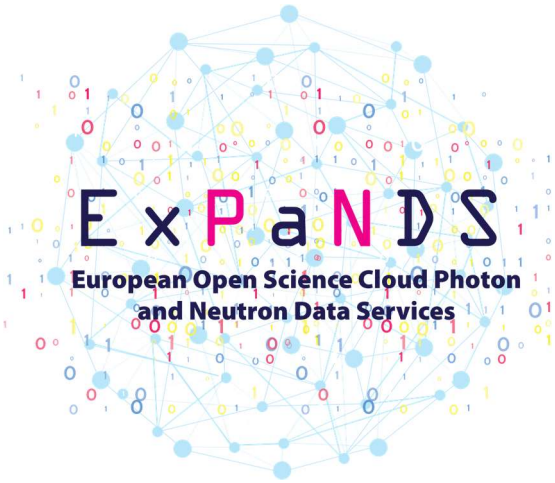

D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management

Document Control Information

Settings	Value
Document Identifier:	D2.2
Project Title:	ExPaNDS
Work Package:	WP2
Work Package Lead	UKRI
Deliverable Lead	ALBA
Document Author(s):	Daniel Salvat (ALBA), Alejandra Gonzalez-Beltran (UKRI), Heike Görzig (HZB), Brian Matthews (UKRI), Abigail McBirnie (UKRI), Majid Ounsy (SOLEIL), Sylvie Da Graca Ramos (DLS), Andrei Vukolov (Elettra)
Document Contributor(s):	Sophie Servan (DESY), Jonathan Taylor (ESS)
Doc. Version:	1.0
Dissemination level:	Public
Date:	14/12/2020

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

Abstract

ExPaNDS WP2 has the objective of Enabling FAIR Data, providing guidelines, recommendation and practical experience to the project and the wider PaN community on the best practise in generating FAIR experimental data for National RIs. In order for the facility to deliver experimental data “FAIR at the point of leaving the facility”, the infrastructure systems and procedures involved in the facilities experimental lifecycle should be coordinated to construct and maintain the FAIRness of that data. In particular, at each stage in the experimental lifecycle, metadata and contextual information should be collected (automatically and manually) to form as complete a record as possible of the experiment so that that data can be FAIR.

In this report, we give an analysis of the metadata that should be generated and recorded at each stage of the lifecycle and consider its contribution to making that data FAIR. This delivers a prioritization which we use to determine which metadata should be collected. We use that prioritization to make a series of recommendations on the types of metadata that should be collected at each stage. These recommendations give a framework for setting standards for common metadata formats and provide a basis for assessing the FAIRness of data generated by facilities. Further work will develop these draft recommendations into our final recommendations for common metadata formats and use of metadata in practice.

Document Log

Version	Date	Comment	Author/Partner
Complete Draft	26/11/2020	Internal review version	Daniel Salvat (ALBA)
1.0	14/12/2020	Final version for submission	Daniel Salvat (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

AAI	Authentication and Authorization Infrastructure
ACILM	Archive Centric Information lifecycle Model
CERIF	Common European Research Information Format
CIF	Crystallographic Information Framework
CIF2	Crystallographic Information Framework II
CLM	Curation Lifecycle Model
DCAT	Data Catalogue Vocabulary
DCC	Digital Curation Centre
DCMI	Dublin Core Metadata Initiative
DESY	Deutsches Elektronen-Synchrotron
DLS	Diamond Light Source
DMP	Data Management Plan
DRUM	Digital Representation of Units of Measure
EC	European Commission
EOSC	European Open Science Cloud
ESS	European Spallation Source
ExPaNDS	European Open Science Cloud Photon and Neutron Data Service
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
HDF	Hierarchical Data Format
HZB	Helmholtz-Zentrum Berlin
HZDR	Helmholtz-Zentrum Dresden-Rossendorf
IUCr	International Union of Crystallography
IUPAC	International Union of Pure and Applied Chemistry
LIMS	Laboratory Information Management System
mmCIF	Macromolecular CIF
NCBI	National Center for Biotechnology Information
NFFA	Nanoscience Foundries & Fine Analysis
NIAC	NeXus International Advisory Committee
ORCID	Open Researcher and Contributor ID
PaN	Photon and Neutron
PaN-data ODI	PaNdata Open Data Infrastructure
PaNOSC	The Photon and Neutron Open Science Cloud
PDB	Protein Data Bank

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

PID	Persistent Identifier
PSI	Paul Scherrer Institute
RDA	Research Data Alliance
RDF	Resource Description Framework
RDM	Research Data Management
RI	Research Infrastructure
UKRI	UK Research and Innovation
W3C	World Wide Web Consortium
WP	Work Package
XML	Extensible Markup Language

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 to bring the experimental data generated by National Photon and Neutron Analytical Research Infrastructures (RIs) within Europe into the scope of the European Open Science Cloud (EOSC), together with services which support the discovery, access and reuse of that data. A key requirement of this aim is to ensure that that data is Findable, Accessible, Interoperable and Reusable (FAIR) so that users across the EOSC can make effective use of that data. Therefore, WP2 of ExPaNDS has the objective of Enabling FAIR Data, providing guidelines, recommendations, and practical experience to the project and the wider PaN community on the best practice in generating FAIR experimental data for National RIs.

Scientific experiments within RIs are on a huge range of topics from across disciplines, using different techniques and instruments. However, they are all within a facility's infrastructure systems and procedures and generally follow a similar overall lifecycle. This lifecycle was detailed in earlier work in the PaNdata-ODI project, in particular, its deliverable D6.1. In order for the facility to deliver experimental data "FAIR at the point of leaving the facility", those infrastructure systems and procedures should be coordinated to construct and maintain the FAIRness of that data. In particular, at each stage in the experimental lifecycle, metadata and contextual information can be collected (automatically and manually) to form as complete a record as possible of the experiment so that that data can be FAIR.

In this report, we provide an initial draft of a common metadata framework to support the generation of FAIR experimental data by facilities. The goal of the common metadata framework is to provide guidance on defining the requirements for FAIR PaN data by defining what information is needed to produce FAIR data, and in which form a research dataset needs to include this information to meet the purpose of future access and reuse. These guidelines should enable data managers and scientists to assess the level of FAIRness of the research data created in PaN experiments by looking at data from different perspectives and by taking potential reuse scenarios into account.

An overview of the requirements of FAIR data is given as a set of baseline principles, and the facilities' research lifecycle described in from previous work in PaNdata-ODI is described and revised in the light of current practice. This lifecycle is compared with other standard approaches to describing data management lifecycles. The current state of the use of metadata and metadata standards within facilities is considered, based on survey results from early within the ExPaNDS project.

The report considers the facilities experimental lifecycle in detail via a gap analysis of the Data Continuum described by the PaN-data ODI D6.1. In particular, it describes the metadata types which can be collected at each step and identifies their role in supporting the FAIR principles, categorizing them what is *essential*, *important* and *useful* under the RDA FAIR Data Maturity Model priority flags, a prioritization which is then reflected in the draft recommendations for the Common FAIR Metadata Framework. Further, for each stage, there is a review of the roles and information systems active at each stage to identify the agents responsible for collecting and maintaining that metadata, and their contribution to the process of making data FAIR throughout the experimental lifecycle.

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

Drawing from this gap analysis, 31 draft recommendations for key elements of a common metadata framework are given, subdivided by their stage in the lifecycle.

Proposal

1. Is the **Principal investigator** declared as part of the metadata fields?
2. Are the **Co-Investigators** declared as part of the metadata fields?
3. Is the **Instrument requested** declared as part of the metadata fields?
4. Is the **Sample description** declared as part of the metadata fields?
5. Is the **Facility** where the proposal is submitted declared as part of the metadata fields?
6. Is the **Proposal Identifier** declared as part of the metadata fields?
7. Is the **Experiment Description** declared as part of the metadata fields?
8. Are the **Proposed Experiment Conditions** declared as part of the metadata fields?

Experiment

9. Is the actual **Visiting Experimental Team** (people who actually participate during the measurement) declared as part of the metadata fields?
10. Are the **Experiment/Measurement dates** declared as part of the metadata fields?
11. Does the **Samples information** provide enough context to understand its structure and characteristics and is declared as part of the metadata fields?
12. Is the **Instrument information** declared as part of the metadata fields?
13. Is the **Calibration information** declared as part of the metadata fields?
14. Is the produced **Dataset information** declared as part of the metadata fields?

Processing

15. Is the resulting **Data Format** declared as part of the metadata fields?
16. Is the **Processing information** declared as part of the metadata fields?
17. Is the **Software package information** used for processing declared as part of the metadata fields?
18. Is the **Original Data** link used for the processing declared as part of the metadata fields?
19. Is the resulting **Dataset information** declared as part of the metadata fields?

Analysis

20. Is the resulting **Data Format** of the Analysis declared as part of the metadata fields?
21. Are the **Files Identifiers** declared as part of the metadata fields?
22. Is the **Software package** used for the analysis declared as part of the metadata fields?
23. Is the **Original Data** link used for the analysis declared as part of the metadata fields?
24. Is the resulting **Dataset information** declared as part of the metadata fields?

Record

25. Is the **Resource Identity** declared as part of the metadata fields?
26. Is the **Creator** of the record declared as part of the metadata fields?
27. Is the **Publisher** of the record declared as part of the metadata fields?
28. Is the **Publication year** declared as part of the metadata fields?
29. Is the **Release date** (Embargo due date) declared as part of the metadata fields?
30. Is the **Title** of the dataset declared as part of the metadata fields?
31. Is the **License** for usage declared as part of the metadata fields?

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

This document represents the first stage in developing a Common Metadata Framework. The recommendations give a framework for setting standards for common metadata formats and provide a basis for assessing the FAIRness of data generated by facilities. In the next stage, we will consult across the project and with stakeholders outside the project, including in the PaNOSC project. We will refine the metadata framework, and in particular, make recommendations on the most appropriate standard metadata formats and supporting practices so that a common interoperable metadata framework for FAIR data can be used across the photon and neutron community.

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.

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

1. Introduction

Photon and Neutron (PaN) facilities have moved from experiment-centric to data-centric activity over the past decade. PaN Data Policies are a response to this shift, while the trend of data volumes and needs for computation are increasing year after year. With this strong focus on data, the need for findable, accessible, interoperable, and reusable (FAIR) data moves up the list of priorities. The purpose of this report is to focus on and provide recommendations in relation to a key element of the research data management process: a Common Metadata Framework for FAIR data in PaN facilities.

1.1 The Shift to Data-centric Photon and Neutron Science

The instruments commonly used today in PaN facilities have become fully computer-based machinery. This has meant a move to data-driven science in the PaN domain. The entire workflow on the instruments is dependent on sophisticated data preprocessing and analysis techniques. Research Infrastructures (RIs) have transformed from being experiment-focused facilities to being data producers with high data flows that vary widely between facilities' instruments and techniques. In such a context, data provenance and workflow transparency are very important to good research data management.

Knowledge can be extracted not only from the analysis and referenced reports and publications but also from the entire data flow, including raw data curation, intermediate treatment, analysis and final derivations. As will be shown below, every stage of the lifecycle may act as both data source and data receiver, i.e. consumer of the data provided by the previous stage. In such a situation, enabling reusability of the data and supporting reproducibility of the analysis are critical tasks for verifying scientific results.

For decades, PaN facilities have put a significant amount of energy into managing this data and metadata efficiently. The FAIR data management model is an attempt to implement the verification and reproducibility of data through opening access in a controlled fashion. This can significantly increase the number of users that collect data from sources known to be FAIR. Thus, the more sources are mature in terms of implementing FAIR principles, the more reusable data will become. This turns the PaN data continuum into a kind of peer-to-peer network, with not only the machines but also the scientists and collaborators acting as peers, in this case, as suppliers and consumers of data.

From a data management point of view, adopting the FAIR principles strongly influences the way in which data should be curated, and the scientific community is taking steps to address this situation. Scientists, i.e. acting as peers, are trying to build a web of trust within which they can share, discuss and manage data. Enabling this activity could be considered as possibly the most challenging task in modern scientific data management.

Sharing the experimental context and data analysis pipeline in a reproducible way alongside results presents a great opportunity to make research FAIR. Of course, scientists may want to do this voluntarily anyway; however, sharing datasets can help to address problems in other areas, for example, within technical or administrative spheres.

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

In practice, every RI has its own regulations, support tools, and processes dedicated to sharing data and results, even between scientists, so from the FAIRness point of view, these need to be the focus of harmonization efforts.

There is also a need to make the data access process itself predictable from the end-user point of view, including, for example, estimated periods of time during which the data can be moved from internal, archival storage to storage accessible from outside the RI. This need encourages RIs to aim to make datasets open access and to provide convenient guidelines for users, thereby implementing accessibility in terms of FAIR.

In 2012, the PaN-data Open Data Infrastructure (PaN-data ODI)[1] project proposed a common PaN facility workflow model,[2] in which data and metadata are produced at various stages of the experimental lifecycle. Along with this workflow definition, comes the need to determine levels of commitment around what will be produced, where it will be kept, for how long, and also the access control principles and recommendations.

1.2 FAIR Data Becomes a Priority

As we saw in the scenario presented in Section 1.1, when considering the FAIR principles, a new set of questions arises around how data are produced, and made findable, accessible, and interoperable so that they can be reused.

Open access policies serve as instruments that help to create the environment in which data can be used by anyone with minimal or no restrictions and legal constraints. However, the implementation of these policies requires transparent principles of data management in accordance with new and existing rules and prescriptions established by dedicated data protection acts such as the General Data Protection Regulation (GDPR)[3] and others.

Open access also requires special efforts to provide data persistence and integrity. One of the most important tasks is to collect and properly support the existence and persistence of the metadata associated with the curated data. The recommendations we make in this report strive to provide convenient guidelines for RIs on how to organize and maintain data that arise from PaN experiments, especially in relation to metadata stewardship and FAIR policy implementation.

1.3 Aims and Purpose of This Report

The goal of the common metadata framework presented here is to provide guidance on defining requirements on FAIR PaN data. Its purpose is to guide PaN data managers and scientists in the

[1] PaN-data Open Data Infrastructure was a FP7 (INFRA-2011-1.2.2; Grant Agreement RI-283556) supported project. Its main goal was the establishment of a data infrastructure, federated among the European Neutron and Photon facilities, to enable the scientific communities access, analysis and sharing of scientific data in a collaborative research environment. http://pan-data.eu/ .
[2] Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910
[3] European Parliament and Council of European Union (2016). Regulation (EU) 2016/679. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

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

process of defining what information is needed to produce FAIR data, and in which form a research dataset needs to include this information to meet the purpose of future reuse. These guidelines should enable data managers and scientists to assess the level of FAIRness of the research data created in PaN experiments by looking at data from different perspectives and by taking potential reuse scenarios into account.

