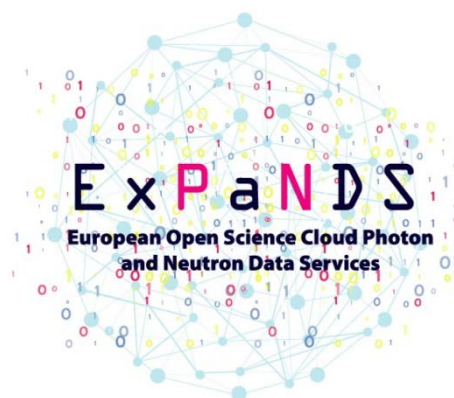


# Report on status, gap analysis and roadmap towards harmonised and federated metadata catalogues for EU national Photon and Neutron RIs



## Document Control Information

Settings	Value
Document Identifier:	D3.1
Project Title:	ExPaNDS
Document Authors:	Alun Ashton (PSI), Silvia Da Graca Ramos (DLS), Alejandra Gonzalez Beltran (UKRI)
Responsible Partner:	PSI
Doc. Issue:	1
Dissemination level:	Public
Date:	03/11/2020

## Abstract

The ExPaNDS project aims at deploying into EOSC Data Catalogues and data analysis services. This document describes the status, a gap analysis, and a roadmap required to achieve harmonised and federated (meta)data catalogues within EOSC of the participating national Photon and Neutron (PaN) Research Infrastructures (RIs).

## License

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA



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

## Table of Contents

Executive Summary	2
1. Introduction	3
2. Current Status of ExPaNDS PaN Facility Metadata Catalogues	3
2.1 Survey Summary	5
3. ExPaNDS architecture – Catalogue Service and Roadmap	5
3.1 Prerequisites for a Metadata Catalogue	6
3.2 Choice – Choosing a Metadata Catalogue	8
3.3 Minimum Product for a Metadata Catalogue	9
3.4 Facility Integration Status	10
3.5 EOSC Integration	10
3.6 PaN Specific Implementations	11
3.7 FAIR Data	13
4. Gap Analysis	13
4.1 Method	13
4.2 Results and Assessment	13
5. Conclusions and Future	14
References	14

## Executive Summary

This deliverable describes and quantifies the status and roadmap being followed by the ExPaNDS National Research Infrastructures (RIs) focussing on the technical implementation and infrastructure required to achieve harmonised and federated metadata catalogues. Please note that these catalogues hold metadata and provide access to the data. The overall infrastructure solution is called the metadata catalogue or data catalogue interchangeably in this document.

The document enumerates the core prerequisites, stages and future features envisioned as part of section 2.2 PaN Data Catalogue Services (Search) of the General Architecture description in relation to the EOSC services [ (Scardaci 2020)]

The first part of the document elaborates on our roadmap, which every facility is undertaking on the journey towards becoming an EOSC contributor and implementing the General Architecture. Each facilities' journey is different due to the heterogeneous nature of PaN facilities. The variations can come from a facilities age, scale, national requirements and priorities, which has led to a mixed landscape of software, implementations, and lifecycle, which we will quantify. We will then present the status of the ExPaNDS PaN facilities and our measure of where we currently are as a gap analysis.



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

## 1. Introduction

The European Open Science Cloud (EOSC) Photon and Neutron Data Services (ExPaNDS) is an ambitious project that will enable the provision of EOSC services to the scientific users of Photon and Neutron (PaN) sources. The need for new federated services to enable data access, analysis and sharing is crucial for advances in science. Metadata catalogues play an important role for finding, accessing and using research data. Furthermore, making data Findable Accessible Interoperable and Reusable (FAIR) will enable facility users to manage their own data, reprocess it in the European Cloud, share it with collaborators and increase transparency in their scientific results. The alignment of WP3 requirements with WP2 (Enabling FAIR data) and WP4 (Data Analysis as a Service), and with complementary work in PaNOSC ([www.panosc.eu](http://www.panosc.eu)) and EOSC can lead to a successful implementation of FAIR data in the metadata catalogues provided by national RIs.

The ExPaNDS project covers a large number of national RIs across Europe: ALBA (Spain), DESY (Germany), Diamond Light Source (UK), Elettra (Italy), HZB (Germany), HZDR (Germany), UKRI-STFC ISIS (UK), Max IV (Sweden), PSI (Switzerland) and Soleil (France). The STFC Scientific Computing Department and EGI are also partners in this project.

The goals of this work package (WP3) are to support the delivery of FAIR data and provide services to the EOSC data analysis services. These goals stand on a number of key pillars:

Each facility must

1. Determine the legal ownership and lifecycle of experimental data at their facilities (Data Policy).
2. Have the technical capabilities to manage the lifecycle of the experiment data.
3. Have the technical capabilities, software and infrastructure to capture, record and publish metadata associated with experimental data at their facilities.

