Grant Agreement No: 101057511

EURO-LABS

EUROpean Laboratories for Accelerator Based Science

HORIZON-INFRA-2021-SERV-01-07 Project EURO-LABS

MILESTONE REPORT

[OPENNP CATALOG PERIMETER, ARCHITECTURE, AND STANDARDS]

MILESTONE: MS35

Document identifier:	EURO-LABS-MS35.docx
Due date of milestone:	End of Month 12 (31/08/2023)
Report release date:	24/08/2023
Work package:	WP5: Open, Diverse and Inclusive Science
Lead beneficiary:	CNRS
Document status:	Final

Abstract:

This document constitutes milestone M35 of the EURO-LABS project. It summarizes the conceptual design for a prototype for the openNP catalog allowing Research Infrastructure and researchers to centralize the metadata of their datasets. This will foster improved scientific practices following the FAIR data principles towards effective Open Science in the community. It also provides them with an opportunity to automatically see their datasets to be findable for the rest of the scientific community and its future services.

EURO-LABS Consortium, 2023
For more information on EURO-LABS, its partners and contributors please see https://web.infn.it/EURO-LABS/

The EUROpean Laboratories for Accelerator Based Science (EURO-LABS) project has received funding from the European Union's Horizon Europe Research Infrastructure (RI) services advancing frontier knowledge under Grant Agreement no. 101057511. EURO-LABS began in September 2022 and will run for 4 years.

Delivery Slip

	Name	Partner	Date
Authored by	A. Matta A. Lemasson	CNRS/LPCCaen CNRS/GANIL	23/07/2023
Edited by	A. Matta A. Lemasson C. Hornung A. Mistry	CNRS/LPCCaen CNRS/GANIL GSI GSI	16/08/2023
Reviewed by	A. Lemasson MJG Borge	CNRS/GANIL CSIC	21/08/2023
Approved by	Navin Alahari [Scientific coordinator]	GANIL	24/08/2023

TABLE OF CONTENTS

Executive summary

The aim of EURO-LABS WP5.2 on Open Data is to provide the tools necessary for the communities to share their science products in a harmonised way respecting the FAIR principles, promoting reproducibility of: the results, open science, and maximising cross-fertilisation by reuse of datasets. One of the key components to achieve this goal is the dataset catalog. The present conceptual design report documents the specification of the openNP dataset catalog.

The foreseen implementation for the service is an open catalog where identified users can create new entries and obtain an associated hash, DOI or a similar persistent identifier, making them effectively traceable, findable and citable. An entry could describe a data set as well as a data format, associated software, an experimental configuration, and different types of auxiliary data, such as a data management plan and an archived electronic log book. The service does not aim at storing data, but rather references their existence, relation, and place of storage.

The document describes the expected user experience when using such a tool, as well as the implication for the experimental workflow of the community. A section of the document presents a set of minimal meta-data that needs to be collected and should be used as the starting point for the development/evolution of the service.

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

The proposed use of the readily available Dataverse framework is discussed and justified. This software framework allows the deployment of a meta-data catalog, as well as offering a complete suite of tools to manage, search and interact with them.

1. INTRODUCTION

The Research Infrastructures (RI) in EURO-LABS are engaged in open science practices and FAIR practices of their data, software and science tools to allow reproducibility of science results and the development of open science analyses across domains and infrastructures.

The aim of EURO-LABS WP5.2 on Open science and Data is to provide the tools necessary for the communities to share their science products in a harmonised way respecting the FAIR principles, promoting reproducibility of the results, open science, and maximising cross-fertilisation by re-use of datasets. One of the key components to achieve this goal is the datasets catalog. The present conceptual design report documents the work needed to build a demonstrator for the openNP catalog.

2. GOALS AND OBJECTIVES

A first step toward open science in the community of accelerator based research in nuclear and particle physics within the EURO-LABS Research Infrastructures is related to the openNP dataset catalog. It aims at developing a central service to reference and access existing datasets and associated software including, for instance, experimental and simulated datasets, theoretical calculations and the associated software.