The framework will build strongly on the findings of the PaN-data project,[4] as well as on recent Research Data Alliance (RDA) recommendations and a survey undertaken within the ExPaNDS[5] project. We structure our report into six sections:

- Section 2 reviews relevant documents and models, beginning with a discussion of key underlying ideas and concepts.
- Section 3 summarises results related to metadata standards, identifiers, and data catalogues drawn from a recent survey of ExPaNDS partner facilities.
- Section 4 considers use case scenarios and related roles and systems.
- Section 5 looks in detail at the metadata types, roles, and information systems that present in each stage of the idealised experimental lifecycle
- Section 6 draws together the content of the earlier sections to propose summary draft Recommendations for a Common FAIR Metadata Framework.

2. Key Foundations and Concepts

This section describes requirements on metadata to meet the FAIR data principles. The goals and requirements on metadata for datasets to be archived in the facilities' repositories are presented. These goals and requirements are used later in the gap analysis in Section 5, which considers the data continuum model presented in the PaN-data ODI Deliverable 6.1.[6]

After the introduction to the FAIR principles in Section 2.1, Section 2.2 looks at the minimal metadata requirements that allow the deposition of datasets in most general purpose data repositories like Zenodo and that are required for OpenAIRE or B2FIND. We then move on to consider requirements for describing the context of the datasets. The final discussions of Section 2.2 are more domain specific. First, we provide a small retrospect on previous efforts around the standardisation of metadata and metadata collections, especially in PaN. Then, we consider additional, related metadata efforts.

[4] PaN-data was supported in two European projects between 2010 and 2014, PaN-data-Europe and PaN-data Open Data Infrastructure. These projects explored standards, tools and methods for a common approach to supporting and developing data infrastructure in European large-scale PaN analytic facilities. http://pan-data.eu/

[5] ExPaNDS is the European Open Science Cloud (EOSC) Photon and Neutron Data Service. The project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641. https://expands.eu/

[6] Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910

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

Section 2.3 introduces the lifecycle perspective on data and metadata creation, exploring domain-specific models of the PaN experimental lifecycle as well as domain agnostic models of research data management (RDM).

2.1 The FAIR Data Principles

The FAIR principles were defined in 2016 by Wilkinson et al.,[7] crystallising guidelines to enhance the findability, accessibility, interoperability and re-usability of research data. These principles apply to both machines (supporting automation) and humans, data and metadata.

In June 2020, the RDA FAIR Data Maturity Model Working Group published the FAIR Data Maturity Model: Specification and Guidelines with the aim to "develop a common set of core assessment criteria for FAIRness, as an RDA Recommendation".[8] This work constituted a follow up from the FAIR principles, which set guidelines without strict rules, and therefore, can support different interpretations and be applied differently by the various stakeholders.

The FAIR Data Maturity Model indicators for assessing adherence to the FAIR principles have been specified in order to normalise assessment. In addition, the model defines priorities for the indicators. The three levels of indicators of importance defined by the FAIR data maturity model are:

- **Essential**: such an indicator 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.
- **Important**: such an indicator 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.
- **Useful**: such an indicator addresses an aspect that is nice-to-have but is not necessarily indispensable to achieve FAIRness.[9]

There are indicators for each of the FAIR principles, described and assessment details provided, as well as classified by importance. However, the FAIR Data Maturity Model has been created by an interdisciplinary group and does not include any domain or usage specific perspective as described earlier.

Other projects and initiatives are also developing tools and frameworks to accompany institutions and facilities on the adoption of FAIR. As an example, the FAIRsFAIR Data Object Assessment

[7] Wilkinson, M. D. et al. (2015). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**:1. https://doi.org/10.1038/sdata.2016.18
[8] RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. https://doi.org/10.15497/RDA00050
[9] Ibid.

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

Metrics[10] have been developed in the FAIRsFAIR project[11] on a European level. They consist of fifteen criteria, which include the FAIR guiding principles.

The FAIR principles play a key role in the European Open Science Cloud (EOSC) ecosystem. For example, the EOSC FAIR Executive Board Working Group has elaborated seven recommendations on implementing FAIR metrics considering the above mentioned outcomes and activities.[12]

A more technical view on FAIR data was introduced in 2018 by the European Commission (EC) Expert Group on FAIR Data (Simon Hodson, Chair): the FAIR Digital Object.[13] The FAIR Digital Object Model describes the technical ecosystem required for the realisation of FAIR data. Central to this model is the description of the FAIR Digital Object in four layers consisting of the digital object, identifiers, standards and code, and metadata (see Figure 1 below).

The FAIR Digital Object should be integrated into a technical ecosystem for FAIR data consisting of essential components such as policies, Data Management Plans (DMPs), identifiers, standards, and repositories. Persistent Identifiers (PIDs) are assigned to the different components of FAIR Digital Objects, e.g. such as data, metadata, code and algorithms, models, licenses, as well as to the FAIR Digital Objects themselves.

[10] Devaraju, A., Huber, R., Mokrane, M. et al. FAIRsFAIR data object assessment metrics. https://doi.org/10.5281/zenodo.4081213

[11] FAIRsFAIR - Fostering Fair Data Practices in Europe - aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle. FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558. https://www.fairsfair.eu/

[12] Genova, F., Aronsen, J. M., Beyan, O. et al. (2020). Recommendations on FAIR metrics for EOSC. Second draft for consultation. http://doi.org/10.5281/zenodo.4106116

[13] EC Expert Group on FAIR data (2018). Turning FAIR into reality. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

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

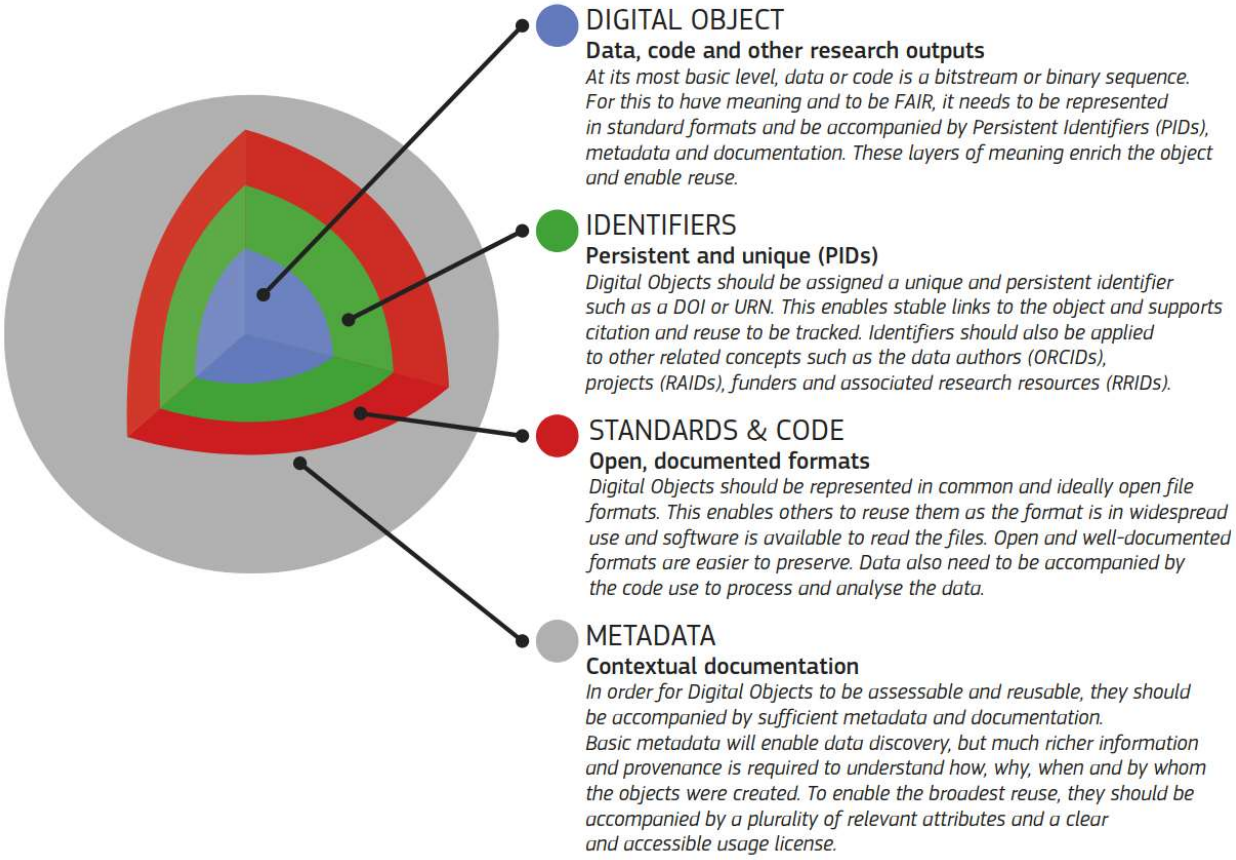

Figure 1: A model for FAIR Digital Objects, noting the elements that need to be in place for data to be Findable, Accessible, Interoperable and Reusable[14]

2.2 User Needs Matter: Metadata Levels

In order to find and reuse research data, one of the reusability FAIR principles demands data's rich description "with a plurality of accurate and relevant attributes".[15] While there may be less debate about what constitutes accuracy, certainly the question of what is 'relevant' depends on context. In other words, for what purpose is the 'rich description' needed or being used?

The importance indicators in the RDA Maturity Model introduced in Section 2.1 do not give any details with regard to the metadata quantity and level of detail needed to describe research data. Different approaches have been undertaken to decide on the level of detail of required metadata. Mostly, they are based on roles and related activities to be performed. For example, Houssos, Jörg & Matthews (2012) present three levels for the consideration of metadata requirements:

[14] Ibid.
[15] Wilkinson, M. D. et al. (2015). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**:1. https://doi.org/10.1038/sdata.2016.18

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

1. General browsing and discovery by a non-domain specialist
2. Enough context for a non-domain specialist researcher to understand
3. Domain specific and enables a domain specialist to answer the questions.[16]

In the next sections, we discuss these three levels further. It is important to recognise that, in practice, the boundaries between the three levels may be blurred and the levels may overlap.

2.2.1 General browsing and discovery by a non-domain specialist

There are several generic or domain-agnostic metadata standards that support describing datasets, each designed with specific use cases or purposes in mind. Given that these standards can be applicable across domains, they are mainly used to provide high-level interoperability, i.e. to support data browsing and data discovery based on common terms that can be performed without specific domain knowledge.

The DataCite organisation focuses on supporting the process of locating, finding and citing research data. The DataCite metadata schema defines core metadata properties that are required for the citation of a resource. The schema requires six mandatory elements that include: identifier, creator, title, publisher, publication year, and resource type. At the time of writing, the latest version of the schema is 4.3.[17] The serialisation of the DataCite schema is in Extensible Markup Language (XML).

The Dublin Core Metadata Standard, also known as the Dublin Core Metadata Element Set, includes fifteen terms and a dozen properties, classes, datatypes, and vocabulary encoding schemes.[18] The terms are widely used for the description of resources as the terms are about general metadata such as title, subject, creators and contributors, language, and formats. The Dublin Core is expressed using the Resource Description Framework (RDF), where each term has a unique and persistent identifier and some descriptions (e.g. such as a label and a definition). These terms can be used for data annotation and to produce Linked Data.

2.2.2 Enough context for a non-domain specialist researcher to understand

Another important generic metadata standard that focuses on the description of datasets and data catalogues is the Data Catalogue Vocabulary (DCAT). DCAT is an RDF vocabulary and a recommendation by the World Wide Web Consortium (W3C). The first version of the vocabulary (DCAT1 from 2014),[19] considers use cases around governmental data catalogues.

[16] Houssos, N., Jörg, B. and Matthews, B. (2012). A multi-level metadata approach for a Public Sector Information data infrastructure. http://purl.org/net/epubs/work/62617
[17] DataCite Metadata Working Group (2019). DataCite metadata schema 4.3. https://schema.datacite.org/
[18] Dublin Core Metadata Initiative (DCMI) (2020). DCMI metadata terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[19] W3C (2014). Data catalog vocabulary (DCAT). https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/

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

The latest recommended version, DCAT 2,[20] made significant changes to the original version by considering many new use cases around research data. In particular, the vocabulary now supports cataloguing any resource and has specializations for data services in addition to datasets. Other additions include covering temporal and spatial characteristics of catalogued resources, provenance information, relationships between datasets and agents. The Dataset eXchange Working Group is still ongoing and working on DCAT version 3, which will include terms for describing dataset series, and identifying dataset versions and their relationships, among other things.

DCAT is based on the Dublin Core metadata standard, and thus, it provides all the minimum metadata that enable general browsing and discovery by non-domain specialists. These metadata cover information such as title, description, relevant dates, language, creators, contributors, and publishers of the dataset or any type of catalogued resource.

In addition, DCAT provides extra information that supports providing more context and provenance, which would allow a non-specialist to understand the data. This includes metadata about the semantic relationship of a dataset with other datasets, relationships with other entities (e.g. persons, organisations, and other agents with different roles), provenance information, the dataset's temporal and spatial characteristics, and support for different distribution formats, including compression packages. DCAT is a domain-agnostic vocabulary, but it can be used in combination with domain vocabularies to provide more information for domain specialists.

2.2.3 Domain specific and enables a domain specialist to answer questions

In the last 20 years, various initiatives have addressed standardisation of data formats, related metadata, data workflows, and data catalogues in the PaN domain. The contributions of these initiatives, which we summarise below, can help us to describe PaN data in order to achieve the third metadata level, which aims to enable domain experts to answer questions.

In the mid 1990s, the NeXus format was created, based on the Hierarchical Data Format (HDF). The NeXus format defines vocabulary, structure, and serialization of data created in neutron, X-ray, and muon sciences, and it is now a common standard for data in those domains.[21]

The NeXus International Advisory Committee (NIAC) supervises the evolution and maintenance of the NeXus common data format and promotes the adoption of specific metadata standards per scientific domain. Members of many facilities worldwide are contributing to the NIAC.[22]

Between 2010 and 2014, two EC-funded projects were organised. PaN-data Europe was the first project from 2010 – 2011, followed by PaN-data ODI, a FP7 Project funded from 2011 – 2014.[23]

The PaN-data Europe and PaN-data ODI projects created two deliverables that are relevant for this present report:

[20] W3C (2020). Data catalog vocabulary (DCAT) –version 2. https://www.w3.org/TR/vocab-dcat-2/
[21] Könnecke, M., Akeroyd, F. A., Bernstein, H. J. et al. (2015). The Nexus data format, *J. Appl. Cryst.* **48**. https://doi.org/10.1107/S1600576714027575
[22] More details available at https://www.nexusformat.org/ .
[23] PaNdata (n.d.). PaNdata about. http://pan-data.eu/about

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

1. PaN-data ODI Deliverable 6.1: Model of the data continuum in photon and neutron facilities[24]

2. PaN-data Europe Deliverable D2.1: Common policy framework on scientific data.[25]

In some disciplines closely related to PaN, institutions and organisations have put some efforts into the standardisation of research data. Here, especially the International Union of Crystallography (IUCr) needs to be mentioned. For crystallography data, the Crystallographic Information Framework (CIF) standard[26] and its extension, the Macromolecular CIF (mmCIF),[27] have been created. In 2016, the new Crystallographic Information Framework II (CIF2)[28] was introduced.

