D2.7: Final Recommendations for FAIR Photon and Neutron Data Management

Document Control Information

Settings	Value
Document Identifier:	D2.7
Project Title:	ExPaNDS
Work Package:	WP2
Document Author(s):	Nicolas Soler (ALBA), Abigail McBirnie (UKRI), Alejandra Gonzalez-Beltran (UKRI), Andrey Vukolov (ELETTRA), Carlo Minotti (PSI), Heike Görzig (HZB), Krisztian Pozsa (PSI)
Document Reviewer(s):	Darren Spruce (MAX IV), Brian Matthews (UKRI)
Doc. Issue:	1.0
Dissemination level:	Public
Date:	07/07/2022

Abstract

The present deliverable pursues the work initiated in ExPaNDS Deliverable 2.2, which established a common metadata framework for FAIR data generated in Photon and Neutron (PaN) facilities. The modalities of implementation of this framework across facilities are examined and current practices and tools for metadata capture, storage and exposure are highlighted. This work is also the occasion to examine the commonalities of the framework with other initiatives developed inside and outside ExPaNDS such as the common search API, the data catalogues (ExPaNDS WP3), the EOSC discovery platforms B2FIND and OpenAire as well as the NeXus data format. It also provides the reader with guidelines on current tools and schemata available to record provenance and digital preservation information. The essence of these discussions is summarised as a list of practical recommendations at the end.

Document Log

Version	Date	Comment	Author/Partner
Complete Draft	13/05/2022	Internal review version	Nicolas Soler (ALBA)
1.0	07/07/2022	Final version for submission	Nicolas Soler (ALBA)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Abbreviations and acronyms

AAP	Authentication, Authorization and Profile
ANSTO	Australian Nuclear Science and Technology Organisation
API	Application Programming Interface
CDMA	Common Data Model Access
DAaaS	Data Analysis as a Service
DCAT	Data Catalogue Vocabulary
DMP	Data Management Plan
DOI	Digital Object Identifier
EBI	European Bioinformatics Institute
EOSC	European Open Science Cloud
ESRF	European Synchrotron Radiation Facility
ESS	European Spallation Source
ExPaNDS	European Open Science Cloud Photon and Neutron Data Service
FAIR	Findable, Accessible, Interoperable, Reusable
FaXToR	Fast X-ray Tomography & Radiography
GDPR	General Data Protection Regulation
HDF	Hierarchical Data Format
HZDR	Helmholtz-Zentrum Dresden-Rossendorf
IUCr	International Union of Crystallography
MCX	Materials Characterisation by X-ray diffraction
mmCIF	Macromolecular Crystallographic Information File
NIAC	NeXus International Advisory Committee
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
PaN	Photon and Neutron
PaNdata ODI	PaNdata Open Data Infrastructure
PaNET	PaN Experimental Technique
PaNOSC	The Photon and Neutron Open Science Cloud
PDB	Protein Data Bank
PI	Principal Investigator
PID	Persistent Identifier
PREMIS	Preservation Metadata Implementation Strategies
RDA	Research Data Alliance
REST	Representational State Transfer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

W3C	World Wide Web Consortium
WP	Work Package
XML	Extensible Markup Language
XRD	X-ray Powder Diffraction
XRF	X-Ray Fluorescence

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Executive Summary

The European Open Science Cloud Photon and Neutron Data Services (ExPaNDS) project aims at bringing the experimental data generated by European National Photon and Neutron Analytical Research Infrastructures (RIs) into the scope of the European Open Science Cloud (EOSC), together with services that support the discovery, access and reuse of that data. A key requirement of this aim is to ensure that the collected data complies with the FAIR principles (i.e. have to be Findable, Accessible, Interoperable and Reusable) at the point of leaving the facility, so that users across the EOSC can make effective use of that data. In order to fulfil this objective, ExPaNDS Work Package 2 (WP2), provides guidelines, recommendations, and practical experience to the project and to the wider Photon and Neutron (PaN) community on the best practices in generating FAIR experimental data for National RIs.

Therefore, the data generated at PaN RIs must be properly annotated and documented via the capture of various metadata records along their lifecycle. ExPaNDS deliverable 2.2 studied and formalised this lifecycle, basing itself on the temporal representation adopted in previous work from PaNdata-ODI. The proposed Common Metadata Framework enumerates the different types of metadata that should be collected at each step of the data lifecycle (from proposal to publication) and identifies their role in supporting the FAIR principles. In addition, the various metadata fields are categorised into what is essential, important and useful under the RDA FAIR Data Maturity Model priority flags.

The present report pursues this initiative by **examining the modalities of implementation of the Common Metadata Framework for FAIR data across PaN facilities**. It reviews current practices and tools for metadata capture, storage and exposure and highlights the commonalities of the framework with other initiatives and tools developed inside and outside ExPaNDS. In relation to all these aspects, we make a series of recommendations, which fall under three broad areas:

1. FAIR metadata within PaN RIs
2. FAIR PaN metadata in the EOSC
3. Sustainability of the FAIR metadata framework

FAIR metadata within PaN RIs

Across the experimental lifecycle within PaN RIs, there are multiple information sources – both human and machine – that play a role in metadata production and collection. In many cases, it is important that these sources interact and integrate within and across the various stages of the experimental lifecycle.

We suggest using the FAIR metadata framework as a basis to carefully design the metadata acquisition plan for a particular instrument. Metadata records heavily depending on users (e.g. calibration, sample, experimental notes) should also be given special attention.

In relation to metadata at PaN RIs, we additionally make several specific recommendations:

- Where possible, favour the use of persistent identifiers for data, people, instruments and samples. Promote as well the use of elogbooks/notebooks and sample databases for user-dependent metadata.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

- We recommend the use of NeXus/HDF5 as a self-contained and self-descriptive format to store data and scientific metadata, facilitating data exchange and reuse. Administrative metadata can to a certain extent also be stored using NeXus.

- In relation to software used in the experimental lifecycle, the FAIR principles for Research software provide a good indication of the minimum metadata that needed to be collected (e.g. to support citation of the software used).

- Regarding provenance of data derived from the processing and analysis steps of the experimental lifecycle, while it is not possible to recommend specific how-to capture metadata guides for all the variations of how data are produced in these stages of the life cycle, the most important factor is that software version, environment (dependencies) and workflows are captured as completely as possible.

FAIR PaN metadata in the EOSC

Beyond PaN RIs, EOSC indexing and discovery services such as B2FIND and OpenAIRE offer the means to make PaN datasets more findable by those outside the PaN domain. To enable these generic tools to harvest their metadata, it is important that PaN RIs provide OAI-PMH endpoints and mappings of their metadata to the metadata schemas used by the EOSC services. Regarding metadata mapping, one should bear in mind that these services are aimed at cross-domain discovery and are not designed to capture the full domain-specific richness that is possible using the ExPaNDS metadata framework.

In order to meet the primary purpose of the EOSC discovery tools, the information made available through B2FIND and OpenAIRE should be: 1.) sufficient for the initial enquiries of a non-domain specialist; and, 2.) able to point that user to where they can find further details.

At present, each PaN provider interacts with B2FIND and OpenAIRE on an individual basis, with the result that different metadata mappings are produced by the different facilities. Over time, we encourage the adoption of the ExPaNDS metadata framework in order to lead to more consistency in these mappings; however, it is likely that some ongoing differences will remain due to local practices and policies at the partner RIs.

Sustainability of the FAIR metadata framework

All of the aspects of the FAIR metadata framework will vary to some extent amongst the different ExPaNDS partner RIs. As well, metadata needs for FAIR will likely evolve over time (e.g. as new techniques are introduced). It is therefore desirable that the metadata framework, its definitions and its modalities of implementation are maintained and updated regularly well beyond the ExPaNDS project. One possible way of supporting this activity would be through the work of a permanent committee integrated into the management of all facilities.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

Introduction

ExPaNDS deliverable D2.2[1] established a metadata framework for FAIR data[2] in PaN facilities using as a starting point the PaNdata ODI D6.1 data continuum,[3] which is the sequence of steps normally followed by an experimental team when being granted access to an instrument: **proposal**, **approval**, **scheduling**, **experiments**, **storage, processing**, **analysis**, and finally **record**/**publication**. This continuum was augmented with updated information about metadata required to improve data FAIRness.

While D2.2 focused on building this metadata framework adopting a data consumer perspective, we propose in this deliverable D2.7 to shift to the data producer perspective by **reflecting on the practical implementation of this metadata framework in PaN facilities**, in alignment with the output originating from WP3 and WP4. From the moment they are collected, metadata will follow a journey and will be expressed in different formats and standards in order to be:

- aggregated from different sources;
- ingested and exposed into a metadata catalogue and in one or several datafiles;
- searched by domain expert and non-domain-expert users as well as harvested from other repositories or databases;
- projected into high-level metadata for ingestion into EOSC repositories (B2FIND, OpenAire|Explore and possibly others);
- projected into high-level metadata for ingestion into generalist repositories (e.g. Google Dataset Search).

After reminding the outlines of the FAIR metadata framework, we depict the journey undertaken by metadata from the moment of their generation, detailing the different **sources that come into play**, the different subsets of metadata exposed at each stage as well as the encoding schemes and mechanisms of aggregation and transmission necessary to achieve exposure in different locations. This report also reviews the state of the art of the **NeXus format** (Chapter 4), and examines its compliance with our metadata framework in different domains of application. It is also the occasion to confront the metadata framework with the different **metadata storage and transmission models used in WP3** (data catalogue, search-APIs and EOSC harvester repositories). Finally in Chapters 5 and 6, we provide practical recommendations on FAIR software and provenance information.

This deliverable is therefore intended for anyone involved in metadata collection and recording/ingestion in photon and neutron facilities, such as data manager, data acquisition engineer, data processing and scientific computing, beamline scientists. Decisions about the

[1] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825
[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. 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
[3] Matthews, B. et al. (2012). *Model of the data continuum in Photon and Neutron Facilities*. Zenodo. https://doi.org/10.5281/zenodo.3897190

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

nature and format of the metadata to collect for any particular instrument should always be the result of detailed discussions between the different stakeholders and we hope that the present recommendations will help facilitate this process well beyond the ExPaNDS project, in compliance with the FAIR data principles.

Finally, it is useful at this stage to remind the reader about the **benefits** to be obtained from a better compliance of the FAIR data principles. For this we can outline a few use cases.

- The Experimental team can find an accurate record of their own work in one place for future use.
- Linking journal publications with the corresponding data allows the traceability and validation of results.
- Future methods for raw data processing and analysis are likely to improve the results obtained with current methods on the same data.
- Method developers need access to raw data in order to build and test their algorithms.
- Reviewers of experimental results may want to assess the quality of the data obtained and their further treatment.
- A team failing in deriving scientific knowledge from raw data may be helped by other contributors (sometimes unexpectedly) if the data are publicly available (e.g. Kaggle competition).
- Publicly available FAIR data can be easily reused for training purposes

Roadmap for the reader

Chapter 1 discusses the nature of the different metadata fields listed in the framework and focuses on the level of control facilities might have on the capture of the corresponding content. It orients the reader towards recent initiatives aimed at facilitating the acquisition of metadata that are directly provided by users (sample information and elogbooks).

Chapter 2 depicts the journey that metadata undergo from their creation to storage in files and data catalogues. It reviews the different types of sources that generate these metadata and provides advice on which data schemata and file formats to use.

Chapter 3 inspects the commonalities between the framework and the data models underlying several tools developed inside or outside ExPaNDS. Mapping tables between the framework and other tools are issued, such as the WP3 common search API (Section 3.1) and EOSC discovery platforms (Section 3.2) and other repositories.

Chapter 4 focuses on the NeXus data format, which aims at being the common standard for self-descriptive containers of data and metadata across PaN techniques. Again a mapping table between the framework and the current NeXus base classes is suggested.

Chapter 5 is the occasion to briefly review how the FAIR principles apply to software artefacts. Advice is provided when it comes to cite software in metadata records.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Chapter 6 concentrates on the concept of provenance information in the context of derived data. Concepts and practical tools for keeping records about software workflows and environment are exposed and good practices are suggested. The topic of digital preservation is also presented.

Chapter 7 condenses the recommendations expressed throughout the previous chapters in a series of bullet points and addresses the topic of GDPR in the context of personal (meta)data collection for research. Final remarks are then issued.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1. The FAIR Metadata Framework

1.1 Reminder of the Framework

The main outcome of ExPaNDS D2.2[4] was the establishment of a set of recommendations leading to a common Metadata Framework for the national Photon and Neutron Research Infrastructures (PaN RIs) in Europe. This framework adopts the temporal representation used in PaNdata ODI D6.1,[5] consisting of steps traditionally leading to the production of scientific data at PaN facilities.

As stated in the introduction section above, this sequence of steps is: **proposal**, **approval**, **scheduling**, **experiments**, **storage**, **processing**, **analysis**, and finally, **record**/**publication**. Each step provides an occasion to collect and store metadata fields of various kinds, thereby increasing one or several of the following aspects: **F**indability, **A**ccessibility, **I**nteroperability and **R**eusability (FAIR).

A visual summary of the framework (based on details extracted from ExPaNDS D2.2) is presented in Figure 1. For each metadata field, the figure indicates which aspects of FAIR are covered, along with their associated prioritisation levels (**P1**-essential, **P2**-important, **P3**-useful). As set out in ExPaNDS D2.2, the definition of these prioritisation levels follows those of the RDA FAIR Data Maturity Model.[6]

- **P1 essential:** addresses an aspect of the utmost importance to achieve FAIRness under most circumstances, or, conversely, FAIRness would be practically impossible to achieve if the indicator were not satisfied.

- **P2 important**: addresses an aspect that might not be of the utmost importance under specific circumstances, but its satisfaction, if at all possible, would substantially increase FAIRness.

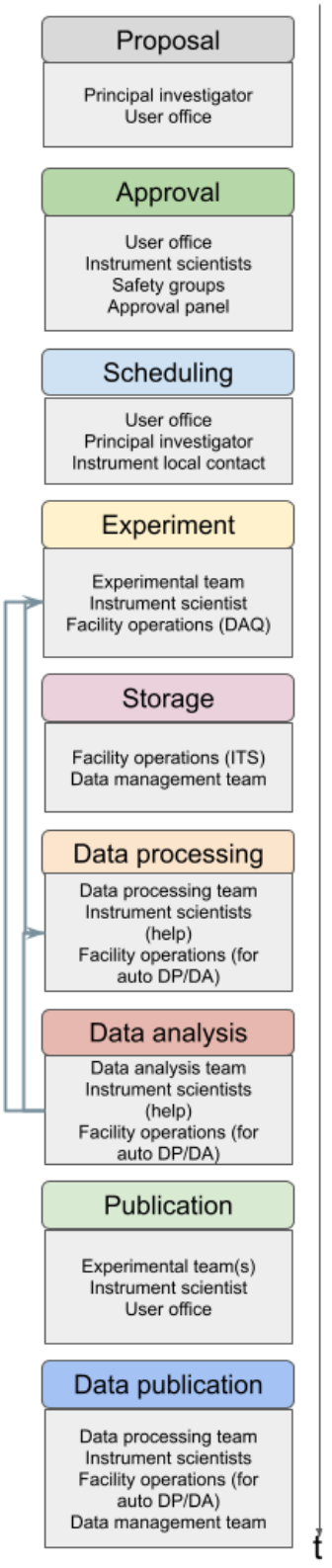
- **P3 useful**: addresses an aspect that is nice-to-have but is not necessarily indispensable to achieve FAIRness

Note that the different fields in the framework correspond to entries that can be very different in terms of definition and content. For example, “PI/Main proposer” usually corresponds to a string allowing to unambiguously identify a person while “sample information” or “processing information” can aggregate a lot of information which definition and scope can vary.

This framework now has to confront and prove useful within the reality and the diversity of setups among the different PaN RIs. Given a particular instrument, different metadata fields might be equally important for FAIRness, yet very different in terms of how easy or difficult it might be to collect them. Therefore, a prioritisation strategy that considers FAIRness alongside other dimensions needs to be established, as discussed in the next section.

[4] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825
[5] Matthews, B. et al. (2012). *Model of the data continuum in Photon and Neutron Facilities*. Zenodo. https://doi.org/10.5281/zenodo.3897190
[6] FAIR Data Maturity Model Working Group. (2020). *FAIR Data Maturity Model. Specification and Guidelines (1.0)*. https://doi.org/10.15497/rda00050

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.



METADATA FIELD	PRIORITY	ASPECT
PI/Main proposer	P1	FA
Co-investigators	P1	FA
Instrument requested	P1	F
Funding source	P2	F
Sample description	P1	F
Proposed experimental conditions	P1	F
[Safety conditions]	P3	F
Experiment description	P1	F
Prior art (related publications, proposals)	P2	F
Facility information	P1	F
Proposal identifier	P1	F
[Approval panel]	P3	/
Sample safety assessment	P2	/
Allocated day & time on instrument	P2	FA
Scheduled visiting experimental team	P2	FA
Safety Training data	P3	/
Detailed experimental planning	P2	F
Sample preparation	P2	FR
[Sample reception]	P3	/
Visiting experimental team (user id)	P1	FA
Experiment date	P1	FA
Sample information	P1	FR
Instrument information	P1	FR
Calibration information	P1	FR
Experimental planning	P2	FR
Environmental parameters	P2	FR
Laboratory notebook	P2	FR
Instrument scientist	P2	F
[Experimental report]	P3	R
Persistent Identifiers (PIDs)	P1	FA
Preservation description information	P1	AR
Dataset information	P1	F
File identifier	P2	AR
[Representation information]	P3	IR
[Instrument parameters]	P3	FR
Processing team (user ID)	P2	AIR
Original data	P1	IR
Data format (after processing)	P1	IR
Dataset information	P2	AIR
Processing information	P1	R
Software package information	P1	R
Analysis team (user id)	P2	AIR
Original data	P1	IR
Software package information	P1	IR
Dependence tracking and workflow	P2	R
Data formats (after analysis)	P1	IR
Dataset information	P2	IR
File identifier	P1	AIR
[Instrument parameters]	P3	IR
[Calibration information]	P3	IR
Authors / Coauthors (user ID)	P1	FA
Proposal information	P1	FA
Publication information	P1	F
persistent Identifier (PID)	P1	F
[Supplementary data information]	P3	F
Resource identity	P1	FI
Related resource	P2	F
Creator	P1	F
Contributor	P2	F
Title	P1	F
Publisher	P1	FI
Publication year	P1	FI
Licence	P1	IR
Release date	P1	IR

Figure 1: A representation of the D2.2 metadata framework for FAIR PaN data. The steps of the PaNdata ODI D6.1and ExPaNDS D2.2 data continuum and associated stakeholders are indicated on the left part. The FAIR prioritisation of the different metadata fields is indicated as follows: P1-essential: bold font, P2-important: regular font, P3-useful: in between brackets. The “ASPECT” columns refers to the aspects covered by the “FAIR” acronym (or ‘/’ when none is covered). Arrows on the left side symbolise the possibility of going back and forth between different steps.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

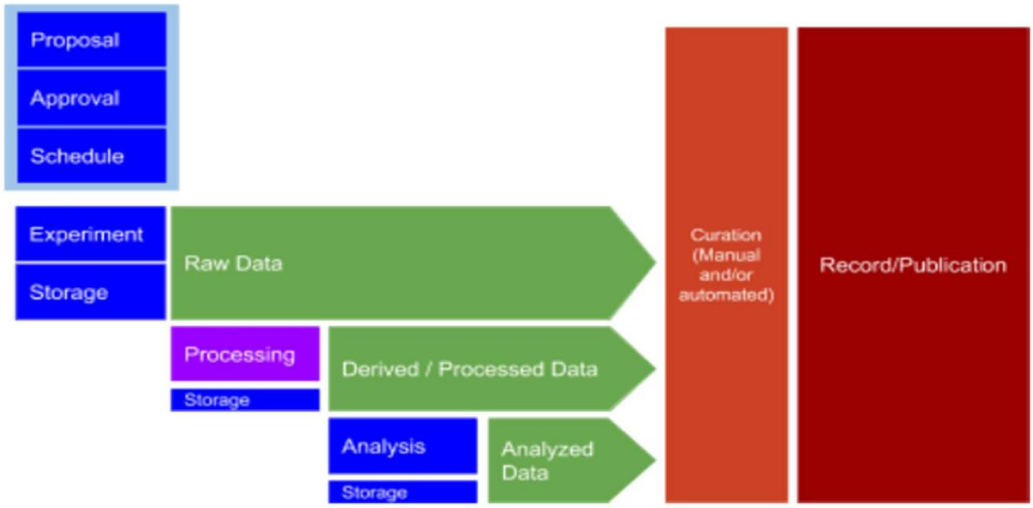


Figure 2: The metadata value stream (Figure 8 extracted from D2.2)[7]

1.2 Choosing which Metadata Fields to Collect

Prioritising which metadata fields to collect should be an informed decision. A large number of metadata fields come from the instrument and are therefore easy to record in a controlled, automated way. Other fields concerning for example the sample, rely on information disclosed by the users, which is hardly controllable at present in terms of accuracy and completeness. In any case, one has to be aware of the fields that are crucial for final (re)users in terms of importance for findability and reusability. Those fields are likely to be often searched in data catalogues.

1.2.1 What is important for data findability?

A data landscape survey[8] was conducted in 2019 for ExPaNDS task 3.2 that aimed at characterising use-cases for data-searching across Photon and Neutron facilities by examining the 'who', 'what', 'how' and 'why' aspects of the search. Keeping in mind that the responses came from beamline/instrument scientists, who are very familiar with the data and may have different search use cases than other types of users, it can be helpful to consider some of the conclusions of this survey in the context of the present report. The different categories of metadata used in this survey are reported in the following table together with examples. We group them as being either "administrative" (i.e. "non-domain specific" or "bibliographic") and "scientific:

[7] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825
[8] https://docs.google.com/document/d/1ytZ0FGzHEAPanKgn5J1L8GeHgGuA8DVe4VfAyx5Rofo/edit

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Group	Category	Example (optional)
Administrative	Admin	Affiliation, proposal title, PID, instrument name
Scientific	Sample/state	Sample name, composition, state, history of preparation, chemical formula, space group
	Physical	Temperature, pressure, beam intensity, electron bunch energy
	Measurement type	Experimental technique
	Beamline setup	Scan command, calibrant file, sample position, monochromator
	Material use/context/project	
	Publication/PDB	
	Data size / resolution	

Table 1: Metadata categories from the WP3 survey on findability

One of the insights of the survey is that **raw data and processed data are likely to be searched with equal importance**. The other aspect that emerges from this study is that users will be most likely searching for either **administrative** (e.g. people names and affiliation, instrument name, proposal title) or **sample-related** metadata (sample name, type, state, composition, preparation, formula, space-group etc.). Physical parameters of the experiment (e.g. temperature, pressure), beamline setup metadata and parameters related to the collected data (size, resolution) can also be searched for, but to a lesser extent. The type of experimental technique might also be of importance in future searches but this aspect was hindered in this study due to the fact that it is often implicitly contained in the sample metadata.

In agreement with the survey, most administrative and sample metadata indicated in our framework (see Figure 1) are considered essential regarding FAIRness. However, while administrative metadata can be collected from the user office or the instrument, proper filling of the sample-related metadata is left under the control of the user, who might often not have the time or will to provide all the necessary details to ensure a FAIR-compliant level of description. These descriptions would also need to be harmonised across datasets within the catalogue to provide valuable information for further data search and re-use. Indeed, sample-related metadata is often a crucial element needed for data analysis and thereby can lead to confidentiality issues if not properly secured. Indeed, scientists are often reluctant to provide a precise description of the sample under study, fearing leakage of scientific information to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

their competitors. This is a critical bottleneck to the migration towards FAIR data in PaN facilities so that proper incentives must be designed to overcome it (see Section 1.2.3).