The foreseen implementation for the service is an open catalog where identified users can create new entries and obtain an associated hash, DOI or similar persistent identifier, making them effectively traceable, findable and citable. An entry could describe a data set as well as a data format, associated software, experimental configuration, and different types of auxiliary data, such as data management plan and electronic log books archive.

Rich and machine readable meta-data associated with each entry will allow an automatic discovery of data-sets and services, making the service effectively inter-operable with other data catalogs. This ensures that all ingredients necessary for future exploitation of data are referenced making them inter-operable and reusable with any future services.

Entries will be regrouped into a collection to build a self-consistent data set, associated with an umbrella identifier. A typical collection being an experiment or a detection setup, it could also be associated with a scientific topic.

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

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

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

These different features will serve as the basis for interfacing future services of data lake & high performance shared computing and analysis platforms, however this will require the development of a common authentication and authorisation infrastructure (AAI).

3. CONCEPTUAL DESIGN

A repository is first and foremost a trusted place where the various actors of the research (Research Infrastructure, researchers, engineers, ..) can safely archive their data and software for long-term preservation and retrieval. But to fully support open science, the catalog must provide more services and be integrated in a future virtual analysis environment.

The catalog will be the entry point for any newcomer looking for datasets through a webportal integrated in the EOSC. It should therefore provide a clear interface, with filtering and search options. Moreover, any contribution should be identifiable without ambiguity, thus allowing citation. The datasets and software and analysis environments must be accompanied with a well defined set of metadata to maximize reusability. Standardized data registration procedure and workflow should be available and adapted for the needs of the Research Infrastructure and Research Collaboration.

A schematic representation of the various stakeholders and topics covered with the openNP environment is provided in Figure 1 for the case of an experimental dataset.

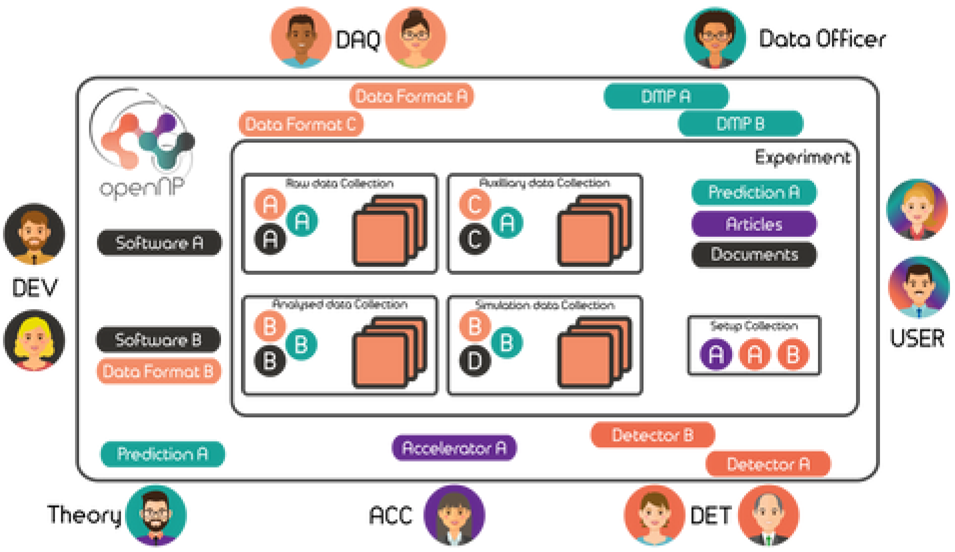

Figure1 : Schematic representation of the openNP catalog

Typical use cases are described below.

1) Experimental dataset produced at a Research Infrastructure :

   The different phases of the dataset cycle in the catalog could be as follows :

   a) Creation of an experiment entry by the Research Infrastructure at the time of scheduling of an experiment. Association of Data Management Plan and Metadata from the experimental proposal (requested experimental conditions, list of collaborators, …). At this stage the entry will undergo embargo and be potentially visible only to direct stakeholders (spokesperson, collaboration, RI).
   b) At the end of the experiment :
      i) The Research Infrastructure will associate the raw dataset and its metadata with the experiment. The dataset storage is handled according to the research

data policy of the infrastructure and the data management plan of the experiment. A potential embargo period can be set accordingly to Data Management Plans.

