Enhancing Data Management in Nuclear Physics through a F.A.I.R.-Compliant Metadata Schema

I. Knežević[1,*] and *A.K.* Mistry[1,2,**]

[1]GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstraße 1, 64291 Darmstadt, Germany
[2]Facility for Antiproton and Ion Research in Europe (FAIR GmbH), Planckstraße 1, 64291 Darmstadt, Germany

Abstract. Within the field of experimental nuclear physics, steps are underway to further advance good Research Data Management practices. At the heart of this is the goal of generating well-described F.A.I.R.-level experimental data and ensuring compliance with these principles. Here, we describe the landscape of Research Data Management in the nuclear physics domain, with a focus on supporting future actions with a F.A.I.R.-compliant metadata schema prototype for the community. In this work, we present a prototype metadata schema and platform (NPPilot), developed to support F.A.I.R.-compliant metadata generation and management workflows in experimental nuclear physics. The schema's nodal, multi-layered structure supports metadata enrichment across domains, and a user-friendly generator facilitates schema customisation and export in various formats. We detail the development process, key features, and potential applications. We also describe how NPPilot will serve as the foundation for further developments through the NAPMIX project, which aims to expand its scope to a broader range of research domains. An outlook is given toward broader expansion of the schema to cross-domain physics fields.

1 Introduction

The increasing complexity and volume of data in experimental nuclear physics necessitate robust Research Data Management (RDM) solutions. The Findable, Accessible, Interoperable, and Reusable (F.A.I.R.) principles [1] emphasise metadata – the data that describes data – as a crucial component. While metadata describing the research process is often produced within experimental workflows, it lacks a systematic structure that aligns with this process and ensures interoperability across institutions and disciplines. Rather than focusing solely on metadata creation, the goal is to develop a structured and scalable metadata model that enables better organization, reuse, and integration into existing research infrastructures.

To enable this, common standards must be developed while avoiding the creation of a plethora of similar schemas that would ultimately hinder the overall goal. Therefore, it is essential to engage the experimental nuclear physics community as a whole in driving the development of these domain-specific schemas and leverage existing standards as far as possible. While the implementation of F.A.I.R. principles in the research process is desired, it's equally important to balance these requirements with researchers' practical needs. Rather than treating F.A.I.R. principles as abstract concepts, they should be integrated in ways that enhance research workflows. The Nuclear Physics European Collaboration Committee (NuPECC) Long Range Plan [2] has identified this balanced approach as a key priority. This aligns with ongoing Open Science initiatives such as ESCAPE [3] and PUNCH4NFDI [4], where the research community actively participates in developing practical solutions.

Such an endeavor requires not only developing a community-approved schema, but also promoting its adoption during data collection and publication. This requires user-friendly infrastructure and simplified data input processes. These improvements increase efficiency, conserve resources, and encourage researchers to publish well-documented datasets. Success depends on maintaining researcher involvement in developing both the schema and its user interfaces. Mechanisms for data access (e.g., Authentication and Authorization Infrastructure, AAI) remains the responsibility of institutions, which must provide the necessary digital infrastructure to manage and store the underlying data while ensuring metadata is properly linked and discoverable.

*e-mail: i.knezevic@gsi.de
**e-mail: a.k.mistry@gsi.de

After consultation with the community, including laboratories involved in the Euro Labs project [5] and the wider NuPECC community, a gap was identified for an experimental nuclear physics schema that a) did not yet exist and b) could not be bridged by any existing schema.

We first discuss some of the challenges in RDM and how they can be solved through metadata generation, before describing the design and implementation of a prototype structure. To address this, we developed a prototype metadata schema and generator – NPPilot [6] – which we introduce in this work. The NAPMIX project will further expand its scope and capabilities across related domains.

2 Long-standing Challenges in Research Data Management

2.1 Towards F.A.I.R. data

The F.A.I.R. data principles have become a foundation stone for modern research data management, providing a framework to guide the creation of valuable and usable resources for the long term. Established in 2016, awareness among researchers continues to rise. However, achieving 'F.A.I.R.-level data' is a shared effort that requires researchers, supported by appropriate infrastructures, to adopt good research data management practices. While the principles provide guidance, their implementation depends on the right combination of technical systems, institutional and inter/national support, and community engagement.

One of the primary hurdles is the inconsistent application of these principles across projects and institutions. Without clear guidelines, adequate tools, and a shared understanding of the importance of F.A.I.R. principles, data may remain siloed, poorly described, and difficult to reuse. This is often compounded by a deficiency of time and resources available to spend on implementation. These gaps hinder the full realisation of F.A.I.R. principles, and limit the impact of research by making data less discoverable and usable across the broader scientific community and beyond.

2.2 Growing Data Volumes and Unstructured Data

As with other fundamental research fields, nuclear physics is experiencing unprecedented growth in data generation, often termed a "Cambrian explosion of data" [7]. Technological advances and increasingly complex experimental methods have led to an exponential increase in data volume over the past 15 years. Modern experiments can generate up to petabytes of data annually, creating significant challenges for storage, organization, and analysis.

These issues are compounded by the fact that, while raw data generation from core instruments can be well structured on the individual experiment level, there is a continued prevalence of data that lacks common description schemas, such as raw sensor outputs, equipment log files, or free text. Unlike structured data, organised in a database of predefined formats, unstructured data lacks inherent organization, making it challenging to manage and access. Without comprehensive metadata, these datasets can become effectively unusable, hampering research and collaboration.

Implementing robust metadata practices at both project and experiment levels is essential. Well-designed metadata provide the necessary context and organization to ensure that even large, unstructured datasets remain discoverable and valuable for scientific research.