The International Union of Pure and Applied Chemistry (IUPAC) has created the Compendium of Chemical Terminology.[29]

For RDA, the International Materials Resource Registries Working Group has developed a vocabulary for materials.[30] The European project Nanoscience Foundries & Fine Analysis (NFFA) has developed a metadata schema for nano materials,[31] building on previous work done by CODATA-VAMAS.[32]

Digital Representation of Units of Measure (DRUM) is a newly established CODATA task group.

In 2020, a Gold Standard for macromolecular crystallography diffraction data was adopted.[33] This agreed standard builds upon the NeXus/HDF5 NXmx application definition[34] and the IUCr imgCIF/CBF dictionary.[35]

2.3 The Lifecycle Perspective on Data and Metadata Creation

In order to determine metadata requirements, different lifecycles of data and metadata creation have been analysed. The lifecycles presented in the following subsections look at data and

[24] Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910
[25] Dimper, R. (2011). Common policy framework on scientific data. https://doi.org/10.5281/zenodo.37384978
[26] Hall, S. R., Allen, F. H. and Brown, I. D. (1991). The Crystallographic Information File (CIF): A new standard archive file for crystallography", *Acta Cryst.*, A**47**. http://ww1.iucr.org/iucr-top/cif/standard/cifstd1.html
[27] IUCr (2005). Macromolecular CIF dictionary. https://www.iucr.org/resources/cif/dictionaries/cif_mm
[28] Bernstein, H. J., Bollinger, J. C., Brown, I. et al. (2016). Specification of the Crystallographic Information File format, version 2.0, *J. Appl. Cryst.* **49**. https://doi.org/10.1107/S1600576715021871
[29] IUPAC (2020). IUPAC gold book. https://goldbook.iupac.org/
[30] RDA International Materials Resource Registries WG (2017). Materials registry vocabulary draft. https://www.rd-alliance.org/system/files/documents/Materials_Registry_vocab_draft_170321.pdf
[31] Bunakov, V., Matthews, B., Jejkal, T. et al. (2018). NFFA Deliverable D11.14 Final metadata standard for nanoscience data. https://www.nffa.eu/media/202831/d1114_final-metadata-standard-for-nanoscience-data.pdf
[32] CODATA-VAMAS Working Group on the Description of Nanomaterials & Rumble, J. (2016). Uniform description system for materials on the nanoscale, version 2.0. http://doi.org/10.5281/zenodo.56720
[33] Bernstein, H. J., Förster, A., Bhowmick, A. et al. (2020). Gold Standard for macromolecular crystallography diffraction data, *IUCrJ*, **7**. https://doi.org/10.1107/S2052252520008672
[34] NIAC (2020). NeXus user manual and reference documentation. 3.3.2.10. NXmx. https://manual.nexusformat.org/classes/applications/NXmx.html#nxmx
[35] IUCr (2005). Image CIF dictionary. https://www.iucr.org/resources/cif/dictionaries/cif_img

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

metadata creation from different perspectives. There is the PaN perspective and the generic RDM lifecycle perspective.

2.3.1 PaN-data ODI D6.1 idealised facilities lifecycle

The idealised facilities lifecycle originates from the PaN-data ODI D6.1 deliverable and models research-related activities as a series of stages, from proposal submission to paper publication.

For each activity in the idealised facilities lifecycle (see Figure 2 below), metadata types and related sources and roles are described:

- The processes of **Proposal**, **Approval** and **Scheduling** produce metadata, which partly is required to describe data produced in the Experiment.
- The first data is produced during the **Experiment** process. The PaN-data Deliverable 6.1 mentions the datasets of raw experimental data associated with each sample, and calibration data as well as metadata.
- In the **Storage** process, only metadata is produced. In the **Analysis** process, again data is produced: processed and derived datasets, and graphical information for visualisation and software code, as well as related metadata.
- In the **Publication** process, the journal article and supplementary data, including related metadata, are produced.[36]

[36] Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910

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

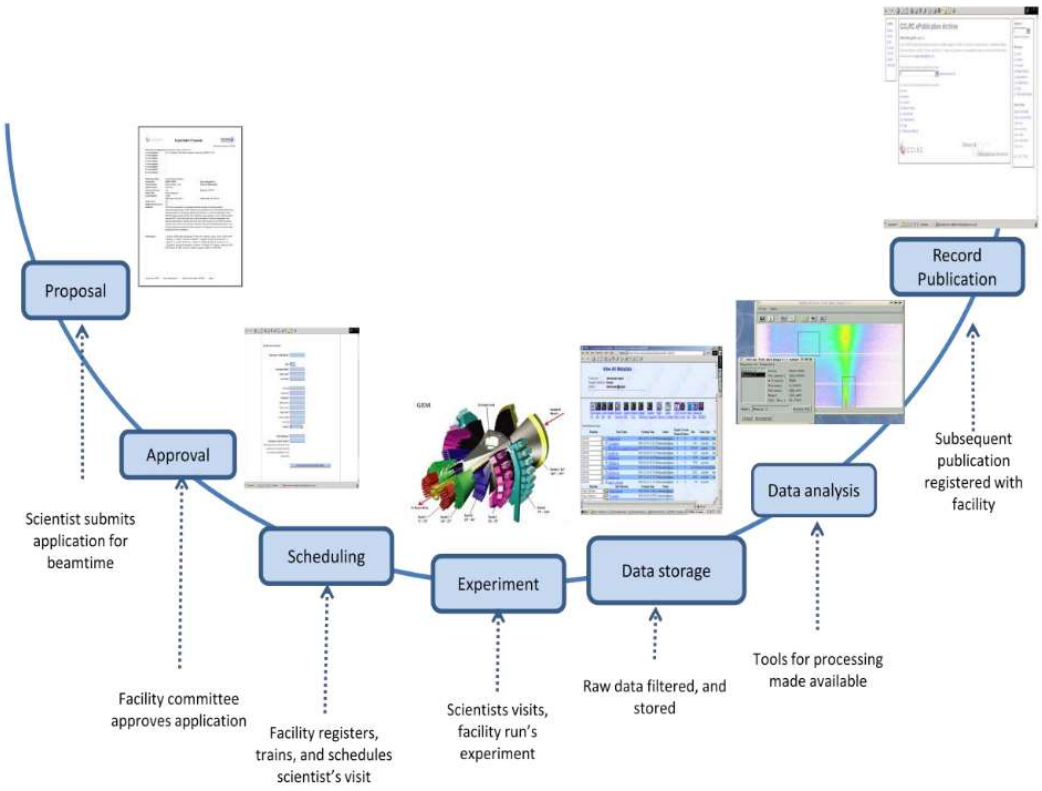

Figure 2: The idealised facilities lifecycle from PaN-data ODI D6.1 [37]

2.3.2 Classes of Experimental Data in the PaN Science Life Cycle from the Soleil Data Policy

In the Simplified illustration of Classes of Experimental Data in the Science Lifecycle from the Soleil Data Policy (see Figure 3) data and metadata production are central elements.[38]

Activities related to data and metadata production during and after an experiment are the focus of the lifecycle.

Although this is a specific example drawn from the PaN domain, it is important to note that this model does not necessarily apply to all experiments. The discussion in Section 4 considers elements of this model further.

[37] Ibid.
[38] Gagey, B. (ed.) (2018). SOLEIL data management policy. https://www.synchrotron-soleil.fr/en/file/11308/

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

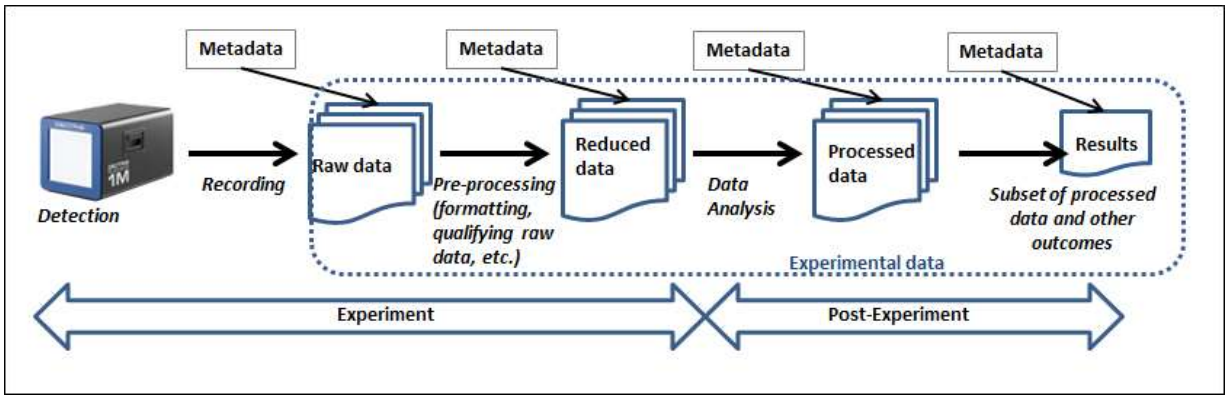

Figure 3: Simplified illustration of Classes of Experimental Data in the Science Life Cycle (from the Soleil Data Policy)

2.3.3 The Research Data Management (RDM) lifecycle

RDM involves many steps. Taken together, these processes are often modelled using a lifecycle approach. Many RDM lifecycle models exist. In this section, we examine three such models:

1. the Digital Curation Centre (DCC) Curation Lifecycle Model (CLM)[39]
2. a very simplified version of an RDM lifecycle, the Archive Centric Information Lifecycle Model (ACILM),[40] which was published in the SHAMAN project[41] in 2011
3. our own PaN facilities RDM model.[42]

The DCC created the CLM to provide a roadmap that ensures that all necessary steps in a curation lifecycle of RDM are covered. It starts from initial conceptualization to either disposal, or selection for reuse and long-term preservation.[43]

[39] DCC (2020). Curation Lifecycle Model. https://www.dcc.ac.uk/guidance/curation-lifecycle-model
[40] Brocks, H., Kranstedt, A., Jäschke, G., Hemmje, M. (2010). Modeling context for digital preservation. In *Studies in Computational Intelligence,* **260**. Szczerbicki, E. and Nguyen, N. T. (Eds.), pp. 197 – 226. https://www.researchgate.net/publication/225757343_Modeling_Context_for_Digital_Preservation
[41] The SHAMAN Integrated Project aimed at developing a new framework for long-term digital preservation (more than one century) by exploring the potential of recent developments in the areas of GRID computing, federated digital library architectures, multivalent emulation and semantic representation and annotation. The project was funded under FP7-ICT Grant agreement ID: 216736. https://cordis.europa.eu/project/id/216736
[42] Gonzalez-Beltran, A. (2020). Large-scale facilities experimental lifecycle & FAIRness. http://doi.org/10.5281/zenodo.4067988
[43] Higgins, S. (2008). The DCC Curation Lifecycle Model, *Int. J. Digit. Curation*, **3**. https://doi.org/10.2218/ijdc.v3i1.48

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

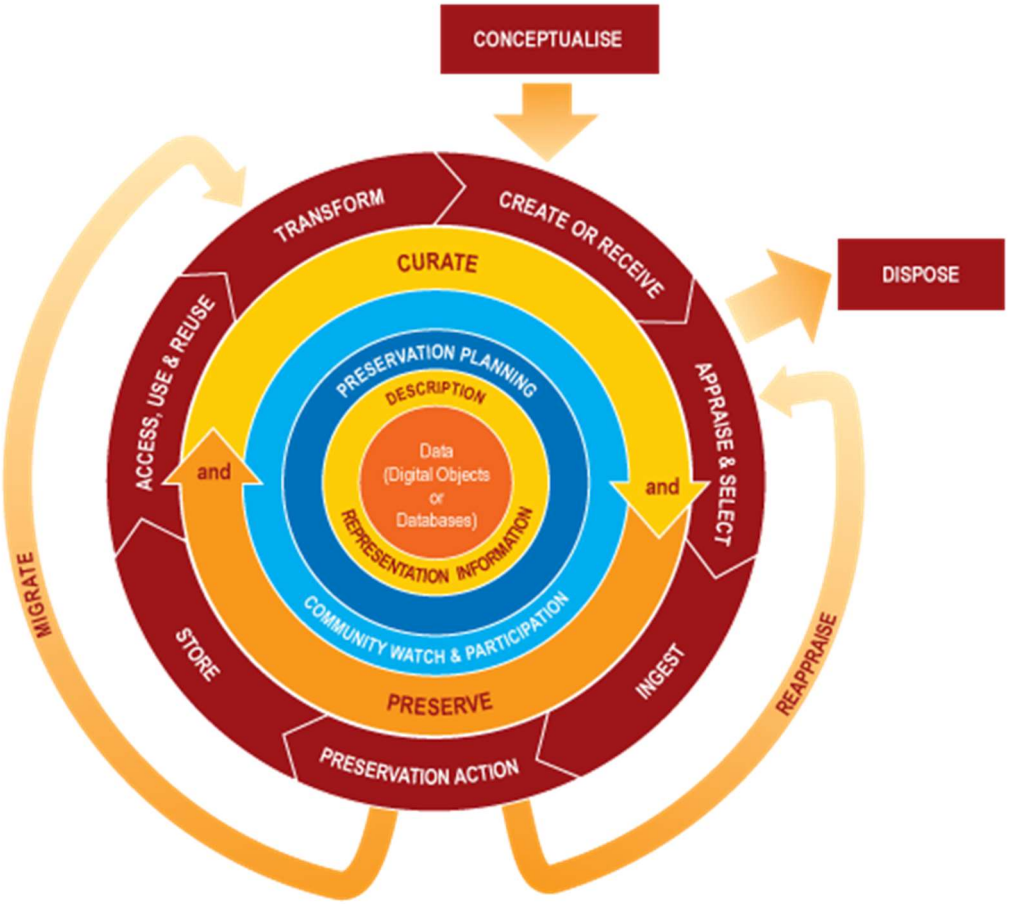

Figure 4: The DCC CLM RDM lifecycle model[44]

As in the ACILM, the CLM of the DCC integrates activities before and after the preservation of the Data Object. While in the ACILM, the focus of these activities is preservation, in the CLM, the activities are specified for RDM.

The phases of the ACILM incorporate activities both before and after archiving. The phases are creation, assembling, archiving, adoption, and reuse, where creation and assembling comprise the pre-ingest phase, and adoption and re-use, the post-access phase.