ii) The data officer and the Research Infrastructure data officer will add required metadata based on realized conditions of the experiment. This includes associations with data format, the associated detector configurations, beams, …

1) Analyzed Dataset from the analysis of an experimental dataset
   a) The data officer can submit an entry with an analyzed data set. This includes metadata related to the link to the raw dataset and software used to produce the data.
   b) Publications could be linked to the analyzed dataset.
2) Description of beams, detection systems, data formats, facilities, data management plans…
   a) In each case, a data officer could submit entries to the catalog describing the different types of apparatus or reference documentation.
   b) These entries could then be linked with any type of dataset entries

In addition, the service should anticipate connection to a future ecosystem involving authentication system, diverse and federated storage resources, so-called data lakes, and analysis platforms similar to what is being developed in the frame of the ESCAPE collaboration [1].

The features will serve as the basis for interfacing future services of data lake \& high performance computing and analysis platforms, but will require the development of a common authentication and authorisation infrastructure (AAI) service which is currently being developed within the WP 5.2 of the EURO-LABS project (https://iam-eurolabs.ijclab.in2p3.fr/).

4. DESCRIPTION OF THE OPENNP CATALOG

4.1. OPENNP CATALOG

4.1.1. Description of the catalog

The openNP catalog will be used to centralize the information on the available datasets and their metadata in a structured manner. The openNP catalog does not intend to be a data repository and dataset storage is handled by the data producers (Research Infrastructure, Collaboration, …). A dataset is in general linked with an ensemble of metadata. These metadata can describe the experimental condition (Data Management Plan, experimental setup, beam, target, facilities, data format, software, logbook, …). The openNP aims at defining collections of experimental dataset (associated with Research Infrastructure) and simulated or analyzed datasets (associated with research groups or collaborations). The catalog will be the entry point for any newcomer looking for a dataset through a webportal integrated in the EOSC.

4.1.2. Implementation in Dataverse framework

The openNP platform handles the need to centralize the metadata of the dataset ensuring the findability and accessibility of the dataset. Various alternatives have been investigated to achieve these goals, in the respect of the FAIR principles, and with a vision beyond the duration of the EURO-LABS project.

Several data catalogs frameworks have already been developed and are used in the research communities. The two main frameworks being used are Zenodo [2] and Dataverse [3].

The Dataverse framework is an open source project widely used in the research ecosystem for data repository and catalogs. The service is able to deal with many kinds of content (software, datasets, documents...) and publication versioning, providing a unique DOI for each version, thus allowing precise citation of the used datasets. The use of collections (“dataverse”) will allow us to filter and organize the content of the repository.

Dataverse framework makes use of DOI persistent identifier (Digital Object Identifier,) to identify uniquely every entry that is uploaded to its database. In other words, a DOI is the fingerprint of a deposit. The deposited entries can be all forms of research outputs: datasets, source code/software, workflows, publication, posters etc. All the deposited entries will contain certain required information (title, description of the entry, authors, publication date, access right and license, etc.). This information will be converted into human and machine-readable metadata within the repository. Once an entry is published, a DOI will be automatically assigned to the publication and the entry will be integrated into the repository. The repository will also have the possibility to harvest related catalogs (existing Research Infrastructure data catalogs, software catalogs, …)

Dataverse framework allows the creation of collections to gather together datasets and entries under a similar activity (Research Infrastructure, detection systems, research group, …). Dataverse allows the harvesting of the whole repository through the Open Archives Initiative Protocol for Metadata Harvesting - OAI-PMH [4]. This protocol is extensively used by other repositories and was developed precisely to harvest metadata descriptions of records (an uploaded entry to a repository).

The harvested metadata for each record are available in various formats based on the DataCite Metadata Schema [5]. When a new entry is uploaded to Dataverse the provided information/description is directly mapped to the DataCite schema representation. In addition, the repository allows the upload of other types of standard metadata schema representations. This means that besides being able to harvest metadata from the repository in the DataCite standard, Dataverse allows the upload of other machine-readable metadata standards files with the goal of improving the datasets metadata in general.