2.3 Reproducibility Crisis

The scientific community continues to face challenges in ensuring the reproducibility of research findings, though the degree of impact varies across disciplines. A survey published in Nature in 2016 revealed that approximately 40 percent of researchers reported being unable to reproduce their own experiments, while over 70 percent had failed to replicate the experiments of others [8]. In the context of experimental nuclear physics, reproducibility can be hindered by incomplete metadata, inconsistent documentation, or the loss of critical experimental parameters.

This issue arises from several factors, including inadequate documentation of experimental procedures, lack of access to original datasets, and insufficient metadata to describe the context and methodology of experiments. Without these critical details, it becomes difficult, if not impossible, to validate or build upon previous work.

Addressing this crisis requires improving data management practices, particularly through the use of comprehensive metadata schemas that ensure all relevant information is properly captured, preserved, and accessible for future research.

The research data management process is often represented as a life-cycle. In the context of achieving F.A.I.R.-compliant data and metadata, if data is reused or forms the basis of a new project, the process may evolve into a series of interconnected lifecycles, as illustrated in Figure 1. Metadata capture progresses through these cycles, with new projects applying metadata from previous ones. Well-described and structured metadata is required at each step in both projects to enable this.

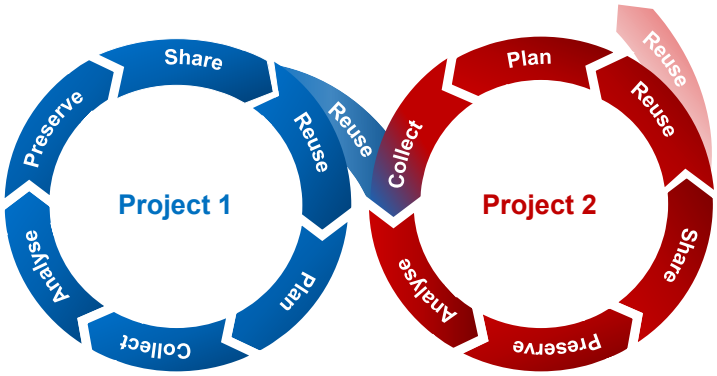

Figure 1. Example of Research Data Management lifecycles. Data generated in Project 1 may be reused as a separate project or interoperated with generated data in Project 2. F.A.I.R.-level data and metadata are required to achieve this in an efficient manner.

2.4 Gaps in awareness and adoption of RDM practices

A significant challenge in achieving effective data management is the varying levels of awareness and understanding of good Research Data Management (RDM) practices among researchers. While advanced technologies and robust metadata schemas can address many technical challenges, their adoption often depends on researchers having the tools and knowledge to implement RDM best practices.

Researchers may lack access to core curricula on RDM principles, such as data organisation, metadata creation, and long-term preservation. This gap can sometimes result in inconsistent file naming conventions, incomplete metadata, and non-standardised workflows, which hinder effective data sharing and reuse.

Without clear incentives or institutional support, researchers may perceive RDM as an added administrative burden, which slows the adoption of F.A.I.R. data practices. This reluctance impacts individual projects and contribute to broader challenges, such as difficulties in reproducing experiments due to missing or incomplete documentation.

Closing this awareness gap requires research institutions, universities, funding agencies, and national bodies to take an active role by providing training, clear guidelines, and practical support (e.g. data stewards) tailored for the researchers needs. Activities in this direction are being actively discussed and implemented – see, e.g., [9, 10]. Building a shared understanding of these practices can help researchers integrate them naturally into their work. Promoting and demonstrating the benefits of adpotion, and incentivising good RDM practices will be crucial.

3 Standard Workflow in Experimental Nuclear Physics

Managing research data effectively requires a systematic approach. This section outlines an example workflow in experimental nuclear physics, based on procedures at the GSI Helmholtz Centre for Heavy Ion Research, and also holds valid for similar facilities. Each workflow stage highlights where metadata is generated. Figure 2 illustrates this workflow, covering all critical steps of a research project, from proposal submission to publication of the scientific outputs.

3.1 Current Workflow Overview

The standard workflow includes the following key stages, including input of metadata describing the experimental data:

1. **Proposal Management System:** The workflow begins with researchers submitting proposal documentation to apply for experimental time using a software tool, such as the General Access Tool to the Experimental facilities of GSI (GATE). This web-based interface is designed to manage applications for both the collaboration and the reviewers. If the proposal application is accepted by the laboratory management, the subsequent phases proceed. This stage marks the first instance of metadata input, such as details about the research team, key objectives, and experimental infrastructure to be used.

2. **Data Management Plan:** If a proposal is approved, researchers prepare a Data Management Plan (DMP) using software such as the Research Data Management Organiser (RDMO) [11]. This interface hosts a DMP

catalogue that can be completed online. The DMP outlines how data will be handled throughout the project and should be revisited and updated periodically throughout the research project. More detailed metadata are collected at this point, including storage and archiving needs, and plans for data publication.

3. **Data Generation:** Raw experimental data is collected through detectors coupled to data acquisition systems(DAQs) for structuring and processing. DAQs typically include timestamping, label, and validation of the incoming data streams. This represents the base metadata layer and includes details such as the data collection infrastructure, conditions, and timing.

4. **Data Storage in a Distributed File System:** Raw data collected during an experiment are stored, for example, to the GSI Lustre distributed file system [12]. At this stage, the metadata includes user-defined file naming conventions, dataset structure, and access control permissions. The metadata helps in identifying, retrieving, and organising datasets efficiently, especially for large-scale projects.

5. **Data Archiving:** In parallel to active storage, data is archived in long-term storage media such as magnetic tapes to ensure preservation. Metadata generated here typically includes archival dates, storage locations, and preservation formats required for ensuring long-term reusability.