1.2.2 Towards a prioritisation strategy

The metadata framework presented in Figure 1 establishes a convenient "recipe" where the importance of each metadata record is evaluated with regard to the FAIR principles. However, when the time comes to apply this recipe to a particular instrument / beamline, other criteria should be considered in order to prioritise the metadata collection. Let's have a closer look at the metadata records from the "Experiment" stage of the framework in order to illustrate this:

Metadata record	Prio.	Asp.	Type	Source
Visiting experimental team (user id)	P1	FA	administrative	user office
Experiment date	P1	FA	administrative	user office
Sample information	P1	FR	scientific:sample	sample DB or user
Instrument information	P1	FR	scientific:beamline setup	data acq. and control system
Calibration information	P1	FR	scientific:beamline setup	beamline scientist or user
Experimental planning	P2	FR	scientific:context	principal investigator
Environmental parameters	P2	FR	scientific:physical	data acq. and control system
Laboratory notebook	P2	FR	scientific:context, sample	user
Instrument scientist	P2	F	administrative	user office
[Experimental report]	P3	R	scientific: context	user

Table 2: Metadata records extracted from the "Experiment" step of Figure 1. The type of record and likely source is presented in the last two columns.

Administrative metadata can be retrieved from the user office while most scientific metadata will be collected while the experiment is running. A **metadata ingestion system** is necessary to compile information from the different sources enumerated in the last column of Table 2. In terms of automation, it is desirable to couple the user office database (**proposal system**) and the **data acquisition system** of the instrument with the data management system.

The difficulty to collect certain **instrument metadata** and **environmental parameters** automatically, regardless of their importance for FAIR, will be very variable depending on the actual setup of each facility and should be the result of careful discussion between the different stakeholders (instrument scientists, users, data acquisition and management, data cataloguing and archival staff).

It is clear from Table 2 that a certain number of metadata records might rely on **direct user input** (sample, calibration, experimental planning, laboratory notebook, experimental report), thereby requiring much effort from users, who would need to see the value of this to be motivated to provide all of the information. Not all metadata elements are equally important in terms of FAIR. The experimental planning for instance (P2) will often have been provided at an earlier stage as a prerequisite for the users to access the facility and is therefore likely to be accurate. The experimental report (P3) is also usually mandatory and will contain redundant information with the laboratory notebook/logbook, which is the subject of Section 1.3.

Sample information is considered essential for FAIR and contains various types of entries (e.g. sample name, provenance, chemical formula, sequence). We should keep in mind that

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

one experiment encompasses several samples and that each sample can in turn be used to collect several datasets. It will be the subject of Section 1.2.3.

Calibration information deserves special attention too since it can be treated as a sub-experiment with its own samples and associated datasets. It is therefore important to keep track of which calibrant product was used for each calibration dataset as well as instrument and context information (e.g. flat-field, dark-field correction, goniometer rotation axis calibration, etc.).

Overall each **dataset** (a dataset being considered the result of a **scan**) should therefore be associated with metadata about the sample in use (or a reference), associated calibration files, instrument parameters (e.g. orientation, exposure time, flux, temperature, pressure etc.) and other contextual and administrative information. This can only be achieved thanks to a carefully designed ingestion of information from several distinct sources (sample database, user office database, data acquisition and controls, elogbook/notebook). The use of **persistent identifiers** for people, samples, instruments and data should be encouraged in order to record unambiguous and stable information. Finally one should mention that the **read/write permissions** of each metadata field should also be under control in order to respect embargo periods and privacy (see Section 6.2).

1.2.3 Sample-related metadata

Sample information is crucial for both findability and reusability of the data. As mentioned before, obtaining an accurate record of **sample provenance**, **state** and **composition** largely relies on the user and therefore poses a particular challenge. Several strategies can be implemented to mitigate this issue:

- **Extract information from the proposal**: Samples and substances must be declared in order to validate the safety of an experiment.

- **Provide a sample-description interface** that helps the user centralise information. This would ideally be coupled to the proposal system and to the sample tracking system to have a unique source of information.

- **Reward users** for entering FAIR sample metadata by helping them storing and formatting the information for publication.

Note that the **LEAPS-STARS**[9] project started in December 2021 is currently investigating solutions to tackle this problem. Its goal is to make sample information FAIR-compliant in order to help users and user offices to manage samples. It relies amongst others on the use of **sample persistent identifiers** and the development of a **set of common standard information items** about samples. It should in the long term provide a commonly adopted solution to the problem of obtaining reliable and complete sample description to achieve identification and handling.

[9] https://leaps-initiative.eu/digital-leaps-is-on-its-way/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Other initiatives in science exist that could provide inspiration for this purpose. Among them, it is worth mentioning the **BioSamples database** developed at the EBI,[10] which outlines the necessary components for efficient sample-metadata management:

- The use of sample persistent identifiers (PID).
- A common, yet extendable data model to describe samples.
- A single entry point for users to enter and curate sample information from a unified interface.
- A MongoDB document database and associated interfaces (web and RESTful APIs)
- An Authentication, authorization and Profile (AAP) system for secure sample submission and retrieval.
- The possibility of restricting read and write access until the embargo period is over.
- A curation system, allowing to track changes.
- A validation system.

Considering PaN facilities, the use of **sample persistent identifiers** (PIDs) pointing to accurate and complete sample records would prove very beneficial. In order to achieve this, a similar unified interface used as a unique entry point would have to be built and coupled with the different data management systems involved in the data continuum steps and the sample tracking system. The same interface should be used for entering/updating sample information during, e.g. the proposal, experiment or data publication steps.

1.3 Maximising the Potential of elogbooks

Compliance with the FAIR principles involves an easy and open access to scientific data after the embargo period. While this is already reality in some PaN facilities, another problem is to convince users to abandon pencil and paper for taking notes during the experiment in favour of an electronic notebook (an 'elogbook') integrated into the metadata catalogue. Indeed, users' annotations are often very rich and contain key metadata to ensure the correct interpretation and reuse of the data. For example, the elogbook can contain data about the sample preparation history (provenance), its state(s) during the experiment and other pieces of information required by the framework that could be automatically parsed.

An example of such elogbook is the one developed around ICAT[11] at the ESRF. The elogbook is organised as a series of events,[12] in which multiple users can introduce text or images. The events can contain annotations from the user but also notifications from the acquisition software (error messages, command lines etc.). Describing in detail these functionalities is out of the scope of this report, nevertheless the system of information tagging[13] that comes with the annotation system could be particularly interesting for parsing information required in the "Experiment" stage of the framework.

[10] Courtot M et al. (2019). *BioSamples database: an updated sample metadata hub*. Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D1172–D1178, https://doi.org/10.1093/nar/gky1061

[11] https://gitlab.esrf.fr/icat/elogbook-standalone

[12] https://data.esrf.fr/investigation/121810241/events

[13] https://data.esrf.fr/tag?investigationId=121810241

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

In the case of SciCat,[14] while there is currently no privileged elogbook implementation, facilities adopting the data catalogue can leverage on the SciCat APIs and the database flexibility to build custom elogbook integrations. The information collected using the logbook will then be persisted in the database and can be investigated using the catalogue UI. It is left, as a consequence, to each integration to extract automatically metadata from elogbooks and inject them in the data catalogue accordingly. An example of elogbook integration with SciCat is SciChat, in use at ESS,[15] a member of the PaNOSC sister project, which is based on the Matrix chat client.[16]

An ergonomic electronic logbook/notebook annotation system, tightly coupled to the user's data stored in facilities, allows rich metadata to be captured (annotations, system logs, images etc.). It reduces the risk of (meta-)data loss and ensures better preservation and reusability. We therefore recommend their implementation and promotion among users. Independently of the elogbook implementation, we propose that facilities implementing electronic notebooks adopt a harmonised approach for tagging both user and acquisition system information based on the "Experiment" part of the framework and automatically extract them in order to inject them into the data catalogue. This system would provide a convenient way of collecting essential information from the users while the experiment is ongoing, thereby releasing them from the burden of having to enter the information later through distinct interfaces. By essence, electronic notebooks are based on time-stamped events, which is not suited to provide complementary information about the experiment as a whole. A better, alternative approach would be to use a sample-focused electronic laboratory notebook, easily editable by the users before, during and after the experiment, (ideally coupled to the sample tracking system). As mentioned in Günther et al. 2021,[17] a system based on Jupyter notebooks could also serve this purpose by allowing to combine code, notes and images in a single file.

[14] https://scicatproject.github.io/
[15] https://europeanspallationsource.se/
[16] https://matrix.org/clients/
[17] Günther et al. (2021). *FAIR meets EMIL: Principles in Practice*. ICALEPCS2021, Shanghai, China. https://doi.org/10.18429/JACoW-ICALEPCS2021-WEBL05

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

2. The Metadata Journey in PaN Facilities

2.1 The Different Sources of Metadata in a PaN Facility

Across the experimental lifecycle and within PaN RIs, there are multiple sources that relate to metadata production and collection. These sources include both humans (i.e. roles) and machines (i.e. systems).

ExPaNDS D2.2 identifies three broad roles that feature within the experimental lifecycle:

1. **Data Producer:** Examples include users and instrument scientists. These roles provide information related to the context of the experiment, such as sample or study description.
2. **Data Consumer:** Examples include users and reusers of data. Such roles rely on access to data in order to validate scientific results, to reproduce analyses, or to derive new scientific results.
3. **Data Manager:** Examples include data managers and librarians. Such roles curate data and act as data custodians.

Likewise, ExPaNDS D2.2 highlights the three main types of information systems that feature across the experimental lifecycle:

1. **Data Production Systems:** Examples include proposal systems and acquisition systems.
2. **Data Consuming Systems:** Examples include systems for accessing, visualising, or downloading data.
3. **Data Management Systems**: Examples include data processing and analysis systems.

Note that these systems are usually in a continual stage of upgrade and development. Drawing on detail provided in ExPaNDS D2.2, Table 2 summarises the roles and information systems that play a role in enabling FAIR data at each stage of the experimental lifecycle. Note that these nine stages include the data processing and data record stages newly proposed in ExPaNDS D2.2.[18]

[18] The original experimental lifecycle as described in PaNdata ODI Deliverable 6.1 included only the first seven stages presented in Table 3.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Experimental Lifecycle Stage	Roles involved *NB: Some roles listed below are a subset of a wider role that is also listed (e.g. instrument scientists are also facility staff).*	Information systems involved *NB: Some systems below may be part of a wider system that is also listed (e.g. the proposal submission system may be a part of the user office system).*
Proposal	· PI · Facility staff	· User office system · User registration and management system · User identity system · Proposal submission system
Approval	· Facility staff · Instrument scientists · User office staff · Safety group · Approval panel (including external scientists) · Experimental team	· User office system · Approval system
Scheduling	· PI · Facility staff · Instrument local contact	· User office system · Scheduling system
Experiment	· Experimental team · Instrument scientists · Facility operations staff (e.g. acquisitions systems staff)	· User office system · User account management system · Facility central control system · Data acquisition and control system · Storage system · Sample database systems (e.g. ISPyB, EPICS archiver)
Data storage	· Experimental team · Data infrastructure team	· Data acquisition system · File writer/generator system · Data management system · Data storage system · Facility repository · Data publication system (i.e. Data catalogue) · Archival systems

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Experimental Lifecycle Stage	Roles involved *NB: Some roles listed below are a subset of a wider role that is also listed (e.g. instrument scientists are also facility staff).*	Information systems involved *NB: Some systems below may be part of a wider system that is also listed (e.g. the proposal submission system may be a part of the user office system).*
Data analysis	• Experimental team • Analysis team • Instrument scientists • User office staff	• Data storage systems (may also be a Data Catalogue) • User office systems • Software catalogue systems (i.e. that locates, links to, or references analysis and/or visualisation software) • Data analysis and computing resource platforms • Visualisation systems
Publication	• Experimental team • Instrument scientists • User office staff • Library staff	• User office system • Research output tracking systems • Library systems • Institutional repository
Data processing	• Experimental team • Processing team • Instrument scientists	• Data storage system (e.g. could include the Data catalogue) • Software catalogue system (i.e. that locates, links to, or references data processing software)
Data record/publication	• Experimental team • Facility staff (i.e. who help with minting PIDs) • Record publishers • PID providers	• PID minting system • Facility information management system • PID provider systems • PID platform (e.g. a searchable and downloadable system where PIDs are accessible and data are downloadable)

Table 3: Roles and systems involved in metadata production and collection

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

2.2 Metadata Aggregation and Exposure

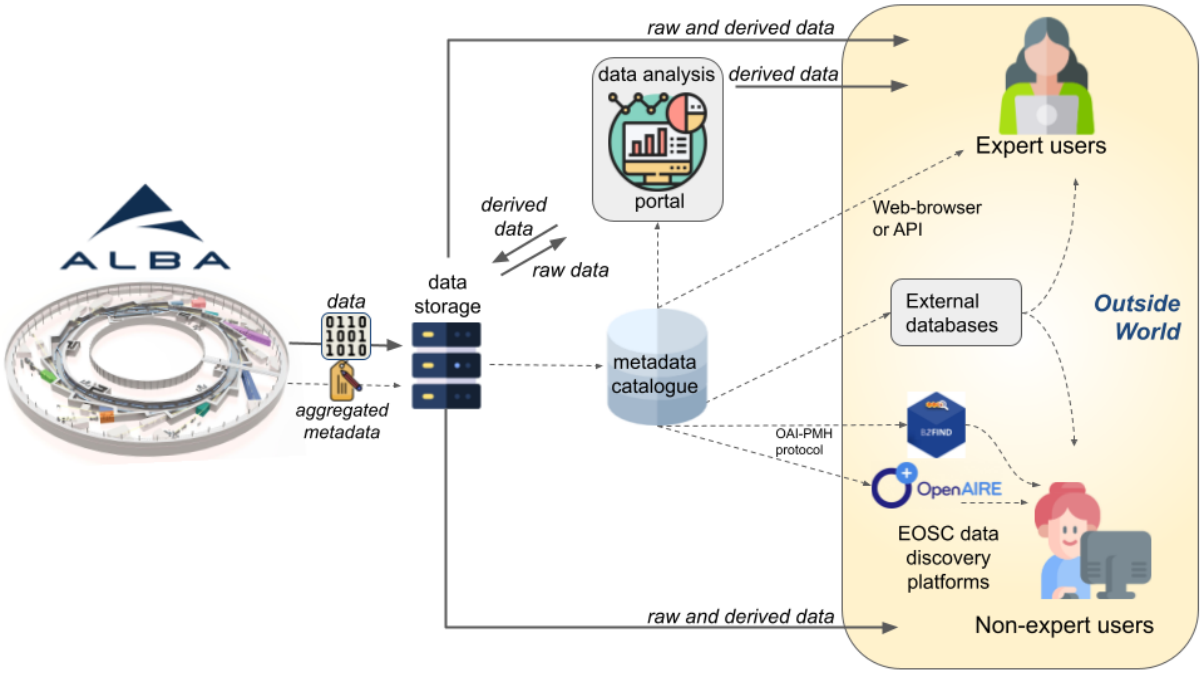


Figure 3: A high-level view of the (meta)data journey at PaN facilities. The flows of data and aggregated metadata are respectively shown using continuous and dotted lines.

Table 3 depicts the multitude of metadata sources in PaN facilities. Among them, we can distinguish those sources that generate administrative metadata, and those sources, such as the Data acquisition and control system (Daq), which generate scientific metadata. Throughout the different PaN facilities, these sources are embodied by very different setups, and it would be impossible to recommend a detailed workflow about how to connect these sources together covering all these cases. However, as illustrated in the schematized view of the (meta)data journey at PaN facilities shown in Figure 3 above, data and metadata should be eventually:

1. Stored using appropriate formats and standards
2. Exposed into a metadata catalogue

Several questions arise when considering these constraints under a FAIR perspective:

- **Findability:**
  - What is the minimal subset of metadata to expose in the data catalogue in order to ensure that the data will be found, in agreement with the search parameters of expert and non-expert users?

- **Accessibility:**
  - Should there be different release-dates according to the confidentiality level of metadata exposed (i.e. sample or person-related metadata fields)?

- **Interoperability:**
  - Which formats and standards should be used to ensure smooth data exchange?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

- **Reusability:**
  - What is the best way to ensure persistence and preservation of both data and the documentation and environment needed to exploit them?

Again, the final responsibility on how to tackle these questions is left to each individual facility; we can, however, highlight a few guidelines in the following sections.

2.2.1 Aggregating metadata

Here, several strategies are possible. After ingestion, metadata can be first assembled in a single file (or in case of the usage of NeXus with HDF external links, one entry or master file and an arbitrary number of files linked by the master file) and then this file can serve itself as a source for exposing metadata in a catalogue or repository. While this approach has the advantage of ensuring a persistent location for storing metadata independently of the existence of a metadata catalogue, it implies a dependency on the storage file format in order for the catalogue to be properly updated. The reverse approach (filling the metadata catalogue first and from then, storing metadata into files) conveys similar disadvantages. A recommended strategy is therefore to have a limited number of file and metadata formats and apply a parallel, mirror, ingestion in both storage (nowadays often NeXus files) and the data catalogue.

2.2.2 File formats and metadata schema

File formats describe how data bits are structured to encode the information in a file while metadata schemata describe how the information in a file is structured and which vocabulary is used. Concerning metadata schemata, as stated before, we can distinguish **administrative** (e.g. project and IT system) and **scientific** metadata associated with the raw data. Other schemata can also be used when it comes to describe **provenance information** for processed data or **preservation** information (these last two points will be covered in Chapter 6). In other words, each file format and metadata schema has its **application profile**.

Administrative metadata

For the discipline-independent (administrative) metadata, there are some recommendations within the EOSC, where **DataCite** and **Dublin Core** are advised. Also the W3C Recommendations[19] offer good choices for metadata formats. For example, **DCAT**[20] can also be used, as it extends Dublin Core terms with specific vocabulary to describe datasets, data catalogues and related entities such as data services. Since W3C Recommendation DCAT version 2, DCAT has included terminology to describe research data more precisely, providing more details for dataset identifiers, licensing, access rights, dataset quality information, spatial and temporal resolution descriptions as well as enabling better relationships with external references. In DCAT3, which at the time of writing is going through the process to become a W3C Recommendation, the main additions cover versioning, checksums and dataset series as well as the treatment of inverse properties and other revisions.[21]

[19] https://www.w3.org/TR/?status=REC
[20] https://www.w3.org/TR/vocab-dcat-3/
[21] Gonzalez-Beltran, A. and Winstanley, P. (2022, February 17). The Data Catalogue Vocabulary (DCAT). Zenodo. https://doi.org/10.5281/zenodo.6142906

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Scientific metadata

In order to find more discipline-specific but also other standards, the metadata standards catalogue of the Research Data Alliance (RDA)[22] might be a good choice as well as the FAIRsharing standards catalogue.[23] Concerning experimental data, the variety of domains using PaN facilities mean that some use formats specialised to particular techniques, such as the Crystallographic Information Framework (CIF)[24] of the International Union of Crystallography (IUCr).[25] NeXus, which will be covered in Chapter 4, aims to be a reference data format for PaN science, by providing specialised metadata fields adapted to each technique.

Other ontologies of interest

Specialised standards exist to address specific aspects. **Provenance** information (understood in this context as knowledge about how data have been created) can be well structured thanks to ontologies like PROV-O[26] and its derivatives. Likewise, **preservation** (information required for long-term usability) can be structured using standards like PREMIS.[27] Other initiatives exist like the sensor network ontology (SSN).[28] Provenance and preservation are discussed in Chapter 6.

File formats

The most important feature of file formats in a data repository is that it is possible for third parties (and after a long period of time) to be able to open the file. Very dominant file formats are therefore ASCII and for binary files: HDF5.[29] HDF5, which is also the main physical format associated with NeXus, is sometimes reported to have some performance issues (informal discussions). One workaround adopted in some cases is that data acquired during a measurement is written in a file format offering higher performance and then converted to HDF5 when deposited in the repository or before usage.

When choosing a file format or a metadata schema it is recommended to review, apart from the direct usage and performance, aspects such as the range of existing tools allowing users to work with the chosen format and metadata (edition, visualisation, analysis etc.), as well as their distribution in the community, and sustainability.

Should we aggregate multiple files and formats?

It is very common to have **a variety of files in a dataset**. Each file can follow its own standard, depending upon what needs to be expressed. It is advisable to have a **manifest** or readme file giving an overview of what is in a dataset and e.g. description of a dataset, what can be expected from each file, checksums of files, with which programs these files can be opened or where related scripts can be found. This can be done in free text, but DataCite, DCAT and (if it has to be very detailed) PREMIS might be good choices (see Chapter 6).

[22] https://rdamsc.bath.ac.uk/subject-index
[23] https://fairsharing.org/search?fairsharingRegistry=Standard
[24] https://www.iucr.org/resources/cif/spec
[25] https://www.iucr.org/
[26] https://www.w3.org/TR/2013/REC-prov-o-20130430/
[27] http://loc.gov/standards/premis/v3/index.html
[28] https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/
[29] https://www.hdfgroup.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3. Alignment of the ExPaNDS Metadata Framework with PaN, EOSC, and Other Data Cataloguing, Indexing and Discovery Services

This chapter aims to explore the correspondence between the D2.2 framework (see Figure 1) and tools developed in ExPaNDS WP3 such as the common search API and the data catalogues. Another important outcome of WP3 is the capability for EOSC discovery platforms like B2FIND and OpenAire to harvest metadata from a facility's catalogue(s).

Any agent searching or harvesting (meta)data in a repository will need to understand the metadata schemata in use in this repository in order to match it to its own internal data model. This chapter will therefore be focused on how to match the metadata records listed in the framework to different metadata schemata used in the aforementioned WP3 tools (Section 3.1). Section 3.2 focuses on metadata harvesting by B2FIND and OpenAire, which both use a metadata schema derived from DataCite.

3.1 Compatibility of the Framework with Metadata Catalogues and the PaN Search API

The implementation of the PaN search-API service follows the requirements sets in the **PaN search-API schema**, which provide a unified means of searching and reading metadata, across different flavours of data catalogues (e.g. ICAT, SciCat). This has the advantage that the users of the PaN search-API are not required to know the details of the different data catalogues. It also facilitates machines' readability of different data catalogues, leveraging a shared interface.

The queryable metadata exposed by the PaN search-API has been agreed upon in the course of the ExPaNDS and PaNOSC projects and can be found in detail in its requirements document.[30] Schematically, the supported fields are shown below in Figure 4. For convenience, a short description of each box follows, as an extract from the PaN search-API Data Model.

[30] https://www.panosc.eu/wp-content/uploads/2020/12/D3.1_API-definition.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

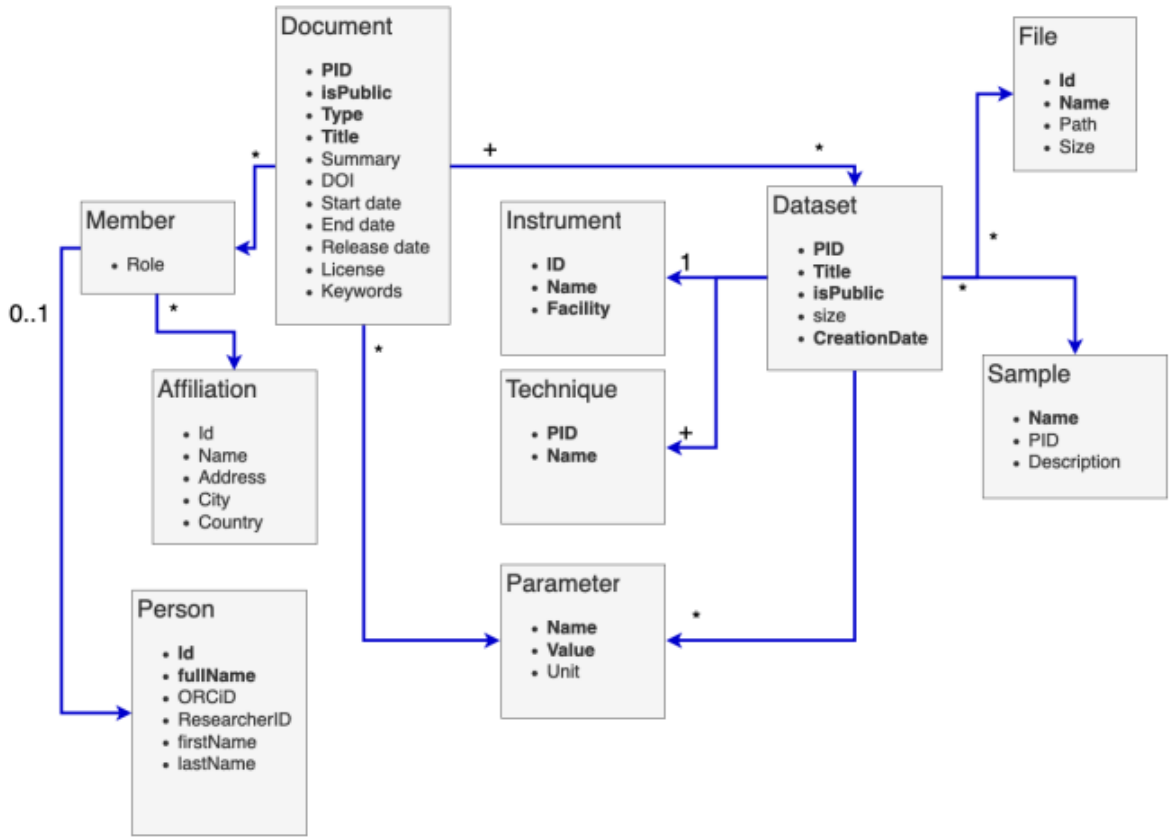


Figure 4: UML class diagram of the PaN search-API data model.