To facilitate FAIR and interoperable data for the PaN domain, the community must

4. Determine and agree standards and descriptions for metadata captured to describe the experiment and its results including its lifecycle.
5. Agree on standards for accessing and publishing the data.
6. Agree on compatible Authentication and Authorisation Infrastructures (AAls).

While ExPaNDS aims to facilitate all these areas, along with its key partner PaNOSC, none of the later ambitions can be realised without the facilities having invested in and provisioned for a metadata catalogue for recording experimental data and taken the steps required to integrate their service as harmonised and federated metadata catalogues in conjunction with other EU national Photon and Neutron RIs.

Below, the current status of the facilities will be outlined, a roadmap to deliver, and a measure of the progress made by ExPaNDS to enable PaN RI's at a national level, to become a part and enhance the EOSC.

## 2. Current Status of ExPaNDS PaN Facility Metadata Catalogues

A survey was conducted in December 2019 to understand the status of the metadata catalogues across the ExPaNDS RIs [ (Ashton 2019)]. The results show that the national RIs present diverse degrees of maturity of their metadata catalogues. At the time of the survey, some facilities did not have any metadata catalogue while others had all their instruments connected to it and were already



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

making their data publicly available with Persistent Identifiers (PIDs). However, almost all RIs already have in place a Data Policy with a Data Embargo period specified. The two main metadata cataloguing tool suites developed and used in X-ray and Neutron facilities are: ICAT (including ICAT+, an extension of ICAT)[ (ICAT 2010)] and SciCat (SciCat 2018)]. Some national RIs are using custom solutions but are investigating alternatives as they could benefit for more collaborative software development.

The survey results have been updated in September 2020. The main differences are that some RIs have started testing new metadata catalogue solutions and one RI has benefited from the COVID-19 situation to prioritise the need to publish data with DOIs and made COVID-19 related results publicly available. According to the new responses, facilities with custom solutions for their metadata catalogue are now installing/testing one of the two main catalogues i.e. ICAT and SciCat. However, generally, the ongoing COVID-19 situation had a negative impact on recruitment and availability of technical staff and accordingly progress.

Question	Response
Data catalogue used:	ICAT: 3 facilities ICAT+: 1 facility (not in production) SciCat: 2 facilities Custom but investigating/testing SciCat: 1 facility Custom but investigating/testing ICAT: 1 facility Under investigation: 2 facilities
Data catalogue URL:	URL provided: 7 facilities ( 2 sites are internal and the others external) <a href="https://scicat.maxiv.lu.se/">https://scicat.maxiv.lu.se/</a> <a href="https://discovery.psi.ch/">https://discovery.psi.ch/</a> <a href="https://gamma-portal.desy.de">https://gamma-portal.desy.de</a> <a href="https://icat.diamond.ac.uk">https://icat.diamond.ac.uk</a> <a href="https://topcat.helmholtz-berlin.de">https://topcat.helmholtz-berlin.de</a> <a href="https://vuo.elettra.eu">https://vuo.elettra.eu</a> <a href="https://data.isis.stfc.ac.uk/">https://data.isis.stfc.ac.uk/</a>
Login required (non Anonymous access):	Yes: 6 facilities (for federated data only) No: 1 facility N/A: 3 facilities
File formats archived:	Only HDF5: 1 facility Only HDF5 and NeXus: 2 facilities Only NeXus, HDF5, ascii, spec: 1 facility Only NeXus, HDF5, ascii, tiff: 1 facility Only NeXus, ascii, isis raw, image formats: 1 facility All formats generated: 3 facilities NA: 1 facility
Database used for metadata of data catalogue:	MongoDB: 2 facilities MongoDB + PostgreSQL: 1 facility Oracle: 3 facilities MariaDB: 1 facility Oracle + PostgreSQL: 1 facility NA: 2 facilities
Primary software language of data catalogue interface:	JavaScript: 2 facilities Java: 2 facilities Java + Javascript: 2 facilities Python + PL/SQL: 1 facility



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

	PL/SQL: 1 facility NA: 2 facilities
<b>Main technology of data catalogue interface:</b>	Angular: 6 facilities React: 1 facility APEX: 1 facility NA: 2 facilities
<b>Number of public dataset/files:</b>	28037: 1 facility 14872: 1 facility 61: 1 facility 10: 1 facility 2: 1 facility 0: 3 facilities NA: 2 facilities
<b>Using OAI-PMH:</b>	Yes: 1 facility No: 8 facility N/A: 1 facility
<b>Minting DOIs/PID:</b>	Yes: 2 facilities No: 3 facilities In Progress: 4 facilities N/A: 1 facility
<b>Data embargo policy:</b>	No embargo (Data is the property of the PI): 1 facility 5 years: 1 facility 3 years: 4 facilities 3 years extensible to 4 years: 1 facility 3 years extensible twice a year with justification: 1 facility Will be PAN-compliant: 1 facility NA: 1 facility
<b>Data embargo policy URL:</b>	URL provided: 6 facilities
<b>Number of instruments connected to the data catalogue:</b>	0: 3 facilities 1: 1 facility 2: 1 facility 15: 1 facility 29: 1 facility 37: 1 facility 45: 1 facility N/A: 1 facility