[44] DCC (2020). Curation Lifecycle Model. https://www.dcc.ac.uk/guidance/curation-lifecycle-model

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

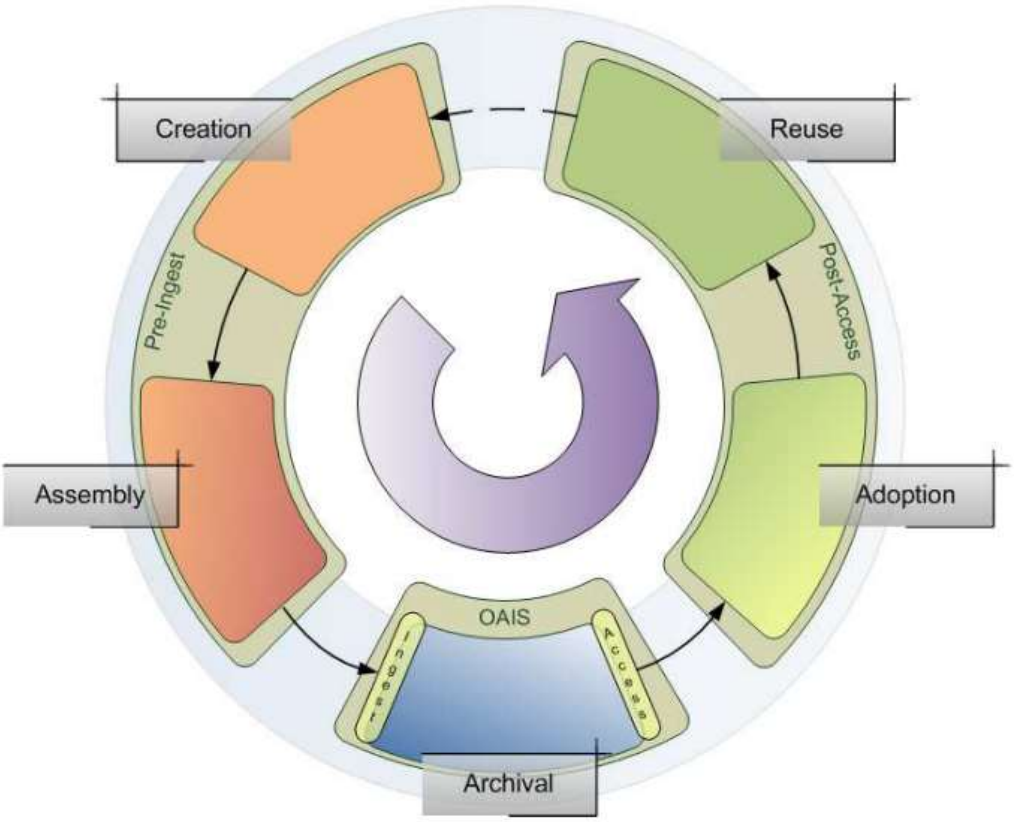

Figure 5: The ACILM RDM lifecycle model[45]

Figure 6 (see below) provides another view of the Research Data Management lifecycle and especially highlights the activities in the three different phases: planning, experiment and post-experiment phases.

The outer cycle shows the different processes in the experimental lifecycle. The following cycle enumerates the different data and metadata being produced at each of the phases. The innermost cycle shows the interaction between three catalogues hosting the different resource outputs: publications, data and software.

These catalogues, which here are represented as conceptual services, provide the metadata for the different research outputs and the links between them. The process of preserving these research outputs is represented as a goal for the outermost cycle, but this is a process that should happen along the experimental lifecycle. Similarly, the process of making the research outputs FAIR is an ongoing process along the experimental lifecycle.

[45] Brocks, H., Kranstedt, A., Jäschke, G., Hemmje, M. (2010). Modeling context for digital preservation. In *Studies in Computational Intelligence,* **260**. Szczerbicki, E. and Nguyen, N. T. (Eds.), pp. 197 – 226. https://www.researchgate.net/publication/225757343_Modeling_Context_for_Digital_Preservation

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

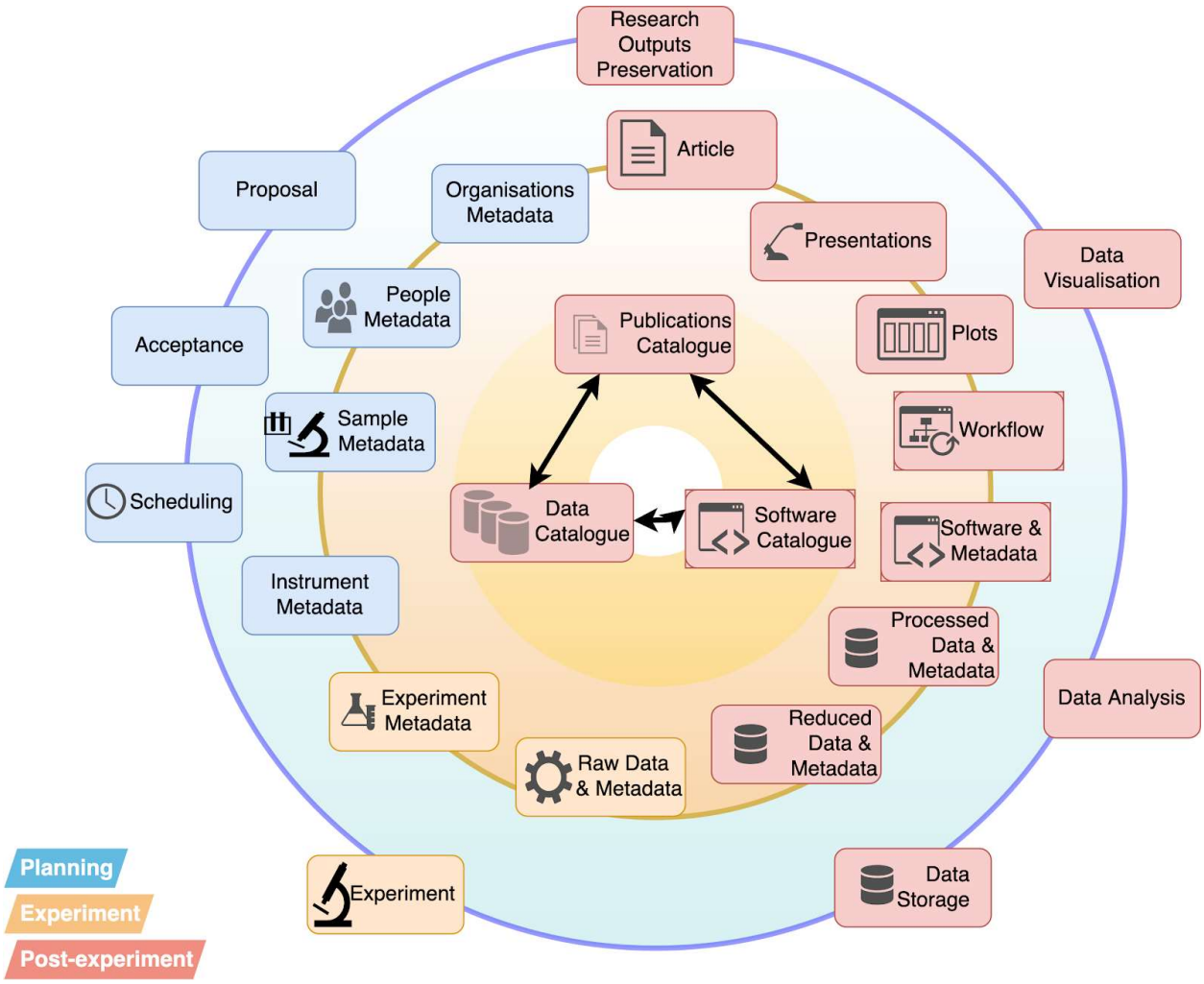

Figure 6: RDM lifecycle model for PaN Science[46]

3. Current Status of FAIR Metadata at National RIs

In December 2019, the ExPaNDS project conducted a landscaping survey of ExPaNDS facilities.[47] The survey aimed to establish a baseline on the current state of FAIR data policies and data management practices at the ten facilities[48] participating in ExPaNDS. Amongst other

[46] Gonzalez-Beltran, A. (2020). Large-scale facilities experimental lifecycle & FAIRness. http://doi.org/10.5281/zenodo.4067988

[47] Ashton, A., Da Graca Ramos, S., Matthews, B. et al. (2019). ExPaNDS data landscaping survey. https://doi.org/10.5281/zenodo.3673811

[48] These ten ExPaNDS facilities/RIs are: Deutsches Elektronen-Synchrotron (DESY), Paul Scherrer Institute (PSI), Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Diamond Light Source, MAX IV, Elettra, ALBA, SOLEIL, Helmholtz-Zentrum Berlin (HZB), and ISIS Neutron and Muon Source. https://expands.eu/partners/

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

information, the survey gathered specific details about the use of metadata standards, PIDs, and data catalogues. The sections below expand on these three elements.

3.1 Metadata Standards

Although all ExPaNDS RI facilities replied that their Data Policy covers metadata explicitly, only two of them have specific metadata standards defined, and another one mentions HDF5 as the standard specified. Some responses justified this lack of explicit standards as a way to provide flexibility to the metadata definition over time.

There is quite a variety of different services provided to users to capture metadata. All RIs claim to provide human-interaction tools for enriching the data capture with metadata, such as Notebooks or Elogbook. Confluence, Google Docs, and Slack are also mentioned in the various responses received.

On the other hand, RIs have also mentioned different automatic metadata ingestion systems such as GDA (data acquisition software),[49] openBIS (institutional data management software – lab notebook),[50] or simple scripting.

3.2 Identifiers

Facilities' policies and practices are increasingly using PIDs. Some facilities are also engaging with efforts to use public identifiers for users, such as the Open Researcher and Contributor ID (ORCID). PIDs are also a key component of the EOSC.[51] Recent work is investigating the possibilities of persistently identifying other elements of the research lifecycle, such as the instrumentation.[52] This is very important because the instrumentation context, when implementing FAIR, provides an opportunity to make data more interoperable by giving additional information on the instrument parameters. Software licensing and other legal factors must be taken into account during the metadata enrichment process. This is still possible because RIs always act under their own internal regulations and contractual obligations guiding instrumentation-related activities. Additionally, there are no widely used standards for sharing metadata associated with instruments.

The PID infrastructure in the context of implementing FAIR can be a source of problems where the data processing workflow should be traceable. In that case, a PID of the same type (for example, DOI) can be issued for datasets generated at each stage of data processing. This challenging problem requires interlinking the PIDs at each stage in the process into a provenance graph to provide traceability.

[49] DLS (2020). GDA software for science. http://www.opengda.org
[50] openBIS (n.d.). openBIS. https://openbis.ch/
[51] EOSC FAIR and Architecture Working Groups (2020). A Persistent Identifier (PID) policy for the European Open Science Cloud. https://op.europa.eu/en-GB/publication-detail/-/publication/35c5ca10-1417-11eb-b57e-01aa75ed71a1/language-en
[52] Stocker, M., Darroch, L., Krahl, R. et al. (2020). Persistent identification of instruments. *Data Science Journal*, **19**. http://doi.org/10.5334/dsj-2020-018

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

Interlinking is also required for implementing interoperability in the context of FAIR because there is a need to recover experimental contexts more accurately. The provenance graph mentioned above can be used for that purpose.

In the storage context, this case requires recording also the metadata about how a given piece of data links with others, i.e. data provenance. The most recent information about PID infrastructure is collected within the draft EOSC PID Architecture.[53]

3.3 Data Catalogues

The ExPaNDS RI's facilities provide cutting edge experiments to both public and industrial users. Data constitutes a key asset to advance scientific research, especially when combined with data sharing and reuse.

In the last few years, two main data catalogue toolsets have been developed and are currently in use in some X-ray and Neutron facilities: ICAT (including ICAT+) and SciCat. Deployment and implementation are the tasks of ExPaNDS WP3.[54] Other facilities are still investigating or currently developing their own data catalogue.

For many facilities, the data catalogue will provide access to data and the ability to make data publicly available after an embargo period. As described above, a lack of generalized metadata naming schema and vocabularies produces the need for the data catalogue to be highly configurable. This makes development more difficult because the given data catalogues, in practice, aim to be integrated all together to a federated data catalogue.

All of the problems described here require joined up efforts from all involved RIs to clarify, agree and redefine the environment for harvesting metadata. The access control policy, in the context of data protection regulations, is also an important issue. The RIs have a potential solution: providing a unified authentication environment for all users, which requires the development of a digital identity infrastructure and corresponding policies. To illustrate that solutions exist, the EOSC is proposing to provide an Authentication and Authorization Infrastructure (AAI) architecture (currently in draft form).[55] Authentication and authorization are relevant matters related to the accessibility (i.e. A) element of FAIR.

The recommendations presented here in this deliverable encourage RIs to agree upon common standards for metadata policies, naming schemas and vocabularies, and digital identity policies and to produce management practice suggestions for the entire PaN community.

[53] Schwardmann, U., Fenner, M., Hellström, M. et al. (2020). PID architecture for the EOSC, version 0.3. https://docs.google.com/document/d/1T-bpNsmuxQewsLq48XTyUJoe0lsV7poaXohpgDo9W34/edit#heading=h.81f971wdg07u
[54] ExPaNDS (2020). Work packages. https://expands.eu/work-packages/
[55] EOSC AAI Architecture Work and Subgroup (2020). EOSC AAI architecture. https://docs.google.com/document/d/12l0xVU9oiXqtVqkwrJijj16L-i_YcM_K6SHkNMf4IWE/edit

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

4. Use case scenarios and related roles and systems

This section introduces some example use cases to be considered in the gap analysis. The vision of the Common Metadata Framework should integrate the experimental workflow, which is PaN specific, with the FAIR guidelines. Having a clear picture of the final product would allow us to identify the gap between what is currently in place in RIs and what still needs to be done.

By comparing the PaN-data ODI research lifecycle with the SOLEIL lifecycle, two types of steps in the lifecycle can be identified. In the steps proposal, approval, scheduling and storage, mainly metadata are produced, while in the steps introduced in the SOLEIL lifecycle data taken in the experiment relevant for subsequent data reduction, processing and analysis and metadata for later archival are produced.

Bringing together the PaN research lifecycle and the RDM lifecycle as presented in Section 2, it becomes clear the datasets produced in each step of the SOLEIL lifecycle will enter their own RDM lifecycle.

By applying the simplified ACILM RDM lifecycle to the produced datasets, we can see that, after creation, the datasets are assembled before being archived. Before reuse, the datasets have to be appropriately adopted. Each reuse of one or more datasets creates a new dataset. **Creation**, **Assembly**, **Archival**, **Adoption** and **Reuse** apply to all phases of the SOLEIL research lifecycle (see Figure 7 below).