6. **Computing Infrastructure and Analysis:** Data can be processed on High Performance Computing (HPC) infrastructure coupled to the shared storage system (see, e.g., [13]). These analyses typically lead to new dataset generation in the form of pre-processed (i.e., semi-derived), and result (tabulated) data. In addition to new dataset metadata, descriptions on the computing environment such as software libraries and workflows employed are also collated.

7. **Code Repository:** Analysis scripts and codes are managed in a version-controlled environment such as Git. These code repositories track changes and dependencies, and enable collaborative development. Metadata here include version history, contributors, licensing, and other software-related metadata (see, e.g., [14]).

8. **Publication:** Final research outputs, such as results and findings are published in scientific journals and repositories. This represents the final of metadata collection in this data lifecycle. Bibliographical metadata such as Persistent Identifiers (PIDs), citations, licenses and links between the research products are collected here.

This workflow serves as a useful guide for determining when and where metadata is generated, helping to identify possible replications, schema requirements, and the ideal stages for metadata capture.

3.2 Metadata and Digital Research Products (DRPs)

The metadata generated at each stage of the experimental workflow forms the foundation of our schema, ensuring traceability and reuse of data. This approach aligns with the concept of Digital Research Products (DRPs) as defined by the PUNCH4NFDI initiative [4]. Rather than storing the data itself, DRPs instead capture metadata about all digital outputs of the research process, such as publications, datasets, software, and workflows and provide structured references to data storage locations. This ensures that data remains findable and accessible without requiring duplication of large-scale datasets, as described in [15].

Building on this concept, we have elevated the idea of DRPs by using them as a common ground between different research areas. Regardless of the specific field – whether nuclear physics, astrophysics, or particle physics – each domain generates similar research outputs, including datasets, software code, journal publications, and workflows. However, metadata alone is insufficient for ensuring full usability of datasets across disciplines. Effective cross-domain data reuse also requires access tools and execution environments that allow researchers to interact with heterogeneous datasets. Without these supporting mechanisms, integrating data across disciplines would impose an excessive burden on both data producers and consumers. Recognizing this, our approach seeks to not only standardize metadata representation but also develop infrastructure that facilitates both metadata interoperability and practical data accessibility across research domains.

The DRPs are categorised into six main types, each representing a key aspect of the research process. These include:

Project

As an extension of the PUNCH4NFDI concept, we introduce the *Project* as a top-level DRP, which serves as an umbrella entity encompassing all other DRPs. The project provides the context, goals, and overarching metadata for all related outputs.

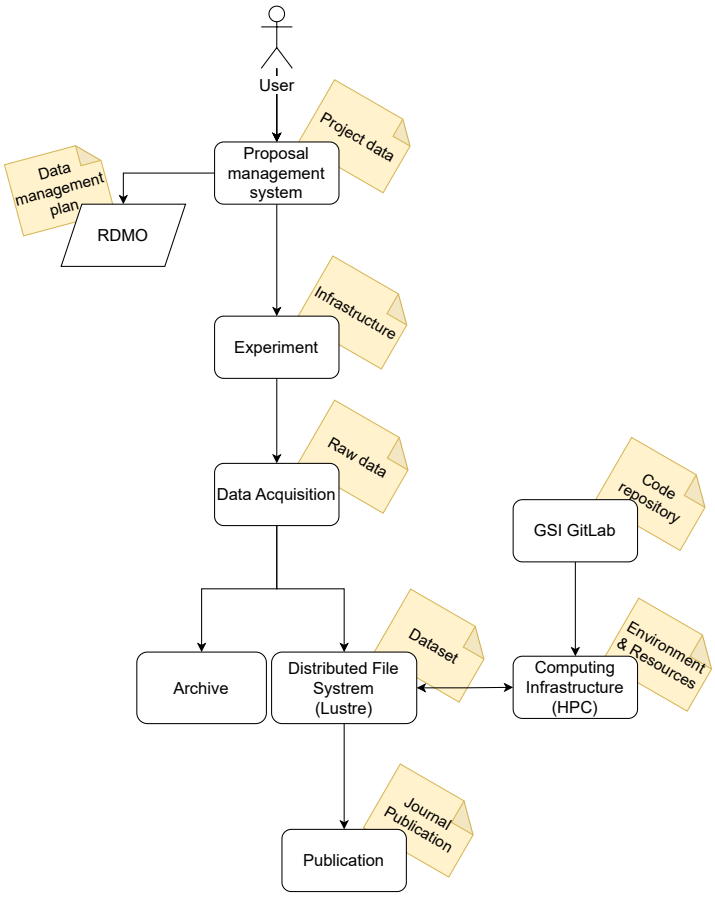

Figure 2. Standard workflow in experimental nuclear physics, highlighting key stages where metadata is generated.

Publication

Publications include journal articles, preprints, and other written outputs. They are identified using persistent identifiers (PIDs) such as Digital Object Identifiers (DOIs) or International Standard Book Numbers (ISBNs).

Software

Software encompasses codebases and computational tools stored in platforms such as Git. These can be linked to general purpose, community supported publication repositories such as Zenodo for release record management, or linked to domain-specific repositories such as CASA [16]. These software DRPs are versioned and often assigned PIDs such as DOIs for long-term traceability.

Dataset

Datasets include raw, pre-processed, calibrated, and result experimental data, and can be published in public repositories such as Zenodo, discipline-specific repositories like HEP-Data, or institutional-level repositories. Each dataset is assigned a PID for unique referencing.

Infrastructure

Infrastructure encompasses metadata records of research facility infrastructures and instruments, such as particle accelerators, ion separators and detection equipment. These records detail hardware and configurations, and can be assigned PIDs and managed using versioning systems.