## 2.1 Survey Summary

The results of the survey highlight the significant synergies that already exist between national Photon and Neutron RIs (e.g. ICAT, SciCat and NeXus). The results also show how significant work and collaboration has already been undertaken to capture and standardise the experiment data at a facility level (over 40,000 files or datasets available open access). ExPaNDS aims to build on these existing initiatives and lead the EU national Photon and Neutron RIs towards harmonised and federated metadata catalogues for FAIR PaN data in the EOSC.

## 3. ExPaNDS architecture – Catalogue Service and Roadmap



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

The ExPaNDS Catalogue Service architecture has been outlined as part of the Deliverable 1.5 for the General Architecture. However, the ExPaNDS facilities have all progressed at differing rates due to the varied levels of national funding and prioritisation. Figure 1 outlines the relationship between the Architecture of the Catalogue Service and the ExPaNDS roadmap to support the infrastructure for harmonised and federated metadata catalogues for EU national Photon and Neutron RIs.

The aim of the roadmap is to give a quantitative measure of the status and later, the progress of ExPaNDS in enabling national EU PaN experimental data to contribute to the EOSC. The difference between the current status and desired complete outcome will form the basis of the gap analysis. The current results also highlight areas where focus can be directed so that ExPaNDS deliverables and activities will make an impact.

Each step in the roadmap is broken down to a defined subset of criteria that can be followed iteratively to move from no metadata catalogue to an EOSC-compliant metadata catalogue solution. Each of the steps in the roadmap can be further broken down to a series of questions used to measure progress, as described below (see also Table 2).

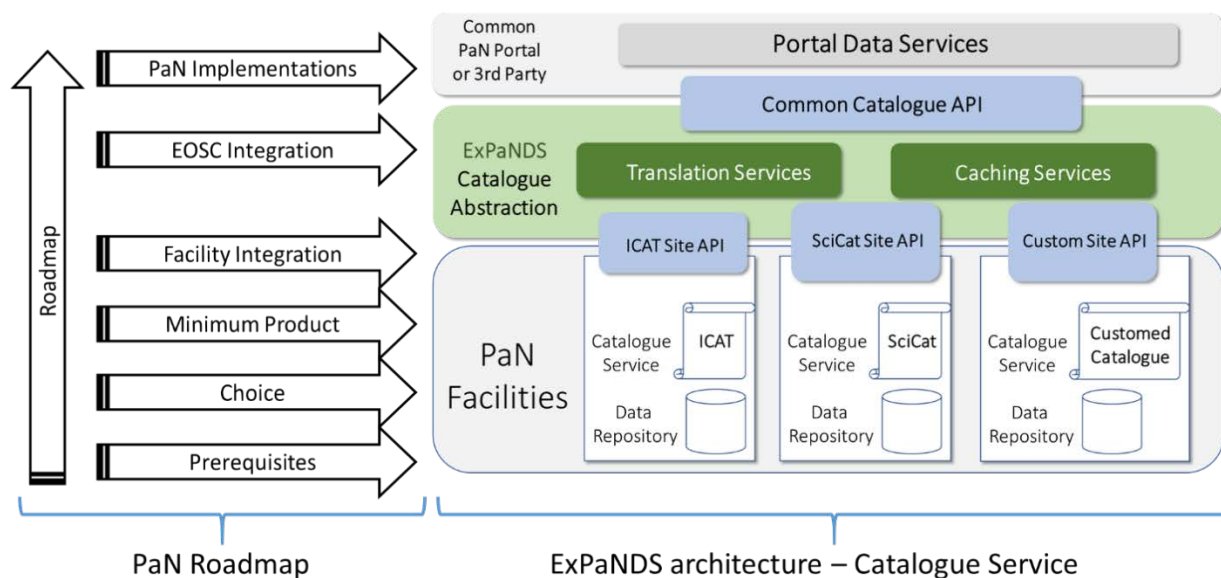


Figure 1: ExPaNDS architecture – catalogue service (right) and roadmap for implementation (left)

## 3.1 Prerequisites for a Metadata Catalogue

The prerequisites described in this section give guidelines to help a facility identify the main components for the catalogue, and to ensure they have the foundations needed to host, support and enable a metadata catalogue within EOSC. Each guideline is introduced below, including a description and a question for data managers, whose answer will indicate the level of fulfilment of the criteria.