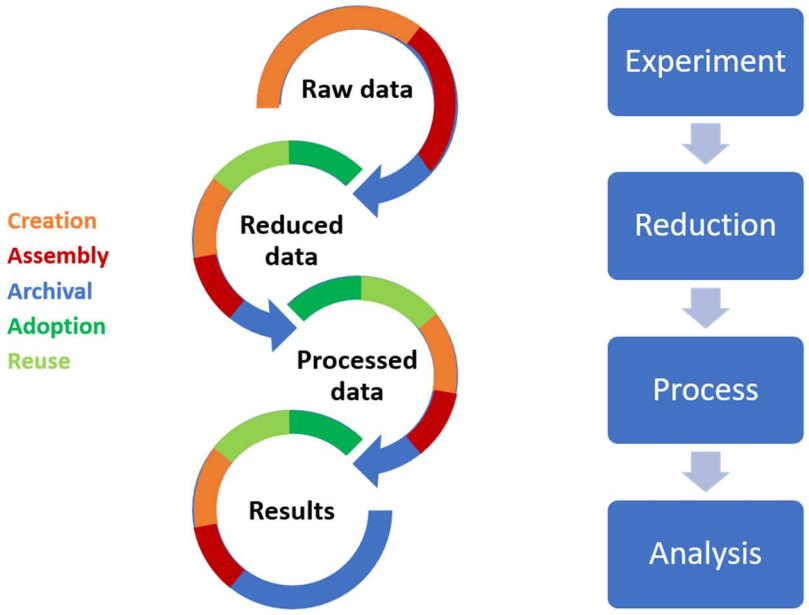

Figure 7: Another model of the PaN RDM lifecycle: focus on the data-centric experiment

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

The Proposal, Approval and Scheduling steps, together with the overall project, provide context to the datasets produced in the facilities and will have to be added in the assembly phase to the data object.

The Storage step of the PaN-data ODI lifecycle comprises the assembly, archival and adoption phase of the ACILM lifecycle. In this step a variety of actions might be performed on the dataset that either need or produce metadata: Contextual metadata extraction; Data access control; Data backup; Data format control; Data retention; Disposition; Notification; Restricted searching; Storage cost; Use agreement.[56]

This section presents example use case scenarios (see Table 1 below) intended to guide the data continuum gap analysis that follows in Section 5. We also consider roles and information systems in relation to the use cases.

4.1 Use Cases

To ensure that the minimal metadata that accompanies the basic publication workflow is sufficient for envisaged future use scenarios, it is necessary to consider what such future scenarios may be. This is important as the metadata that provides the context in a basic publication is often minimal.

In relation to FAIR, metadata needs to provide sufficient context information, including provenance. However, the question of what constitutes the minimum metadata set required is very dependent on user purpose and need (see section 2.2). As such, the needs of different stakeholders involved in the data lifecycle should be considered.

The following scenarios are selected examples presented as role-centric use cases to facilitate understanding and to easily identify roles and activities in which FAIR should contribute to enhance the experience.

User type	**Use case**
User Scientist	Data Analysis immediately after the experiment
User Scientist	Validating measurement
User Scientist	Finding and re-analysing measurement data with other algorithm after a long period of time or third-party
User Scientist	Reusing data for simulations (i.e. comparing experiment with theory)
User Scientist	Reproducing experiments (with similar sample)

[56] Moore, R., Stotzka, R., Cacciari, C., Benedikt, P. (2015). Practical policy (Version 1.0). http://doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC

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

Instrument Scientist	Preparing and improving instrument and instrument setup to allow better experimental measurement.
Instrument Scientist	Advising/Supporting visiting scientists in relation to their experiments and setting up an instrument.
Reviewer	Peer-reviewing data in relation to a research paper
Data Managers	Registering a Persistent Identifier (PID) for long-lasting reference of the data for future links to publications.
Data Managers	Tracking data usage using both global and internal persistent identifier (PID)
Data Managers	Archiving data for long term usage
Data Managers	Making research data available to PaN Search API
Data Managers	Making research data available to EOSC
Project Administrators	Needing to write reports
Publishers	Acting as agents in assessing completeness of metadata, verification and peer review of results.
Publishers	Controlling the authorship of the last derivatives of the data
Funders	Controlling the RIs policies of financial activities.
Funders	Acting as arbiters from the RI side against Principal Investigators.

Table 1: Selected example use case scenarios

4.2 Roles

Within all the use cases, we have identified three types of roles which can be played by different people within the lifecycle, depending on the step we are actually analysing. In all the cases, we identify the following generic roles:

- **Data Producer (e.g. User Scientists, Instrument Scientists):** should provide information that will help define the context of the experiment (e.g. description of the study, definition of the sample). In many cases, this context information is available before the experiment is performed.

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

- **Data Consumer (e.g. Scientific Validator, Someone who reuses the data):** Whoever uses the data to either validate the scientific results, reproduce the analysis, or reprocess the dataset for new scientific results.
- **Data Manager (e.g. Data Management role: Archivist, Librarian):** validates the integrity of the data (e.g. curation, custody).

4.3 Information Systems

Today, information systems are crucial to achieving the objective of any data management endeavour. In this context, we identify as many information system types as roles in the previous sub-section.

- **Data Production Systems:** Proposal systems, which are used to submit experimental cases; acquisition systems, which are used to collect data and metadata from measurements.
- **Data Consuming Systems:** Systems for data accessing, visualization or downloading.
- **Data Management Systems**: Systems for data processing, analysis, etc.

5. Data Continuum Gap Analysis for FAIR

Given the use case scenarios set out in Section 4, Section 5 aims to analyse what is missing from the initial presentation of the Data Continuum described by the PaN-data ODI D6.1 deliverable.[57] In particular, for the metadata types, we aim to identify what is essential, important and useful, as this prioritization will help us to make our draft recommendations for the Common FAIR Metadata Framework that will follow in Section 6.

In the sections below, we analyse the metadata types, roles, and information systems proposed in each stage of the PaN-data ODI D6.1 Data Continuum, considering these specifically in relation to what is needed to enable FAIR data.

As a first step in our gap analysis, we prioritize the metadata type fields, employing the RDA FAIR Data Maturity Model (see Section 2.1) priority flags of **P1- ESSENTIAL**, **P2-IMPORTANT**, and **P3-USEFUL**.[58] For each field, we also specify which aspect(s) of FAIR (as identified by the four letters of the acronym) is/are supported by that metadata field.

Next, we examine the roles related to each stage of the Data Continuum by considering two related questions:

1. Who provides the information necessary to document the data (and, therefore, to create the metadata)?

[57] Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910
[58] RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. 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.

2. Who is the custodian of that information?

As a final step, we comment on the information systems associated with each stage of the PaN-data ODI D6.1 Data Continuum. As with the metadata types and roles, we focus on how these information systems may (or may not) directly or indirectly (e.g. through interdependencies with other systems) contribute to the process of making data FAIR throughout the experimental lifecycle.

5.1 Proposal

In the Data Continuum process, the Proposal is identified as the starting phase where the user submits a proposal applying to use a particular instrument at the facility and for time to undertake experiments/measurements on particular material samples. This is lodged with the Facility.

At this early stage, no data is produced, but the information provided will serve as contextual metadata for the subsequent stages, and is considered essential for reproducibility.

5.1.1 Proposal: metadata types

At this step, facilities are storing valuable data which will become the contextualized metadata of the later published data, accessible in the Data Catalogues.

The analysis conducted in the PaN-data ODI 6.1 deliverable identified the following metadata types:

- **Principal Investigator/Main Proposer:** 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 organization. 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]**

- **Instrument requested:** Instrument of the facility that will be used in the event of approval. **[P1-ESSENTIAL- F]**

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

- **Sample:** 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.

- **Proposed Experimental Conditions:** Declaration of the environment and instrument setup required to perform the experiment proposed. **[P1-ESSENTIAL-F]**.

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

- **Safety conditions:** Information about the level of safety risk regarding the samples or the environments needed in order to perform the measurement or the experiment **[P3-USEFUL-F]**.

- **Experiment Description:** 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]**

- **Prior art:** Related publications or proposals linked to either the proposal itself or the participants. **[P2-IMPORTANT-F]**

The Proposal step must enable any scientist to present the scientific case to a specific facility and request concrete techniques and equipment. Given this assumption, and the fact that some of the fields might be considered as implicit during the metadata gathering process, these are the fields which should complement the process:

- **Co-Investigators:** 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]**.

- **Facility information**: 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.

- **Proposal identifier:** the facility being identified within the metadata fields, and assuming that each facility manages unique identifiers for its proposals, it is **[P1-ESSENTIAL-F]** that the Proposal identifier is added as part of the metadata fields.

5.1.2 Proposal: roles

At the proposal step, two main roles have been identified, from the point of view of data and metadata production:

- **Information provided by:** the Principal Investigator interacts with the submission system and provides a set of fields and information related to her/his proposal. Of course, and as has been mentioned in the previous subsection 5.1.1, she/he also declares who is contributing to the scientific case, i.e. as part of the Experimental Team.

- **Custodian of the information:** the Research facility will receive the submitted proposal and move it to the internal approval processes to, at the end of the workflow, grant or reject the amount of time to perform the experiment. After the submission, it is up to the facility to keep the information in its systems as a source of metadata to be linked to or incorporated with the later produced data.

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

5.1.3 Proposal: information systems

- **User Office systems:** Platforms used by the facility to communicate with users willing to come to perform the experiment. These systems are named after the User Office team, which serves as a liaison from the organization to the user. Amongst features of the system are user registration, proposal submission, visit management, and/or experiment report.
- **User registration and management:** System used by user scientists to identify themselves and give contact information for future communication with the facility. This specific platform, often included as part of the User Office System, shall trigger authorization mechanisms to make use of specific facility services when the submitted proposal is approved.
- **User identity:** Again, often included as part of the User Office systems, this specific system serves as a link from the facility with any integrated identity provider, such as UmbrellaID.
- **Proposal submission systems:** the platform, also usually part of the User Office system, in which scientists submit their experimental case to be evaluated by the facility to aim for beam time.

5.2 Approval

The application goes to an approval committee who judges the scientific merits and technical feasibility of the proposal and makes a recommendation to approve or reject the proposal.

The User Office team collates all submissions and convenes the scientific panel for evaluation. The scientific panel recommends the approval or rejection of the submitted proposals. The User Office then informs the Principal Investigator about the resulting decision.

The relevance of approval or rejection of the proposal is not perceived as significant in terms of FAIRness.

5.2.1 Approval: metadata types

- **Approval Panel:** Identifies the people involved in evaluation of the proposals. This panel shall not be considered relevant for FAIR, but the panel is relevant for internal evaluation tracking. **[P3 - USEFUL]**

The Approval step of the process is internal and does not provide additional value to the future data produced.

It might be relevant, though, that during the evaluation, the panel provides additional information and assessment on how the experiment would be more successful in terms of instrument set up or parameters. If these suggestions are followed by the PI and the Experimental Team, the scientific case might undergo some changes that might be of interest to track for provenance.

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

Additional steps, such as **Sample Safety Assessment [P2 - IMPORTANT],** might also be relevant to the FAIR purpose.

5.2.2 Approval: roles

- **Information provided by:** The facility staff, including the User Office for formal review, Instrument Scientists as part of the feasibility review (aka Technical Review) and the Safety Group shall complement the Approval Process information. The Approval Panel, which is usually made up of external scientific experts, shall provide relevant feedback and assessment on how the experiment shall be performed.
- **Custodian of the information:** The research facility will be the custodian of the information, and provide the results of the assessment to the PI and the Experimental Team.

In case any of the above mentioned teams need more information, it will be the User Office team that will ask for more information to complete the whole process.

5.2.3 Approval: information systems

- **Approval System**: usually part of the User Office system, as it shall be linked to the proposal and all the related information. The Panel, with the help of local facility staff, will provide the information and submit it to the system to keep for tracking reasons.

5.3 Scheduling

Time on the instrument is allocated to successful proposals to determine when the experiment will be scheduled to take place.

The Principal Investigator agrees the dates with the instrument scientist, and shares the scheduling with the experiment team members who will attend the execution of the experiment/measurement.

At this step, no data is collected yet. All data included is used for the preparation and planning of the experiment to come. Last minute changes shall be considered.

The preparation of the visit may also involve the sample tracking: techniques used for growing the samples; laboratories, outside and inside the facilities, where the samples were treated.

5.3.1 Scheduling: metadata types

- **Allocated Time on Instrument:** Day and time when the experiment is planned. Note, however, that this may change and need to be updated at the Experiment stage. For coordination and service providing purposes, this information is considered as important, but might not be essential. **[P2-IMPORTANT-FA]**

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

- **Scheduled Visiting Experimental Team**: In the Scheduling context, the Experimental Team will assign who actually will be present, either on-site or remotely, during the experiment. It might differ from the Experimental team declared during the submission of the proposal, and should definitely be confirmed during the Experiment step. **[P2-IMPORTANT-FA]**

- **Training**: Either the Instrument Team, the Safety Team, or the User Office itself, representing the facility policies, will require the Experimental Team, which will be present during the experiment, to pass a Training course before proceeding to the experiment. The flow of information related to this specific request is not considered as relevant for FAIR purposes, and this is why it is classified as **[P3-USEFUL]**

- **Detailed Experimental Planning**: Additional information about the experimental technique to perform, specific samples to measure and their details, etc. The amount of information provided at this point might not be as extensive as it would be during the Experiment step. This is why this specific field is considered as **[P2-IMPORTANT-F]** for FAIR.

- **Sample Preparation**: Information related to the sample preparation process, which may include techniques, specific laboratory information. Since the range of possibilities for Sample preparation is wide, but might be very relevant in some domains for FAIR, this field is considered as **[P2-IMPORTANT-FR]**.

- **Sample Reception**: Type of handling needed for sample transportation (e.g. temperature, high pressure, toxic or radio-active material). **[P3-USEFUL]**

The main objective of the Scheduling step of the data continuum lifecycle is planning for the experiment of samples' measurement. No actual data has been produced at this step, and only contextual metadata is collected. Taking this into account, none of the fields listed above are considered **[P1-ESSENTIAL]**, and it will be up to each facility to decide if, at this point, storing this kind of metadata is relevant.

No additional fields were identified for inclusion in this stage.

5.3.2 Scheduling: roles

- **Information provided by:** The Principal Investigator, who is the main contact of the Experimental Team, will work with the facility to find a suitable date and time to bring their samples (under which conditions) and to determine who will be present (on-site or remotely) during that specific measurement. From the facility, the Instrument Local Contact will provide support during the whole scheduling process.

- **Custodian of the information:** The research facility, using the Scheduling System, will store the information.

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

5.3.3 Scheduling: information systems

- **Scheduling System**: usually part of the User Office system, will serve not only as a scheduling system but also as the main communication tool for the facility and the scientists.

5.4 Experiment

After the time has been allocated to an instrument and the tasks related to the scheduling performed, the Experimental Team should be able to perform an experiment either on site or remotely. In relation to the Experiment stage, it is important to note that, while practices vary across facilities, instruments and experiments, we are trying here to capture a generalized process; in practice, there may be variations which a particular FAIR data process will need to take into account.

Before the session, the Experimental Team will need to prepare their samples. This could be performed at either the User Institution, the facility, or an external laboratory. Details on sample tracking could be recorded in the User Office or another Laboratory Information Management System (LIMS) or Sample Tracking System such as ISpyB.[59]

The beginning of the experimental session will be generally used for calibration, beam and optics alignment before a user can effectively take data in the instrument. During the experiment, a user will place the sample in the corresponding sample stage, configure the instrument hardware (i.e. optics components, detectors), and perform detector calibration, when required.