Workflow

Workflows represent procedural and computational frameworks used during research, including execution environments such as REANA[17], containerised applications, and distributed compute resources. These workflows are referenced through container IDs (e.g., Docker/OCI), and can optionally include build recipes or environment specifications to ensure reproducibility.

3.2.1 The Role of DRPs in Research Transparency and Reproducibility

By organising DRPs with consistent metadata and unique identifiers, researchers ensure that all outputs remain findable, accessible, and reusable. This enhances transparency by providing a complete and structured record of the research process. Furthermore, the alignment of DRPs across different research domains creates opportunities for interdisciplinary collaboration, as the standardised framework simplifies integration and sharing of data and methods.

4 Metadata Stratification

The metadata schema is designed with a stratified, nodal, multi-layered structure to facilitate data sharing, findability, and customisation. This approach organises metadata into three distinct levels: project-specific, discipline-specific, and experiment-specific metadata as illustrated in Figure 3. The stratification ensures that metadata common across projects and disciplines is placed at the top, while discipline-specific and more highly specialised metadata are layered below. To ensure that these levels remain connected and interoperable, structured relationships are maintained between the strata through shared identifiers, linking project-level metadata with relevant discipline-specific and experiment-specific information. This allows metadata elements at different levels to reference one another, while preserving their distinct roles within the schema.

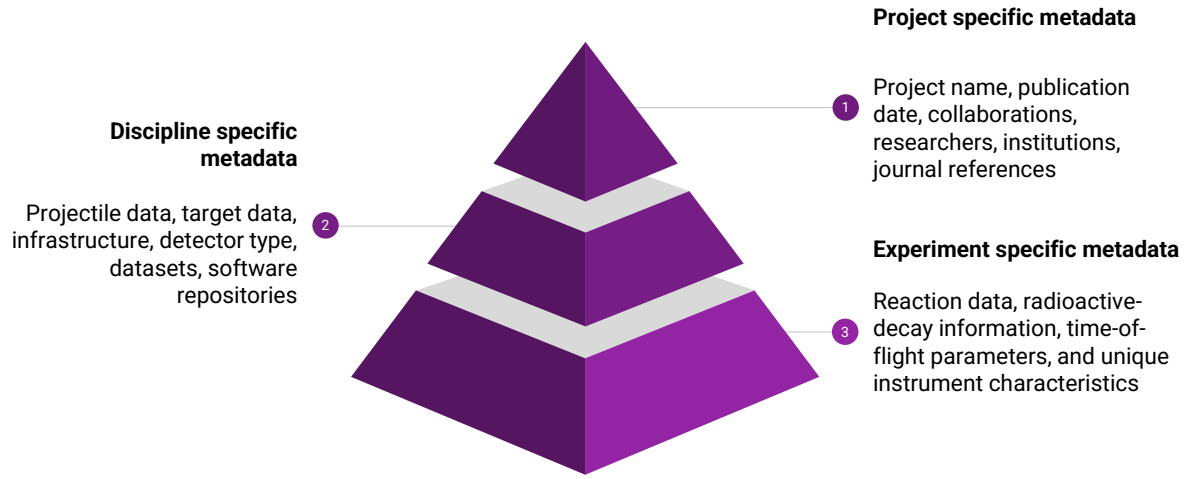

Figure 3. Stratification of metadata into project-specific, discipline-specific, and experiment-specific levels.

4.1 Stratification Levels

Project-Specific Metadata (Top Level)

At the apex of the hierarchy are project-specific metadata, which include broadly applicable and easily shareable information such as the project name, publication date, collaborations, researchers, institutions, and relevant journal references. These metadata fields are widely used in standard, well established schemas such as DataCite and DublinCore.

Discipline-Specific Metadata (Medium Level)

The middle layer contains metadata specific to the discipline. For nuclear physics, this includes details such as projectile and target information, infrastructure, detector type, datasets, and software repositories. While these metadata fields capture the unique requirements of nuclear physics, they maintain relevance across related fields, and can have commonalities with NeXus [18, 19] and OpenPMD [20], for example. Discipline-specific metadata are associated with project metadata through standardised dataset references, while also providing direct key links to experiment-specific metadata, ensuring data integrity across different research contexts.

Experiment-Specific Metadata (Base Level)

The base layer captures metadata unique to specific experiments, such as measured reaction data, radioactive-decay information, time-of-flight parameters, and unique instrument characteristics. These detailed parameters are important for reproducing and analysing experimental results and provide flexibility for tailoring the schema to specific experimental needs.

4.2 Benefits of Stratification

This stratified approach provides several key advantages. By isolating experiment-specific metadata at the base, researchers can customise the schema to meet their needs without affecting higher levels. The clear relationships between layers enhance findability, making it easier to locate and retrieve relevant data. The shared project-specific layer ensures interoperability across disciplines, while the structured organisation facilitates collaboration by enabling differing research domains to understand and effectively use each other's data.
Additionally, the schema supports metadata enrichment by allowing new tables or fields to be added to the experiment-specific cluster. This flexibility ensures that domain-specific needs can be accommodated without altering the core structure. Similarly, other research domains can define their own discipline-specific clusters while leveraging the shared project and dataset layers.

5 The Metadata Schema Design

The metadata schema prototype is implemented as a database structured to organise and manage metadata efficiently. The database is divided into clusters, where each cluster represents one segment equivalent to the Digital Research Product (DRP). Initially, the schema is structured around three primary clusters: *Project*, *Dataset*, and *Experiment*. This clustering approach ensures that metadata is logically grouped based on its role within the research process, and at the same time providing data stratification, from common to more specific.