### 3.1.1 Infrastructure

*Has your facility provisioned the required infrastructure for a metadata catalogue and associated services?* The need for infrastructure, i.e. appropriate hardware and services, to host a reliable



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

metadata catalogue and associated data cannot be underestimated. With a PaN facility capable of producing petabytes of data on an annual basis with references to millions of files and datasets the infrastructure required must be established and sourced.

## 3.1.2 Facility Proposal Info

*Does your facility have a UOS and make it accessible to be integrated to your metadata catalogue services?* When a user applies to use a PaN instrument at a facility through its proposal system, the facility starts collecting useful information that can be added to the catalogue. This includes user information (potentially including authentication and authorisation details), proposal and scheduling. The facility system in charge of this is frequently referred to as a user office system (UOS) and is seen as an essential source of information which must be accessible, according to data access policies, to be included in a metadata catalogue.

## 3.1.3 Data Standards

*Has your facility agreed on and started to migrate to a Standard Data format?* PaN experiment data sources can vary significantly depending on the age and source of detectors being utilised and the software in control of the experiment. There are a number of initiatives for the standardisation of the data being recorded and archived. The most prevalent at PaN facilities is the NeXus data format [ (M. Könecke 2015)]. Standardising will make the data more accessible and ultimately easier to record in a metadata catalogue and utilising open specifications like NeXus will ensure compliance with the EOSC draft interoperability framework.

## 3.1.4 PID Issuer

*Has your facility secured a PID Issuer?* To enable a dataset to be uniquely identified, a facility must ensure they have the legal agreements in place to host and issue a persistent identifier compliant with the EOSC PID policy as outlined in EOSC draft interoperability framework document [ (Hellström 2020)].

## 3.1.5 Data Policy

*Does your facility have a Data Policy?* An institutional data policy will cover the data ownership, the institutions legal responsibility to data stewardship and the data retention and publication lifecycle.

## 3.1.6 Resources

*Has your facility put in place/identified the roles to support/maintain and develop your metadata catalogue services?* As previously eluded to in 3.1.1, as well as significant provision for hardware for a metadata catalogue and supporting services, either on premise or at a third-party provider, there is a need for clear roles and responsibilities within the facility to support, maintain and develop the infrastructure.

## 3.1.7 Summary of Prerequisite Questions

Prerequisite	Question
Infrastructure:	<i>Has your facility provisioned the required infrastructure for a metadata catalogue and associated services?</i>
Facility proposal info:	<i>Does your facility have a UOS and made it accessible to be integrated to your metadata catalogue services?</i>
Standard data:	<i>Has your facility agreed on and started to migrate to a Standard Data format?</i>
PID issuer:	<i>Has your facility secured a PID Issuer?</i>



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



<b>Data policy:</b>	<i>Does your facility have a Data Policy?</i>
<b>Resources:</b>	<i>Has your facility put in place/identified the roles to support/maintain and develop your metadata catalogue services?</i>

TABLE 2: SUMMARY TABLE OF PREREQUISITES QUESTIONS

## 3.2 Choice – Choosing a Metadata Catalogue

The choice of a metadata catalogue can have a significant impact on ease and time to delivery. There are now a number of choices that can be selected and the items below is aimed at outlining and measuring the progress in a selection. The summary of questions for each aspect is given in Table 3.

### 3.2.1 Clear and Agreed Requirements

*Has your facility agreed and prioritised the requirements?* An essential step in choosing a metadata catalogue is to agree on the needs and requirements with the key stakeholders such as facility users, management and support staff. On a practical level, one requirement could be the need to harvest enough information with the data to allow an independent researcher to process and analyse the data without contacting the data creator. There are other considerations that need to be prioritised such as the need to meet funding sources and national data retention and publication policies for data (see ExPaNDS deliverable 2.1 [ (Matthews 2020)]) as well as those of international funding bodies such as the European Union.

### 3.2.2 Solutions Investigation

*Has your facility undertaken an investigation of the currently available solutions?* With more than one potential solution for a metadata catalogue available within the PaN and wider community, an element of research into the solutions is required. This area is advancing at a significant pace and currently lacks a good online resource to refer to. However, within ExPaNDS work package three, a later deliverable will help with the demonstration of the SciCat and ICAT metadata catalogues' and training material for a reference implementation to.

### 3.2.3 Test Candidate Solutions

*Has your facility tested candidate solutions?* One or more solutions from the results of 3.2.2 will need to be tested for suitability at a facility.

### 3.2.4 Decision

*Has your facility agreed a way forward/decision on the choice of a metadata catalogue?* Once the research has been completed, a decision needs to be made and conveyed to the stakeholders. The decision will vary from facility to facility depending on the agreements reached in 3.2.1 and will vary from joining an existing collaboration, outsourcing or purchasing a commercial solution to developing a new solution in house.