Classes marked with an asterisk (*) are optional, but can be present multiple times. What is marked with a plus sign (+) is required at least once. Numbers (1) or ranges (0…1) indicate a single required instance or zero or one instance respectively. Short description of each of the classes:

"... **Document**: Represents a proposal, beamtime, measurement campaign, a (data) publication, or groups of datasets for a specific sample. A curated list of document types may come out of the ontology task in Work Package 3.

Dataset: A dataset combines information about an experimental run, including optional File, Sample, Instrument and Technique. The granularity is so that this should be normally **the smallest unit that can be analysed sensibly**. That may require including multiple files, for example multiple images from a tomography run.

File: Reference to a data file, which can be used for further PaNOSC services, like analysis, visualisation, transfer etc.

Instrument: Experimental station where an experiment took place. In order to provide a consistent view of the data, the choice was to only allow **a single entity per dataset**. Some facilities wish to express finer granularity and could distinguish between end stations, beamlines, sample environment equipment or detectors. But from the user point of view a consistent level of detail in the information returned is desirable.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Technique: The experimental method used. At least one value is required. The list currently being created in the ontology task […] will include sufficient generic choices, like “neutron” or “scattering”, that should match any data from a particular instrument or beamline. This way, legacy data where the specific intentions of the experiment may not have been recorded, can also receive a technique label that still provides the user benefit over having no such information. The aim of the ontology task is to come up with a hierarchical or inclusive scheme that would allow matching related techniques, for example recognise “absorption spectroscopy” as a “spectroscopy” technique.

Sample: Substance, material or object probed by neutrons or photons in the experiment.

Parameter: The sample and the technique will be some of the **most frequently used search terms**, according to the use cases that have been sampled. In addition the dataset, instrument, file or sample can have a number of parameters that may be useful to further filter on. How these parameters are associated in the individual data catalogues at partner sites depends on the choices made there. For simplicity **this API attaches parameters exclusively to datasets and documents**, as these are the main search endpoints (see below). The ontology task will curate a list of parameter keywords that the partners will then map onto what is stored in their catalogues (now and in future). Typical examples would be:

- sample temperature
- sample size or thickness
- photon energy
- neutron wavelength
- total number of counts

In the ontology discussions these will receive a single unique name. What is currently in use by the API demonstrator and test cases is for illustration only, like for the technique, roles, etc. Parameter values are **scalar measurement values with units**. **Strings are also permitted**. We rely on JSON using double quotes for strings, for example { "name": "detector1_name", "value": "incoming_beam" } versus { "name": "detector1_data", "units": "A", "value": 3.38e-05 }) to distinguish either.

Member: An individual associated with the data in a role defined by the role property, for example the principal investigator of the experiment, or a person involved in the data analysis. The allowed values of the role property also come out of the ontology task

Person: An individual associated with the data in a role defined by the Member class, for example the principal investigator of the experiment, or a person involved in the data analysis.

Affiliation: Home institution of a member…”[31]

It is noted that **all the fields in Figure 4 should support querying** in order to be returned by PaN search-API. Some of the ontologies mentioned in the description are not yet implemented. For more details, the section “Documentation of the API” in the referenced document provides more details, along with the API endpoints.

[31] Ibid.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

The PaN search-API services support some **plugins** (two at the time of writing). One, the **scoring service**,[32] which enables ranking of the results based on their pertinence to the query of the user; a second, the **PaN ontologies API service**,[33] takes care of the logic defined in ExPaNDS Deliverable 3.2[34] and enriches the search capabilities of the API. This means that, as long as datasets are labelled with the PID of the technique from the PaNET ontology, the PaN ontology API plugin applies the PaNET ontology to the user's search and returns the results accordingly. This corresponds to the ontology task mentioned in the description of the Technique class of Figure 4.

Table 4 (see below) drafts a mapping between the fields exposed by the PaN search-API and the metadata framework (see Figure 1) at the time of writing. Note that data catalogue implementations (e.g. ICAT, SciCat) often have richer data, and thus, a user can find more information if properly redirected by the PaN search-API service to the data catalogue where the richer metadata resides.

[32] https://github.com/panosc-eu/scoring
[33] https://github.com/ExPaNDS-eu/pan-ontologies-api
[34] https://doi.org/10.5281/zenodo.4806026

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	Metadata framework field	PaN search-API field	Occurrence	Allowed Values	Comments and Issues
Proposal	PI/Main Proposer	Member with role from ontology	Zero or one	Object	In the case the document type is a proposal
	Co-Investigators	Member with role from ontology	Zero or many	Textual	In the case the document type is a proposal
	Instrument requested	Instrument	1	Object	In the case the document type is a proposal
	Sample description	Sample	Zero or many	Object	In the case the document type is a proposal
	Proposed experimental conditions	Parameter	Zero or many	Object	In the case the document type is a proposal
	Safety conditions	Parameter	Zero or many	Object	In the case the document type is a proposal
	Experiment description	Document.summary	One or many	Textual	It can include the experiment description.
	Facility information	Affiliation	One or many	Object	Member's affiliation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.
 DOI: 10.5281/zenodo.6799106

Metadata Type	Metadata framework field	PaN search-API field	Occurrence	Allowed Values	Comments and Issues
Scheduling	Allocated day and time on instrument	Document.start/end Date	1	Date	In the case the document is a proposal or beamtime
	Scheduled visiting experimental team	Member with role from ontology	Zero or many	Object	In the case the document type is a proposal or a beamtime
	Sample preparation	Sample.description	Zero or many	Textual	The sample description can contain info about the sample preparation
Experiment	Visiting experimental team	Member with role from ontology	Zero or one	Object	
	Experiment date	Dataset.creationDate	1	Date	Date of the collection of the dataset
	Sample information	Sample.description	Zero or many	Textual	
	Instrument information	Instrument.name	1	Textual	It overlaps with row 3. It is currently not possible to distinguish between the instrument requested during proposal and the one at experiment time
	Calibration information	Parameter	Zero or many	Object	Parameters can contain calibration data. A list of common agreed parameters can be found in the referenced PaNOSC deliverable

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	Metadata framework field	PaN search-API field	Occurrence	Allowed Values	Comments and Issues
	Instrument scientist	Member with role from ontology	Zero or many	Object	
Storage	Persistent identifier	dataset.pid	1	Textual	
	Dataset information	Dataset	One or many	Object	
	File identifier	File	One or many	Object	
	Instrument parameters	Parameter	Zero or many	Object	Parameters can contain calibration data. A list of commonly agreed parameters can be found in the referenced PaNOSC deliverable
Data processing	Processing team	Member with role from ontology	Zero or one	Object	
Data publication	Resource identity	Document.DOI	1	Textual	In the case the document type is a data publication
	Creator	Member with role from ontology	Zero or one	Object	In the case the document type is a data publication
	Contributor	Member with role from ontology	Zero or many	Object	In the case the document type is a data publication

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	Metadata framework field	PaN search-API field	Occurrence	Allowed Values	Comments and Issues
	Title	Publication.title	1	Textual	In the case the document type is a data publication
	Publisher	Member with role from ontology	Zero or one	Object	In the case the document type is a data publication
	Publication year	Document.release date	1	Date	In the case the document type is a data publication
	Licence	Publication.licence	1	Textual	In the case the document type is a data publication
	Release date	Publication.release date	1	Date	In the case the document type is a data publication

Table 4: Metadata mapping between D2.2 framework and the PaN search-API

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.1.1 Metadata mapping between the PaN search-API and its ICAT and SciCat implementations

As the PaN search-API requires an implementation for each data catalogue flavour (e.g. ICAT, SciCat), we reference here links that include mappings, either explicit or implicit, between the PaN search-API and its ICAT[35,36,37] and SciCat[38,39,40] implementations. We think the references serve the purpose better, as they dynamically incorporate any future change and keep track of the changes that might occur in the mapping. More details are available in the ExPaNDS deliverable D3.3.[41]

This mapping is fundamental for the users to query different flavours of data catalogues, since ICAT and SciCat search-API are needed to translate the fields queried from the user to the actual metadata fields in the data catalogue.

The reader, given Table 4 and the provided links, can recursively map the PaN search-API fields to the ones of D2.7.

3.2 EOSC Data Indexing and Discovery Services

Researchers are likely familiar with key data repositories that cover their own field of expertise, but at times, they may need to broaden their scope to explore heterogenous data coming from other techniques and/or areas of study related to their work. An example would be a researcher interested in the latest technologies for building batteries who also requires data on the impacts of these technologies on the environment. This researcher will have to find data ranging from the structure and stability of various types of batteries to societal and global environmental data. Finding, understanding and browsing the different domain specific repositories is potentially a very time-consuming task, which can be greatly simplified by the use of discovery tools, whose purpose is to make disparate scientific data findable in a single place.

3.2.1 B2FIND

EUDAT-B2FIND (B2FIND)[42] is a multidisciplinary indexing service and data portal managed by the European Data Infrastructure (EUDAT)[43] that can harvest metadata from providers (e.g. PaN RIs) in order to achieve the aforementioned goal. Metadata is harvested via provider

[35] https://github.com/ral-facilities/datagateway-api/#mapping-between-panosc-and-icat-data-models
[36] https://github.com/ral-facilities/datagateway-api/blob/main/datagateway_api/search_api_mapping.json.example
[37] https://github.com/ral-facilities/datagateway-api/blob/main/datagateway_api/src/search_api/panosc_mappings.py
[38] https://github.com/SciCatProject/panosc-search-api/blob/master/common/mappings.js
[39] https://github.com/SciCatProject/panosc-search-api/blob/master/common/filter-mapper.js
[40] https://github.com/SciCatProject/panosc-search-api/blob/master/common/response-mapper.js
[41] Gonzalez-Beltran, A., Minotti, C., Davies, L. et al.. (2022). Demonstrate ICAT and SciCat released with APIs compatible with ExPaNDS federated EOSC services (1.1). https://doi.org/10.5281/zenodo.6363591
[42] See http://b2find.eudat.eu/
[43] See https://eudat.eu/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

endpoints using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[44] With the B2FIND **OAI-PMH**, there is the **option to provide metadata in either** Dublin Core[45] **or the** DataCite[46] metadata schemas.

B2FIND has its own metadata schema. The schema derives from DataCite, and thus, this makes B2FIND's metadata schema broadly compatible with that of OpenAIRE (see Section 3.2.2 for more on OpenAIRE). A notable difference, however, is that B2FIND incorporates the additional elements of Discipline, Instrument (i.e. especially relevant to PaN RIs) and TemporalCoverage.

For convenience, Table 5 below reproduces the B2FIND metadata schema.[47] As the information in the table indicates, there are 26 elements in the B2FIND metadata schema, gathered under four broad categories:

1. General Information
2. Identifier
3. Provenance
4. Representation

Each element has a specified number of allowed occurrences and values, as well as a level of obligation. These levels of obligation include:

- Mandatory (M): properties must be provided.
- Mandatory if applicable: (M/A): if your metadata contains this value, you must provide it.
- Recommended (R): properties are optional, but strongly recommended for interoperability and higher quality of the metadata.
- Optional (O): properties are optional and provide richer description.

Where possible, providers are strongly encouraged to supply both mandatory and all recommended and optional metadata, as this increases the chances of metadata being found by those searching B2FIND.

[44] Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/
[45] Dublin Core Metadata Initiative (DCMI) (2020). DCMI metadata terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[46] DataCite Metadata Working Group (2019). DataCite metadata schema 4.3. https://schema.datacite.org/
[47] EUDAT-B2FIND (2019). EUDAT-B2FIND Metadata Schema. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrence	Allowed Values	Comments and Issues
General Information	Community (M)	The scientific community, research infrastructure, project or data provider from which B2FIND harvests the metadata.	1	Textual	
	Title (M)	A name or a title by which a resource is known	1-n	Textual	
	Description (R)	All additional information that does not fit in any of the other categories. May be used for technical information. Could be an abstract, a summary or a table of content. It is good practice to supply a description.	0-1	Textual	
	Keywords (R)	Subject, keyword, classification code, or key phrase describing the resource.	0-n	List of strings	Try to use keyword thesauri from community-specific vocabularies.
Identifier	DOI (M/A)	A persistent citable identifier that uniquely identifies a resource.	0-1	Must be resolvable URI, registered at DataCite as DOI.	**At least one resource identifier is mandatory.**
	PID (M/A)	A persistent identifier that uniquely identifies a resource.	0-1	Must be resolvable URI, registered at a handle server.	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrence	Allowed Values	Comments and Issues
	Source (M/A)	An identifier that uniquely identifies a resource. It may link to the data itself or a landing page that curates the data.	0-1	Should be resolvable URI.	
	RelatedIdentifier (O)	Identifiers of related resources.	0-n	Should be resolvable URI.	
	MetadataAccess (R)	Link to the originally harvested metadata record.	0-1	Should be resolvable URI.	Automatically generated by B2FIND script (GetRecord request for OAI-PMH).
Provenance	Creator (R)	The main researchers involved working on the data, or the authors of the publication in priority order. May be a corporate/institutional or personal name.	0-n	The personal name format should be: family, given. Non-roman names may be transliterated according to the ALA-LC schemes.	Examples: Smith, John; Miller, Elizabeth.
	Publisher (M)	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role.	1-n		Examples: World Data Center for Climate (WDCC); GeoForschungsZentrum Potsdam (GFZ); Geological Institute, University of Tokyo, GitHub

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrence	Allowed Values	Comments and Issues
	Contributor (O)	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.	0-n	List of names	
	Instrument (O)	The technical instrument(s) used to generate, observe or measure the data.	0-n	Could be instrument ID (or name) and hosting facility name.	
	PublicationYear (M)	Year when the data is made publicly available. If an embargo period has been in effect, use the date when the embargo period ends.	1	UTC Year format (YYYY)	
	FundingReference (O)	Information about financial support (funding) for the resource.	0-n	Could be funder name or grant number.	
	Rights (R)	Any rights information for this resource.	0-n	Textual	
	OpenAccess (M/A)	Information on whether the resource is openly accessible or not.	1	Boolean	Automatically generated by B2FIND script based on the information given in "Rights" element. Default value is "True" unless stated otherwise.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrence	Allowed Values	Comments and Issues
	Contact (O)	A reference to contact information for this resource.	0-n	List of names	
Representation	Language (R)	Language(s) of the resource.	0-n	Allowed values are ISO 639-1 or ISO 639-3 language codes or text.	Examples: en; eng; English
	ResourceType (R)	The type(s) of the resource.	0-n	Free text	Examples: Dataset; Image; Audiovisual
	Format (R)	Technical format of the resource.	0-n	Textual	Use file extension or MIME type where possible, e.g. PDF, XML, MPG or application/pdf, text/xml, video/mpeg.
	Size (O)	Size information about the resource.	0-n	Free text	Examples: 15 pages; 6 MB; 45 minutes.
	Version (O)	Version information about the resource.	0-n	Suggested practice: track major_version.minor_version.	Example: v1.02
	Discipline (M)	The research discipline(s) the resource can be categorised in.	1-n	Controlled vocabulary, see b2find_disciplines.json.	If not applicable, add community specific discipline term.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrence	Allowed Values	Comments and Issues
	Spatial Coverage (O)	The spatial coverage the research data is related to. Content of this category is displayed in plain text. If a longitude/latitude information is given it will be displayed on the map.	0-1	Geographical coordinates • lat/lon for point • [min_lat,min_lon, max_lat, max_lon] for bounding box • or free text.	Recommended, in accordance with DataCite: Use WGS 84 (World Geodetic System) coordinates. Use only decimal numbers for coordinates. Longitudes are -180 to 180(0 is Greenwich, negative numbers are west, positive numbers are east), Latitudes are -90 to 90 (0 is the equator; negative numbers are south, positive numbers north).
	Temporal Coverage (O)	Period of time the research data itself is related to. Could be a date format or plain text.	0-1	YYY,YYYY-MM-DD, YYYY-MM-DDThh:mm:ssTZD or any other format or level of granularity described in W3CDTF24.	Use RKMS-ISO860125 standard for depicting date ranges. Example: 2004-03-02/2005-06-02.Years before 0000 must be prefixed with a - sign, e.g. -0054 to indicate 55 BC. You can also use plain text, e.g. Viking Age.

Table 5: B2FIND Metadata Schema[48]

[48] EUDAT-B2FIND (2019). EUDAT-B2FIND Metadata Schema. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.2.2 OpenAIRE

OpenAIRE Explore (OpenAIRE)[49] is similar to B2FIND in terms of being an indexing and discovery tool that harvests metadata via OAI-PMH; however, unlike B2FIND, OpenAIRE indexes more than just data (e.g. it also indexes publications, software, and other research outputs). As with B2FIND, it is possible to provide metadata to OpenAIRE via a provider OAI-PMH endpoint; however, the DataCite metadata schema must be used, and Dublin Core is not supported.

Additionally, OpenAIRE harvests metadata from B2FIND,[50] meaning that if a provider's metadata is harvested by B2FIND, then it will also be harvested by OpenAIRE (although not necessarily exposed by the service).[51] OpenAIRE provides a Content Providers Dashboard to entities registered with it service that offers providers front end access to back end services, including metadata enrichment and usage statistics, that are not available if providers do not register directly but instead rely on indirect harvesting of their metadata into OpenAIRE from B2FIND.[52]

OpenAIRE uses the DataCite metadata schema v3.1[53] with a few minor adjustments.[54] The complete DataCite metadata schema v3.1 is too extensive to reproduce here; however, it is useful to highlight where variations from the DataCite schema are made by OpenAIRE.

Table 6 below lists the DataCite metadata schema properties and supplies a comment where adjustments to these have been made in the OpenAIRE application profile.[55] Note that, similarly to the B2FIND elements, the OpenAIRE properties have allowed numbers of occurrences and values (not shown in Table 6) and each property is associated with a level of obligation:

- Mandatory (M): the field must always be present in the metadata record. An empty element is not allowed.
- Mandatory when applicable (MA): when the value of the field can be obtained it must be present in the metadata record.
- Recommended (R): the use of the field is recommended.
- Optional (O): the property may be used to provide complementary information about the resource

[49] See https://explore.openaire.eu
[50] OpenAIRE (2022). B2FIND. https://explore.openaire.eu/search/dataprovider?datasourceId=re3data_____::730f562f9efe8a3b3742d2da510d4335
[51] For further information on dataset exposure in OpenAIRE, see https://guidelines.openaire.eu/en/latest/data/use_of_datacite.html , Section 'Related publications and dataset information'.
[52] OpenAIRE (2022). PROVIDE - How to validate and register your data source. https://www.openaire.eu/validator-registration-guide
[53] DataCite Metadata Working Group (2014). DataCite Metadata Schema 3.1. https://schema.datacite.org/meta/kernel-3.1/
[54] These minor adjustments are detailed in the OpenAIRE guidelines: OpenAIRE (2022). Use of DataCite. Section 'What's different'. https://guidelines.openaire.eu/en/latest/data/use_of_datacite.html
[55] OpenAIRE (2022). Application Profile Overview. https://guidelines.openaire.eu/en/latest/data/application_profile.html

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Property	Comment
1. Identifier (M)	
1.1 identifierType (M)	Unlike DataCite, OpenAIRE allows for DOIs and other types of identifiers.
2. Creator (M)	
2.1 creatorName (M)	
2.2 nameIdentifier (R)	OpenAIRE recommends including a nameIdentifier such as an ORCID or a ISNI if available.
2.2.1 nameIdentifierScheme (R)	
2.2.2 schemeURI (R)	
2.3 affiliation (R)	
3. Title (M)	
3. titleType (O)	
4. Publisher (M)	
5. PublicationYear (M)	
6. Subject (R)	
6.1 subjectScheme (O)	
6.2 schemeURI (O)	
7. Contributor (MA/O)	OpenAIRE uses this property and sub-properties to allow unique and persistent identification of the funder who has funded wholly or partly the dataset described. This does not exclude also using this property for additional contributors as defined by DataCite Metadata Schema v3.1. See Funding information.
7.1 contributorType (MA/O)	
7.2 contributorName (MA/O)	
7.3 nameIdentifier (MA/O)	
7.3.1 nameIdentifierScheme (MA/O)	
7.3.2 schemeURI (O)	
7.4 affiliation (O)	
8. Date (M)	*Mandatory* property in OpenAIRE instead of *recommended* in DataCite. See Embargo date information.
8.1 dateType (M)	
9. Language (R)	
10. ResourceType (R)	
10.1 resourceTypeGeneral (R)	
11. AlternateIdentifier (O)	
11.1 alternateIdentifierType (O)	
12. RelatedIdentifier (MA)	*Mandatory when applicable* property in OpenAIRE instead of *recommended* in DataCite. See Related publications and datasets information.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Property	Comment
12.1 relatedIdentifierType (M)	
12.2 relationType (M)	
12.3 relatedMetadataScheme (O)	
12.1 schemeURI (O)	
12.1 schemeType (O)	
13. Size (O)	
14. Format (O)	
15. Version (O)	
16. Rights (MA)	*Mandatory when applicable* property in OpenAIRE instead of *recommended* in DataCite. See Access rights and license information.
16.1 rightsURI (MA)	
17. Description (MA)	*Mandatory when applicable* property in OpenAIRE instead of *recommended* in DataCite.
17.1 descriptionType (MA)	
18. GeoLocation (O)	
18.1 geoLocationPoint (O)	
18.2 geoLocationBox (O)	
18.3 geoLocationPlace (O)	

Table 6: OpenAIRE Application Profile (metadata schema)[56]

3.2.3 Metadata mapping between the Dublin Core and DataCite schemas and the ICAT and SciCat data catalogues

Every flavour of a data catalogue, in this context, ICAT and SciCat, need to translate the fields defined by the Dublin Core and DataCite schemas to match their internal storing formats. This is conceptually very similar to Section 3.1.1 of this document, and, as before, we reference here the links, either explicit or implicit, to the two main data catalogues implementations, ICAT and SciCat, by providing, in order, the mappings between Dublin Core and ICAT;[57] DataCite

[56] OpenAIRE (2022). Application Profile Overview. https://guidelines.openaire.eu/en/latest/data/application_profile.html
[57] https://github.com/icatproject/icat.oaipmh/blob/master/src/main/config/oai_dc_transformer.xsl.example

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

and ICAT;[58] Dublin Core and SciCat;[59] and between DataCite and SciCat.[60] As aforementioned, we take advantage of the property of these links to reflect future changes.