5.1 Use cases

The prototype structure was based on use cases developed at a small number of large and smaller research institutes representing a diverse selection of accelerator-based experimental techniques and devices: GSI, GANIL – Linear and sychrotron accelerator experiments, HZDR – Pelletron, CNA-Seville – Accelerator Mass Spectrometry (AMS) and tandem accelerator, HUN-REN ATOMKI – cyclotron, and IST-Lisbon – Ion-Beam Analysis (IBA). These use cases had distinct characteristics, while offering some degree of overlapping elements and remaining within the scope of the prototype study.

5.2 Clustering at the Experiment Level

Beyond the primary clusters, the schema takes a more granular approach within the *Experiment* cluster by further subdividing it into specialised sub-clusters. These sub-clusters include *Projectile*, *Target*, *Reaction*, and *Product*, among others. Atomising the database in this way allows metadata to be highly modular and reusable across different use cases and research domains.

For instance, the *Projectile*, *Target*, and *Reaction* sub-clusters within the experimental cluster can be referenced in other experimental contexts, such as IBA or AMS. This design promotes consistency and interoperability, as researchers can share and reuse well-defined metadata components across various fields without duplication or modification.

5.3 Benefits of Clustered Design

The clustered database design offers several key advantages:

- **Modularity:** Sub-clusters within the experimental cluster allow metadata to be reused across different experiments and disciplines.
- **Scalability:** The atomised structure supports the addition of new fields or clusters as research needs evolve.
- **Interoperability:** Shared clusters, such as *Project* and *Dataset*, ensure compatibility across disciplines and research use cases.
- **Flexibility:** Researchers can tailor metadata organisation to the specific requirements of their field while maintaining consistency with the overall schema.

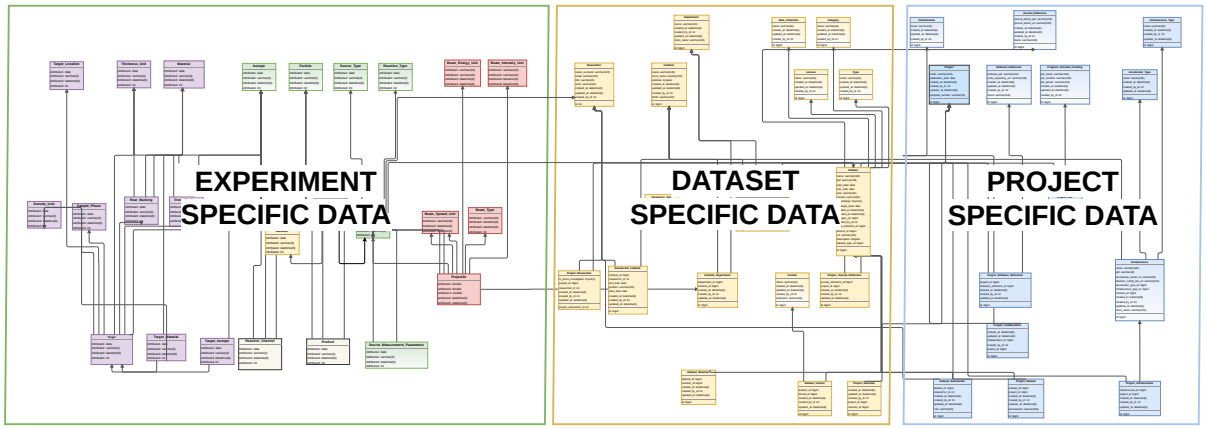

Figure 4. Database design illustrating the separation of *Project*, *Dataset*, and *Experiment* clusters, with further atomisation of the *Experiment* cluster into sub-clusters such as *Reaction*, *Projectile*, and *Target*.

5.4 Application to Diverse Research Fields

The modular design of the schema is particularly advantageous for interdisciplinary research. By organising metadata into reusable sub-clusters, the schema can be applied in diverse experimental contexts. For example, the *Projectile* and *Reaction* sub-clusters can be integrated into workflows for both nuclear physics experiments and Ion Beam Analysis, which ensures consistency and reduces redundancy. Similarly, the flexibility of the schema allows researchers to add domain-specific sub-clusters without affecting the core structure.

6 Project Implementation and Platform Prototype

The prototype structure of NPPilot was implemented using Django, a high-level Python framework, with MySQL as the database backend. This section describes the development process, technical implementation, and key features of the prototype platform.

6.1 Development Process

The development began with a conceptual design phase, during which the necessary attributes, cardinalities, data types, and relationships were defined for the use cases. The initial use cases were selected from two accelerator-based nuclear physics experiments at GSI. A simple form was used to collect experiment-relevant metadata items, in addition to project information. In a second phase, additional input was gathered from other laboratory partners. All use cases were then cross-checked to identify overlaps and commonalities.

This conceptual model served as the foundation for an Entity-Relationship (ER) diagram, which was subsequently used to design the database schema.

Django was selected for its ability to facilitate rapid prototyping. Its model abstraction allowed for straightforward translation of the conceptual schema into database tables, while its migration system automated schema updates. This streamlined approach enabled the team to focus on refining the schema and ensuring its adaptability to different research scenarios.

6.2 Prototype Platform Features

The prototype platform leverages Django's integrated features to manage metadata effectively. Each cluster, such as *Project*, *Dataset*, and *Experiment*, is represented as a dedicated data entry form within the Django admin interface. These forms allow users to input, edit, and manage metadata in a modular and organised way, maintaining relationships between clusters through foreign key fields. For example, the *Dataset* cluster can reference multiple experiments, enabling comprehensive data linking and reuse.