### 3.2.5 Summary of Choice Questions

Choice	Question
<b>Clear and agreed requirements:</b>	<i>Has your facility agreed and prioritised the requirements?</i>
<b>Solutions researched</b>	<i>Has your facility undertaken an investigation of the currently available solutions?</i>
<b>Solutions tested</b>	<i>Has your facility tested candidate solutions?</i>



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



Decision	Has your facility agreed a way forward/decision on the choice of a metadata catalogue?
----------	--

TABLE 3: SUMMARY TABLE OF CHOICE QUESTIONS

## 3.3 Minimum Product for a Metadata Catalogue

A minimal product is the first stage of deployment an integration of a functional metadata catalogue at a PaN facility.

### 3.3.1 File Ingest

*Is your metadata catalogue deployed and capable of ingest data file metadata into metadata catalogue and be correctly associated to experiments?* The vast majority of measurements at a PaN facility leads to the creation of a file on disk media. This data (raw or derived) must subsequently be associated with the experiment parameters agreed to be recorded in the metadata catalogue. This may lead to duplication of the metadata that resides both in the file and the metadata catalogue, however the metadata catalogue metadata will render itself searchable and easily accessible across experiments.

### 3.3.2 Authenticated and authorised data access

*Has your deployed metadata catalogue been integrated with an authentication and authorisation layer?* As metadata and data associated with the experiment will be subject to the experiment lifecycle as described in the data policy, the metadata catalogue will need a federated role-based access to allow for e.g. specified embargo periods where only the data owner would be able to access the data.

### 3.3.3 Domain agnostic metadata

*Is your facilities deployed metadata catalogued recording Domain agnostic metadata?* One of the aims of ExPaNDS and EOSC is to have experiment data available to a wider audience of researchers. To facilitate this goal a number of domain agnostic metadata standards have been published e.g. Dublin Core [ (Core 2020), Datacite [(Datacite 2020)] or similar bibliographic data. This in turn can be used to produce a PID.

### 3.3.4 PaN API

*Does your facility deployed metadata catalogue has the PaN Common Catalogue API?* As part of the PaNOSC and ExPaNDS collaboration, a Common Catalogue API facilitates a harmonised search between different PaN RIs metadata catalogues.

### 3.3.5 Stability

*Has your metadata catalogue achieved the expected level of stability/service resilience?* The services of a metadata catalogue can be utilised in a number of different ways e.g.

- A real time critical component of an experiment.
- A less time crucial service for data discovery potentially with a service level agreement and expectation by data consumers.

Although guaranteed to evolve, the stability of the service will need to be established as part of the prerequisites and requirements and will vary from facility to facility.

## 3.5.6 Summary of Minimal Product Questions

Minimum Product	Question
-----------------	----------



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

<b>File ingest:</b>	<i>Is your metadata catalogue deployed and capable of ingest data file metadata into the chosen catalogue and be correctly associated to experiments?</i>
<b>Authenticated data access:</b>	<i>Has your deployed metadata catalogue been integrated with an authentication and authorisation layer?</i>
<b>Domain agnostic data:</b>	<i>Is your facilities deployed metadata catalogued recording Domain agnostic metadata?</i>
<b>PaN API:</b>	<i>Does your facilities deployed metadata catalogue have the PaN Common Catalogue API?</i>
<b>Stability:</b>	<i>Has your metadata catalogue achieved the expected level of stability/service resilience?</i>

TABLE 4: SUMMARY TABLE OF MINIMUM PRODUCT QUESTIONS

## 3.4 Facility Integration Status

The rollout of a metadata catalogue at a PaN facility is most commonly iterative taking into account of one or more instrument or experiment at a time. Although this is partially due to resource limitations, it is also good practice to ensure scalability and stability of the operations as well as being able to account for any customisations needed.

Facility Integration	Question
<b>10% &lt; 25%</b>	<i>Have you integrated between 10 – 25% of your facilities instruments?</i>
<b>25% &lt; 50%</b>	<i>Have you integrated between 25 – 50% of your facilities instruments?</i>
<b>50% &lt; 75%</b>	<i>Have you integrated between 50 – 75% of your facilities instruments?</i>
<b>75% &lt;= 100%</b>	<i>Have you integrated between 75 – 100% of your facilities instruments?</i>

TABLE 5: SUMMARY TABLE OF THE FACILITY INTEGRATION STATUS QUESTIONS

## 3.5 EOSC Integration

A number of the ExPaNDs facilities already have advanced metadata catalogues but none have yet been able to fully integrate them into the EOSC community. A single facility cannot unilaterally be part of the EOSC-hub as it has to achieve an integrated environment, and standards and agreements must be reached with other PaN facilities and the wider EOSC community. Below are the key implementations that will be addressed by ExPaNDs.