A scan will represent one or multiple motors moving while one or more detectors will be acquiring data.

Steps performed when recording metadata and data during a scan:

1. At the start of a scan, some metadata can already be recorded such as motor positions that are fixed, information from the sample, Experimental Team members, and scan identifiers.

2. During a scan, the motor positions, timestamps, detector data, and monitors values can be recorded. In some cases, reduced data is also provided by control and data acquisition systems during the scan, especially for long running scans or a queue of scans. This reduced data is a first evaluation of the experiment to evaluate if the scan(s) should be continued or stopped; generally, it does not take into account the overall geometry of the experimental set-up. This process allows facilities to make good use of the experimental time.

3. At the end of or during a scan, automated data processing pipelines with more sophisticated data analysis software are run. This gives valuable information about the quality of the data acquired and aids decisions on next experimental steps.

[59] ispyb (2020). ISPyB. https://ispyb.github.io/ISPyB/

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

There is an ever-growing demand for performing more sophisticated data processing during a scan rather than at the end. During the experiment, the data and metadata for each scan will be recorded.

5.4.1 Experiment: metadata types

- **Visiting Experimental Team (User Identity):** 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 is their affiliation. For Findability purposes, this field is considered **[P1 - ESSENTIAL - FA]**.
- **Sample Information:** 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.
- **Instrument Information:** 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.
- **Experiment Planning:** 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]**.
- **Environmental Parameters:** As part of the instrument setup, the data stored in this field/s is considered as **[P2-IMPORTANT-FR]** but not essential due to the fact that, in some cases, no changes are applied during the instrument set up for a particular experiment.
- **Calibration Information:** 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.
- **Laboratory Notebooks:** Annotations, either automatic or manual, are an essential part of the information for Reproducibility of any experiment and provenance of the resulting data. **[P2 - IMPORTANT - FR]**

Additional fields, not explicitly mentioned but implicit as they are included in other parts of the document, are the following fields:

- **Instrument Scientist:** 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].**
- **Experiment Date:** the actual date when the experiment/measurement is performed. **[P1 - ESSENTIAL - FA]**

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

- **Experimental Report:** If available, this adds contextual information about the measurement or experiment. [**P3 - USEFUL - R**]

5.4.2 Experiment: roles

- **Experimental Team**: who informs about the experiment requirements and configures/set-up the experiment. Note that the Experimental Team may include the Instrument Scientist (see below).
- **Instrument Scientist**: who will help the experimental team during the experiment with the set-up, software, calibration, etc.
- **Facility Operations**: Support people (e.g. Acquisition systems staff) who will facilitate the collection of the data and metadata.

5.4.3 Experiment: information systems

- **User Office:** provides information on the experiment, schedule, samples
- **User Account Management System:** Authorization system of each facility which shall provide read and write permissions to manage the collected data and add context if needed.
- **Data Acquisition and Control System:** Usually integrated by the Central Control system of the facility, the Data Acquisition system is on the front line of the measurement, recording motor positions and detector data.
- **Storage System:** permissions, storage related metadata (see below).
- **Sample Database Systems:** Other databases such as ISPyB (sample tracking), EPICS archiver.

5.5 Data Storage

The PaN-data ODI deliverable suggests that the Storage stage only covers the archiving, cataloguing and publication of the raw data, i.e. the data which are collected from the instrument: “Data is aggregated into datasets associated with each experiment, stored in secure storage, within managed data stores in facility, and systematically cataloged.”[60]

For interoperability and reusability reasons, we are proposing here to include Storage as not a stage of the facility lifecycle but rather as the part of any process that may produce additional data or metadata derived from the raw data.

[60] Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910

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

Our Recommendations aim to encourage RIs to develop and share storage specifications that, in turn, could guide the wider PaN community on how to attribute **any** data placed into storage enclosures. This attribution should, among other information, include the purpose of the storage for the given piece of data (for example, Archival, Long-term, Short-term, Open-access, Embargoed etc.). These should be connected not to the technical implementation of the storage enclosure but rather to the specific documented **procedures** that grant the user access rights to the data. For example, when placing a request to access a dataset stored as an Archival unit, the user should know that s/he will receive the data via any possible transportation mechanism within a specified period of time under persistent credentials (for examples, see the draft EOSC AAI Architecture[61]) and via a link with a known, limited lifetime.

Our Recommendations encourage RIs to define and publish their storage attribution specifications in accordance with their Data Policies and DMPs. Another task that forms part of Storage is the aggregation of metadata harvested from very different sources inside the RI. These aggregated data should be attributed, so, as much as possible, unification of field names and generalized harvesting processes is strongly encouraged to support such attribution. Creation and publication of generalized metadata naming conventions for system administrators and data stewardship engineers of partner RIs are also encouraged.

A special aspect of data storage related to PaN data is the application of regulations. Here is the very special situation when the regulations should be applied 'between' the stages of the data lifecycle. For example, under governance of the given DMP and Data Policy, data could be regulated with the known set of rules during acquisition. But when the data are placed into the externally accessible storage enclosure (irrespective of whether there is authorized access or not), they are regulated by different sets of rules.

Aggregation of the metadata, excluding internally required service fields, is also a regulated operation; however, in this case, the data exchange agreements should be applied in addition before the metadata are registered as stored entities. These described peculiarities require consideration of storage as a prolonged subprocess integrated into the other processes within the data continuum.

Looking forward, the persistent identification of scientific instruments is also a problem in terms of implementing FAIR. This identification also falls under storage activity because the instrument context is represented by constantly stored data that should be accessible externally to be FAIR-compliant. The PIDINST and DataCite projects have developed approaches to identify and represent the instrument as an entity within the metadata continuum.[62]

From the storage perspective, any dataset acquired for long-term storage should be linked with related instrument metadata to be FAIR-compliant. Also, there is a gap in the PaN-data ODI Data Continuum concerning which metadata should be stored for a given instrument as the totally persistent representation of that instrument. This is important, offering the possibility of identifying an instrument and capturing a generalized set of parameters related to it that can then be

[61] EOSC AAI Architecture Work and Subgroup (2020). EOSC AAI architecture. https://docs.google.com/document/d/12l0xVU9oiXqtVqkwrJijj16L-i_YcM_K6SHkNMf4IWE/edit

[62] Stocker, M., Darroch, L., Krahl, R. et al. (2020). Persistent identification of instruments. *Data Science Journal*, **19**. http://doi.org/10.5334/dsj-2020-018

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

disseminated across the community on the storage level. This could even provide opportunities to recover attributions for corrupted raw datasets, drawing on shared parameters to identify the instruments on which the data were produced.

Data storage should, therefore, be considered as a reusable subprocess of the whole data continuum.

5.5.1 Data storage: metadata types

- **Dataset Information**: 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]**
- **File Identifiers:** Identifies which files correspond to a specific dataset. File identifiers might be relevant for internal processes within the facility but not as essential for FAIR. [**P2 - IMPORTANT - AIR**]
- **Instrument Parameters:** Although the Instrument Parameters appear in the PaN-data ODI D6.1 Data Storage stage, they essentially represent a combination of the Instrument Information and Collaboration Information. Thus, they are not considered as a must during the Storage stage, i.e. because they are already captured during the Experiment stage. **[P3 - USEFUL - FR]**
- **Preservation Description Information:** Describes how the dataset is preserved within the facility (e.g. Storage or Archiving). **[P1 - ESSENTIAL - AR]**
- **Representation Information:** Format and structure of the files linked to the datasets. **[P3 - USEFUL - IR]**
- **Persistent Identifiers**: Unique identifier within or outside the organization that is linked to the data files or datasets. **[P1 - ESSENTIAL - FA]**

5.5.2 Data storage: roles

- **Experimental Team**: arranging to take data off site.
- **Data Infrastructure Team**: managing the data storage, transfer and publication process.

5.5.3 Data storage: information systems

- **File Writer/Generator System:** Integrates with the Data Acquisition Systems of the Experiment step and writes the data gathered, and often reduced, into the correct format files.
- **Data Management Systems:** Automated procedures that apply each facility's Data Policy will be also centralized in the Data Management System, starting from the acquisition to the storage and the eventual archiving of the data.

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

- **Data Storage Systems**: All acquired data shall be initially stored, with immediate access, in the facility's repository. Otherwise called "hot" data, there should be no waiting time between the request and the actual access to these data.
- **Data Publication Systems:** The data are made remotely accessible through cataloguing and publishing using Data Publication Systems, also known as Data Catalogues. The Data Catalogue interacts with the Data Storage and the Data Management Systems together to comply with the different aspects of the Data Policy, such as accessibility of the data and the application of the Embargo Period.
- **Archival Systems:** Complying with the Facility Data Policy, the archival system should be prepared to, either automatically or manually, move data from the storage to external devices, e.g. tapes, turning 'hot data' into 'cold data'. A retrieval procedure should also be in place if needed.

5.6 Data Analysis

The analysis is the step in the research where results are created. From the viewpoint of the RDM lifecycle, this is the step where the data that has been created, curated/adopted, and archived/stored is used.

In an ideal world, the data is retrieved from a data repository downloaded or passed to a remote analysis environment. The person doing the analysis is now taking advantage of the previous preparing steps.

The data analysis is, in most cases, a planned step when the data has been produced, and, with the rise of data repositories, unforeseen analysis steps might be performed. Therefore, deposited data with a wide range of metadata and data might be reused in unforeseen analysis steps.

In the three planning steps and in the experimental step, all required information has been created, while in the storage step, this information is being assembled and transformed.

The person wanting to use the data relies on these preparing steps that lead to the following:

1. Enough information in the data catalogue to find the data.
2. Clear terms of usage.
3. Access.
4. Formats that allow opening of the file.
5. Understanding the contents of the file.
6. Appropriate contents of the file for data analysis or other usage.
7. Clear provenance.

It has to be noted that, depending on the type of experiment, the Data Analysis is not done on the raw data but on other derived data, which we term 'Processed data'. In this sense, data processing

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

identifies steps performed on the raw data for transformation, integration or extraction in an appropriate output form for further analysis.

The data retrieved from the repository might have already passed some of these steps that ideally are described in the provenance. However, very often, these steps are performed after retrieving the data from the repository. Each of the created datasets enters the RDM lifecycle. Depending on the reproducibility of these steps and storage costs of resulting data, these data are ingested into the repository.

The flow between the steps described in PaN-data ODI 6.1 of initial post-processing, analysis of derived data, visualisation of data, and combination with other data should be made as 'easy' as possible. Easy means that the hurdles for users measuring in different facilities are as low as possible and that their analysis software can be used without major efforts on the datasets, including considering the usage of automated processing pipelines.

The scientist takes the raw data of the experiments, or even the processed data derived from those data, and carries out further analysis. The data from the instruments is typically in terms of counts of particles at particular frequencies or angles and needs highly specialized interpretation to derive the required end result, typically a 'picture' of a molecular structure, or a 3-D image of a nanostructure.

The analysis process is typically very unpredictable, and much of it takes place within the user scientists' institution(s) and under their control; again, much of the intellectual input of the scientists is involved in this part of the process, and the services of the facility staff have limited input. Here we give an outline of the general types of stages that are carried out in this step of the scientific process:

- **Initial post-processing:** Initial post-processing of raw data may be relatively standardized, generating processed data. For example a "reduced" dataset may be generated which is the result of comparing raw with calibration data and with background noise removed. This stage is often undertaken in the facility, using standardized methods and software.
- **Analyse derived data:** further analysis steps are undertaken by applying analysis software packages to the data to extract particular features or characteristics, or to fit it to a model, for example, to derive a molecular structure.
- **Visualise data:** data is transformed into a graphical form, which can be visualized and explored to provide a communication mechanism to the user scientists, as well as more widely.
- **Combine with other data:** the data is merged or compared with other data taken from other instruments or from modelling and simulations.
- **Interpret and analyse results:** the results are assessed by the scientific team to determine whether the results achieved so far are scientifically significant enough to warrant publication. If not, further analysis steps may be required.

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

- **Experimental report:** At some point after the experimental data has been taken, the experimental team is requested to produce an experimental report on the results of the use of the facility, which should be lodged with the facility.

The data analysis stage involves a range of data types, including processed and derived datasets, graphical information for visualisation, and software code.

5.6.1 Data analysis: metadata types

- **Analysis Team (User Identity):** 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.

- **Data Formats:** The format of the data is considered [**P1** - **ESSENTIAL - IR**] for interoperability and reproducibility as this field implicitly provides information about what kind of software can be used to work with the data format and what is the expected structure of the information inside.

- **Dataset Information:** Of the three identified levels of information for a given collection of data (Visit, Dataset and File), there are common contextual metadata which are relevant for the Analysis (e.g. Instrument set up). Nevertheless, most information should be available at the collection time, and not during the Analysis. In some cases, aggregation of information at the dataset level might be interesting. That is why this metadata field is considered [**P2** - **IMPORTANT - IR**].

- **File Identifiers**: Any given experiment or measurement can produce dozens of files which are used later on for the analysis. Therefore, it is [**P1** - **ESSENTIAL - AIR**] to identify the files used during this process to make sure the reproducibility can happen.

- **Instrument Parameters:** At the Analysis stage, the instrument parameters should not be considered more than [**P3** - **USEFUL - IR**] since this information should have already been captured at the Experiment stage and does not explicitly belong to any analysis procedure.

- **Calibration Information**: Similarly as with the Instrument parameters, the calibration information does not belong to this stage. It is, thus, considered only as [**P3** - **USEFUL - IR**], being that this classification is the less important of all of them.

- **Software Package Information:** 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**]

- **Dependence tracking and workflow**: Any dependence or relevant additional information to enable the resulting analysed data to be as consistent as possible is considered [**P2** - **IMPORTANT - R**] for reproducibility.

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

For this stage, we are proposing to include the following field:

- **Original Data:** identifying the original data is [**P1 - ESSENTIAL - IR**] for reproducibility and provenance of the data itself. Whether the Original Data is also Derived data or Raw data, any user should be able to track inversely where the data is coming from.

5.6.2 Data analysis: roles

- **Analysis Team:** As mentioned in the previous subsection, the Analysis Team, as part of the experimental team, is directly involved in the derivation of analysed results from the collected data. Some, or all, of the Analysis (or reduction) done by the Analysis Team may be performed automatically by a machine, in which case configurations for that process would need to be captured.
- **Instrument Scientist:** is likely to be involved in giving scientific advice and input on how to proceed with the interpretation and analysis of the data.
- **User Office**: accepting the Experimental Report.

5.6.3 Data analysis: information systems

- **Data Storage Systems**: where the data is located and made available for users to download or access it through a remote data access system (e.g. Data Catalogue).
- **User Office Systems:** The User Office will store the Experimental Reports, as part of the analysis results.
- **Analysis Software**: Software used for the Data Analysis. We are suggesting to use a Software Catalogue System in which all Analysis Software Packages should be located, referenced or linked.
- **Visualisation Systems**: Similarly as the Analysis Software Packages, the visualization systems would also include Data Analysis platforms, and Computing Resources in general, to visually enable the reconstruction of complex structures.