The platform supports complex filtering and search operations, allowing users to analyse interconnected metadata. Researchers can combine filters to trace relationships between projects; for example, identifying datasets shared across experiments or determining which experiments contributed to specific publications. These capabilities enhance both the discoverability and traceability of metadata.

User management is handled by Django's built-in authentication system, which allows for role-based access control and conditional visibility of metadata. For instance, datasets under embargo can be configured to display only a subset of their metadata to external users, while full access is reserved for authorised personnel. This ensures compliance with data sharing policies while respecting data ownership preferences.

6.3 Selection of the environments

The use of Django offers several benefits that have proven critical during the prototype phase. The combination of Django models, migrations, and the admin interface supported a seamless transition from the conceptual design to the prototype platform. Relationships between clusters, maintained directly through the database schema, enable modularity and scalability, making it easy to extend the platform with additional clusters or fields as required. Furthermore, the flexibility of Django ensures that the platform can be customised to meet specific requirements while maintaining consistency with the core schema.
The decision to use a relational database over a NoSQL alternative, such as MongoDB, was driven by several key considerations. While the early adoption of NoSQL alternatives would facilitate the design phase thanks to its adaptability, relational databases excel at managing structured data with well-defined relationships, making them ideal for a schema that relies on interconnected clusters. Additionally, the use of SQL ensures compatibility with a wide range of existing tools and systems, while the built-in support for complex queries facilitates the detailed search and filtering capabilities required for this platform's future development. These advantages align with the goal of creating a scalable and interoperable metadata schema.

7 The Metadata Generator

The Metadata Schema Generator is designed to facilitate the creation and export of metadata. The generator serves as a critical component of the metadata management platform by enabling users to compile comprehensive metadata for their projects, datasets, and experiments.

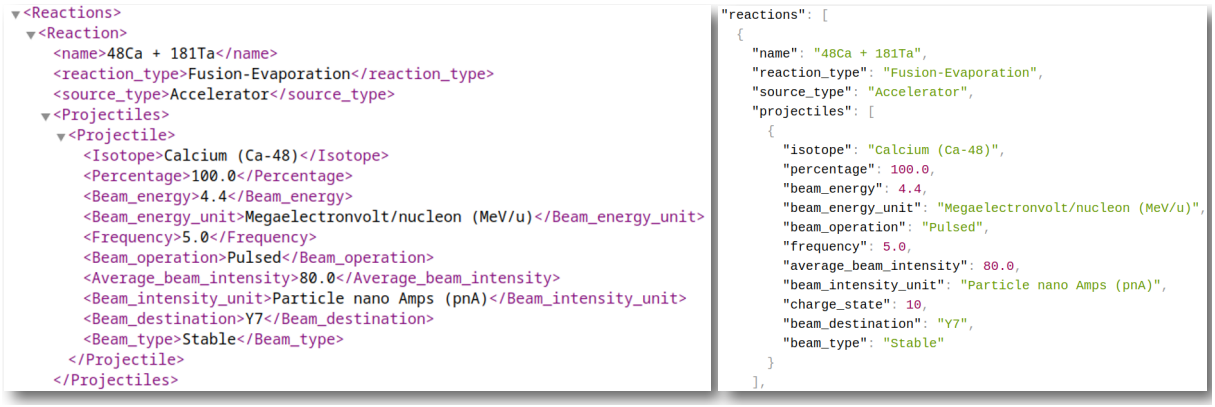

Figure 5. Comparison of JSON and XML export formats, showcasing their consistent nomenclature and structure.

7.1 Functionality and Features

The generator provides an interface for selecting and exporting metadata. From the Projects page, users can choose a specific project and then configure their export options. Supported export formats include JSON and XML, both of which adhere to a consistent nomenclature and structure, ensuring interoperability between systems. Figure 5 illustrates the uniformity between JSON and XML exports.

To accommodate diverse institutional and disciplinary standards, the generator offers two encoding options:

- **GSI encoding:** Designed primarily for the description of GSI-specific research procedures, data structures, and experimental metadata.
- **DataCite encoding:** Aligns with internationally recognized metadata standards for research data, providing a structured format for cross-community metadata exchange.

It is important to note that DataCite encoding captures a subset of the full metadata schema, as its focus is on high-level project metadata and dataset descriptions, rather than detailed experimental procedures. Conversely, the 'GSI' encoding provides a more granular representation of data structures specific to the research environment at GSI, as well as incorporating elements collected from the use cases. While mapping to DataCite serves as a proof-of-principle for interoperability with external metadata standards, not all elements of the GSI schema can be fully mapped to DataCite due to their differing scopes. Future developments will explore expanding these mappings to additional metadata schemas to enhance interoperability while preserving experiment-specific details.

7.2 Data Serialisation and Adaptability

A key feature of the generator is its reliance on serialisation to compile metadata from interconnected clusters, such as *Project*, *Dataset*, and *Experiment*. By serialising each cluster separately, the schema achieves a high degree of modularity and flexibility. This design ensures the preservation of metadata relationships, enabling integration with external systems or standards.

For example, metadata from an experiment can reference its associated datasets or projects, with all relationships maintained in the exported file. This comprehensive packaging of metadata enhances usability and ensures that no critical information is lost during the export process.

7.3 Customisation Capabilities

Recognising the need for adaptability, the generator includes functionality for attribute customisation. Users can modify attribute names to align with specific standards or institutional guidelines by providing key-value pairs of the original and new attribute names. While this method currently requires coding expertise, it offers significant flexibility. Future development plans aim to implement an interactive interface for attribute customisation, making this feature accessible to a broader range of users.

7.4 Future Development

In addition to enhancing customisation features, future updates to the generator will include support for additional export formats and standards, as well as a feature that enables automatic data import, from an existing metadata schema.