### 3.5.1 Federated Access to Data

*Does your metadata catalogue support the Umbrella federated identity system?* In section 3.3.3 Authenticated Data Access, it was highlighted that a need for authentication and authorisation layer was needed to enforce the data publication policy. However, to facilitate this in a cross RI manner the Umbrella federated identity system has been agreed to be utilised. This will enable ExPaNDs to define a common security and privacy framework to ensure secure and trustworthy data exchange and All process within the community as stipulated in the EOSC draft interoperability framework document.

### 3.5.2 OAI-PMH Implementation

*Does your facility metadata catalogue have an open OAI-PMH interface?* “The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP.” [<https://www.openarchives.org/pmh/>]. OAI-PMH has been agreed, within the European PaN facilities, to be utilised to aggregate the metadata in EOSC-compliant



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

services supporting data findability across facilities. Partner site have already developed an OAI-PMH plug-in for the two main metadata catalogues at PaN facilities.

### 3.5.3 Harvesting by EOSC Services

*Is the data on your OAI-PMH channel being harvested into an EOSC catalogue service?* Once a facility has an OAI-PMH interface that is correctly publishing data, the metadata can be harvested into EOSC catalogue services e.g. OpenAire [ (OpenAire 2020)] and B2Find [ (B2Find 2020)].

### 3.5.4 EOSC Service Access

*Can your metadata catalogue be consumed by/host/ utilised by EOSC services through the EOSC Portal?* As the ultimate aim of open access data is to increase its value by reusability, the service must be utilised by other EOSC services. Such a service is being developed as work package 4 in ExPaNDS.

### 3.5.5 EOSC Statistics

*Can your metadata catalogue produce and publish access statistics?* A part of 3.5.4 will be to monitor the usability and reusability of the original data and the overall service as a potential quality and value indicator.

### 3.5.6 Summary of EOSC Integration Questions

EOSC Integration	Question
Federated access to data:	<i>Does your metadata catalogue support the Umbrella federated identity system?</i>
OAI-PMH implemented:	<i>Does your facility metadata catalogue have an open OAI-PMH interface?</i>
Harvesting by EOSC services:	<i>Is the data on your OAI-PMH channel being harvested into an EOSC catalogue service?</i>
EOSC services access:	<i>Can your metadata catalogue be consumed by/host/ utilised by EOSC services through the EOSC Portal?</i>
EOSC statistics:	<i>Can your metadata catalogue produce and publish access statistics?</i>

TABLE 6: SUMMARY TABLE OF EOSC INTEGRATION QUESTIONS

## 3.6 PaN Specific Implementations

Open and FAIR data (see WP2 tasks future deliverables D2.2 and D2.7) must be addressed for every science domain and below are outlined the key stages for PaN experiment data.

### 3.6.1 Domain Metadata Ingest

*Does your metadata catalogue include domain specific metadata?* Domain metadata pertaining to the measurement and its results are currently being agreed as part of an ontology in ExPaNDS 3.2 task and associated deliverable. The extent of the data that will be captured within the metadata catalogue will have been agreed both as part of the facility metadata catalogue requirements and as part of the ontology agreement.

### 3.6.2 RAW and Derived Data Catalogued

*Does your metadata catalogue contain and distinguish between raw and derived experimental data?* The definitions associated with raw and derived data are addressed elsewhere in WP2 and should also be stated as part of the facilities data policy.



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

## 3.6.3 Data Provenance

*Does your metadata catalogue contain the provenance of the data?* The metadata and file records contained within a metadata catalogue should contain the history of the information and how it was generated and collected.

## 3.6.4 Calibration Data

*Are there links to the required calibration data?* Experiment measurements at PaN facility sites frequently need references to calibration data to allow the results to be processed and freed from any instrument or sample artefacts.

## 3.6.5 Link to external PIDs

*Are there links to external PIDs information (such as Instrument PID, ORCID, Sample PID) associated to the data?* Linking data with other resources, via the resources' PIDs, such as Organisations (i.e ROR or GRID), Users (i.e OrCid or ResearcherId ), Sample and Instrument will facilitate interoperability [ (Fenner, Powering the PID Graph: announcing the DataCite GraphQL API 2020)]. In particular, work on instrument PID has already been performed by the Research Data Alliance Working Group on Persistent Identification of Instruments (PIDINST) [ (Markus Stocker, Persistent Identification of Instruments 2019)]. As well as the calibration data, information to the experimental setup and the instrument itself is valuable metadata for any reviewing or consuming researcher. Furthermore, the ExPaNDS task 2.4 (Persistent Identifier Infrastructure) will provide guidelines for its implementation.

## 3.6.6 Link to Publications

*Are there links to the eventual publications associated with the data?* Although not known at the point of collecting data, a measure of the quality of the data will be a link to its publication either in a journal or similar publication database, especially if it is reviewed.