4.1.3. Collections

Various collections could be created for the Research Infrastructures and for the collaborations. A typical collection will be the Research Infrastructures and the collaborations (around an experiment and/or a detection setup). The organization and integration of data collected from various sources in a collection will be ensured by data curators in the various Research Infrastructures. The data curators may not be the same person as the data officers of individual datasets.

4.1.4. Data curation and workflows in the openNP catalog

Research Infrastructure, spokespersons of experiments or detectors could submit data sets to their collections in the catalog. The workflows for submitting the datasets will depend on the type of dataset and the policies of Research Infrastructure. Data curation policies should be released with the creation of a collection and data officers will be identified.

4.2. METADATA TYPES

In this section, the various types of metadata that are intended to be collected are described. Only the common fields described in 4.2.1 are required. The other types are optional and can be associated with the dataset depending on the case.

In parallel to the work performed within the EURO-LABS project, a global reflection on the metadata of interest in the field of accelerator based nuclear and particle physics was started among the Research Infrastructure of the EURO-LABS project and will be used to upgrade the metadata information based on emerging needs.

4.2.1. Common Fields

Description : These are the minimal fields common to all entries in the catalog.

Field	Description	Type	required / optional
doi	Digital Object Identifier associated with the entry	doi	required
Owners and maintainers	List of owner and maintainer of the entry. This person is a priori different from the author associated with the entry, e.g. the data officer of the collaboration or laboratory.	Array of string or ORCID number	required (ORCID number optional)

License	License under which the dataset is released	Array of string	required
Authors/Contributors	A list of person and collaboration that participated in the data creation process associated with the entry.	Array of string or ORCID number	required (ORCID number optional)
Start of data taking	Start of data taking	Timestamp	optional
End of data taking	End of data taking	Timestamp	optional
Publication date	Date at which the entry is made public.	Timestamp	required
Type	Nature of the entry: Beam, source, Detector, Data format, Data set, ...	integer	required
Status	Status of the entry within its life cycle: Draft, Pending, Embargo, Released or Archive).	integer	required
Time line	Realized and predicted timeline for the entry (e.g. embargo release date, ...)	array of date and integer	required

4.2.2. DATA FORMATS

Field	Description	Type	required / optional
Version	Version identifier.	doi/swhid	required
Specification	Documentation with description of the matching sample file.	Array of string	required
Software	Software implementation and link to source code.	doi / swhid	optional
Sample file	A sample file matching the description of the specification document.	link to dataset	required

4.2.3. BEAM DELIVERING FACILITY

A beam delivery facility. The facility delivers a beam to the end user or to another apparatus to produce another beam.

Field	Description	Type	required / optional
Name	Name of the installation or apparatus. e.g. SPIRAL1, IGISOL, BigRIPS	string	required
Type	Type of delivered beam : ISOL, Fragmentation, Stable	string of string	required

4.2.4. BEAMS

Nature of the beam used for the experiment. More than one beam could be defined for a given experiment.

Field	Description	Type	required / optional
Beam Delivery facility	Beam delivery facility	Doi to beam delivering facility	required
Usage	Production or Delivered	string	required
Particle	Particle type in the beam. e.g.: 11Li, pion, neutron, electron, gamma, molecular beam	Array of string	required
Energy	Energy of the beam MeV	double	required
Energy Spread	Energy spread in keV	double	required
Charge State	Charge State of the beam	integer	optional
Intensity	Intensity of the beam in pps	double	required
Cocktail beam	Register if the beam is part of a cocktail beam, and if yes, which one w/r to the current collection	int	optional

4.2.5. SOURCES

Description: Radioactive sources.

Field	Description	Type	required / optional
Type	Identify the emitter	string	required
Activity	Activity of the source in Bq.	float	required

Calibration Date	Date at which the source activity has been measured	timestamp	required
Specification	Description of the source	string	optional

4.2.6. TARGETS / SAMPLES

Description: A target or sample is a piece of matter, which is radiated by a beam or with particles from a source. In the case of a target e.g. a secondary beam can be produced for further experimental purposes. In the case of a sample is this piece of matter already the object of investigation.

Field	Description	Type	required / optional
Material	Description of the material	string	required
Thickness	Material thickness units	string	required

4.2.7. DETECTOR

Description: A detector entry consists of a device that can record any form of interaction with a particle.