8 From prototype to full-scale demonstrator

The development of the metadata schema and accompanying platform represents a significant step toward addressing the challenges of Research Data Management (RDM) in experimental nuclear physics. By aligning with the F.A.I.R. principles, the schema enables enhanced metadata, addressing long-standing issues in data management. NPPilot was a first step, however, several challenges and opportunities remain for advancement.

The development of the NPPilot prototype has acted as an essential springboard to broaden the reach and scope of the concept. To further advance the project, a broader community has been formed, including the original use cases from the prototype development, along with members from the communities of hadron physics, high-energy physics, and astroparticle physics. The community is made up of members who support Open Science initiatives such as the ESCAPE, PUNCH4NFDI, and DAPHNE4NFDI projects. Funding was recently obtained through the OSCARS framework for a dedicated project: the Nuclear, Astro, and Particle Metadata Integration for eXperiments (NAPMIX) [21].

The key goals of NAPMIX are described in this section.

8.1 Schema Expansion

The schema will be expanding to include the other experimental fields listed, as well as further development on the nodes for the prototype use cases. This will act as a driver for interoperability between domains, as well as establishing overlapping criteria. The nodal-based structure of the prototype minimises the work in this direction, and leaves the schema open to further developments beyond the NAPMIX project.

8.2 Standards Compliance and Data Exchange

In addition to the current DataCite mapping of the prototype, the schema will be mapped to other discipline-specific schemas such as NeXus and OpenPMD. These mappings will primarily apply to the deeper, experiment-specific layers of the schema (corresponding to the NAPMIX schema), where detailed metadata related to experimental configurations, detector parameters, and raw data structures must be described. This will broaden its application to a wider range of communities and ensure compliance with a diversity of data repositories.

Planned support for the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is also foreseen, which will simplify metadata synchronisation and harvesting between systems.

8.3 Dedicated API for System Integration

To enhance integration capabilities and facilitate interoperability, a dedicated API will be developed as part of the metadata platform. This API will allow external systems to interact directly with the metadata platform, enabling automated data exchange, synchronisation, and management. In particular, the API will enable the retrieval of metadata in machine-readable formats such as JSON and XML. These formats define how metadata is structured and serialized for transmission, while the metadata schema itself is specified through either the DataCite standard or the NAPMIX schema, which describes entities and their associated properties. Additionally, the API will allow the submission of new metadata entries, while supporting and linking clusters such as Project, Dataset, and Experiment. Customisable querying and filtering features will enable systems to extract only the required metadata. Authentication mechanisms will be implemented to ensure secure access and compliance with data sharing policies. Finally, the API will be designed to minimise integration overhead, making it easier for external institutions and repositories to adopt the platform as part of their workflows.

8.4 Integration of the Metadata Platform into Experimental Workflows

Aligned with the API development, the metadata platform will be embedded in the experimental workflow, linking multiple systems to ensure consistent metadata management throughout the research lifecycle. Figure 6 illustrates how the standard workflow at GSI in experimental nuclear physics, previously depicted in Figure 2, incorporates the metadata platform. Key integration points include:

Proposal Management System: The metadata platform gathers top-level metadata, such as project details, scientists involved, funding information, and data size, through an API connection to the Research Data Management Organiser (RDMO) and the proposal management system.

Experiment Execution: During the experiment, the metadata platform operates in parallel, continuously collecting and structuring metadata related to the experimental process.

Data Storage: Metadata is exported in JSON/XML format and accompanies the experimental data during storage. Storage systems integrate with the metadata platform to ensure consistent metadata management.

Open Data Access: The experimental data, along with its metadata schema, is made accessible via HTTP links utilising XRootD and SciToken-based access.

Publications Repository: The metadata platform forwards the data through the Storage to the MyCORE-MIR[22] Publication Repository, to support long-term data preservation and accessibility. MyCORE-MIR is an open-source framework designed for managing and publishing digital content, offering features such as flexible metadata schema support, hierarchical content organisation, and role-based access control.

8.5 Front-End deployment

A front-end will be coupled to the database to ease the end user interface. This will enable manual input of metadata combined with the API capabilities to auto-fill metadata items, where possible.

8.6 Community Uptake

To ensure widespread adoption of metadata generation in combination with data publication, workshops and training material targeted at researchers will be prepared. This will include hands-on events focused on promoting both the platform and its capabilities, as well as working directly with researchers on metadata for their projects.

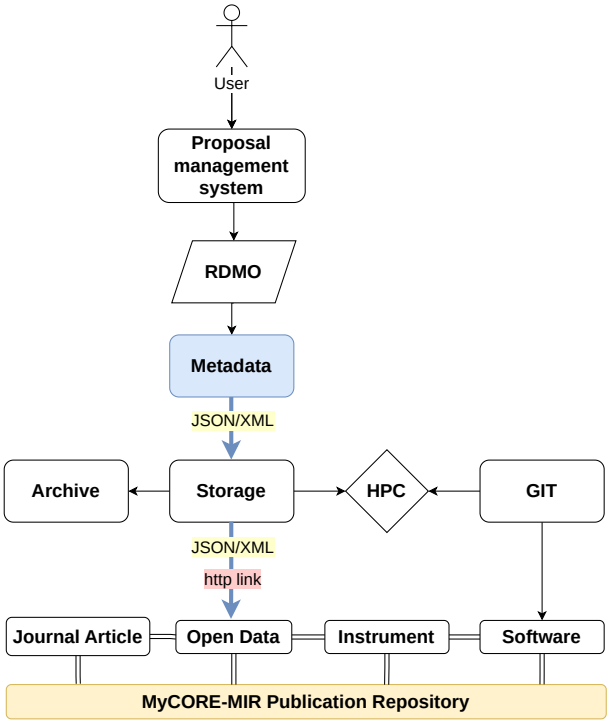

Figure 6. Integration of the metadata platform within the experimental workflow, highlighting connections with RDMO, storage systems, and open data repositories.