3.3 Mapping the ExPaNDS Metadata Framework to the Metadata Schemas of B2FIND and OpenAIRE

3.3.1 The purpose of B2FIND and OpenAIRE and the implication of this for mapping the metadata framework

As described in Section 3.2, provider endpoints are one necessary aspect for harvesting via OAI-PMH. The other key requirement is a mapping of the metadata used by the provider to the metadata schema used by the indexing service (i.e. that is doing the harvesting). As reviewed in Section 1.1, ExPaNDS has produced a metadata framework for PaN RIs that covers the collection of metadata across the experimental lifecycle to enable FAIR data.[61] The framework prioritises metadata types according to their relevance to FAIR (essential, important, useful) and also sets out which metadata types relate to which aspect(s) of FAIR (F, A, I, R).

The aim of the ExPaNDS metadata framework is to promote the collection of as rich and as complete metadata as possible to support the production of FAIR PaN data. In particular, availability of this metadata allows expert PaN users to have the best opportunity not only of finding data of the most use to them (for example, through specific searches on parameters or sample info) but also of reusing data (i.e. because they have sufficient contextual information to understand the data fully). This said, the framework is also designed to capture the metadata that would be most useful in more generic contexts and for those beyond the PaN domain, including researchers in other disciplines. It is precisely this non-domain specific context and heterogeneous user group that B2FIND and OpenAIRE seek to target with their indexing services and discovery portals.

Thus, in this sense, B2FIND and OpenAIRE are fundamentally not designed to cater for the domain specialist expert. And, indeed, their metadata schemas reflect this point, in that they contain fewer elements/properties and therefore, less detail overall in the metadata than, for example, the ExPaNDS metadata framework. As a result, PaN scientists are unlikely to be able to find the same extent of information through these generic tools than they might find through the PaN-specific tools (e.g. facility metadata catalogues, PaN Search API). Nonetheless, the information made available through B2FIND and OpenAIRE should 1.) be sufficient for, at least, the initial enquiries of a non-domain specialist; and, importantly, 2.) be able to point that user to where they can find further details. When we consider the ExPaNDS metadata framework in light of these two points, then it makes sense to recognise that not every metadata field in our framework will or needs to map to a corresponding field in the B2FIND or OpenAIRE metadata schemas.

[58] https://github.com/icatproject/icat.oaipmh/blob/master/src/main/config/oai_datacite_transformer.xsl.example
[59] https://github.com/SciCatProject/oai-provider-service/blob/master/src/providers/scicat-provider/repository/scicat-dc-mapper.ts#L33-L67
[60] https://github.com/SciCatProject/oai-provider-service/blob/master/src/providers/scicat-provider/repository/openaire-mapper.ts#L32-L134
[61] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.3.2 Must all PaN RIs map their metadata the same way?

If we accept as a given the point that we need only map parts of the ExPaNDS framework to the metadata schemas of B2FIND/OpenAIRE, then another question arises: should each individual PaN RI be expected to map their metadata to the metadata schemas of B2FIND and OpenAIRE in the exact same way? After all, the ExPaNDS metadata framework is intended to act as a guideline across all of the national European PaN RIs.

To address this question, it is helpful to reflect on the prioritisation element of the metadata fields in the ExPaNDS framework. Even for those fields prioritised as 'essential', we know from feedback from ExPaNDS partners that, due to a range of practical concerns (resources, other plans, RI strategy), ExPaNDS partner facilities do not all move forward with implementations in the same order or to the same timeline.[62,63] Thus, even for the 'essential' fields of the metadata framework, there may well be considerable variation around what is collected, certainly at present and in the short to medium term.

In the longer term, however, as implementations are put in place at the different ExPaNDS partner facilities, we might expect to see more consistency in terms of how PaN experimental lifecycle metadata are mapped to the metadata schemas of B2FIND and OpenAIRE. Nonetheless, some differences will very likely remain due to local practices at facilities; for example, a facility might have a policy of including organisational affiliations rather than personal names in their metadata, and that facility may discuss this approach with B2FIND and agree a mapping accordingly.[64]

3.3.3 Differences and possibilities in mapping PaN metadata (examples from B2FIND)

As such, in answer to the question posed above about whether all PaN RIs must map their data in the same way, while the argument that we should aim for as much consistency as possible seems fair, we cannot necessarily expect every PaN RI to follow **exactly** the same mapping. And indeed, this is exactly what we see in practice.

For example, consider the following three metadata records from PSI (Figure 5),[65] HZDR (Figure 6),[66] and ESS [i.e. a PaNOSC[67] partner] (Figure 7)[68] that are currently available in B2FIND. Note that for brevity, we omit the titles and abstracts (equating to the B2FIND metadata schema elements 'Title' and 'Description') from the three B2FIND landing page screen grabs shown in the figures below.

---

[62] McBirnie, A., Matthews, B., Gagey, B. et al. (2021). ExPaNDS D2.3: Final Data Policy Framework for Photon and Neutron RIs. https://doi.org/10.5281/zenodo.5205825

[63] ExPaNDS (2022). ExPaNDS workshop on FAIR metadata for PaN RIs. 2 March 2022.

[64] In the case of B2FIND, the mapping can result from a discussion between the provider and B2FIND. See B2FIND (2019). Mapping onto EUDAT-B2FIND Metadata Schema. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema

[65] EUDAT-B2FIND (2022). Partial screen grab for the record: Ihli, J. (2022). Dataset: Sparse ab initio x-ray transmission spectrotomography for nanoscopic compositional analysis of functional materials. http://b2find.eudat.eu/dataset/fb1bce5e-ee34-5a26-90d3-de9c2f6f6417

[66] EUDAT-B2FIND (2022). Partial screen grab for the record: Neumann-Kipping, M. and Hampel, U. (2019). Ultrafast X-ray tomography image data of bubbly two-phase pipe flow around a ring-shaped constriction. http://b2find.eudat.eu/dataset/45890b2f-a6d6-5719-8981-5283554f39a0

[67] Photon and Neutron Open Science Cloud (PaNOSC) project (2018-2022) funded by the European Commission under the H2020-EU.1.4.1.1. programme Grant Agreement 823852 www.panosc.eu

[68] EUDAT-B2FIND (2022). Partial screen grab for the record: Pfeiffer, D. (2018). Sample data from NMX. http://b2find.eudat.eu/dataset/5e887124-5cad-5d57-ba91-208862b5f82e

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

PaN

Identifier	
DOI	https://doi.org/10.16907/65d49f7e-578d-4b20-815d-71d845fb9dfe
Metadata Access	https://doi.psi.ch/oaipmh/oai?verb=GetRecord&metadataPrefix=oai_dc&identifier=10.16907/65d49f7e-578d-4b20-815d-71d845fb9dfe

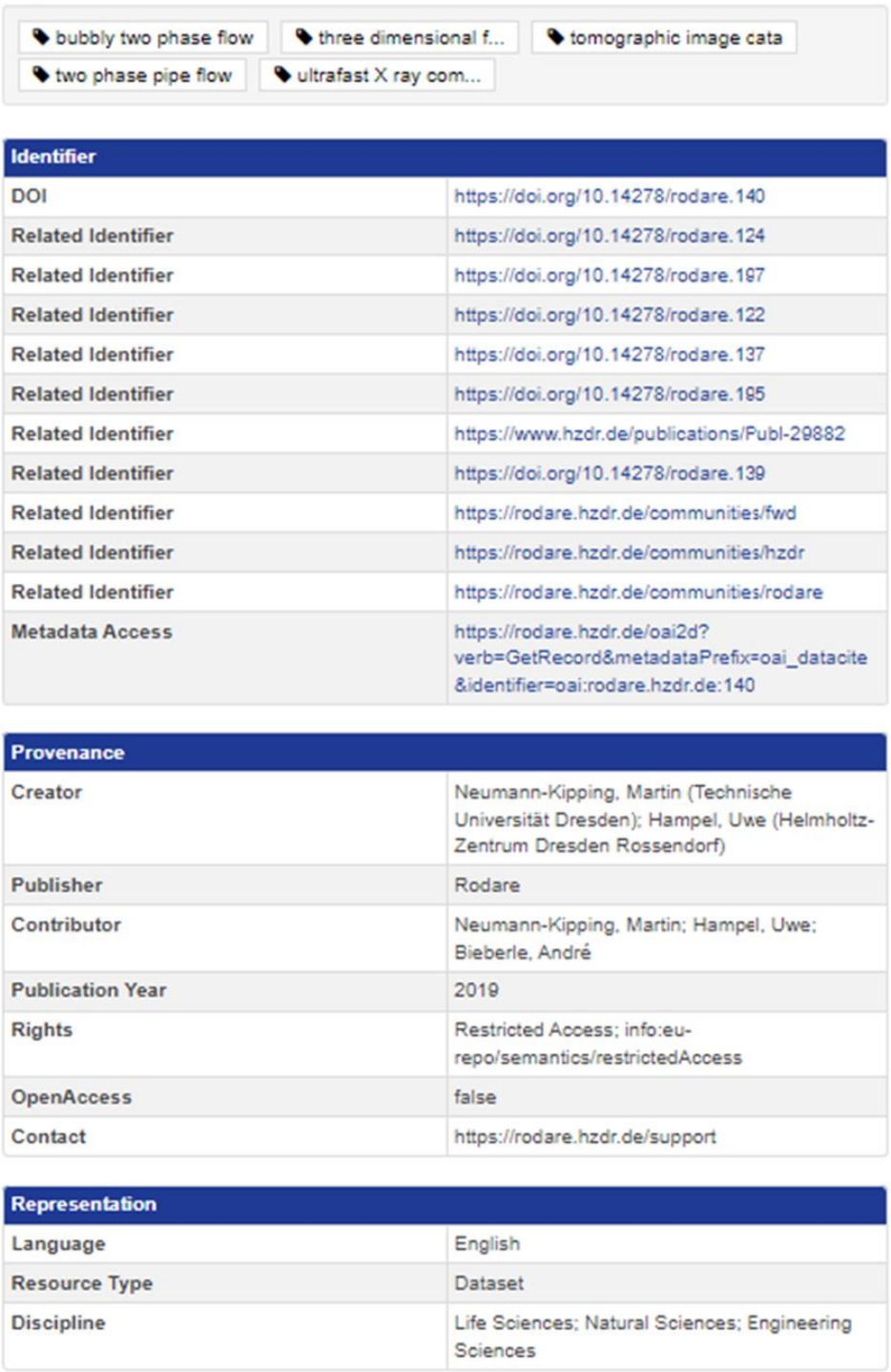
Provenance	
Creator	Johannes Ihli
Publisher	PSI
Publication Year	2022
Rights	Available to the public.
OpenAccess	true
Contact	PSI

Representation	
Resource Type	dataset
Discipline	Life Sciences; Biology; Basic Biological and Medical Research

Figure 5: Screen grab of the PSI metadata record in B2FIND for the dataset "Sparse ab initio x-ray transmission spectrotomography for nanoscopic compositional analysis of functional materials" http://b2find.eudat.eu/dataset/fb1bce5e-ee34-5a26-90d3-de9c2f6f6417[69]

[69] EUDAT-B2FIND (2022). Partial screen grab for the record: Ihli, J. (2022). Dataset: Sparse ab initio x-ray transmission spectrotomography for nanoscopic compositional analysis of functional materials. http://b2find.eudat.eu/dataset/fb1bce5e-ee34-5a26-90d3-de9c2f6f6417

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.



bubbly two phase flow | three dimensional f... | tomographic image cata | two phase pipe flow | ultrafast X ray com...

Identifier	
DOI	https://doi.org/10.14278/rodare.140
Related Identifier	https://doi.org/10.14278/rodare.124
Related Identifier	https://doi.org/10.14278/rodare.197
Related Identifier	https://doi.org/10.14278/rodare.122
Related Identifier	https://doi.org/10.14278/rodare.137
Related Identifier	https://doi.org/10.14278/rodare.195
Related Identifier	https://www.hzdr.de/publications/Publ-29882
Related Identifier	https://doi.org/10.14278/rodare.139
Related Identifier	https://rodare.hzdr.de/communities/fwd
Related Identifier	https://rodare.hzdr.de/communities/hzdr
Related Identifier	https://rodare.hzdr.de/communities/rodare
Metadata Access	https://rodare.hzdr.de/oai2d?verb=GetRecord&metadataPrefix=oai_datacite&identifier=oai:rodare.hzdr.de:140

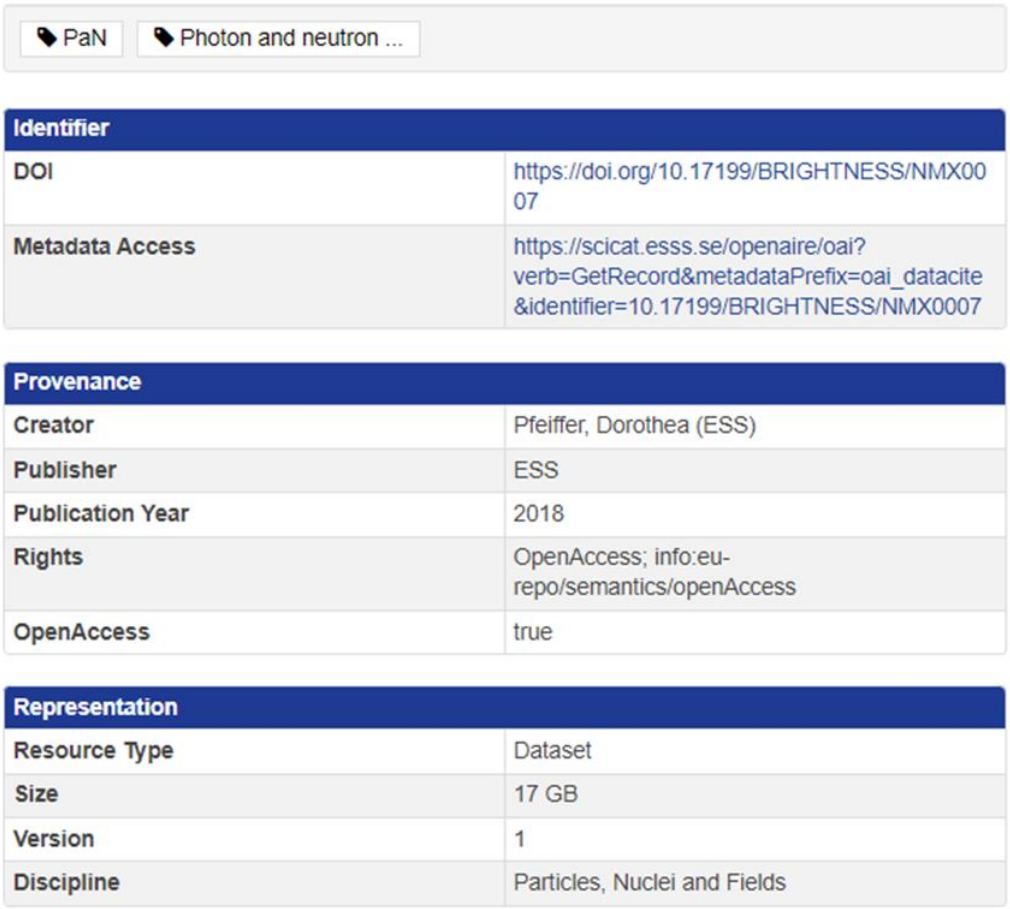
Provenance	
Creator	Neumann-Kipping, Martin (Technische Universität Dresden); Hampel, Uwe (Helmholtz-Zentrum Dresden Rossendorf)
Publisher	Rodare
Contributor	Neumann-Kipping, Martin; Hampel, Uwe; Bieberle, André
Publication Year	2019
Rights	Restricted Access; info:eu-repo/semantics/restrictedAccess
OpenAccess	false
Contact	https://rodare.hzdr.de/support

Representation	
Language	English
Resource Type	Dataset
Discipline	Life Sciences; Natural Sciences; Engineering Sciences

Figure 6: Screen grab of the HZDR metadata record in B2FIND for the dataset "Ultrafast X-ray tomography image data of bubbly two-phase pipe flow around a ring-shaped constriction" http://b2find.eudat.eu/dataset/45890b2f-a6d6-5719-8981-5283554f39a0[70]

[70] EUDAT-B2FIND (2022). Partial screen grab for the record: Neumann-Kipping, M. and Hampel, U. (2019). Ultrafast X-ray tomography image data of bubbly two-phase pipe flow around a ring-shaped constriction. http://b2find.eudat.eu/dataset/45890b2f-a6d6-5719-8981-5283554f39a0

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.



PaN | Photon and neutron ...

Identifier	
DOI	https://doi.org/10.17199/BRIGHTNESS/NMX0007
Metadata Access	https://scicat.esss.se/openaire/oai?verb=GetRecord&metadataPrefix=oai_datacite&identifier=10.17199/BRIGHTNESS/NMX0007

Provenance	
Creator	Pfeiffer, Dorothea (ESS)
Publisher	ESS
Publication Year	2018
Rights	OpenAccess; info:eu-repo/semantics/openAccess
OpenAccess	true

Representation	
Resource Type	Dataset
Size	17 GB
Version	1
Discipline	Particles, Nuclei and Fields

Figure 7: Screen grab of the ESS metadata record in B2FIND for the dataset "Sample Data from NMX" http://b2find.eudat.eu/dataset/5e887124-5cad-5d57-ba91-208862b5f82e [71]

Even a cursory glance across these three records shows a difference in the number and types of metadata elements included in each. Recall from Section 3.2.1 that the inclusion of certain metadata elements is mandatory (either 'mandatory' or 'mandatory if applicable') within the B2FIND metadata schema. The 'mandatory' elements are limited to five: 'Community', 'Title', 'Publisher', 'Publication Year', and 'Discipline'. Of these, in the B2FIND user interface 'Title' is found at the top of the record landing page (and thus, is not shown in the figures above) and 'Community' is found under a separate tab on the landing page (also not shown in the figures above). However, we can see that the remaining three mandatory elements — 'Publisher', 'Publication Year', 'Discipline' — are indeed included in all three of the metadata records above.

Where differences in the three records appear, these are in relation to the 'mandatory if applicable' elements (i.e. because not all facilities collect the same metadata types, and therefore, have these subsequently available for harvesters) and the 'recommended' and 'optional' elements. While facilities do not have a choice about the first of these, i.e. in that they must include it if the element exists, they do have a choice about the 'recommended' and 'optional' elements. It is in relation to these elements that the ExPaNDS metadata framework

[71] EUDAT-B2FIND (2022). Partial screen grab for the record: Pfeiffer, D. (2018). Sample data from NMX. http://b2find.eudat.eu/dataset/5e887124-5cad-5d57-ba91-208862b5f82e

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

may be able to provide best practice guidelines about mapping PaN RI metadata to metadata schemas such as those of B2FIND and OpenAIRE.

Put simply, it is likely that, where possible, mapping to all of the 'optional' and 'recommended' elements/properties will provide the richest and most complete information possible to users of the B2FIND and OpenAIRE discovery portals. And indeed, we can see this playing out in practice in the three example records illustrated above:

- a **Related Identifier** offers searchers the opportunity to look further and follow a trail of information
- knowing the **Size** of a datafile allows a searcher to better determine how long it might take to download from the facility
- a searcher can determine whether or not a dataset is **Open Access**
- a **DOI** provides an easy way to cite the dataset
- a **Keyword** (i.e. 'Tag' in the example figures above) gives information about the topic of the dataset
- and so on.

Additionally and importantly, some of the 'optional' and 'recommended' elements offer potential for the use of controlled vocabularies. Indeed, B2FIND encourages exactly such an approach for the 'Keyword' element, for which it recommends, "Try to use keyword thesauri from community-specific vocabularies."[72] To this end, PaN providers could agree amongst themselves to draw on terms from the PaNET ontology[73] — for example, to provide keywords related to technique — or from other PaN community ontologies.[74] There is also the option to employ community-specific terms, which could again come from a controlled vocabulary, for the 'Discipline' metadata element.

The B2FIND 'Related Identifier' element can enable more rich information to be incorporated into the metadata record. For example, if PIDs for instruments come into common usage, PaN providers could ensure they always include the link to the instrument PID in the related identifier element of the B2FIND metadata. The same would hold true for any sample PIDs. Likewise, raw datasets could be linked to results datasets and related journal publications, assuming these all have PIDs.

Finally, the 'Description' element should not be overlooked in terms of its ability to supply rich information to searchers. Certainly, an abstract can be included. Additionally, the free text nature of the 'Description' element gives considerable range to include specifics of the method used, including details of the sample and measurement parameters, where providers can provide this information. As the DataCite (i.e. on which the B2FIND metadata schema is based – see Section 3.2.2) guidelines note: "It cannot be emphasized enough how valuable …

[72] B2FIND (2019). EUDAT-B2FIND Metadata Schema. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema
[73] See https://github.com/ExPaNDS-eu/ExPaNDS-experimental-techniques-ontology
[74] Collins, S. P., da Graça Ramos, S., Iyayi, D. et al. (2021). ExPaNDS ontologies v1.0. https://doi.org/10.5281/zenodo.4806026

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Description [is] to other scholars in finding the resource and then determining whether or not the resource, once found, is worth investigating further, re-using or validating."[75]

3.3.4 Initial guidelines for mapping the ExPaNDS metadata framework

With this overarching recommendation to map to 'optional' and 'recommended' elements where possible in mind (see Section 3.3.3), Table 7 below proposes our initial suggestions for mapping metadata types from the ExPaNDS metadata framework to the elements of the B2FIND metadata schema. In this example, we propose a mapping in relation to what is often considered the most primary type of dataset produced in PaN research: a 'raw' dataset. While each facility might formally define this type of dataset slightly differently,[76] it is to this type of dataset that, at present, the records found in many facility ICAT/SciCat metadata catalogues relate.

However, it is important to note that this is not always the case; for example, RODARE (HZDR) and edata (STFC/ISIS) include 'results' data. In many ways, these types of datasets are similar to raw datasets, so it should be possible to take the raw dataset example presented in Table 7 below and apply a similar approach to a results dataset without too much difficulty. Bear in mind, however, that it is likely that more metadata types from the processing and analysis stages of the ExPaNDS metadata framework will feature in the resulting mapping. For example, the inclusion of metadata types related to the analysis software and analysis methods used will likely be important to include in the mapping to the 'Description' element - whereas, for the raw dataset example illustrated below, such metadata types will not be relevant.

Other types of relevant datasets beyond raw data and results data can also be envisaged — for example, datasets relating to samples used in PaN research. At present, though, formal metadata records of such datasets are not a common reality in PaN, so we do not focus on them in the example we provide here. Finally, it is also important to note the example of data publications, which can relate to any type of dataset. Some facilities (e.g. HZB) do produce these along with the relevant metadata records. Again, as with mappings for raw, results, and other types of datasets, the ExPaNDS framework metadata types should be similarly applicable to the case of data publications.

The suggested mappings of the ExPaNDS metadata types (found in column 7 of Table 7) to the B2FIND metadata schema elements (found in column 2 of Table 7) are rarely one to one. Often, there are several possible ExPaNDS metadata types that could be mapped. Where the occurrence rules in the B2FIND metadata schema (see column 4 of Table 7) allow multiple occurrences, the multiple suggestions of ExPaNDS metadata types are included as AND/OR, i.e. to indicate that any number (sometime including zero) are possible to include. However, where only a single occurrence is allowed in the B2FIND schema yet there are multiple possible ExPaNDS metadata types, these metadata types are included as OR, i.e. to indicate that a choice must be made and only one of the types can be mapped.

