FAIR** data promote openness and accessibility in experimental data, while **FAIR** data also focus on making experimental data more discoverable, interoperable and reusable. Apart from enhancing the development of data-driven technologies that have already been discussed in the previous sections [links to previous subsections], these principles also encourage experimental groups to make available, in an accessible, interoperable and reusable way, a much larger amount of measured data, including raw data, analysed but unpublished data, as well as analysis software and metadata describing the measurements, detector calibration, target preparation, and the details of the data analysis. This tremendous wealth of experimental information will lead to a new approach to data curation based on new modern tools for retrieving experimental data, not limited to the traditional extraction from tables and figures in published articles. The availability of these experimental details in a direct, findable, and easily-retrievable form would facilitate the evaluation process, and conversely, feedback from the evaluation and validation process could be used to reanalyse the raw data without necessitating a new measurement. Creating adequate and open repositories to store this excessive volume of data, along with the needed metadata to search and find is a necessary condition for this data to be findable and usable in data curation. Another benefit to nuclear

data curation from findable, interoperable, and reusable experimental data is that the entire analysis procedure could be verified by independent experimentalists thus potentially uncovering bugs or errors prior to publication or evaluation. The push to develop new data-driven technologies will also affect the data curation technologies introducing automation where possible and Artificial Intelligence/Machine Learning (AI/ML) techniques to compensate for scarce or missing experimental data.

To be able to absorb all these developments into nuclear data curation effectively and seamlessly, to meet the demands of both basic and applied research, the current nuclear data databases need to be modernized and the nuclear data experts need to collaborate with data scientists, database programmers, and experts in AI/ML. Training the next generation of nuclear data experts to develop and implement these technologies and approaches is of paramount importance for the future of nuclear data curation.

Recommendations :

To rise to the challenges in nuclear data curation and seize the emerging opportunities, it is recommended that :

- the European nuclear physics research and applications communities combine forces to establish a comprehensive European nuclear data program with well-defined priorities defined by stakeholders and sustainable funding to secure nuclear data career paths.
- dedicated efforts to train the next generation of nuclear data experts in data evaluation and the use of AI/ML methods and modern data-driven technologies are supported.
- the cooperation between nuclear data curators, data scientists, database programmers and AI/ML experts, and international organisations (IAEA, OECD/NEA) is strengthened.

REFERENCES

1. Horizon Europe funding programme Open science policy
2. Towards a reform of the research assessment system
3. Wilkinson et *al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
4. *Open science–related policies in Europe https://doi.org/10.1093/scipol/scac082*
5. Second French plan for Open science (2021-2024)
6. The Italian "Piano nazionale scienza aperta (2022)"
7. National Strategy for Open Science (Estrategia Nacional de Ciencia Abierta, 2023-2027)
8. Open Science as Part of Research Culture. Positioning of the German Research Foundation
9. Policies of open science and research in Finland

GLOSSARY

AI - Artificial Intelligence
CERN - Conseil Européen pour la Recherche Nucléaire
CC-BY - Creative Commons By
ENSDF - Evaluated Nuclear Structure Data File
EOSC - European Open Science Cloud

ESCAPE - European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures
EXFOR - Exchange Format (Experimental nuclear reaction database)
FAIR (Data) - Findable Accessible Interoperable Reusable (Data)
FAIR - (Facility) - Facility for Antiproton and Ion Research
FPGA - Field Programmable Gate Arrays
F-UJI
GANIL - Grand Accélérateur National d'Ions Lourds
GPU - Graphics Processing Unit
GSI - Gesellschaft für Schwerionenforschung mbH
HPC - High-performance computing
IAEA - International Atomic Energy Agency
ILL - Institut Laue Langevin
JEFF - Joint Evaluated Fission and Fusion file
ML - Machine Learning
OECD/NEA - Nuclear Energy Agency of the Organisation for Economic Co-operation and Development
OIDC - OpenID Connect
PANOSC - Photon and Neutron Open Science Cloud

Recommendations TWG9 - Open Science and Data

Open science and FAIR data offer an important opportunity for the nuclear physics community to uphold the highest research standards and enhance its societal impact, by treating the scientific production process as a strategic asset. The nuclear physics community should vigorously endorse and adopt open science practices, be actively involved in shaping the necessary policies, and lead the way in their implementation.

The richness and diversity of the community presents major challenges in applying common FAIR data principles across the variety of data (theoretical, experimental, software, and scientific publication) and research ecosystems (small to large experiments, short to long live collaborations, multi-site travelling detector facilities). To reach this goal of establishing a scalable standard widely adopted by the community, a substantial investment in resources for its implementation and training of the next generation of data specialists is essential.

Progress in fundamental nuclear physics research, which encompasses nuclear structure, dynamics, and nuclear astrophysics, relies on the availability of quality-assured nuclear data characterizing basic nuclear structure and reaction properties. Additionally, advancements in nuclear technologies and their various applications are greatly influenced by the prompt integration of cutting-edge scientific knowledge into these databases. To achieve both objectives, it is essential to recognize that the expertise, research facilities, and best practices required for nuclear data development extend beyond the capabilities of any single field or application. Therefore, a coordinated and collaborative effort at both the national and international levels accompanied by significant investment is imperative.

Open Science development in Nuclear Physics

- The creation and adoption of open science policies and guidelines addressing the pillars of open science such as open data, open source software, open hardware, as well as promoting best practice within individual institutes and research infrastructures should be strongly encouraged.

- Funding for open access publishing, providing strategies and infrastructure for data and software publication, and training of researchers in open science practices is essential.

Towards open Data Life Cycle in Nuclear Physics

- Strongly support the application of the FAIR principles and common scientific computing frameworks : encourage training and investment in human resources for data management (data officers, data curators,…) at the various levels (institutions, labs, collaborations) to effectively advance the open data practices.

- The creation of coordination bodies to pursue standardization of the Data Life Cycle to ensure data FAIRness should be supported. The development of guidelines and

tools for researchers and collaborations should be engaged towards an effective implementation of these practices.

Open software and analysis workflows

The primordial role of software in the reproducibility of scientific results should be formalized through systematic software publications and software and computing sessions in workshops and conferences. Software collaborations should be encouraged to improve structuration, oversight, and enable institutional recognition through awards and career advancement.

Infrastructures for and effective open science

A strong investment in federated infrastructures should be pursued. Relying on technology standards being promoted in other disciplines for data cataloging and data management, data access, data preservation, user analysis and reproducibility. The various stakeholders (researchers, collaborations, research infrastructures, institutes) should actively contribute in joint activities with the scientific community and follow the technical developments in the field. The implication of the nuclear physics community in existing cross-domain European initiatives should be strengthened and future activities within JENAA should be initiated with the goal to implement scalable infrastructures, favoring economies of scale, and adapted to the nuclear physics community specific and diverse needs.

Nuclear Data (Databases) and Evaluation

- Combine forces of the European nuclear physics research and applications communities to establish a comprehensive European nuclear data program with well-defined priorities defined by stakeholders and sustainable funding to fulfill the needs in nuclear structure and dynamics, astrophysics and applications.

- Support dedicated efforts to train the next generation of nuclear data experts in data evaluation and the use of AI/ML methods and modern data-driven technologies.

- Strengthen the cooperation between nuclear data curators, data scientists, database programmers and AI/ML experts, and international organizations (IAEA, OECD/NEA).