9 Conclusion

Research communities are actively working to tackle complex Research Data Management challenges. A key element of these efforts is creating specialized metadata frameworks and user interfaces tailored to specific scientific domains. This project has presented a metadata schema prototype tailored to the nuclear physics community, with a nodal, multi-layered structure that promotes adaptability and cross-domain application. The outcomes of this NPPilot project lay the foundation for improving metadata practices across disciplines, enabling better data sharing, reuse, and collaboration. The prototype allows users to generate structured metadata that describes data associated with experiments. This metadata is stored in standardized serialization formats such as JSON and XML, following both DataCite and GSI-specific encoding to ensure interoperability.

While the initial outcomes are promising, its broader impact has yet to be fully realised. The platform's development will be advanced under the NAPMIX project. Collaborations with a broader range of research infrastructures, alongside Open Science project partners in EOSC and ESCAPE will play a pivotal role in refining the schema and ensuring its broad adoption. Through ongoing efforts, this work aims to establish a robust metadata management platform and associated schema that not only addresses the immediate needs of the nuclear physics community, but also contributes to the broader goals of F.A.I.R.-compliant data benefiting the wider research community in the long term .

The software described in this work is published under an Open Source license and can be found here: https://doi.org/10.5281/zenodo.14770678

We thank GSI and all external collaborators who contributed to this project. In particular, we would like to thank Mohammad Al-Turany, Yvonne Leifels, Radoslaw Karabowicz, Harry Enke, and Gisela Schmidt for careful reading of the manuscript. The authors acknowledge the OSCARS project, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751. A.K.M. acknowledges the EURO-LABS Project, which has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101057511.

References

[1] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, P. Bourne et al., The FAIR Guiding Principles for scientific data management and stewardship, Sci Data **3**, 160018 (2016). 10.1038/sdata.2016.18

[2] NuPECC Long Range Plan 2024, https://nupecc.org/lrp2024/Documents/nupecc_lrp2024.pdf, last accessed: December 11, 2024

[3] The ESCAPE Project, https://projectescape.eu/, last accessed: December 11, 2024

[4] PUNCH4NFDI, Digital Research Products (DRP) - Service Class 4, https://www.punch4nfdi.de/services/service_classes/service_class_4/, last accessed: December 9, 2024

[5] EURO-LABS, EUROpean Laboratories for Accelerator Based Sciences, https://web.infn.it/EURO-LABS/, last accessed: December 11, 2024

[6] I. Knežević, A.K. Mistry, Pilot study: Experimental nuclear physics - dataset metadata generator (2024), pilot implementation of the NPPilot metadata platform for experimental nuclear physics, https://doi.org/10.5281/zenodo.14770678

[7] P. Taylor, Amount of data created, consumed, and stored 2010-2023, with forecasts to 2028, https://www.statista.com/statistics/871513/worldwide-data-created/ (2024), retrieved December 11, 2024

[8] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature **533**, 452 (2016). 10.1038/533452a

[9] GSI Helmholtzzentrum für Schwerionenforschung GmbH, FAIR GmbH, GSI/FAIR Guidelines on Research Data Management (2023), available at: https://doi.org/10.15120/GSI-2023-00916

[10] Helmholtz Metadata Collaboration Training, https://helmholtz-metadaten.de/en/hmc-trainings, last accessed: January 30, 2025

[11] RDMO consortium, Research Data Management Organiser, https://rdmorganiser.github.io/, last accessed: December 10, 2024

[12] V. Penso, M. Dessalvi, F. Uhlig, Lustre - GSI Shared Storage, https://hpc.gsi.de/virgo/user-guide/storage/lustre.html, last accessed: February 10, 2025

[13] V. Penso, M. Dessalvi, F. Uhlig, Virgo - GSI Cluster Computing Infrastructure, https://hpc.gsi.de/virgo/, last accessed: February 10, 2025

[14] The codeMeta project, https://codemeta.github.io/, last accessed: December 10, 2024

[15] T. Schörner, H. Enke, P. Bechtle, The PUNCH4NFDI Science Data Platform (SDP) and Digital Research Products (DRPs), https://doi.org/10.5281/zenodo.8358549 (2023)

[16] The CASA Team, CASA, the Common Astronomy Software Applications for Radio Astronomy, Publications of the Astronomical Society of the Pacific **134**, 114501 (2022). 10.1088/1538-3873/ac9642

[17] REANA: REproducible research data ANAlysis platform, https://reanahub.io/, last accessed: December 16, 2024

[18] NeXus data Format, http://www.nexusformat.org/, last accessed: December 11, 2024

[19] M. Könnecke, F.A. Akeroyd, H.J. Bernstein, A.S. Brewster, S.I. Campbell, B. Clausen, S. Cottrell, J.U. Hoffmann, P.R. Jemian, D. Männicke et al., The NeXus data format, Journal of Applied Crystallography **48**, 301 (2015). 10.1107/S1600576714027575

[20] Open Particle Mesh Database, https://www.openpmd.org/#/start, last accessed: December 16, 2024

[21] NAPMIX collaboration, Nuclear, Astro, and Particle Metadata Integration for eXperiments, https://oscars-project.eu/projects/napmix-nuclear-astro-and-particle-metadata-integration-experiments, last accessed: December 11, 2024

[22] MyCORE-MIR, Publication repository, https://github.com/MyCoRe-Org/mir, last accessed: December 10, 2024