[75] DataCite Metadata Working Group (2014). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 3.1. https://schema.datacite.org/meta/kernel-3.1/doc/DataCite-MetadataKernel_v3.1.pdf

[76] McBirnie, A., Matthews, B., Gagey, B. et al. (2021). ExPaNDS D2.3: Final Data Policy Framework for Photon and Neutron RIs. https://doi.org/10.5281/zenodo.5205825

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Since it is important to have a clear definition of each ExPaNDS metadata type listed in the table to really understand how and why these particular metadata types are suggested in the mapping, the full definitions of the included metadata types are provided for reference at the end of this deliverable in Appendix A. The definitions included are copied over exactly as they are written in the metadata framework presented in ExPaNDS D2.2.[77]

Accompanying each ExPaNDS metadata type listed in column 7 of Table 7 are the prioritisation for FAIR assigned to that metadata type, the aspects of FAIR (i.e. F,A,I,R) to which the metadata type is relevant, and the stage of the experimental life cycle during which the metadata type appears.[78] Note that, because the example mapping in Table 7 takes a raw dataset as its example, we see only the occasional metadata type from the data processing stage of the framework and none at all from the data analysis stage. The prioritisation and FAIR aspects related to each ExPaNDS metadata type are included to allow reflection on these in relation to the levels of obligation (i.e. M, M/A, R, O) found in the B2FIND schema.[79] Of course, it is not the intention of the levels of obligation to relate directly to FAIR; however, they do tell us something about which metadata elements B2FIND considers most important for searching and browsing (i.e. Findability) and how it prioritises the additional richness of metadata (i.e. 'recommended' elements versus 'optional' elements).

As always with such guidelines, there is likely to be an ongoing discussion and potential for disagreement about the details of the mapping presented in Table 7. Thus, the suggestions provided should not be seen as final in any way. Rather, our hope is that they will serve as a working starting point for future development. In particular, for cases where there is no equivalent ExPaNDS metadata type to map to a B2FIND element (i.e. a 'N/A- not included as a metadata type' entry in column 7 of Table 7), this raises the possibility that the ExPaNDS framework is missing metadata types that could be beneficial to include. For example, it might well make sense to add an equivalent to the B2FIND 'Resource Type' element to our framework. Most likely, this metadata type was overlooked in the development of the framework because it seemed too obvious, i.e. the entire framework focuses on data. Yet, in a generic context — for example, such as OpenAIRE, which includes a range of research output types — it does become important to have the capability to include the resource type as part of the metadata.

To avoid the present deliverable becoming overly long, we do not include a similar example mapping from the ExPaNDS framework to the OpenAIRE metadata schema. Instead, Table 8 provides detail on the compatibility between the schemas of B2FIND and OpenAIRE, as well as with Dublin Core and DataCite[80]. Using the information in Table 8, it should be possible to translate the mapping presented in Table 7 (i.e. ExPaNDS framework to B2FIND) to a mapping to OpenAIRE. This task is also potentially made easier for any facility already using DataCite by the fact that OpenAIRE uses the DataCite metadata schema, with only a few minor changes (see Section 3.2.2).

[77] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825
[78] Ibid.
[79] B2FIND (2019). EUDAT-B2FIND Metadata Schema. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema
[80] EUDAT-B2FIND (2019). Concordance with Other Standards. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
General Information	Community (M)	The scientific community, research infrastructure, project or data provider from which B2FIND harvests the metadata.	1	Textual		N/A – not included as a metadata type However, it is likely that Publisher [P1-FI; data publication/record stage] OR Facility Information (Name component only) [P1-F; proposal stage] can serve to provide this information in most cases.
	Title (M)	A name or a title by which a resource is known	1-n	Textual		Title [P1-F; data publication/record stage]
	Description (R)	All additional information that does not fit in any of the other categories. May be used for technical information. Could be an abstract, a summary or a table of content. It is good practice to supply a description.	0-1	Textual		There are many metadata types that could contribute to the Description element. Key metadata types not possible to include under other B2FIND metadata elements might include: Sample Information [P1-FR; experiment stage] AND/OR Sample [P1-F; proposal stage] AND/OR Experiment Date [P1-FA; experiment stage] AND/OR Experiment Description [P1-F; proposal stage]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
						AND/OR Experiment Planning [P2-FR; experiment stage] AND/OR Calibration Information [P1-FR; experiment stage] Note that the Description element is textual, it is possible to include many other additional metadata types found within the ExPaNDS metadata framework, especially for datasets other than raw data; for example, Software Package Information [P1-IR; analysis stage] for results datasets.
	Keywords (R)	Subject, keyword, classification code, or key phrase describing the resource.	0-n	List of strings	Try to use keyword thesauri from community-specific vocabularies.	N/A – not included as a distinct metadata type However, drawing out keywords from some metadata types in the ExPaNDS framework could be very helpful, for example, from the: Experiment Description [P1-F; proposal stage] – for technique, methods, sample keywords AND/OR Sample [P1-F; proposal stage] – for sample keywords AND/OR Instrument Information (Name, Organisation, ID components only) [P1-FR; experiment stage] – for instrument name keywords

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
						AND/OR Facility Information (Name component only) [P1-F; proposal stage] – for facility name keywords AS B2FIND indicates, it is preferable to use a controlled vocabulary for keywords, if possible.
Identifier	DOI (M/A)	A persistent citable identifier that uniquely identifies a resource.	0-1	Must be resolvable URI, registered at DataCite as DOI.	**At least one resource identifier is mandatory.**	Resource Identity [P1-FI; data publication/record stage] OR Persistent Identifiers [PI-FA; data storage stage]
	PID (M/A)	A persistent identifier that uniquely identifies a resource.	0-1	Must be resolvable URI, registered at a handle server.		Resource Identity [P1-FI; data publication/record stage] OR Persistent Identifiers [PI-FA; data storage stage]
	Source (M/A)	An identifier that uniquely identifies a resource. It may link to the data itself or a landing page that curates the data.	0-1	Should be resolvable URI.		Resource Identity [P1-FI; data publication/record stage] OR Persistent Identifiers [PI-FA; data storage stage]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
	RelatedIdentifier (O)	Identifiers of related resources.	0-n	Should be resolvable URI.		Related Resource [P2-F; data publication/record stage]
	MetadataAccess (R)	Link to the originally harvested metadata record.	0-1	Should be resolvable URI.	Automatically generated by B2FIND script (GetRecord request for OAI-PMH).	N/A – generated automatically by B2FIND
Provenance	Creator (R)	The main researchers involved working on the data, or the authors of the publication in priority order. May be a corporate/institutional or personal name.	0-n	The personal name format should be: family, given. Non-roman names may be transliterated according to the ALA-LC schemes.	Examples: Smith, John; Miller, Elizabeth.	Principal Investigator/Main Proposer [P1-FA; proposal stage] AND/OR Co-Investigators [P1-FA; proposal stage] AND/OR Instrument Scientist [P2-F; experiment stage] AND/OR Visiting Experimental Team [P1-FA; experiment stage] AND/OR Creator [P1-F; data publication/record stage] Note that there are several other metadata types in the framework that could also be included under the Creator element, especially for datasets other than raw data; for example, Analysis Team [P2-AIR; analysis stage] for results datasets.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
	Publisher (M)	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role.	1-n		Examples: World Data Center for Climate (WDCC); GeoForschungsZentrum Potsdam (GFZ); Geological Institute, University of Tokyo, GitHub	Publisher [P1-FI; data publication/record stage]
	Contributor (O)	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.	0-n	List of names		Contributor [P2-F; data publication/record] Note that, as with the Creator element there is the potential to include other metadata types under the Contributor element
	Instrument (O)	The technical instrument(s) used to generate, observe or measure the data.	0-n	Could be instrument ID (or name) and hosting facility name.		Instrument Information (Name, Organisation, ID components only) [P1-FR; experiment stage] AND/OR Facility Information (Name component only) [P1-F; proposal stage]
	PublicationYear (M)	Year when the data is made publicly available. If an embargo period has been in effect, use the	1	UTC Year format (YYYY)		Release Date (Year component only) [P1–IR; publication/record stage]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
		date when the embargo period ends.				
	FundingReference (O)	Information about financial support (funding) for the resource.	0-n	Could be funder name or grant number.		Funding Source [P2-F; proposal stage]
	Rights (R)	Any rights information for this resource.	0-n	Textual		License [P1-IR; publication/record stage]
	OpenAccess (M/A)	Information on whether the resource is openly accessible or not.	1	Boolean	Automatically generated by B2FIND script based on the information given in "Rights" element. Default value is "True" unless stated otherwise.	N/A – not included as a metadata type However, as B2FIND generates this element automatically from the Rights element, it is important to ensure that if the dataset is still under embargo, the Release Date (Year component only) [P1 –IR; publication/record stage] metadata type is included under the Rights element. Bear in mind that if only the License [P1-IR; publication/record stage] is included under the Rights element, and the License type is open access, then if no Release Date information is present, B2FIND will assume the dataset is open access.
	Contact (O)	A reference to contact information for this resource.	0-n	List of names		N/A- not included as a metadata type In practice and with appropriate consent, it may make sense to map:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
						Principal Investigator/Main Proposer [P1-FA; proposal stage] to the Contact element. AND/OR Creator [P1-F; data publication/record stage]
Representation	Language (R)	Language(s) of the resource.	0-n	Allowed values are ISO 639-1 or ISO 639-3 language codes or text.	Examples: en; eng; English	N/A – not included as a metadata type
	ResourceType (R)	The type(s) of the resource.	0-n	Free text	Examples: Dataset; Image; Audiovisual	N/A – not included as a metadata type
	Format (R)	Technical format of the resource.	0-n	Textual	Use file extension or MIME type where possible, e.g. PDF, XML, MPG or application/pdf, text/xml, video/mpeg.	Representation Information [P3-IR; data storage stage] AND/OR Data Format [P1-IR; data processing stage]
	Size (O)	Size information about the resource.	0-n	Free text	Examples: 15 pages; 6 MB; 45 minutes.	N/A – not included as a distinct metadata type However, it is possible that information about size may be encompassed under the broader metadata types of (or some combination of these):

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
						Representation information [P3-IR; data storage stage] AND/OR Dataset information [P1-F; data storage]
	Version (O)	Version information about the resource.	0-n	Suggested practice: track major_version.minor_version.	Example: v1.02	N/A – not included as a distinct metadata type
	Discipline (M)	The research discipline(s) the resource can be categorised in.	1-n	Controlled vocabulary, see b2find_disciplines.json.	If not applicable, add community specific discipline term.	N/A – not included as a distinct metadata type Some of the comments under the Keyword element above may be relevant here, especially the point about controlled vocabularies.
	Spatial Coverage (O)	The spatial coverage the research data is related to. Content of this category is displayed in plain text. If a longitude/latitude information is given it will be displayed on the map.	0-1	Geographical coordinates • lat/lon for point • [min_lat, min_lon, max_lat, max_lon] for	Recommended, in accordance with DataCite: Use WGS 84 (World Geodetic System) coordinates. Use only decimal numbers for coordinates. Longitudes are -180 to 180(0 is Greenwich,	N/A – not included as a metadata type

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Metadata Type	B2FIND Name	Description	Occurrences	Allowed Values	Comments and Issues	ExPaNDS Framework: Metadata Type
				• bounding box • or free text.	negative numbers are west, positive numbers are east), Latitudes are -90 to 90 (0 is the equator; negative numbers are south, positive numbers north).	
	Temporal Coverage (O)	Period of time the research data itself is related to. Could be a date format or plain text.	0-1	YYY,YYYY-MM-DD, YYYY-MM-DDThh:mm:ssTZD or any other format or level of granularity described in W3CDTF24.	Use RKMS-ISO860125 standard for depicting date ranges. Example: 2004-03-02/2005-06-02.Years before 0000 must be prefixed with a - sign, e.g. -0054 to indicate 55 BC. You can also use plain text, e.g. Viking Age.	N/A – not included as a metadata type However, if a dataset is part of a continuous, ongoing long term experiment, it may be relevant to include relevant temporal coverage information from Experiment Description [P1-F; proposal stage] OR Experiment Date [P1-FA; experiment stage] Otherwise, for one-off experiments, there could also be an argument for using the Temporal element to include the actual experiment date, i.e. when the dataset was generated, although this is not the 'intended' use of the Temporal Coverage element. Experiment Date [P1-FA; experiment stage]

Table 7: Initial suggestions for mapping metadata types from the ExPaNDS framework to the B2FIND metadata schema.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Accompanying notes for Table 7: The example mapping presented in this table assumes the resource being mapped is a raw dataset. The first 6 columns are reproduced from a published table summarising the B2FIND metadata schema[81]. The metadata types listed in column 7 come from the ExPaNDS metadata framework presented in ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management, which provides the full definition for each metadata type.[82] The full definitions for the ExPaNDS metadata types listed in column 7 are also included at the end of the current deliverable as Appendix A.

[81] B2FIND (2019). EUDAT-B2FIND Metadata Schema. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema

[82] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

DataCite 4.3	B2FIND	OpenAIRE	Dublin Core	Comments and Issues
1. Identifier	Identifier [DOI or PID or Source (URL)]	1. Identifier	Identifier	While for DataCite a DOI is mandatory as identifier, B2FIND requires "only" at least an URL linked to the underlying data resource.
2.1 creatorName	Creator	2.1 creatorName	Creator	
3. Title	Title	3. Title	Title	
4. Publisher	Publisher	4. Publisher	Publisher	
5. PublicationYear	PublicationYear	PublicationYear	Date	
6. Subject	Keywords and/or Discipline	6. Subject	Subject	
7.1 contributorName	Contributor	7. Contributor	Contributor	
8. Date	PublicationYear or TemporalCoverage	8. Date	Date	The DataCite definition here is a bit vague (*Different dates relevant to the work*). B2FIND has the element *PubicationYear*, i.e. the year the dataset is published or when its embargo period ends. Another temporal element of B2FIND would be *TemporalCoverage*, i.e. the interval of time that the underlying data of the resource covers, with a useful 'Filter by time' search option associated on the B2FIND GUI.
9. Language	Language	9. Language	Language	
10. ResourceType	ResourceType	10. ResourceType	Type	
11. AlternateIdentifier	N/A	11. AlternateIdentifier	N/A	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

DataCite 4.3	B2FIND	OpenAIRE	Dublin Core	Comments and Issues
12. RelatedIdentifier	RelatedIdentifier	12. RelatedIdentifier	Relation or Source	
13. Size	Size	13. Size	N/A	
14. Format	Format	14. Format	Format	
15. Version	Version	15. Version	N/A	
16. Rights	Rights	16. Rights	Rights	
17. Description	Description	17. Description	Description	
18. GeoLocation	SpatialCoverage	18. GeoLocation	Coverage	In B2FIND *SpatialCoverage*, i.e. the geospatial coverage, is associated with a 'Filter by location' map search interface.
19. FundingReference	FundingReference	7. Contributor, 7.1 contributorType="Funder"	N/A	

Table 8: Compatibility and Mappings between B2FIND, OpenAIRE, DataCite v4.3, and Dublin Core metadata schemas.

Accompanying notes for Table 8: Reproduced from a B2FIND table entitled 'Concordance with Other Standards'.[83]

[83] EUDAT-B2FIND (2019). Concordance with Other Standards. http://b2find.dkrz.de/guidelines/mapping.html#b2fmdschema

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.4 Other Repositories

Another example of the advantage of common metadata is the crawling of datasets by Google Dataset search.[84] Google crawls periodically the landing pages of published datasets, leveraging on markups on the landing pages containing the metadata following schemas like DCAT[85] and schema.org.[86] The findings are indexed and made searchable via Google, along with the other crawled datasets and serve a heterogeneous audience. Guidelines on the required and optional metadata are available on the Google guides webpage.[87]

We point here, following the same file route of Sections 3.1.1 and 3.2.3, the mapping between DCAT and ICAT and schema.org and SciCat.[88]

Relations to other repositories are mainly made using PIDs in the metadata records that are linking to other published items like datasets or articles, e.g. what should be seen in the DOI DataCite metadata record of a published experiment/raw data collected in a measurement and later analysed should be:

- reference to processed/analysed – assigned with a PID e.g. DOI
- reference to journal article where the analysed data is published – assigned with a PID e.g. DOI

The processed/analysed dataset would need a reference to software that was used for data processing/analysis and the experiment/raw data it originated from.

As yet another example, the **Open Databases Integration for Materials Design** (OPTIMADE)[89] consortium aims to make materials databases interoperable by developing a specification for a common REST API. Therefore, they have developed an OPTIMADE REST API specification and some Python tools. Here,[90] a list of providers can be found. A Gateway allows querying various repositories listed in the provider list using their REST API.

[84] https://datasetsearch.research.google.com/
[85] https://www.w3.org/TR/vocab-dcat-3/
[86] https://schema.org/
[87] https://developers.google.com/search/docs/advanced/structured-data/dataset
[88] https://github.com/SciCatProject/LandingPageServer/blob/develop/src/app/publisheddata-details/publisheddata-details.component.ts#L66-L77
[89] https://www.optimade.org/index
[90] https://www.optimade.org/providers-dashboard/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

4. The NeXus Format and its Application Definitions

NeXus[91] is a data format for neutron, x-ray and also muon sources. It is promoted and accepted as a standard format for data storage in more and more PaN facilities. Its scope is to describe the instrumental setup and conditions of a measurement, as well as to register the measurement data and even to describe the basic processing steps of the data. NeXus also provides so-called **application definitions** to standardise the data for specified experimental techniques. In order to comply with its scope NeXus has a defined vocabulary and structure.

Why have a chapter dedicated to NeXus here? First, because NeXus is an attempt to establish a common data format that covers all PaN techniques, solving the problem of having a multitude of heterogeneous formats to manage for software tools having to handle PaN data. Second, because the current NeXus/HDF5 implementation allows the aggregating of vast amounts of data (e.g. images) over a small number of files, and most importantly for this report, the storing of **associated metadata**. The aim of Nexus is therefore to allow the storing of data and metadata in a **self-contained and self-descriptive** fashion, thereby increasing interoperability and reusability. This section explores the possibilities currently offered by NeXus in terms of metadata storage, in the light of the framework presented in Figure 1.[92]

4.1 Background on the NeXus/HDF5 Format

4.1.1 NeXus data and file formats

The NeXus data format is independent of the file format. Nevertheless, HDF5 is the dominant format for serialising NeXus files and various utilities for reading, writing, browsing, and using NeXus files have been created (see below). Recently other file formats have been suggested due to performance issues in HDF5. Using the linking mechanisms of HDF5, datasets applying the NeXus standard don't have to be composed of only one file. For structuring metadata and data, a masterfile having all metadata in NeXus and linking to the data files is a common approach. If the data files are using HDF5, a seamless integration with master files allowing browsing through the whole dataset is possible.

The NeXus architecture is organised in a way that is compatible with the underlying hdf5 structure:

- **Groups** (folder-like which have a type descriptor and a NeXus base class name associated with them)
- **Fields** (file-like, can be scalar or multidimensional arrays)
- **Attributes**: extra-information associated with particular groups or fields
- **Links** (pointers to existing data somewhere else)

[91] Konnecke, M. et al., J. Appl. Cryst. 48, 301-305. 2015. The NeXus data format https://doi.org/10.1107/S1600576714027575

[92] Note that in the PaNOSC project, a deliverable dedicated to NeXus is in preparation (D3.5 NeXus Metadata Schema).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Data type and units

The NeXus definition language (NXDL)[93] defines its own data types for fields or attributes, similar to common types found in programming languages (e.g. NX_CHAR for string, NX_UINT for unsigned integer etc.).[94] NXDL does not impose restrictions on which units should be used[95] but provides unit categories entered as strings (NX_CHAR), e.g. NX_ENERGY is a category that refers for instance to “J” or “keV”.

Pros and cons of implementing NeXus in PaN facilities

There exists several motivations for facilities to switch to NeXus among the different techniques and instruments covered. The main advantages can be summarised as follows:

- Avoiding having multiple formats and conversions for both human inspection and analysis software
- Having an open versatile, self-contained and self-descriptive format
- NeXus files can act as containers for e.g. image series
- Providing quick default visualisation
- NeXus aims to encompass all PaN techniques by providing flexible application definitions.

NeXus constitutes therefore a vast effort of data format standardisation beneficial for data storage, exchange and reuse. At the moment, there is however sometimes a certain resistance to move towards Nexus in PaN facilities, due to the complex nature of the data format and its coupling with HDF5. This resistance is however leveraged nowadays by the fact that more tools become available to read, write and inspect NeXus data, as well as to analyse them. The physical coupling of the data format (NeXus) with the file format (HDF5) is not absolute and other file formats might be envisaged in the future.

Architecture

NeXus base classes can be seen as dictionaries of field names and their meanings which are permitted in a particular NeXus group implementing the NeXus class. Application definitions[96] exist for several techniques, which define the minimum required information necessary to satisfy data analysis or other data processing.

By design, a NeXus file is portable and self-descriptive, from the fact that it can contain a broad range of experimental (scientific) metadata scattered in between the different fields and attributes. We will discuss in Section 4.2 the descriptive level reached by NeXus metadata in current application definitions and compare it with the experimental metadata required from our framework.

[93] https://manual.nexusformat.org/nxdl.html
[94] https://manual.nexusformat.org/nxdl-types.html#nx-char
[95] https://manual.nexusformat.org/datarules.html#design-units
[96] https://manual.nexusformat.org/classes/applications/index.html

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

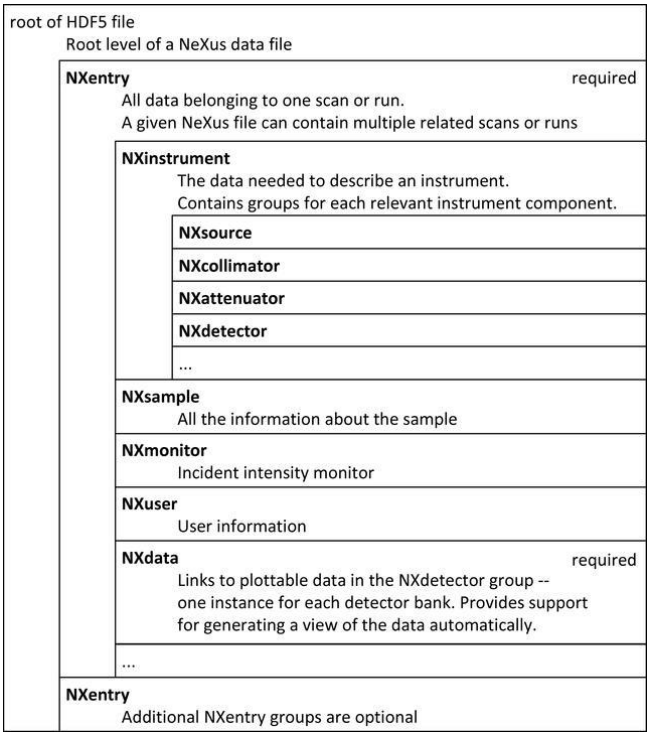


Figure 8: Typical organisation of a NeXus file[97]

Figure 8 shows the typical organisation of a NeXus entry, which corresponds to one scan. The metadata about instrumentation and sample are stored in different groups (of class NXinstrument and NXsample respectively) while the “data" group, of class NXdata, contains links to what is stored in the group of class NXdetector.

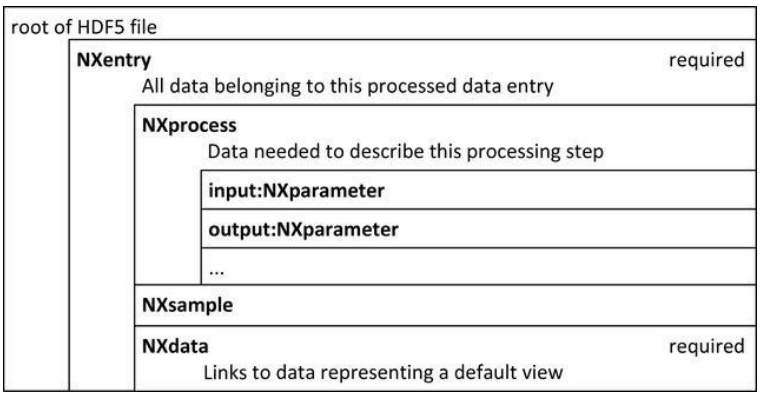


Figure 9: Structure of the NXprocess group[98]

A NXprocess group also exists to store details about **data processing** (see Figure 9), amongst others about the program (or programs) used and the corresponding version, date of processing and other metadata. Additional NXparameter subgroups can be added to a NXprocess group. NXparameter subgroups are containers for storing the input and output parameters of the program used for processing.