5.7 Publication

After collecting the data, through data curation and processing, the analysis results might produce new scientific results for the Experimental Team. At this stage, these results are made publicly available through journal articles.

The facility would want to be acknowledged, and the instrument cited to track the impact of the science produced from the use of the facility.

Much of the work in this stage involves the Experimental Team at their home institutions and does not involve facility support staff directly.

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

Data types involved in this stage include: the peer-reviewed article and related supplementary data.

5.7.1 Publication: metadata types

- **Authors/Coauthors (User Identity):** In the publication context, the Experimental Team prepares the journal and publication, linking to the data collected (raw data) or processed (derived data). This metadata field is considered as **[P1 - ESSENTIAL - FA]**
- **Proposal Information:** Proposal information or a link to the Proposal Identifier is also **[P1-ESSENTIAL - FA]** to understand the initial scientific case.
- **Publication Information:** Information about publisher and year of publication. **[P1 - ESSENTIAL - F].** It is highly recommended to give these pieces of information a separated metadata field.
- **Supplementary Data Information:** Additional information about the publication. **[P3 - USEFUL - F]**

For this stage, we are proposing to include the following field:

- **Persistent Identifier (PID):** Long-lasting reference to a digital resource, in this case the Publication. It reliably points to the digital resource. **[P1 - ESSENTIAL - F]**

5.7.2 Publication: roles

- **Experimental Team:** will prepare papers.
- **Instrument Scientist:** often involved in writing the paper as an author or co-author of the paper to be published.
- **User Office:** record the association of a paper with an experiment
- **Library:** The Library service, or library team, will lodge a metadata record and an appropriate copy of the publication.

5.7.3 Publication: information systems

- **User Office Systems**: Will contain the references of the Experimental Proposal of the Publications. In some of the cases, the User Office systems also store the information of the publication itself, i.e. acting as a Library System.
- **Research Output Tracking Systems:** Will track the impact of the publication for the facility.
- **Library Systems**: Stores the publication information and links it to the User Office systems information on proposal, experimental team and results.

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

- **Institutional Repository**: Holds the metadata record for the publication and, eventually, stores the publication itself, and the references to the journal to enable its access.

5.8 Additional Data Continuum Stages to Consider

The original PaN-data ODI D6.1 data continuum does not cover sufficiently some stages of the experimental lifecycle that have since emerged as both key and distinct. These include data processing and the depositing, archiving and curating of experimental data in publicly searchable data catalogues. The sections below address these additional stages of the present day facilities experimental data continuum.

5.8.1 Data processing

Initially considered as part of the Analysis step, we are proposing Data Processing as an independent stage of the Data Continuum due to its impact on the production of the data.

In most of the cases, there is a processing step involving the collected data before doing any kind of analysis. It has been widely understood that the processing, or even curation, of this data was an offline step of the whole research process.

For reproducibility reasons, there is an ever-growing need to identify derived data as an entity eligible to be published in data catalogues. Understanding where the data comes from and the changes that have occurred has become essential for FAIR.

5.8.1.1 Data processing: metadata types

- **Processing Team (User Identity):** the team performing the processing should be identified. This team might differ from that involved in the Experiment step or the Analysis step. Identifying who is doing the processing is [**P2** - **IMPORTANT - AIR**] but not essential for FAIR.
- **Data Format:** The format of the data is considered [**P1** - **ESSENTIAL - IR**] to get to know the structure of the information inside.
- **Original Data:** identifying the original data is [**P1** - **ESSENTIAL - IR**] for reproducibility and provenance of the data itself. Whether the Original Data is also Derived data or Raw data, any user should be able to track inversely where the data is coming from.
- **Dataset Information:** Same as in the Analysis process. [**P2** - **IMPORTANT - AIR**].
- **Processing Information:** Description of the algorithms used against the datasets and datafiles for processing. An [**P1** - **ESSENTIAL - R**] metadata field for provenance and reproducibility.
- **Software Package Information:** Similarly to the Analysis step, the processing step must include the software used for processing, as well as the version, as an [**P1** - **ESSENTIAL - R**] element for reproducibility.

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

5.8.1.2 Data processing: roles

- **Processing Team**: the Experimental team performs the processing of the data, either right after the collection of the raw data collection or from its own laboratory or facility. Some, or all, of the Processing done by the Processing Team may be performed automatically by a machine, in which case configurations for that activity would need to be captured.
- **Instrument Scientist:** She/He helps with the processing of the data, as part of the support provided during the visit of the Experimental Team.

5.8.1.3 Data processing: information systems

- **Data Storage Systems**: where the data is located and made available for users to download or access it through a remote data access system (e.g. Data Catalogue).
- **Processing Software Packages**: Software used for the Data Processing. Similarly to the Analysis step, we are suggesting to use a Software Catalogue System in which all Processing Software Packages should be located, referenced or linked.

5.8.2 Data record and/or publication

Funders increasingly recognise a wide range of research outputs, including research data, as valuable research outcomes in their own right. Historically, this was not the case: the focus was much more exclusively on research publications, especially the journal article, hence the reason that the final stage of the PaN-data ODI D6.1 data continuum deals only with publications and the supplementary data related to these publications.

Of course, publications are still an important research output, and facilities do still have the need to record these as outcomes of experiments. However, it is also important now for facilities to have a formal record of datasets. These data records may relate to any of the classes of experimental data (see Section 2.3.2), with some facilities archiving only raw data and others archiving additional types of data. Through data catalogues, these data can be made available, normally after an embargo period, for general searching and reuse, a similar process to the approach many facilities already employ for their publication catalogues.

While drawing on metadata collected throughout the experimental lifecycle, the data recording or data publishing process sits in the final stage of the experiment lifecycle, similarly to the recording of a publication. As well as a minimum set of bibliographic/descriptive metadata (i.e. which would be required for any publication process), the data record or data publication should contain information needed for the data to be FAIR.

5.8.2.1 Data record and/or publication: metadata types

- **Resource Identity:** should include the type of identifier, the identifier itself, and any related resource linked to it. **[P1 - ESSENTIAL - FI]**

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

- **Related Resource:** it would be either publications, proposals, other datasets. **[P2 - IMPORTANT - F]**
- **Creator:** Person or organization creating this resource. **[P1 - ESSENTIAL - F]**
- **Contributor:** Any Person or organization which contributed to the creation of the resource. **[P2 - IMPORTANT - F]**
- **Title:** Public name for the dataset. **[P1 - ESSENTIAL - F]** for data citation.
- **Publisher:** person or organization publishing this record. **[P1 - ESSENTIAL - FI]**
- **Publication Year:** Year dataset is published. **[P1 - ESSENTIAL - FI]**
- **License:** Will inform any data consumer about what can be done with the data and how authorship must be treated. **[P1 - ESSENTIAL - IR]**
- **Release Date:** Embargo period due date. The day when the dataset becomes Open Data. **[P1 - ESSENTIAL - IR]**

5.8.2.2 Data record and/or publication: roles

- **Experimental Team:** the Experimental Team will do the actual publication action, either automatically or manually, depending on the type of services provided. The Data Record and Publication might be attached to a persistent identifier minting action in which datasets and files to be linked together as part of the scientific result of the research.
- **Facility:** might provide the infrastructure and tools to enable the Persistent Identifier minting of the selected datasets.
- **Record Publishers**: In many cases, the same facility acts as record publisher after the PID is minted. It will publish a location where metadata is accessible and data is downloadable, Data Policy applied.
- **Persistent Identifier Provider:** third-party and authorized organization which, in agreement with the facility, will provide a unique identifier, linked to selected datasets, upon request (e.g. DataCite)

5.8.2.3 Data record and/or publication: information systems

- **Persistent Identifier Minting System:** Information Management System which will be integrated with the Persistent Identifier Provider and the Facility Information Management System to generate persistent identifiers upon request.
- **PIDs Platform:** a searchable and downloadable system where PIDs are accessible, as well as data downloadable. Any link or reference to any of the PIDs shall lead the searcher to this platform.

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

6. Moving Towards a Common FAIR Metadata Framework

The elaboration of recommendations is an ongoing process and will require feedback on its implementations.

6.1 The Data/Metadata Value Stream

By switching our perspective from the process point of view, to the data perspective, we want to highlight the impact each process has on the data and how it enriches and gives more value to the metadata, and thus there is a stream of added value through the experimental process (see Figure 8 below).

Our assumptions start with the idea that there is no value until the data is produced. Any derived data, from the Experiment step and beyond, should be considered a candidate to be published as a record, adding value to the aggregation of data associated with the experiment.

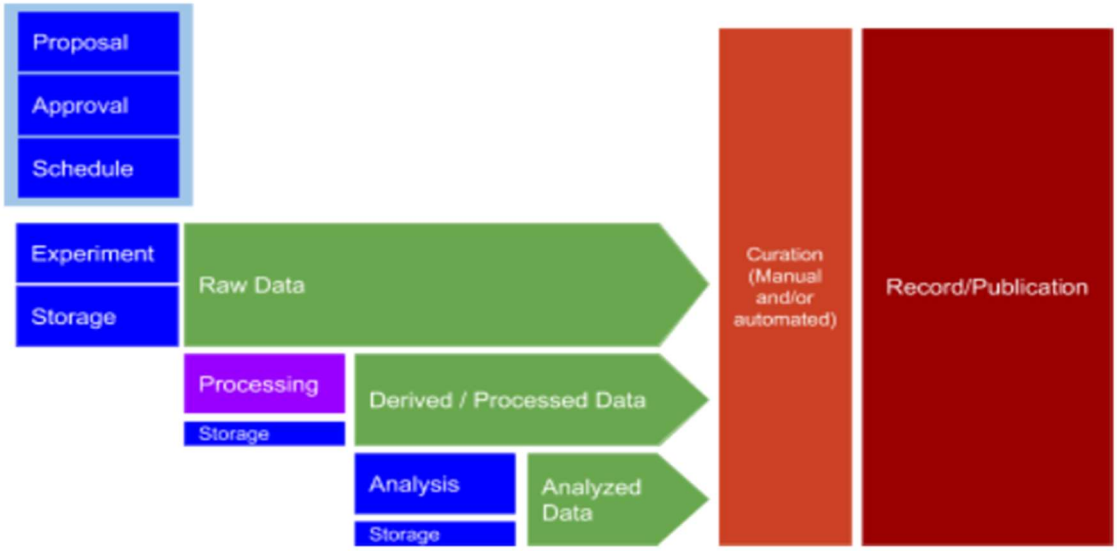

Figure 8: The data/metadata value stream

Within this context, and to identify the major blocks of data/metadata in the value stream, the generic meaning of these are as follows:

- **Raw data:** the data collected from experiments performed on facility instruments that are considered as the source for any further analysis / processing.
- **Derived / Processed Data:** data as a result of Reduction or Processing (new stage) the data during the Experimental lifecycle. Note: some may identify the Processing Data step as part of Analysis Data.

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

- **Analysed Data:** data results from the analysis of either raw, processed or derived data. This newly created data is also a candidate for the record/publication process.
- **Curated data:** Process of making data FAIR.

Taking this newly created intermediate steps, in which data and metadata are enriched, and applying this value stream concept to the Simplified illustration of Classes of Experimental Data in the Science lifecycle mentioned back in Section 2, we are proposing to apply the RDM lifecycle from the DCC after each step (see Figure 9 below).

With this, and presenting the FAIR quality verification process as a continuous validation process, and not an additional one at the end of the pipe, we understand FAIRness is much more likely to be accomplished in order to make the data and metadata exposable from a cataloguing tool.

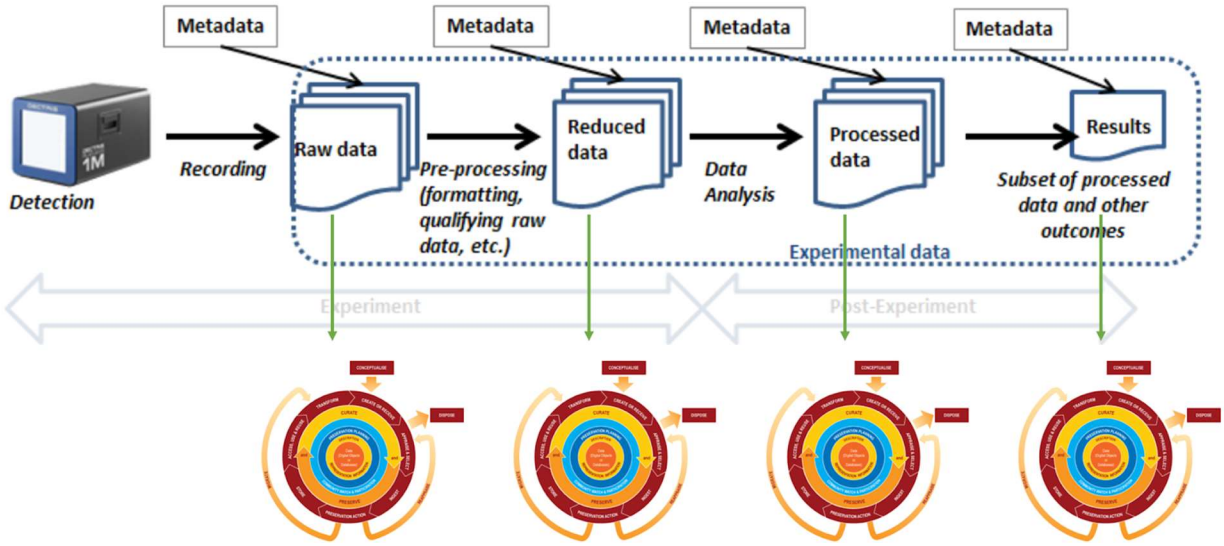

Figure 9: The value stream concept to the Simplified illustration of Classes of Experimental Data in the Science lifecycle with the RDM lifecycle from the DCC applied after each step

6.2 Proposal for FAIR Digital Objects in Photon and Neutron Science

In Section 2.1, the FAIR Digital Object was introduced, and in the gap analysis in Section 5, the phases of the PaN-data ODI research lifecycle were described. Here, we bring these two sections together and describe what information is required to build a FAIR Digital Object and at which stage of the experimental lifecycle the information is available.

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

As described in Section 2.1, a FAIR Digital Object has three layers around the data: PIDs, standards and code, and metadata. These layers will be described in the following and brought together with the research lifecycle and information sources' phases.

Relevant PIDs for PaN are mainly PIDs for datasets, instruments, samples, persons, organisations, funders, and papers. Other items such as formats, data types, metadata schema / semantic artefacts, and units PIDs could also be relevant.

Most PIDs are often minted long before proposal submission but need to be collected at some point in the research lifecycle. Exceptions are the datasets and the sample.