## 3.6.7 eLogbook Data

*Do you harness and capture freeform information in your metadata catalogue?* Due to the vast range of experiments, experiment parameters and potential interpretation and analysis of the samples and data being collected at PaN facilities, it is not always possible to electronically and automatically capture all the metadata associated with the experimental measurements. To facilitate the capture of this data, electronic logbooks (eLogbooks) can be employed.

## 3.6.8 Summary of PaN Specific Implementation Questions.

EOSC Integration	Question
Domain metadata ingest:	<i>Does your metadata catalogue include domain specific metadata?</i>
RAW and derived data catalogued:	<i>Does your metadata catalogue contain and distinguish between raw and derived experimental data?</i>
Data provenance:	<i>Does your metadata catalogue contain the provenance of the data?</i>
Calibration data:	<i>Are there links to the required calibration data?</i>
Link to instrument reference information:	<i>Are there links to instrument reference information with a UID (PID, ORCID, Sample PID...)?</i>
Link to publication:	<i>Are there links to the eventual publications associated with the data?</i>
eLogbook data	<i>Do you harness and capture freeform information in your metadata catalogue?</i>

TABLE 7: SUMMARY TABLE OF PAN SPECIFIC IMPLEMENTATIONS QUESTIONS



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

### 3.7 FAIR Data

The FAIR principles [ (Wilkinson 2016)] have become increasingly important in the implementation of good data management practices in research institutions. The main goal is to enable both humans and machines to Find, Access, Interoperate and Re-use research data and metadata. More recently, the Research Data Alliance (RDA) tried to address divergent interpretations and ambiguities of the FAIR principles by publishing the FAIR Data Maturity Model [ (Willems 2020)]. This model presents specifications and guidelines for FAIR assessment across research disciplines with the definition of indicators and their level of importance (Essential, Important and Useful).

The ExPaNDS deliverable D2.2 will provide a draft set of recommendations for the implementation of FAIR in Photon and Neutron facilities, revised in D2.7. In particular, it will define a common metadata framework for use in both WP3 and WP4.

## 4. Gap Analysis

The roadmap was used to quantitate (though sometimes subjective) a measure to assess the current status of ExPaNDS facilities between their current implementation and the idealised harmonised and federated metadata catalogues for EU national PaN RIs. This measure will also be used to assess progress.

### 4.1 Method

Each facility rated their current status against the questions within the roadmap. Although each step has a unique factor in challenge, complexity and priority, to keep the model simple every question has a rating of 1. The facility integration status had an accumulative rating of 1 to a maximum of 4.

### 4.2 Results and Assessment

The results of the gap analysis are collated in Figure 2. While progress in the first four stages are not directly addressed in ExPaNDS, the community being created will help every facility progress. The results also show how differing priorities enable and result in facilities addressing different stages of roadmap (seen by the spread of completed or uncompleted tasks for each stage). ExPaNDS directly addresses new functionality and the integration of the PaN facilities into EOSC. As capabilities and agreed standards have yet to be delivered, the gap analysis shows where ExPaNDS will make the biggest impact and progress as few facilities have yet to address the EOSC Integration stage (with a total accumulative score of 3/50) and the PaN specific implementations (with a score of 16/70).