[97] Konnecke, M. et al., J. Appl. Cryst. 48, 301-305. 2015. The NeXus data format https://doi.org/10.1107/S1600576714027575

[98] Ibid.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

4.1.2 Maintenance and evolution of the standard

NeXus is an evolving standard and decisions about the application definitions are taken by the NIAC. Each facility can send a nominee to the NIAC to represent the facilities requirements. Changes to the standard are decided through a voting procedure after discussing them. Votes take place in official meetings. NIAC members and meetings are listed here.[99]

Engagement in discussions of requirements and support

The NeXus standard lives through the contributions and commitments from the users and facilities. Having a NIAC member that passes the requirements and other discussions in the facility to the NIAC and other NeXus experts in the facility is a very important contribution. These people are most likely to give support via the NeXus mailing list.[100]

Important discussions and sorting out where NIAC votes are required are also taking place in the monthly telcos.[101] In these telcos github issues[102] are discussed (if they cannot be resolved via github).

Other discussions are going on in the community. Some facilities have working groups around NeXus. There are also communities around specific techniques where discussions and creation of application definitions are organised, for example:

- https://fairmat-experimental.github.io/nexus-fairmat-proposal/
- https://gitlab.hzdr.de/em_glossary/em_glossary/

Implementation of tools and services

NeXus provides an overview of tools and services (termed utilities) to **read**, **write**, **browse**, and **use** NeXus files.[103] Many of these utilities have been created and are maintained by the community depending on their requirements. Most utilities are using HDF5 as a file format, as it complies with current needs. Contributions supporting other file formats are welcome.

NeXus terminology in data catalogues

Some facilities (HZB, ESRF, Diamond) are starting to use the NeXus terminology in their data catalogues. Here, the path to the term used in the NeXus file is commonly registered in the catalogue as a **parameter** type. The parameter type in the data catalogue of the mentioned facilities is the key to the values that can be searched for in the catalogues. These parameter types correspond to the ontology classes in the NeXus ontology. Each class in the NeXus ontology has an IRI e.g. http://purl.org/nexusformat/definitions/NXsensor-value. By using this IRI in the data catalogue, the entries can be directly linked to the ontology and search terms can be easier used across facilities.

[99] https://www.nexusformat.org/NIAC.html
[100] https://manual.nexusformat.org/mailinglist.html
[101] https://www.nexusformat.org/Teleconferences.html
[102] https://github.com/nexusformat/definitions/issues
[103] https://manual.nexusformat.org/utilities.html

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

4.2 How Can the Metadata Framework Be Embedded in NeXus?

In this section, we explore how the different fields of the metadata framework summarised in Figure 1 can be encoded in a NeXus file. NeXus itself is designed as a format that primarily stores scientific (experimental) metadata. Chapter 2 recommended to encode the different types of metadata collected according to their type (i.e. administrative, experimental) in separate, specialised data format. However in the case where no such aggregation is possible, we examine in Table 9 below which of the NeXus base classes can be used to store the different administrative (and scientific) records listed in the framework. In the case where several NeXus base classes are possible, they are separated by 'OR' in the last column of Table 9. When several fields of a base class could be used to store information about a particular metadata record of the framework, brackets are used (e.g. NXsample/[preparation_date, description]). Note that when no suitable NeXus class exist, a free-text description could be used (**NXnote**, defined as "Any additional freeform information not covered by the other base classes").[104]

4.2.1 Raw data

	Framework field	Possible NeXus class/field
Proposal	**PI/Main proposer**	NXuser/[name, ORCID, role="principal_investigator"]
	Co-investigators	NXuser/[name, ORCID, role=""co_investigator"]
	Sample description	NXsample
	Experiment description	NXentry/experiment_description
	Facility information	NXsource/name
	Proposal identifier	NXentry/experiment_identifier
Scheduling	Sample preparation	NXsample/[preparation_date, description]
Experiment	**Visiting experimental team (user id)**	NXuser/[facility_user_id, ORCID]
	Experiment date	NXentry/[start_time, end_time]
	Sample information	NXsample
	Instrument information	NXinstrument, OR NXnote (software etc.) OR NXentry/program_name

[104] https://manual.nexusformat.org/classes/base_classes/NXnote.html#nxnote

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

	Framework field	Possible NeXus class/field
	Calibration information	NXdetector/[calibration_date, angular_calibration_applied, angular_calibration[i, j], calibration_method]
	Experimental planning	NXentry/experiment_description OR NXnote
	Environmental parameters	NXinstrument
	Laboratory notebook	NXnote
	Instrument scientist	NXuser/role="local_contact"
	[Experimental report]	NXentry/[experiment_documentation, notes]
Storage	**Persistent Identifiers (PIDs)**	NXentry/entry_identifier_uuid
	Preservation description information	NXnote
	Dataset information	NXentry/[collection_identifier, collection_description]
	File identifier	NXentry/entry_identifier
	[Representation information]	NXnote
	[Instrument parameters]	NXinstrument

Table 9: Mapping between the metadata framework and NeXus base classes for raw data

The NeXus base classes mentioned in Table 9 can be used to **complement application definitions** (minimal "recipes" dedicated to particular PaN techniques) if they are not already part of them or to build new ones.

Of these application definitions NXarchive is a special one.[105] The description says "This is a definition for data to be archived by ICAT". It gives a suggestion of required metadata and their structure to preserve a minimum of provenance information and physical metadata of the measurement. In the NeXus github, there is an issue about reviewing this application definition.[106]

[105] https://manual.nexusformat.org/classes/applications/NXarchive.html
[106] https://github.com/nexusformat/definitions/issues/1049

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

4.2.2 Processed data

Application definitions for processed data also exist but are scarce at the moment of writing this report. As an example, the **NXtomoproc** application definition[107] allows storing a basic set of information about processed tomographic data (a reconstructed volume in this case). Apart from the processed data, the other fields only allow to retain information about the instrument, sample name and which program was used, together with its version and a link to the original data. Some of the fields of this application definition are taken from the **NXprocess** base class.[108]

The different fields contained in the NXprocess base class are:

- **program**: Name of the program used
- **sequence_index** of processing, for determining the order of multiple NXprocess steps. Starts with 1.
- **version** of the program used
- **date**: date and time of processing
- **parameters,** including a link a reference to **raw_data**
- **note:** will contain information about how the data was processed or anything about the data provenance. The contents of the note can be anything that the processing code can understand, or simple text.

	Framework field	Possible NeXus class/field
Data processing	Processing team (user ID)	NXuser/[name, ORCID, role=““data processing”]
	Original data	NXprocess/parameters/raw_data
	Data format (after processing)	NXprocess/NOTE
	Dataset information	NXentry/[collection_identifier, collection_description]
	Processing information	NXprocess/NOTE
	Software package information	NXprocess/[program, version]
Data processing /analysis	Analysis team (user id)	NXuser/[name, ORCID, role=““data processing”]
	Original data	E.g. NXtomoproc/raw_data
	Software package information	NXprocess/[program, version]

[107] https://manual.nexusformat.org/classes/applications/NXtomoproc.html
[108] https://manual.nexusformat.org/classes/base_classes/NXprocess.html#nxprocess

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

	Framework field	Possible NeXus class/field
	Dependence tracking and workflow	NXprocess/[NOTE, sequence_index]
	Data formats (after analysis)	NXprocess/NOTE
	Dataset information	NXentry/[collection_identifier, collection_description]
	File identifier	NXentry/[entry_identifier_uuid, entry_identifier]
	[Instrument parameters]	NXinstrument
	[Calibration information]	NXdetector/[calibration_date, angular_calibration_applied, angular_calibration[i, j], calibration_method]

Table 10: Mapping between the metadata framework and NeXus base classes for processed data

As we see in Table 10, the base class NXprocess alone would be far from sufficient to store the different metadata and other base classes are needed. Even when using the NXprocess base class, a lot of the records fall into free text description using “NXprocess/NOTE”. Finally the software package information only allows recording a single step of what can be a large data processing workflow (one step driven by one software). Another important point to highlight here is that often, data acquisition and data processing are not sequential but intertwined (characterization in MX) which would be also difficult to record here.

Due to the properties mentioned in Section 4.1, it is likely that processed data will be stored and annotated using NeXus/HDF5 in the future. However the provenance of this data might be difficult to track using the current base classes available. Indeed, it is often the case in PaN facilities that many input dataset are combined together with other sources of scientific input for data processing and more specialised tools might be needed to keep a trace of complicated workflows (Chapter 5). In some disciplines, the end-product (e.g. a protein structure, is the result of many steps of data processings and analysis so that the resulting final data is very “far” from the instrument that was used for raw-data collection. It does not make sense in these cases to stick to NeXus/HDF5 for storing this final data when other standards (i.e. PDB or mmCIF) are required by the community.

Finally, it also makes little sense to try to store metadata information about **journal publication** or **data publication** inside a Nexus raw or processed data set, i. e. These are by nature extrinsic information since a dataset can be cited by several publications and placed into several repositories, and the situation can change over time.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

4.3 Example of NeXus/HDF5 Implementation at PaN Facilities

4.3.1 Implementation notes on NeXus in heterogeneous infrastructure at Elettra

Elettra RI hosts a synchrotron storage ring and free-electron laser FERMI with more than 30 beamline endstations and numerous supporting labs. The infrastructures on the endstations and labs are different regarding data treatment. Elettra was involved in the design of custom HDF5 structures back in 2009 for XRF applications. In 2010 was exposed to the technologies of NeXus in the PaNdata Europe project which was promising wide adaptation of the format. From the start it was very hard to convince beamline scientists to start using it. Then as Python started becoming more popular, the h5py[109] module allowed for an easy way to store mixed type data structures in binary files with random access. NeXus seemed like a superset of HDF5 defined within the more determined ontology, with complex application definitions, and a less friendly generalised Python API. Eventually when the beamlines of the free-electron laser were in development, a specialised list-based file structure with time/bunch-number enumeration was chosen and it was based on a bare metal HDF5. Thus, due to the differences in naming, definitions and ontologies. Nowadays, only two beamlines are using NeXus-based data formatting with ready TANGO integrations (SYRMEP tomography beamline and XRD on MCX) but there are plans to include more. In fact, NeXus often makes data categorization stricter than the existing legacy formats, and snaps it to the metadata. It increases the complexity of software pipelines and consequent development time.

As the main problem of integration of NeXus format at Elettra RI, the absence of involvement of the beamline scientists and staff into NeXus App definition. With other issues such as the legacy programs, existing metadata harvesters and commercial software products that require their own data format and does not support NeXus, the current situation reduces the additive value of the RI workflows integrating NeXus in comparison with the current existing ones. This can make it better to create the packaging clients and second-layer harvesters for the NeXus format above the existing data processing pipeline. This may be a longer process than the centralised implementation since it involves development-on-demand. Elettra's direct and indirect involvement in projects like PaNdata, ExPaNDS, PaNOSC and LEAPS-Innov promotes the adoption of NeXus, but a closer dialogue is needed with end-user team and staff to discover the added value in their everyday work, so the adoption process has higher inertia than expected from the beginning of elaboration of the NeXus format in the project.

4.3.2 Implementation notes on NeXus at SOLEIL

SOLEIL has adopted NeXus as its standard data format since the early beginning. As the data acquisition systems at SOLEIL are based on the Tango software bus which carries all the important information using a client server paradigm, Nexus/HDF5 file handling is provided as a Tango Control System Service. This service is handled through two main Tango devices (written in C++):

- **RecordingManager** which is a front-end high level configuration client of the service
- **TangoRecorder** which takes in charge the harvesting and recording of data/metadata during the acquisition

[109] https://www.h5py.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Auxiliary Tango devices are available to complete the service:

- **ProjectManager** which provides beamline user project choosing, authentication and file access rights per user/project
- **FileTransfer** which acts as a daemon moving Nexus files from local data storage to the central archiving system ruche

The Tango core recording devices were based in the beginning on the C Nexus API, but due its threading issues, a homemade C++ library *libNexusCPP* has been developed directly on top of the HDF5 library yet continuing to use the Nexus Data Format. To ease life of beamline scientists and their users, dedicated wrappers around libNexusCPP have been developed for the main client environments used at SOLEIL: Python, Igor and Matlab.

The client applications developed by the SOLEIL IT team for Nexus file reading, data visualisation and reduction are based on the Java language. They all have been developed to be independent of the data acquisition processes at the beamlines. The main issue was then to cope with the variety of internal naming of entities (beamline energy, etc.) and the overall organisation of data in the files. To tackle this problem, SOLEIL, in collaboration with ANSTO has developed the CDMA (Common Data Model Access) abstract API[110] which can be implemented as a unified layer to access data from a data visualisation/analysis point of view.

The CDMA is a core API that accesses data through a data format plug-in mechanism and scientific application definitions (sets of keywords) coming from a consensus between scientists and institutes. Using an innovative “mapping” system between application definitions and physical data organisations, the CDMA allows data reduction application development independent of the data file container AND schema. Each institute can develop a data access plug-in for its own data file formats along with the mapping between application definitions and its data files. Thus data visualisation/reduction applications can be developed from a strictly scientific point of view and are immediately able to process data acquired from several institutes. The CDMA HDF5 plugin is still maintained by SOLEIL.

4.3.3 Implementation notes on Nexus at Alba

There are ten beamlines in activity at the Alba synchrotron at the time of writing of this document. So far only one is using NeXus as the main storage format for raw data (LOREA beamline, specialised in Angle-resolved photoemission spectroscopy). Note that future beamlines like FaXToR (µ-tomography) and XAIRA (serial-crystallography) will be using NeXus from the start. A progressive transition to NeXus is also foreseen in other beamlines.

LOREA is a very recent beamline so the incorporation of NeXus could be planned from the design phase. At the moment, a custom script (Sardana macro)[111] is in charge of aggregating and converting from the native image format to Nexus and the metadata ingestion strictly follows the NXarpes application definition. The beamline scientists and users are satisfied with this setup, NeXus-formatted data are visualised using *silx view* 0.14.0,[112] and are processed using a Jupyter notebook making use of h5py. As previously stated, an application definition constitutes a minimal set of information associated with a particular type of technique. Dialoging with beamline scientists leaves no doubt on the fact that a richer level of metadata

[110] https://accelconf.web.cern.ch/icalepcs2011/papers/thchaust03.pdf
[111] https://sardana-controls.org/
[112] https://zenodo.org/record/5761269

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

should be reached. One of the points is that more metadata about the sample itself, and in particular its preparation history (provenance) and a log of its physical parameters during the measurement (temperature, pressure) would be necessary to ensure reusability.

4.4 Final Recommendation for Storing Metadata in NeXus Files

The stories collected in the previous sections highlight the fact that NeXus/HDF5 implementation is still at the beginning phase in various facilities and that a substantial amount of work has to be done in order to standardise its use for data ingestion, visualisation and analysis.

There are two aspects that will be reviewed in this section:

1. Which (meta)data to store
2. How to structure what is to be stored

About 1: It is mostly impossible to store all (meta)data in a NeXus file. Here possible and direct usage scenarios for scientific applications, taking into account the creation context and preservation aspects of the dataset have to be considered. In the previous draft deliverable,[113] a list of **priorities** had been created. Also, the level of difficulty to obtain and integrate metadata into a file should be considered. Even if a metadata is of lower priority but easy to integrate it should be part of the file.

About 2: the structure of the (meta)data in Nexus files is guided by possible application definitions. The usage of NeXus application definitions for specific techniques is recommended but bearing in mind that these only represent a “minimal” recipe of metadata records to collect and that more items can and should be added for better reusability.

In case there exists no application definition for a given technique, a suggestion can be made to the NIAC. Suggestions to augment base classes with new fields should also be considered, indeed propositions from the community for NeXus base classes or application definitions are listed in the Contributed Definitions.[114] Finally, the option of storing all (meta)data in one file or having one master NeXus file that links to the measurement data.

We propose that a minimal requirement on a NeXus file from the perspective of FAIR data can be a reviewed archive application definition (NXarchive).[115]

[113] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825
[114] https://manual.nexusformat.org/classes/contributed_definitions/index.html#contributed-definitions
[115] https://manual.nexusformat.org/classes/applications/NXarchive.html

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

5. The FAIR Principles for Research Software

One of the key information (P1) for reusability mentioned in the data processing and analysis tabs of the metadata framework (see Figure 1) is "**software package information**". Applying the FAIR principles for software implies a re-examination of the principles themselves. We summarise thereafter the main initiatives in that direction, starting by recalling the work achieved by the RDA for research software.

In 2021, the FAIR for Research Software Working Group released a document termed "FAIR Principles for Research Software".[116] This work is an attempt to establish a list of recommendations related to each of the four aspects of FAIR principles in order to make software comply with them. These recommendations are structured in each case with a global condition (e.g. F, A, I, R) followed by nested sub principles (e.g. F1, F1.1). It is therefore of great relevance to expose these recommendations in the present report. The reader will refer to the original document for a more detailed overview.

5.1 Software Findability

F: *Software, and its associated metadata, is easy for both humans and machines to find*.

- F1. Software is assigned a globally unique and persistent identifier.
  - F1.1. Components of the software representing levels of granularity are assigned distinct identifiers.
  - F1.2. Different versions of the software are assigned distinct identifiers.
- F2. Software is described with rich metadata.
- F3. Metadata clearly and explicitly include the identifier of the software they describe.
- F4. Metadata are FAIR, searchable and indexable.

5.2 Software Accessibility

A: *Software, and its metadata, is retrievable via standardised protocols.*

- A1. Software is retrievable by its identifier using a standardised communications protocol.
  - A1.1. The protocol is open, free, and universally implementable.
  - A1.2. The protocol allows for an authentication and authorization procedure, where necessary.
- A2. Metadata is accessible, even when the software is no longer available.

[116] https://rd-alliance.org/group/fair-research-software-fair4rs-wg/outcomes/fair-principles-research-software-fair4rs-0

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

5.3 Software Interoperability

I: *Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.*

- I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards.

- I2. Software includes qualified references to other objects.

5.4 Software Reusability

R: *Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).*

- R1. Software is described with a plurality of accurate and relevant attributes.
  - R1.1. Software is given a clear and accessible licence.
  - R1.2. Software is associated with detailed provenance.

- R2. Software includes qualified references to other software.

- R3. Software meets domain-relevant community standards.

5.5 Referencing Software

In case the FAIR principles have to be applied to a piece of software, the software can and should be referenced in the metadata record of a dataset when used for its creation. Referencing the software contributes to R of the FAIR data principles of the dataset.

Best practices on how to cite software have gained some momentum in the last years, notably through working groups such as FORCE11,[117] who designed the **software citation principles** (2016),[118] which have been reproduced here for convenience:

1. **Importance**: Software should be considered a legitimate and citable product of research. Software citations should be accorded the same importance in the scholarly record as citations of other research products, such as publications and data; they should be included in the metadata of the citing work, for example in the reference list of a journal article, and should not be omitted or separated. Software should be cited on the same basis as any other research product such as a paper or a book, that is, authors should cite the appropriate set of software products just as they cite the appropriate set of papers.

[117] https://force11.org
[118] Smith A., Katz D., and Niemeyer K., FORCE11 Software Citation Working Group (2016). Software Citation Principles. PeerJ Computer Science 2:e86.https://peerj.com/articles/cs-86/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

2. **Credit and attribution**: Software citations should facilitate giving scholarly credit and normative, legal attribution to all contributors to the software, recognizing that a single style or mechanism of attribution may not be applicable to all software.

3. **Unique identification**: A software citation should include a method for identification that is machine actionable, globally unique, interoperable, and recognized by at least a community of the corresponding domain experts, and preferably by general public researchers.

4. **Persistence**: Unique identifiers and metadata describing the software and its disposition should persist—even beyond the lifespan of the software they describe.

5. **Accessibility**: Software citations should facilitate access to the software itself and to its associated metadata, documentation, data, and other materials necessary for both humans and machines to make informed use of the referenced software.

6. **Specificity**: Software citations should facilitate identification of, and access to, the specific version of software that was used. Software identification should be as specific as necessary, such as using version numbers, revision numbers, or variants such as platforms.

Many of these principles are indeed highly contributing to one or more aspects of FAIR. In the context of software metadata, the ideal record would be a PID that would point to a repository's landing page containing precise references to software name, versions and authors. The level of granularity of such citation is discussed in the aforementioned Smith et al., 2016 paper.

At this point we should mention initiatives like Software Heritage,[119,120] whose mission is to collect, curate and preserve all publicly available software through attribution of unique identifiers, store their full development history and source code in a referenceable way. Software references are particularly important when recording **provenance information**, which is the subject of the next chapter.

[119] https://www.softwareheritage.org
[120] https://hal.archives-ouvertes.fr/hal-01590958

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

6. Provenance

The data processing and analysis tabs of the metadata framework presented in Figure 1 mentions "**dataset information**" as well as "**dependence tracking and workflow**". Indeed, data reuse implies having knowledge about how derived datasets have been produced, a topic also known as provenance.

6.1 What Is Provenance?

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.[121] It therefore addresses the question of **how an artefact has come into existence**. In this chapter, we focus on provenance information of **derived data**, i.e. data that has been produced as the result of one or several processing/analysis steps. In this context, provenance is therefore centred on strategies to capture information about software environment (dependencies), parameters and workflows. Discussing other types of provenance information would render this report excessively long but note that hints about sample provenance have been provided in Section 1.2.3.

6.1.1 Data processing versus analysis

This section focuses on the steps of the metadata framework that cover data processing and data analysis. These two terms are sometimes used interchangeably depending on the technique covered. However, as already mentioned in ExPaNDS D2.2, it is commonly understood that data processing is anterior to data analysis and often only implies a transformation (merging, reduction, etc.) of the raw data before data analysis takes place (this time through the combination of software and scientific input).

Note that the process is not strictly sequential since there can be multiple rounds of data processing and data analysis before obtaining a satisfying outcome. Data processing and analysis are often performed in the user's home institution, in which case it becomes extremely difficult to keep a record of the actors and of the different steps performed. However most PaN facilities nowadays offer to process and analyse data on premises, either via automatic software pipelines or Data Analysis as a Service (DAaaS) instances via a dedicated portal.

From a (meta)data management perspective, these two steps, data processing and analysis, are identical in that the objective is to keep a trace of the different elements of software used, their version, dependencies and order of execution (workflow).

6.1.2 Why capture provenance information: use cases

Why is software provenance information important? The overall benefit is of course data reusability and reproducibility. Very often in PaN facilities, data processing/analysis services are present on premises and allow obtaining a first scientific outcome while the experiment is running. An **optimised scientific outcome** is often the result of rerunning the data processing/analysis while varying the software used and /or the input parameters. It should be added that the output data may not be retained in the facilities storage. For these reasons, it

[121] https://www.w3.org/TR/prov-overview/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

is essential for users and re-users to understand in detail the different steps that were performed as well as to know which programs, versions and input parameters were in use. A couple of use-cases are listed afterwards:

- Allow a scientist or a computational method developer to recreate the results shown in a paper.
- Allow a scientist or a computational method developer to understand how some dataset was created.
- Allow a scientist to use an intermediate result from an existing analysis as a starting point for their own investigation.
- Allow a scientist to survey existing analysis techniques, to catalogue different approaches.
- Allow a research software developer to gauge to what extent their software is used.
- Allow metadata catalogue developers to conform to a standard about sharing metadata information.
- Inform scientists about which practices they need to follow when writing/developing research software in order to ensure software FAIRness.
- Allow a research institution to audit their software usage (e.g., to drop unused packages).
- Allow an impact assessment if a certain version of the research software is found to produce incorrect results.
- A repository might want to store only the raw data when derived datasets are too big but easy to reproduce given enough information.