Field	Description	Type	required / optional
Version	Version identifier of a specific configuration of the detector	string	required
Description	Description of the setup.	string	required
Configuration file	Link to configuration files for this setup.	doi of dataset	optional
CAD	Link to design files (CAD, step, …) associated with the device.	doi of dataset	optional

4.2.8. DATASET

Description: Entry of a dataset that could arise from experiment or analysis.

Field	Description	Type	required / optional
Dataformat	Associated dataformat.	doi to Data Format	required

Dataset	Associated other datasets (Auxiliary Data)	doi to Dataset	required
Detector	Associated detector(s) version	doi to Detectors	required
Software	Associated software(s) version	doi to Softwares	required
Sources	Associated sources(s) entry	doi to Sources	optional
Beams	Associated beam(s) entry	doi to Beams	optional
Beams production	Associated beam(s) production	doi to Beams Production	optional

4.2.9. EXPERIMENTS

Description: Entry of an experiment. It can be related to several type of data sets

Field	Description	Type	required / optional
Dataset	Associated dataset(s). This includes experimental dataset, auxiliary data, Proposal, Elog, …)	doi to Dataset	required
Facility	Facility where the experiment is done	ROR of facility	required
Detector	Associated detector(s) version	doi to Detector	required
Software	Associated software(s) version	doi to Software	required
Beams	Associated beam(s) entry	doi to Beams	optional
Beams production	Associated beam(s) production	doi to Beams	optional
Accessible system	Systems that can be studies with the dataset	string	required

4.2.10. PUBLICATIONS

Description: Entry of a publication linked to a dataset entry. It can be related to any type of data sets.

Field	Description	Type	required / optional
Publication	Associated a publication	doi to Publication	optional
Experiment	Associated experiment	doi to Experiment	optional
Experimental dataset	Associated dataset	doi to Dataset	optional

Detector	Associated detector version	doi to Detector	optional
Software	Associated software version	doi to Software	optional

4.2.11. DATA STORAGE FACILITY

Description: Entry to define where the dataset is stored if not stored on the openNP repository

Field	Description	Type	required / optional
Name	Storage infrastructure	string	required
Access point	Access point url	url of access point	optional
Protocol	Access point protocol (Ruccio, http, …)	string	optional

5. CONCLUSION AND FUTURE WORK

The present report constitutes the conceptual design for a prototype for the openNP catalog, allowing Research Infrastructure and researchers to centralize the metadata of their datasets, the possibility to correlate various metadata environmenting the datasets and software used in the data curation and analysis with persistent identifiers. This will foster improved scientific practices following the FAIR data principles towards effective Open Science in the community. It also provides them with an opportunity to automatically see their datasets to be findable for the rest of the scientific community and its future services.

Future work will focus on the implementation of a prototype of the service and the co-construction of standardized metadata schemes. The dataset submission workflows will have to be defined by the various parties involved. In addition, the integration of the catalog with centralized authentication and authorization platform and data-access platform will be required.

The openNP could be the entry point in the future for any researcher looking for datasets, softwares metadata, either as a user or as a developer or Research Infrastructure.

6. REFERENCES

[1] [Online] ESCAPE collaboration - https://projectescape.eu

[2] [Online] Zenodo, European Organization For Nuclear Research, and OpenAIRE.(2013). https://doi.org/10.25495/7gxk-rd71

[3] [Online] Dataverse https://dataverse.org/

[4] [Online] http://www.openarchives.org/pmh/.

[5] [Online] https://schema.datacite.org/.

[6] [Online] Research Organization Registry (ROR) - https://ror.org

ANNEX: GLOSSARY

Acronym	**Definition**
AAI	Authentication and Authorization Infrastructure
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
ESCAPE	European Science Cluster of Astronomy & Particle physics ESFRI Research Infrastructure
FAIR	Findable, Accessible, Interoperable, Reusable
ORCID	Open Researcher and Contributor ID (https://orcid.org)
RI	Research Infrastructure
ROR	Research Organization Registry (ROR)
SWHID	Software Heritage IDentifier
WP	Work Package