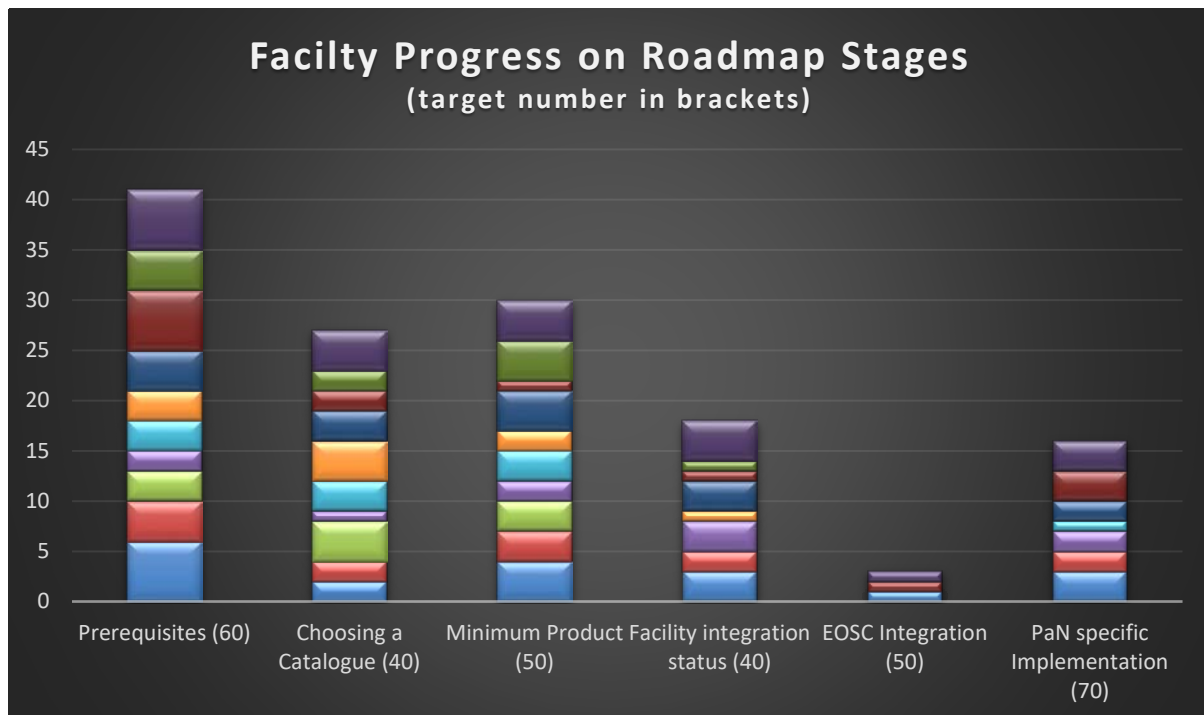


Figure 2: Plot of the accumulative score for ExPaNDS facilities (Y) on each stage of the roadmap, the maximum possible score is in brackets next to the stage name. Each colour represents a facility.

## 5. Conclusions and Future

Whilst progress is being made by the national European Photon and Neutron facilities, the implementation of the infrastructure required to deliver harmonized and federated metadata catalogues varies significantly. However, the overall requirement has now become a fundamental and strategic goal for the facilities.

Metadata catalogues are crucial in the implementation of FAIR data principles. In this document, a roadmap was proposed and a gap analysis across RIs was performed. This will provide a baseline for both tasks 3.4 and 3.5 in the integration of metadata catalogues: (i) to manage the data lifecycle at EU national RIs and (ii) as a service on EOSC-hub. Furthermore, the present roadmap considers requirements from WP2 and data/metadata required for the provision of data analysis services as defined in WP4.

## References

- Ashton, Alun, Da Graca Ramos, Silvia, Matthews, Brian, Salvat, Daniel, & Sander, Knut. 2019. *ExPaNDS Data Landscaping Survey*. <http://doi.org/10.5281/zenodo.3673811>, Zenodo.
- B2Find. 2020. *B2Find*. <https://www.eudat.eu/services/b2find>.
- Core, Dublin. 2020. *Dublin Core Metadata Initiative*. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>.
- Datacite. 2020. *DataCite*. <https://datacite.org/>.
- Fenner, Martin. 2020. *Powering the PID Graph: announcing the DataCite GraphQL API*. <https://doi.org/10.5438/yfck-mv39>.



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



Hellström, Maggie, André Heughebaert, Rachael Kotarski, Paolo Manghi, Brian Matthews, Raphael Ritz, Anders Sparre Conrad, Tobias Weigel, Peter Wittenburg, and Mario Valle. 2020. *Second draft Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC)*. 10.5281/zenodo.3780423.

ICAT. 2010. *ICAT project*. <https://github.com/icatproject>.

M. Könnecke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, S. I. Campbell, B. Clausen, S. Cottrell, J. U. Hoffmann, P. R. Jemian, D. Männicke, R. Osborn, P. F. Peterson, T. Richter, J. Suzuki, B. Watts, E. Wintersberger and J. Wuttke. 2015. *The Nexus data format*. doi:10.1107/S1600576714027575.

Markus Stocker, Louise Darroch, Rolf Krah, Ted Habermann, Anusuriya Devaraju, Ulrich Schwardmann, Claudio D'Onofrio, Ingemar Haggstrom. 2019. *Persistent Identification of Instruments*. <http://doi.org/10.5334/dsj-2020-018>.

Matthews, Brian, Abigail McBirnie, Andrei Vukolov, Alun Ashton, Stephen Collins, Sylvie Da Graca Ramos, Brigitte Gagey, et al. 2020. *Draft extended data policy framework for Photon and Neutron RIs*. 10.5281/zenodo.4014811.

OpenAire. 2020. *OpenAire*. <https://www.openaire.eu/>.

Scardaci, Diego, Daniel Salvat, Patrick Fuhrmann, Anton Barty, Alun Ashton, and Sophie Servan. 2020. *ExPaNDs General Architecture description in relation to the EOSC services*. 10.5281/zenodo.3697704.

SciCat. 2018. *SciCat project*. <https://scicatproject.github.io/>.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data* 3, 160018. doi:<https://doi.org/10.1038/sdata.2016.18>.

Willems, Marieke. 2020. *FAIR Data Maturity Model: specification and guidelines*. 10.15497/rda00045.



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