6.2 Workflows

As mentioned earlier, knowing which software was used, together with associated versions and input parameters is necessary but not sufficient. The interrelations between the different inputs and outputs, specifying the order in which the steps were taken, are of prime importance for reproducibility and reusability. The PROV ontology (PROV-O)[122] has been designed to address this question and is described in the next sections. Complementary to software provenance, digital preservation is briefly discussed in Section 6.2.4

6.2.1 The W3C PROV standard

PROV[123] is a W3C standard supporting the interchange of provenance information. It relies on PROV-O, specifying the three following classes:

[122] https://www.w3.org/TR/2013/REC-prov-o-20130430/
[123] https://www.w3.org/TR/prov-overview/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

- **Entities**: representing an "object": i.e. document, data, website
- **Agents**: person or software performing an activity
- **Activities**: process that makes use of an entity and/or generates a new entity

The relations (properties) between these classes are illustrated in Figure 10:

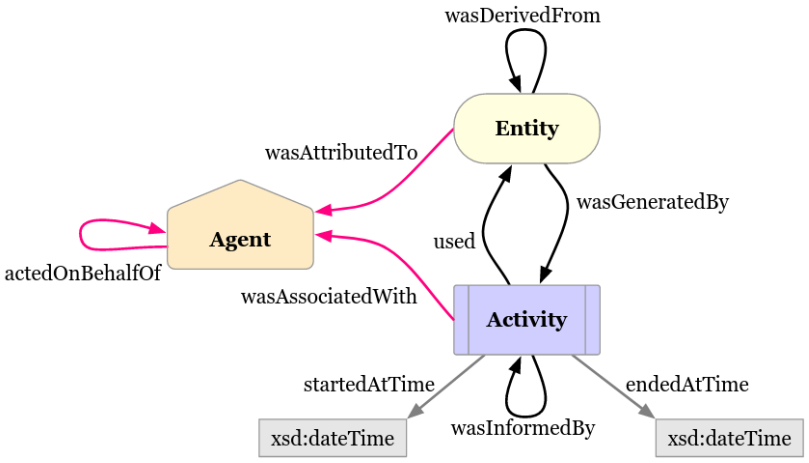


Figure 10: Diagram of the three PROV-O classes and their relations (properties).[124] The responsibility properties are shown in pink.

For convenience, we can keep in mind the following associations when considering data processing/analysis in PAN facilities.

- **Entity** → data (raw or derived), list of input parameters.
- **Agents** → software (and its version) or person.
- **Activities** →process performed by a particular software or a person.

Many projects are under development that focus on how to best capture software workflows and associated information. Some of them use PROV-O as a basis, together with other ontologies in order to capture more information than what is possible using PROV-O alone, including for example documentation, annotations, example data and execution traces (see for instance Preserving workflow-centric research objects).[125]

The following section attempts to illustrate how the PROV data model can be used in practice thanks to a dedicated Python library in order to capture a software workflow typically used in macromolecular crystallography (MX). The goal here is not to be exhaustive but rather to give the reader an idea of the possibilities offered by current tools in this domain.

[124] https://www.w3.org/TR/2013/REC-prov-o-20130430/
[125] https://www.sciencedirect.com/science/article/pii/S1570826815000049

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

6.2.2 Example: capturing a MX software workflow

The steps following the acquisition of raw data in macromolecular crystallography (MX) are very well established and many specialised software, often included as part of suites (e.g. CCP4[126], Phenix[127]) have been developed to perform them.

Starting from raw data (images from area detector containing diffraction spots), the first steps undertaken can be summarised as follows:

- **Integration**: the intensity of each diffraction spot is quantified. All intensities are indexed using miller indices and are output to an intermediate file (integrated data). The software XDS[128] is used in our example.

- **Data reduction**: different correction factors are applied to the measured intensities, which are then converted to amplitudes. Averaging occurs between amplitudes of reflections that must theoretically have the same value. Here we chose to use the program Aimless.[129]

- **Phasing**: Amplitudes are not sufficient to recreate the 3D structure of the molecule(s) under investigation: the phases of each reflection must also be retrieved by a complementary method. The most common one is molecular replacement, which borrows phases from another molecule whose structure is known. The program used in our example is Phaser,[130] which needs an input molecule to borrow the phases from in addition to the reduced data.

Note: Integration and data reduction are considered **data processing** here while the phasing step is considered **data analysis** since it requires the input scientific knowledge about the sequence and structure of the target molecule.

Figure 11 (see below) was generated using the prov[131] library for Python. The corresponding Python 3.7 code (executed in a Jupyter notebook in this case) is provided in Appendix B.

[126] Winn, M. D. et al. *Acta. Cryst.* **D67**, 235-242 (2011) "Overview of the CCP4 suite and current developments"https://doi.org/10.1107/S0907444910045749
[127] Liebschner, D. et al., Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. Acta Crystallogr D Struct Biol 75, 861–877 (2019). https://doi.org/10.1107/S2059798319011471
[128] Kabsch, W. *XDS. Acta Cryst.* D**66**, 125-132 (2010)
[129] Evans, P. and Murshudov, G. 'How good are my data and what is the resolution?' Acta Cryst. D69, 1204-1214 (2013)
[130] McCoy A., Grosse-Kunstleve R., Adams P. et al. J Appl Cryst (2007). 40, 658-674.
[131] https://github.com/trungdong/prov

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

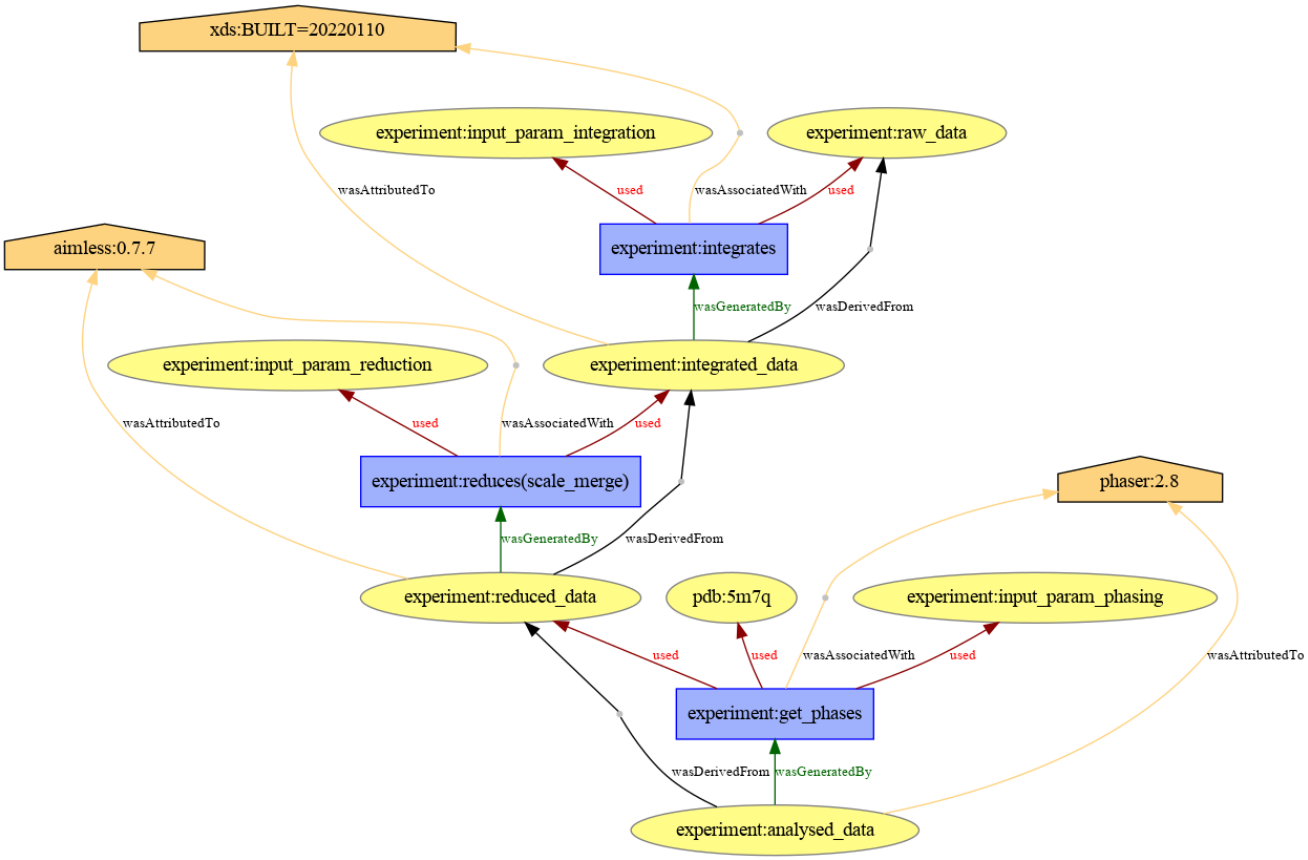


Figure 11: Provenance diagram of a MX workflow

Figure 11 is only a visual representation of a provenance object generated thanks to the library. Other representations are possible such as PROV-N,[132] which is a human-readable format (see Appendix B).

Such a graph would allow users to better understand the different steps performed and therefore allow them to rerun the entire pipeline or only certain selected steps while changing some input parameters or software. We can therefore imagine having such a provenance graph generation **routinely integrated to automatic data processing pipelines** run at PaN facilities. It implies outputting not only the graph itself but also information file(s) containing details about the input parameters and software used (version, environment).

6.2.3 Some tools to record provenance information and example

A non-exhaustive list of references is given thereafter for the reader to explore this thematic:

- Python prov library:[133] An implementation of the W3C PROV Data Model in Python (used to make Figure 11).
- Provneo4j:[134] a Python client for storing PROV documents in Neo4j to use with the prov Python library.

[132] https://www.w3.org/TR/prov-n/
[133] https://github.com/trungdong/prov
[134] https://provneo4j.readthedocs.io/en/latest/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

- Prov-db-connector:[135] a Python module that provides a general interface to save W3C-PROV documents into databases (currently supports the Neo4j graph database). Allows transforming a PROV document into a graph structure.

- Git2prov:[136] allows exposing Version Control System Content as W3C PROV.

- NoWorkflow:[137] Python module that allows scientists to benefit from provenance data analysis even when they don't use a workflow system.

6.2.4 A note about long-term digital preservation

Since the storage tab of the metadata framework presented in Figure 1 mentions "**preservation description information (P1)**", an introduction to this vast subject is relevant in this report.

While provenance is centred on how an artefact (a digital object, e.g. some data) has come into existence, preservation is more concerned about how to ensure that this artefact will still be usable in the long term. Formally speaking the goal is to ensure that certain qualities of the digital object are preserved:

- **Viability**
- **Renderability** (translation of a bit stream into a form that can be viewed by human users)
- **Understandability**
- **Authenticity**
- **Identity**

PREMIS (Preservation Metadata Implementation Strategies)[138] is one of the standards that has been designed to tackle this topic, designed to be easily encoded in XML. The associated **data dictionary** defines **semantic units and subunits**, each of which corresponds to a preservation metadata element (e.g. objectIdentifier, storageMedium, format, etc.) These semantic units can be seen as **properties of the five following entities** forming the base of the PREMIS data model (see Figure 12):

- **Objects** (discrete units of digital information, e.g. a file)
- **Intellectual entities** (e.g. some chunk of data)
- **Events** (something that happens at a point in time, e.g. file creation)
- **Agents** (person, organisation that perform events, thereby affecting objects)
- **Rights** (permissions, copyright etc.)

[135] https://prov-db-connector.readthedocs.io/en/latest/readme.html
[136] http://git2prov.org/
[137] https://pypi.org/project/noworkflow/
[138] https://www.loc.gov/standards/premis/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

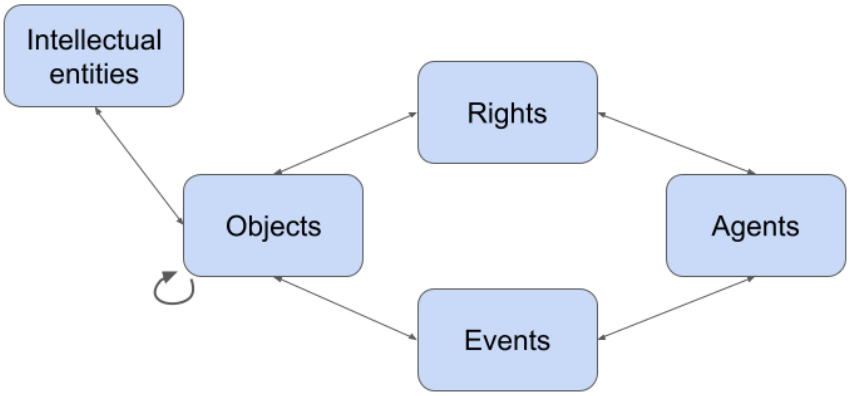


Figure 12: The PREMIS data model

The relations (mostly bidirectional) between these entities are expressed by the following semantic units: linkingEventIdentifier, linkingRightsStatementIdentifier, linkingAgentIdentifier, linkingObjectIdentifier and linkingEnvironmentIdentifier.

What can be achieved with PREMIS?

Once integrated into a repository, preservation metadata could be stored to inform about:

- Which storage medium should be used
- Which operating system or dependencies are needed to use a file
- If a file has been modified or deleted
- Etc.

More information

P. Caplan, Understanding PREMIS. Library of Congress Network Development and MARC Standards Office, 2017[139]

P. Caplan and R. Guenther. Practical Preservation: The PREMIS Experience. "Digital Preservation: Finding Balance," pp. 111–124[140]

6.3 Practical Recommendations for Capturing Data Processing/Analysis Metadata

It would be difficult to recommend specific how-to capture metadata guides for all the possible variations of how scientific data and output are produced, so instead the following will be a series of generic recommendations for what information could be captured.

- For experiments, document:
  - the **firmware version** of the instrument(s)

[139] https://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf
[140] Caplan, P., & Guenther, R. 2005. Practical Preservation: The PREMIS Experience. *Library Trends*, 54(1): 111–124.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

  - any **processing activities** that might have happened before writing the raw data to disk (e.g. data reduction)

- For any kind of processing or analysis, document:

  - the **software environment** (if virtual machines or containers were used this could be a link/reference to download the image that was used)
  - the **exact version of any 3rd party software/code** (i.e. anything that is not written by the researcher) that was used (if open-source a link/reference to the version of the source code is recommended)
  - A representation of the data processing/analysis **workflow** if applicable
  - for code that is written by the researcher, consider **applying the FAIR4RS** principles

- If provenance data is to be recorded, consider using the **W3C PROV standard**. When deciding the level of granularity for documenting the activities, we recommend documenting non-human interrupted/involved steps as a single activity. Steps where human interaction is involved, are recommended to be documented as separate activities.

  Examples:

  - Calling a Python/R/Julia (etc.) script could be shown as a single activity (no need to show individual function calls as separate activities)
  - A Jupyter notebook that is made up of multiple cells that are called by a user individually is best represented as a series of activities

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

7. Final Recommendations

7.1 Summary

The recommendations issued in the different chapters of this report are summarised here.

Chapter 1

- Using the FAIR metadata framework as a basis, carefully design the metadata acquisition plan for a particular instrument through dialogue between the different stakeholders (instrument scientists, users, data acquisition and management staff). Metadata records heavily depending on users (e.g. calibration, sample, experimental notes) should be given special attention.
- Implement a robust sample metadata database system (as a unique source) integrated in the other data management resources of the facility.
- Implement and promote the use of electronic logbooks / notebooks. A system of information tagging would facilitate the parsing of the different types of metadata mentioned by machines.
- Favour the use of persistent identifiers for data, people, instruments and samples when possible.
- Agree on and control the visibility level of each metadata record, in agreement with the embargo period duration.

Chapter 2

- Across the experimental lifecycle within PaN RIs, there are multiple information sources – both human and machine – that play a role in metadata production and collection. In many cases, it is important that these sources interact and integrate within and across the various stages of the experimental lifecycle.
- Each step in the experimental lifecycle produces specific metadata.
- Metadata will be stored in files and exposed in the metadata catalogue. Exposition and storage should ideally be made using their appropriate metadata schemata.
- Avoid dependency on a particular data format when aggregating (meta)data in files and data catalogue. One metadata standard or file format might not be enough to express the information to be stored and exposed.

Chapter 3

- The common search API offers the possibility to query a series of fields that can be directly mapped to the metadata framework presented in Figure 1. However not all fields can be queried yet and the framework can thus serve as a basis to extend the number of fields searchable by the API.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

- EOSC indexing and discovery services such as B2FIND and OpenAIRE offer the means to make PaN datasets more findable by those outside the PaN domain. To enable these generic tools to harvest our metadata, it is important that PaN RIs provide: 1.) OAI-PMH endpoints; and 2.) mappings of our metadata to the metadata schemas used by the EOSC services.

- Regarding metadata mapping, it is important to bear in mind the purpose of services such B2FIND and OpenAIRE. Such tools are aimed at cross-domain discovery; they are not designed to capture the full domain-specific richness that is possible using the ExPaNDS metadata framework.

- It is not necessary nor should we expect to map every aspect of our framework to the metadata schemas of the cross-domain EOSC services. However, to meet the primary purpose of the EOSC discovery tools, it is important that the information we make available through B2FIND and OpenAIRE is: 1.) sufficient for the initial enquiries of a non-domain specialist; and, 2.) able to point that user to where they can find further details.

- Given the nature of the B2FIND and OpenAIRE metadata schemas, it is not possible to map every metadata type found in the more extensive ExPaNDS metadata framework to these schemas. Additionally, multiple ExPaNDS metadata types may map to a given element/property in the B2FIND and OpenAIRE metadata schemas, meaning a one to one mapping is not always possible.

- As an example, we provide some initial guidelines on how ExPaNDS metadata types for a raw dataset could be mapped to the metadata schema of B2FIND. Several key recommendations emerge from this example mapping exercise:

    - To achieve as much richness in the metadata mapping as possible, our recommendation is that PaN providers should seek to map not only mandatory elements but also recommended and optional metadata elements.

    - To increase interoperability and consistency, controlled vocabulary should be used wherever possible, especially for metadata types such as keyword and discipline.

    - The use of PIDs and the related identifier metadata type offers the opportunity to improve findability and to link to other resources, perhaps including sample and instrument PID, if these were to be adopted widely by the PaN community in future.

    - The description metadata type should not be overlooked for its potential to provide valuable additional information, including about the methods and sample.

    - The process of mapping the ExPaNDS framework to the B2FIND schema suggests that some potentially important metadata types may be missing from the ExPaNDS framework. As such, the PaN community may wish to add to and improve upon the current version of the ExPaNDS metadata framework, where experience and practice indicate this could be beneficial.

- At present, each PaN provider interacts with B2FIND and OpenAIRE on an individual basis. The result is that different metadata mappings are produced by the different

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

facilities. While over time, we might expect the adoption of ExPaNDS metadata framework to lead to more consistency in these mappings, it is likely that local practices and policies will still result in some ongoing differences.

Chapter 4

- We recommend the use of NeXus/HDF5 as a self-contained and self-descriptive format to store data and scientific metadata, facilitating data exchange and reuse.
- Administrative metadata can to a certain extent also be stored using NeXus using suggestions from the mapping Tables 9 and 10.
- Section 4.4 provides additional recommendations on NeXus.

Chapter 5

- Refer to the FAIR principles for Research software whenever having to cite software.
- According to the Software Citation Principles,[141] several quality criteria have to be fulfilled whenever citing software:
  - Credit and attribution
  - Unique identification
  - Persistence
  - Accessibility
  - Specificity

Chapter 6

- Please refer to Section 6.3.

7.2 A Note about Metadata Privacy and GDPR

As mentioned in most ExPaNDS partner RIs' data policies, GDPR compliance of the (meta)data (whether human-or machine-generated) obtained and analysed within a project is assessed by the facilities. Although GDPR jurisdiction applies worldwide, it can only be enforced in practical terms within the scope of the European Union (EU). The reader is referred to the ExPaNDS data policy deliverable (D2.3),[142] which highlights some important exceptions for research, which vary from country to country within Europe.

In some special cases of machine-generated metadata, intrinsic identifiers (ID) can legally be considered personal data, linking the user's identity to a data collection activity. As a countermeasure example, France established a special regulation on these cases explicitly investigating the personal identification possibilities built within Google Analytics ID.[143] Thus

[141] Smith A., Katz D., Niemeyer K., FORCE11 Software Citation Working Group. (2016) Software Citation Principles. PeerJ Computer Science 2:e86. https://peerj.com/articles/cs-86/
[142] McBirnie, A., Matthews, B., Gagey, B. et al. (2021). Final data policy framework for Photon and Neutron RIs. https://doi.org/10.5281/zenodo.5205825
[143] https://www.cnil.fr/fr/utilisation-de-google-analytics-et-transferts-de-donnees-vers-les-etats-unis-la-cnil-met-en-demeure

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

IDs should be considered as personal data and their treatment must be legally GDPR-compliant.

GDPR requires derogations to continue to be in line with explicitly declared transparency requirements and privacy by design and default.[144] So, during the development of metadata policies and data management regulations, it is important to give the efforts to identify the data and metadata that would lead to this situation.[145]

Covering the cases of GDPR-eligible metadata coming from non-European jurisdictions is a challenging special task for DMPs, data policies and assessment teams in a project. The main point is to explore the possibility for European RIs to be considered co-responsible with non-European researchers and institutions, for metadata management. It is needed because the regulations for non-European and European researchers may differ so much that data obtained from ExPaNDS partner RIs can be mixed with data that are not subject to GDPR (or known to be GDPR-violating by explicit declaration, like the personal data in Russia eligible for so called "Experimental regulation regime for the data used in innovative infrastructures"[146]). Here a recommendation could be made for the data management plan to include a statement declaring that any derived data/metadata coming from non-European side should not be in any case considered as the property or curated entity of ExPaNDS because the enforcement of GDPR in fact cannot be established and assessed for non-European partner RIs.

7.3 Final Remarks

ExPaNDS Deliverable 2.2[147] established a Common FAIR Metadata Framework, basing itself on a representation of metadata flow created during an experiment designed in the PaNdata ODI D6.1 data continuum.[148] The present report reflects on different practical aspects of the implementation of this framework such as prioritisation, standards, file formats, tools and good practices available to achieve FAIR metadata collection and storage.

All these aspects indubitably vary among PaN facilities and will evolve over time. It is therefore desirable that the framework, its definitions and modalities of implementation get maintained and updated regularly well beyond the ExPaNDS project. An option to consider would be the creation of a permanent committee integrated in the management of all facilities, in charge of addressing these questions in depth so as to promote and adopt common practices among PaN RIs regarding FAIR metadata.

144 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6777499/
145 https://www.acpjournals.org/doi/10.7326/M18-2854
146 https://regulation.gov.ru/projects#npa=106119 (under legislative commissioning now, can be inaccessible from outside Russia)
147 Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825
148 Matthews, B. et al. (2012). *Model of the data continuum in Photon and Neutron Facilities*. https://doi.org/10.5281/zenodo.3897190

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

References

Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., et al. 2015. *Using a suite of ontologies for preserving workflow-centric research objects*. Journal of Web Semantics, 32: 16–42. https://doi.org/10.1016/j.websem.2015.01.003.

Caplan, P. 2009. *Understanding PREMIS*. Library of Congress Network Development and MARC Standards Office. Revised in 2017, Library of Congress Washington DC, USA.

Caplan, P., & Guenther, R. 2005. *Practical Preservation: The PREMIS Experience*. Library Trends, 54(1): 111–124. https://doi.org/10.1353/lib.2006.0002.

Collins, S. P., da Graça Ramos, S., Iyayi, D., Görzig, H., González Beltrán, A., et al. 2021. *ExPaNDS ontologies v1.0.* (ExPaNDS D3.2). https://zenodo.org/record/4806026.

Cosmo, R. D., & Zacchiroli, S. 2017, September 25. *Software Heritage: Why and How to Preserve Software Source Code*. Archive ouverte HAL. https://hal.archives-ouvertes.fr/hal-01590958.

Courtot, M., Cherubin, L., Faulconbridge, A., Vaughan, D., Green, M., et al. 2018. *BioSamples database: an updated sample metadata hub*. Nucleic Acids Research, 47(D1): D1172–D1178. https://doi.org/10.1093/nar/gky1061.

DataCite - *DataCite Schema*. https://schema.datacite.org/.

DCMI Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

Evans, P. R., & Murshudov, G. N. 2013. *How good are my data and what is the resolution?* Acta Crystallographica Section D Biological Crystallography, 69(7): 1204–1214. https://doi.org/10.1107/S0907444913000061.

FAIR Data Maturity Model Working Group. 2020, June 25. *FAIR Data Maturity Model.* Specification and Guidelines. https://zenodo.org/record/3909563#.YGRNng8za70.

FAIRsharing. https://fairsharing.org/search?fairsharingRegistry=Standard.

Gao, Z., Odstrcil, M., Böcklein, S., Palagin, D., Holler, M., et al. 2021. *Sparse ab initio x-ray transmission spectrotomography for nanoscopic compositional analysis of functional materials*. Science Advances, 7(24). https://doi.org/10.1126/sciadv.abf6971.

Gonzalez-Beltran, A., & Winstanley, P. 2022, February 17. *The Data Catalog Vocabulary (DCAT)*. https://zenodo.org/record/6142906.

Gonzalez-Beltran, Minotti, Davies, Leorato, Richards, et al. 2022. *Demonstrate ICAT and SciCat released with APIs compatible with ExPaNDS federated EOSC services*. Zenodo. (ExPaNDS D3.3). https://zenodo.org/record/6363591.

Günther, G., Bär, M., Greve, N., Krahl, R., Kubin, M., et al. 2022, March 1. FAIR *Meets EMIL: Principles in Practice*. https://jacow.org/icalepcs2021/doi/JACoW-ICALEPCS2021-WEBL05.html.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

HDF5 for Python. https://www.h5py.org/.

The HDF Group - Ensuring long-term access and usability of HDF data and supporting users of HDF technologies. https://www.hdfgroup.org/.

Hong, C., Katz, Barker, Lamprecht, Martinez, et al. 2022, May 24. *FAIR Principles for Research Software (FAIR4RS Principles)*.https://zenodo.org/record/6623556#.YqCJTJNBwlw.

Kabsch, W. 2010. *XDS*. Acta Crystallographica Section D Biological Crystallography, 66(2): 125–132. https://doi.org/10.1107/S0907444909047337.

Könnecke, M., Akeroyd, F. A., Bernstein, H. J., Brewster, A. S., Campbell, S. I., et al. 2015. *The NeXus data format*. Journal of Applied Crystallography, 48(1): 301–305. https://doi.org/10.1107/S1600576714027575.

League of European Accelerator-based Photon Sources. *DIGITAL LEAPS is on its way*. https://leaps-initiative.eu/digital-leaps-is-on-its-way/.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., et al. 2019. *Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix*. Acta Crystallographica Section D Structural Biology, 75(10): 861–877. https://doi.org/10.1107/s2059798319011471.

Mascalzoni, D., Bentzen, H. B., Budin-Ljøsne, I., Bygrave, L. A., Bell, J., et al. 2019. *Are Requirements to Deposit Data in Research Repositories Compatible With the European Union's General Data Protection Regulation?* Annals of Internal Medicine, 170(5): 332. https://doi.org/10.7326/m18-2854.

Matthews, B., Kourousias, G., Yang, E., & Griffin, T. 2012, September 30. *Model of the data continuum in Photon and Neutron Facilities*, (PaNdata ODI D6.1). https://zenodo.org/record/3897190.

McBirnie, A., Matthews, B., Gagey, B., Minotti, C., Salvat, D., et al. 2021, August 20. *Final data policy framework for Photon and Neutron RIs,* (ExPaNDS D2.3). https://zenodo.org/record/5205825.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., et al. 2007. *Phaser crystallographic software*. Journal of Applied Crystallography, 40(4): 658–674. Metadata Standards Catalog. https://doi.org/10.1107/s0021889807021206.

Neumann-Kipping, M., & Hampel, U. 2019, August 1. *Ultrafast X-ray tomography image data of bubbly two-phase pipe flow around a ring-shaped constriction* - Publications Repository - Helmholtz-Zentrum Dresden-Rossendorf, HZDR. https://www.hzdr.de/publications/Publ-29882.

The NeXus Data Format. https://www.nexusformat.org.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/.

OPTIMADE. https://www.optimade.org/index.

The PaNET ontology. https://github.com/ExPaNDS-eu/ExPaNDS-experimental-techniques-ontology.

Poirier, S., Buteau, A., and Ounsy, M. (2011). Common Data Model Access: A unified layer to access data from data analysis point of view. https://accelconf.web.cern.ch/icalepcs2011/papers/thchaust03.pdf

PREMIS Data Dictionary for Preservation Metadata, Version 3.0 (Library of Congress). . http://loc.gov/standards/premis/v3/index.html.

PROV-O: The PROV Ontology. https://www.w3.org/TR/2013/REC-prov-o-20130430/.

PROV-Overview. https://www.w3.org/TR/prov-overview/.

Richter, T., Schrettner, L., Caunt, S., Hall, J., & Turner, W. 2020. *API Definition (common search API)*, (PaNOSC D3.1). https://www.panosc.eu/deliverables/deliverable-3-1-api-definition-common-search-api/.

Salvat, D., Gonzalez-Beltran, A., Görzig, H., Matthews, B., McBirnie, A. et al. 2020. *Draft recommendations for FAIR Photon and Neutron Data Management,* (ExPaNDS D2.2). https://zenodo.org/record/4312825.

Semantic Sensor Network Ontology. https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/.

Smith, A. M., Katz, D. S., & Niemeyer, K. E. 2016. *Software citation principles*. PeerJ Computer Science, 2: e86. https://doi.org/10.7717/peerj-cs.86.

Staunton, C., Slokenberga, S., & Mascalzoni, D. 2019. *The GDPR and the research exemption: considerations on the necessary safeguards for research biobanks*. European Journal of Human Genetics, 27(8): 1159–1167. https://doi.org/10.1038/s41431-019-0386-5.

Vincent, T., Valls, V., payno, Kieffer, J., Solé, V. A., et al. 2021, December 6. *silx-kit/silx: 1.0.0: 2021/12/06*. Zenodo. https://zenodo.org/record/5761269.

W3C - *Data Catalog Vocabulary (DCAT) Version 3*. May 2022. https://www.w3.org/TR/vocab-dcat-3/.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., et al. 2016. *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data, 3(1). https://doi.org/10.1038/sdata.2016.18.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., et al. 2011. *Overview of the CCP4 suite and current developments*. Acta Crystallographica Section D Biological Crystallography, 67(4): 235–242. https://doi.org/10.1107/s0907444910045749.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Appendix A: ExPaNDS Metadata Type Definitions

Column 7 of Table 6 in Section 3.3.4 includes several different metadata types drawn from the ExPaNDS metadata framework presented in ExPaNDS deliverable D2.2[149]. In Table 6, the name of the metadata type is provided, along with its prioritisation for FAIR, the aspects of FAIR (i.e. F,A,I,R) to which it is relevant, and the stage of the experimental life cycle during which it appears. However, for brevity, we do not include the full definition of the metadata type in the information contained in the table.

While the name of an ExPaNDS metadata type may be sufficient for understanding why that metadata type has been selected for use in the mapping, the definitions supplied in ExPaNDS D2.2 can help to provide additional clarity. For ease of reference, we reproduce below in alphabetical order the full definitions for the ExPaNDS metadata types that are included in the present deliverable in column 7 of Table 6 found in Section 3.3.4 of D2.2:

Calibration Information [P1-FR; experiment stage]: As the results of a measurement can be affected by changes of instrument characteristics over time, the calibration information is considered as [P1 - ESSENTIAL - FR] to validate the data produced during that particular measurement.

Co-Investigators [P1-FA; proposal stage]: The primary members of the whole Experimental Team. Only one person is identified as Principal Investigator of the proposal; however, in most cases, the proposal is built by a group of people, also known as Co-Investigators. At this stage, and for Findability (e.g. experiments any scientist may have been involved in), we are considering this field, a multi-field, as [P1-ESSENTIAL-FA].

Contributor [P2-F; data publication/record]: Any Person or organisation which contributed to the creation of the resource. [P2 - IMPORTANT - F]

Creator [P1-F; data publication/record stage]: Person or organisation creating this resource. [P1 - ESSENTIAL - F]

Data Format [P1-IR; data processing stage]: The format of the data is considered [P1 - ESSENTIAL - IR] to get to know the structure of the information inside.

Dataset information [P1-F; data storage]: Data files might be part of one or multiple datasets. This field keeps the relationship between the file and the dataset it belongs to. [P1 - ESSENTIAL - F]

Experiment Date [P1-FA; experiment stage]: the actual date when the experiment/measurement is performed. [P1 - ESSENTIAL - FA]

[149] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). ExPaNDS D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312825

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Experiment Description [P1-F; proposal stage]: Provides the experimental information and context for the proposal. It shall include information on the overall objectives, a summary of the experimental method, and expected outcomes. [P1-ESSENTIAL-F]

Experiment Planning [P2-FR; experiment stage]: Experiment planning at this stage aims to complement the [Detailed Experimental Planning] already listed in the previous stage (Scheduling). Nevertheless, unforeseen changes may arise in between. That is why this field is considered [P2-IMPORTANT-FR].

Facility Information (Name component only) [P1-F; proposal stage]: the name of the facility and its information must be explicitly added to the metadata fields. As the data might "travel" from one facility to another, or might be exposed on an EOSC platform, identifying the facility is an [P1-ESSENTIAL-F] step that was not foreseen in previous policies but must be explicitly made.

Funding Source [P2-F; proposal stage]: Legal entity or project funding the proposal submitted by the Principal Investigator. [P2-IMPORTANT-F]

Instrument Information (Name, Organisation, ID components only) [P1-FR; experiment stage]: Details of the Instrument and its status is [P1 - ESSENTIAL - FR] for understanding an experiment performed in the past. This information may also incorporate the software (and versions) that were used for data acquisition. Again, the details provided by the facility will be decided by the facility.

Instrument Scientist [P2-F; experiment stage]: provides support to the Experimental Team while the experiment is performed and serves as instrument expert to ensure the best outcome of the measurement time. [P2-IMPORTANT-F].

License [P1-IR; publication/record stage]: Will inform any data consumer about what can be done with the data and how authorship must be treated. [P1 - ESSENTIAL - IR]

Persistent Identifiers [PI-FA; data storage stage]: Unique identifier within or outside the organisation that is linked to the data files or datasets. [P1 - ESSENTIAL - FA]

Principal Investigator/Main Proposer [P1-FA; proposal stage]: Scientist who will act as the representative of the scientific group which is applying for experiment/measurement time at the facility. The principal investigator is considered either a person (user identity) or an organisation. In the case of a person, the submission system will also store the user institution as an attribute of the user identity. [P1-ESSENTIAL - FA]

Publisher [P1-FI; data publication/record stage]: person or organisation publishing this record. [P1 - ESSENTIAL - FI]

Related Resource [P2-F; data publication/record stage]: it would be either publications, proposals, other datasets. [P2 - IMPORTANT - F]

Release Date (Year component only) [P1 –IR; publication/record stage]: Embargo period due date. The day when the dataset becomes Open Data. [P1 - ESSENTIAL - IR]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Representation Information [P3-IR; data storage stage]: Format and structure of the files linked to the datasets. [P3 - USEFUL - IR]

Resource Identity [P1-FI; data publication/record stage]: should include the type of identifier, the identifier itself, and any related resource linked to it. [P1 - ESSENTIAL - FI]

Sample [P1-F; proposal stage]: Declaration of the samples which will be measured during the experiment/measurement. This field will contain at least the description of the sample as an attribute of the Sample itself. [P1-ESSENTIAL-F]. In the proposal phase, the declaration of the sample will only contain basic information. In most cases, the sample does not exist at this point, and additional details (e.g. structure, or shape - if considered important) will be added at the Experiment stage.

Sample Information [P1-FR; experiment stage]: The information about the sample and its features must be stored in this field. The metadata linked to the sample information field can cover its formula, its characteristics, or even the laboratory where it has been grown. This field is considered [P1 - ESSENTIAL - FR], but the amount of detail provided by each facility may vary.

Title [P1-F; data publication/record stage]: Public name for the dataset. [P1 - ESSENTIAL - F] for data citation.

Visiting Experimental Team [P1-FA; experiment stage]: In the Experiment stage context, the Experimental team refers to the group of people who actually participate during the measurement or experiment. This field or fields identify who they are and what their affiliation is. For Findability purposes, this field is considered [P1 - ESSENTIAL - FA].

As well as the metadata types listed above, other metadata types are mentioned in Table 6 as examples of how it would be necessary to map other metadata types for other types of datasets beyond the raw dataset example that is presented. Thus, we also provide the relevant metadata type definitions:

Analysis Team [P2-AIR; analysis stage]: the team performing the analysis should be identified because this process can be done by a different group from the one who is collecting the data or even was part of the proposal. Although it might be relevant in case of needing to contact them for clarification when reproducing the analysis, this information is considered [P2 - IMPORTANT - AIR] but not essential for FAIR.

Software Package Information [P1-IR; analysis stage]: Any result of the analysis stage must contain the software package, or packages, used to analyse the data, as well as its version and the software configurations used, if possible. This metadata field is considered as [P1 - ESSENTIAL - IR]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Appendix B: Using the 'prov' Python Library to Generate a Software Provenance Graph

Provenance data with Python: simple example for MX

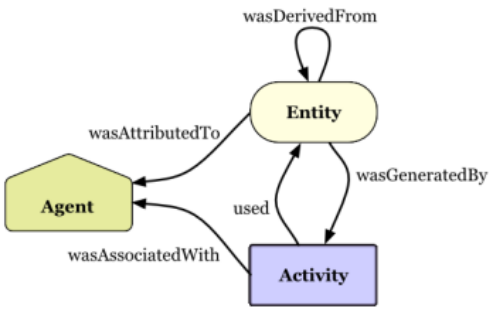
Sources:

- https://trungdong.github.io/prov-python-short-tutorial.html
- https://nbviewer.org/github/trungdong/notebooks/blob/master/PROV%20Tutorial.ipynb

Nicolas Soler, 21 April 2021

Done in a Conda environment using **Python 3.7.10** (default, Feb 26 2021, 18:47:35)

Note: we'll use only the **key concepts** from PROV-O Illustrated in the following picture:



In [1]:
```
# Using the prov library: https://github.com/trungdong/prov
# installed with: conda install -c conda-forge prov
from prov.model import ProvDocument

# Visualization tools
from prov.dot import prov_to_dot
from IPython.display import Image
```

In [2]:
```
# Create a new empty provenance document
d1 = ProvDocument()
```

Namespaces

This is where you provide DOIs or URLs to the entities, agents and activities you are referring to

In [3]:
```
# Declaring namespaces for various prefixes used in the example (for agents and entities
d1.add_namespace('xds', 'https://xds.mr.mpg.de/')
d1.add_namespace('aimless', 'https://www.ccp4.ac.uk/html/aimless.html')
d1.add_namespace('phaser', 'https://www.phaser.cimr.cam.ac.uk/index.php/Phaser_Crystallo
```

Out[3]:
```
<Namespace: phaser {https://www.phaser.cimr.cam.ac.uk/index.php/Phaser_Crystallographic_
Software}>
```

In [4]:
```
d1.add_namespace('experiment', 'some DOI for this experiment')
```

Out[4]:
```
<Namespace: experiment {some DOI for this experiment}>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

In [5]:
```
# Note, the data analysis used another artefact, which is the PDB model for molecular re
# We have to include it here
d1.add_namespace('pdb','https://www.ebi.ac.uk/pdbe/entry/pdb/')
```

Out[5]:
```
<Namespace: pdb {https://www.ebi.ac.uk/pdbe/entry/pdb/}>
```

Entities, agents, activities

In [6]:
```
# Entities (data)
raw_data = d1.entity('experiment:raw_data')
integrated_data = d1.entity('experiment:integrated_data')
reduced_data = d1.entity('experiment:reduced_data')
analysed_data = d1.entity('experiment:analysed_data')

# Entities (input parameters)
input_param_integration = d1.entity('experiment:input_param_integration')
input_param_reduction = d1.entity('experiment:input_param_reduction')
input_param_phasing = d1.entity('experiment:input_param_phasing')
```

In [7]:
```
# pdb_model
pdb_model = d1.entity('pdb:5m7q')
```

In [8]:
```
# Agents (software), with their version
integration_sw = d1.agent('xds:BUILT=20220110')
reduction_sw = d1.agent('aimless:0.7.7')
phasing_sw = d1.agent('phaser:2.8')
```

In [9]:
```
# Activities (processes)
data_integration = d1.activity('experiment:integrates')
data_reduction = d1.activity('experiment:reduces(scale_merge)')
data_phasing = d1.activity('experiment:get_phases')
```

Now establish the relationships between entities (data), agents(sw) and activities

In [10]:
```
# Was derived from (raw to processed data)
d1.wasDerivedFrom(integrated_data, raw_data)
d1.wasDerivedFrom(reduced_data, integrated_data)
d1.wasDerivedFrom(analysed_data, reduced_data)
```

Out[10]:
```
<ProvDerivation: (experiment:analysed_data, experiment:reduced_data)>
```

In [11]:
```
# Activity associated to agent
d1.wasAssociatedWith(data_integration, integration_sw)
d1.wasAssociatedWith(data_reduction, reduction_sw)
d1.wasAssociatedWith(data_phasing, phasing_sw)
```

Out[11]:
```
<ProvAssociation: (experiment:get_phases, phaser:2.8)>
```

In [12]:
```
# Entity (output data) generated by Activity (process)
d1.wasGeneratedBy(integrated_data, data_integration)
d1.wasGeneratedBy(reduced_data, data_reduction)
d1.wasGeneratedBy(analysed_data, data_phasing)
```

Out[12]:
```
<ProvGeneration: (experiment:analysed_data, experiment:get_phases)>
```

In [13]:
```
# Conversely, we have the 'used' relationship between activities and entities (input dat
d1.used(data_integration, raw_data)
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

```
d1.used(data_reduction, integrated_data)
d1.used(data_phasing, reduced_data)

# input parameters
d1.used(data_integration, input_param_integration)
d1.used(data_reduction, input_param_reduction)
d1.used(data_phasing, input_param_phasing)
```

Out[13]:
```
<ProvUsage: (experiment:get_phases, experiment:input_param_phasing)>
```

In [14]:
```
# The pdb model was also used for phasing
d1.used(data_phasing, pdb_model)
```

Out[14]:
```
<ProvUsage: (experiment:get_phases, pdb:5m7q)>
```

In [15]:
```
# Finally, we link agents and entities with "wasAttributedTo"
d1.wasAttributedTo(integrated_data, integration_sw)
d1.wasAttributedTo(reduced_data, reduction_sw)
d1.wasAttributedTo(analysed_data, phasing_sw)
```

Out[15]:
```
<ProvAttribution: (experiment:analysed_data, phaser:2.8)>
```

PROV-N output

PROV-N is a human-readable format for this kind of sw provenance data

In [16]:
```
print(d1.get_provn())
```

```
document
  prefix xds <https://xds.mr.mpg.de/>
  prefix aimless <https://www.ccp4.ac.uk/html/aimless.html>
  prefix phaser <https://www.phaser.cimr.cam.ac.uk/index.php/Phaser_Crystallographic_Sof
tware>
  prefix experiment <some DOI for this experiment>
  prefix pdb <https://www.ebi.ac.uk/pdbe/entry/pdb/>

  entity(experiment:raw_data)
  entity(experiment:integrated_data)
  entity(experiment:reduced_data)
  entity(experiment:analysed_data)
  entity(experiment:input_param_integration)
  entity(experiment:input_param_reduction)
  entity(experiment:input_param_phasing)
  entity(pdb:5m7q)
  agent(xds:BUILT=20220110)
  agent(aimless:0.7.7)
  agent(phaser:2.8)
  activity(experiment:integrates, -, -)
  activity(experiment:reduces(scale_merge), -, -)
  activity(experiment:get_phases, -, -)
  wasDerivedFrom(experiment:integrated_data, experiment:raw_data, -, -, -)
  wasDerivedFrom(experiment:reduced_data, experiment:integrated_data, -, -, -)
  wasDerivedFrom(experiment:analysed_data, experiment:reduced_data, -, -, -)
  wasAssociatedWith(experiment:integrates, xds:BUILT=20220110, -)
  wasAssociatedWith(experiment:reduces(scale_merge), aimless:0.7.7, -)
  wasAssociatedWith(experiment:get_phases, phaser:2.8, -)
  wasGeneratedBy(experiment:integrated_data, experiment:integrates, -)
  wasGeneratedBy(experiment:reduced_data, experiment:reduces(scale_merge), -)
  wasGeneratedBy(experiment:analysed_data, experiment:get_phases, -)
  used(experiment:integrates, experiment:raw_data, -)
  used(experiment:reduces(scale_merge), experiment:integrated_data, -)
  used(experiment:get_phases, experiment:reduced_data, -)
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

```
  used(experiment:integrates, experiment:input_param_integration, -)
  used(experiment:reduces(scale_merge), experiment:input_param_reduction, -)
  used(experiment:get_phases, experiment:input_param_phasing, -)
  used(experiment:get_phases, pdb:5m7q, -)
  wasAttributedTo(experiment:integrated_data, xds:BUILT=20220110)
  wasAttributedTo(experiment:reduced_data, aimless:0.7.7)
  wasAttributedTo(experiment:analysed_data, phaser:2.8)
endDocument
```

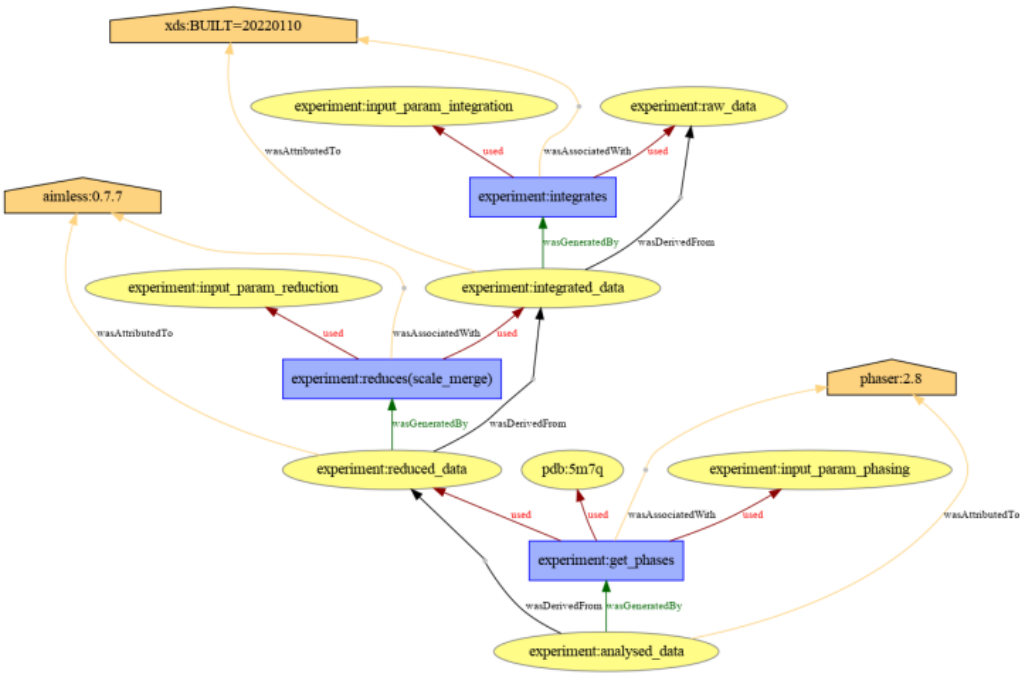
Graphical output

In [17]:
```
# Note: I had a bug here: module 'os' has no attribute 'errno'
# Fixed by installing graphviz on my Ubuntu (sudo apt install graphviz)

out_name = "MX-workflow.png"

dot = prov_to_dot(d1) # generates a graph object
dot.write(out_name, format='png') # generates an inage from the graph
Image(out_name) # displays the image
```

Out[17]:



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.