Before the proposal is submitted, PIDs for instruments, sometimes samples, persons, organisations, funders, and related publications already exist. PIDs can also be created for software and hardware environments.

On **proposal submission**, PIDs of samples if already existing, persons esp. the principal investigators and PIDs for related organisations, funders of the overall project and its funding ID, as well as PIDs related documents as safety documents and documents about prior art to the proposal need to be collected.

In the proposal, a specific instrument might be requested, and related configuration described, in this case, the instrument PID can be linked to the proposal. When the proposal is also submitted the proposal ID is created. Here also a first proposal for the experimental team is being submitted, and the related PIDs need to be collected.

On **proposal approval**, the instrument PID can finally be determined and linked to the approval. This instrument PID will later be linked to the datasets produced in the measurements. Related to the proposal approval are also the principal investigator and the approval committee.

On **proposal scheduling** the PIDs of the experimental team as well as the scheduled time needs to be collected.

During the proposal phase, various documents are produced which can be registered and receive a PID and referencing related PIDs.

When the **experiment** is executed, the collection of a multitude of metadata describing the measurement data is required to achieve a high degree of FAIR. In this step traceability of the experiment needs to be achieved to ensure reusability. Among these are data describing the instrumental setup (configuration and scan parameters) and the sample environment or setup are created. Also, calibration data, as well as the actual measurement data, are produced. Related to this information also is the required software and hardware environment. Very often, the data is not only produced at one source, but other sources such as electronic lab books or different measuring devices take part in the data acquisition.

In the **processing** and **analysis** phase, an existing dataset is used, and a new dataset is derived from the existing dataset. The existing dataset can have its origin from measurement data or data previously processed. Here sometimes in various computational steps, data is adopted and transformed to meet the requirements of the next processing step. It is required to record these processing and analysis steps to achieve reproducibility or at least traceability. Required

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

information here is among other the original data, software and hardware environment including configuration, and algorithms.

The **storage** phase is divided into various steps. These steps might overlap and not be totally sequential. First, if not already done during the experiment, processing or analysis, the data has to be aggregated and adopted. The aggregation step, data from the different sources will be brought together. These data include data collected during the proposal, approval and scheduling phases. Here context and the actual experimental data from the different sources are aggregated. The dataset will have to include the relevant PIDs.

In the next step, interoperability needs to be achieved by applying a standardised format. The required format depends on the repository or databank where the dataset is stored. Results of the analysis phase very often are stored in disciplinary repositories as the Protein Data Bank (PDB) or the GenBank.[63,64] Thus, the datasets have to accomplish the specific requirements. Relevant formats for measurement data have been introduced in Section 2.2. Despite the existence of very established formats in some communities, e.g. the .cbf files using the CIF standard in the crystallographic community, NeXus is a standard that is more and more prevailing as a standard for the archival of measurement data in the PaN community. This standard combines the requirement of interoperability and facilitates reusability by enabling traceability of the experiment execution by standardised rich metadata.

There is still some information required in order to make the experimental data reusable. This information is related to preservation description. This information can be partly stored inside the measurement file, but some information has to be outside. Requirements on metadata for this file are described in Section 2.2. In the specifications of the PREMIS Format,[65] these requirements are described. PREMIS is used to describe formats, creation and usage environment (software, hardware), and the provenance of the dataset from the perspective of general data management.

Another type of description is required to achieve reusability. Information about the experimental team and the research context has already been collected during the proposal phase. This information can be serialised in this step again either by creating an additional file or serialising it inside the experimental dataset. Applicable standards here are DataCite, DublinCore, and the Common European Research Information Format (CERIF).[66,67,68]

Most of the measurement data is normally transferred into the facilities' data repository and metadata ingested into the facilities data catalogue. If not already happened, this is the moment when the datasets PID is minted and related to the dataset. Here again, preservation specific

[63] Berman, H., Westbrook, J., Feng, Z. et al. (2000). The Protein Data Bank, *Nucleic Acids Research*, **28**. https://doi.org/10.1093/nar/28.1.235
[64] National Center for Biotechnology Information (NCBI) (2020). GenBank overview. https://www.ncbi.nlm.nih.gov/genbank/
[65] Library of Congress (2020). PREMIS home. https://www.loc.gov/standards/premis/
[66] DataCite Metadata Working Group (2019). DataCite metadata schema 4.3. https://schema.datacite.org/
[67] Dublin Core Metadata Initiative (DCMI) (2020). DCMI metadata terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[68] EuroCRIS (n.d.). CERIF in brief. https://www.eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html

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

information is required e.g. access rights or open access, end of embargo time, replica, e.g. for faster access and their retention time. This, again, can be described using PREMIS.

6.3 Summary Draft Recommendations for a Common FAIR Metadata Framework

The purpose of this document is to provide a set of recommendations for a common Metadata framework for the Photon and Neutron National Research Institutions. Although National RIs, and ESFRIs, have been collaborating closely over the past decade all facilities have been executing their own individual strategies, also in terms of data management, in many cases due to national regulations.

The first step of the transition to become FAIR was to make sure that the essential pieces of information as proposed in Section 5 above, and captured in metadata fields, are available to set the grounds for a FAIR integration among the different facilities in Europe.

Once the essential metadata fields are identified, facilities should be able to follow their journey to, together, move to a FAIR scenario, identifying Semantics, Structure, Syntax, and Systems along the process.

At each identified step, and before data and metadata are both published on any publicly available catalogue, the FAIR data manager of the RI must validate metadata on the following terms and consider this questions as ESSENTIAL to comply.

6.3.1 Proposal

1. Is the **Principal investigator** declared as part of the metadata fields?
2. Are the **Co-Investigators** declared as part of the metadata fields?
3. Is the **Instrument requested** declared as part of the metadata fields?
4. Is the **Sample description** declared as part of the metadata fields?
5. Is the **Facility** where the proposal is submitted declared as part of the metadata fields?
6. Is the **Proposal Identifier** declared as part of the metadata fields?
7. Is the **Experiment Description** declared as part of the metadata fields?
8. Are the **Proposed Experiment Conditions** declared as part of the metadata fields?

6.3.2 Approval

It is assumed that no metadata fields are relevant for FAIR at the Approval step, being an internal process for the facility.

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

6.3.3 Scheduling

Scheduling the measurement or experiment is considered as planning and does not correspond to any actual information which is relevant for FAIR.

6.3.4 Experiment

9. Is the actual **Visiting Experimental Team** (people who actually participate during the measurement) declared as part of the metadata fields?
10. Are the **Experiment/Measurement dates** declared as part of the metadata fields?
11. Does the **Samples information** provide enough context to understand its structure and characteristics and is declared as part of the metadata fields?
12. Is the **Instrument information** declared as part of the metadata fields?
13. Is the **Calibration information** declared as part of the metadata fields?
14. Is the produced **Dataset information** declared as part of the metadata fields?

6.3.5 Processing

15. Is the resulting **Data Format** declared as part of the metadata fields?
16. Is the **Processing information** declared as part of the metadata fields?
17. Is the **Software package information** used for processing declared as part of the metadata fields?
18. Is the **Original Data** link used for the processing declared as part of the metadata fields?
19. Is the resulting **Dataset information** declared as part of the metadata fields?

6.3.6 Analysis

20. Is the resulting **Data Format** of the Analysis declared as part of the metadata fields?
21. Are the **Files Identifiers** declared as part of the metadata fields?
22. Is the **Software package** used for the analysis declared as part of the metadata fields?
23. Is the **Original Data** link used for the analysis declared as part of the metadata fields?
24. Is the resulting **Dataset information** declared as part of the metadata fields?

6.3.7 Record

25. Is the **Resource Identity** declared as part of the metadata fields?

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

26. Is the **Creator** of the record declared as part of the metadata fields?
27. Is the **Publisher** of the record declared as part of the metadata fields?
28. Is the **Publication year** declared as part of the metadata fields?
29. Is the **Release date** (Embargo due date) declared as part of the metadata fields?
30. Is the **Title** of the dataset declared as part of the metadata fields?
31. Is the **License** for usage declared as part of the metadata fields?

6.3.8 Publication

Publication information might not be part of the metadata fields of the data used of that particular publication.

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

References

Ashton, A., Da Graca Ramos, S., Matthews, B. et al. (2019). ExPaNDS data landscaping survey. https://doi.org/10.5281/zenodo.3673811

Berman, H., Westbrook, J., Feng, Z. et al. (2000). The Protein Data Bank, *Nucleic Acids Research*, **28**. https://doi.org/10.1093/nar/28.1.235

Bernstein, H. J., Bollinger, J. C., Brown, I. et al. (2016). Specification of the Crystallographic Information File format, version 2.0, *J. Appl. Cryst.* **49**. https://doi.org/10.1107/S1600576715021871

Bernstein, H. J., Förster, A., Bhowmick, A. et al. (2020). Gold Standard for macromolecular crystallography diffraction data, *IUCrJ*, **7**. https://doi.org/10.1107/S2052252520008672

Brocks, H., Kranstedt, A., Jäschke, G., Hemmje, M. (2010). Modeling context for digital preservation. In *Studies in Computational Intelligence,* **260***.* Szczerbicki, E. and Nguyen, N. T. (Eds.), pp. 197 – 226. https://www.researchgate.net/publication/225757343_Modeling_Context_for_Digital_Preservation

Bunakov, V., Matthews, B., Jejkal, T. et al. (2018). NFFA Deliverable D11.14 Final metadata standard for nanoscience data. https://www.nffa.eu/media/202831/d1114_final-metadata-standard-for-nanoscience-data.pdf

CODATA-VAMAS Working Group on the Description of Nanomaterials & Rumble, J. (2016). Uniform description system for materials on the nanoscale, version 2.0. http://doi.org/10.5281/zenodo.56720

DataCite Metadata Working Group (2019). DataCite metadata schema 4.3. https://schema.datacite.org/

DCC (2020). Curation Lifecycle Model. https://www.dcc.ac.uk/guidance/curation-lifecycle-model

Devaraju, A., Huber, R., Mokrane, M. et al. FAIRsFAIR data object assessment metrics. https://doi.org/10.5281/zenodo.4081213

Dimper, R. (2011). Common policy framework on scientific data. https://doi.org/10.5281/zenodo.37384978

DLS (2020). GDA software for science. http://www.opengda.org

Dublin Core Metadata Initiative (DCMI) (2020). DCMI metadata terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

EC Expert Group on FAIR data (2018). Turning FAIR into reality. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

EOSC AAI Architecture Work and Subgroup (2020). EOSC AAI architecture. https://docs.google.com/document/d/12l0xVU9oiXqtVqkwrJijj16L-i_YcM_K6SHkNMf4IWE/edit

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

EOSC FAIR and Architecture Working Groups (2020). A Persistent Identifier (PID) policy for the European Open Science Cloud. https://op.europa.eu/en-GB/publication-detail/-/publication/35c5ca10-1417-11eb-b57e-01aa75ed71a1/language-en

EuroCRIS (n.d.). CERIF in brief. https://www.eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html

European Parliament and Council of European Union (2016). Regulation (EU) 2016/679. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

Gagey, B. (ed.) (2018). SOLEIL data management policy. https://www.synchrotron-soleil.fr/en/file/11308/

Genova, F., Aronsen, J. M., Beyan, O. et al. (2020). Recommendations on FAIR metrics for EOSC. Second draft for consultation. http://doi.org/10.5281/zenodo.4106116

Gonzalez-Beltran, A. (2020). Large-scale facilities experimental lifecycle & FAIRness. http://doi.org/10.5281/zenodo.4067988

Hall, S. R., Allen, F. H. and Brown, I. D. (1991). The Crystallographic Information File (CIF): A new standard archive file for crystallography", *Acta Cryst.*, A**47**. http://ww1.iucr.org/iucr-top/cif/standard/cifstd1.html

Higgins, S. (2008). The DCC Curation Lifecycle Model, *Int. J. Digit. Curation*, **3**. https://doi.org/10.2218/ijdc.v3i1.48

Houssos, N., Jörg, B. and Matthews, B. (2012). A multi-level metadata approach for a Public Sector Information data infrastructure. http://purl.org/net/epubs/work/62617

ispyb (2020). ISPyB. https://ispyb.github.io/ISPyB/

IUCr (2005). Image CIF dictionary. https://www.iucr.org/resources/cif/dictionaries/cif_img

IUCr (2005). Macromolecular CIF dictionary. https://www.iucr.org/resources/cif/dictionaries/cif_mm

IUPAC (2020). IUPAC gold book. https://goldbook.iupac.org/

Könnecke, M., Akeroyd, F. A., Bernstein, H. J. et al. (2015). The Nexus data format, *J. Appl. Cryst.* **48**. https://doi.org/10.1107/S1600576714027575

Library of Congress (2020). PREMIS home. https://www.loc.gov/standards/premis/

Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910

Moore, R., Stotzka, R., Cacciari, C., Benedikt, P. (2015). Practical policy (Version 1.0). http://doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC

National Center for Biotechnology Information (NCBI) (2020). GenBank overview. https://www.ncbi.nlm.nih.gov/genbank/

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

NIAC (2020). NeXus user manual and reference documentation. 3.3.2.10. NXmx. https://manual.nexusformat.org/classes/applications/NXmx.html#nxmx

openBIS (n.d.). openBIS. https://openbis.ch/

RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. https://doi.org/10.15497/RDA00050

RDA International Materials Resource Registries WG (2017). Materials registry vocabulary draft. https://www.rd-alliance.org/system/files/documents/Materials_Registry_vocab_draft_170321.pdf

Schwardmann, U., Fenner, M., Hellström, M. et al. (2020). PID architecture for the EOSC, version 0.3. https://docs.google.com/document/d/1T-bpNsmuxQewsLq48XTyUJoe0lsV7poaXohpgDo9W34/edit#heading=h.81f971wdg07u

Stocker, M., Darroch, L., Krahl, R. et al. (2020). Persistent identification of instruments. *Data Science Journal*, **19**. http://doi.org/10.5334/dsj-2020-018

W3C (2014). Data catalog vocabulary (DCAT). https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/

W3C (2020). Data catalog vocabulary (DCAT) –version 2. https://www.w3.org/TR/vocab-dcat-2/

Wilkinson, M. D. et al. (2015). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**:1. https://doi.org/10.1038/sdata.2016.18

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

Annex 1: Metadata FAIR Prioritization Summary Table

Following the same structure as proposed by the RDAFAIR Data Maturity Model,[69] we are proposing a list of indicators which will assess the level of FAIR compliance for photon and neutron facilities, according to the priorities proposed in this document.

The proposed new metadata fields for FAIR compliance are in blue in the Table below.

Stage	FAIR	Indicator	Priority
Proposal	FA	Principal Investigator is identified	P1 - E
Proposal	F	Instrument requested is identified	P1 - E
Proposal	F	Funding Source is identified	P2 - I
Proposal	F	Sample are declared	P1 - E
Proposal	F	Experimental Conditions are requested	P1 - E
Proposal	F	Safety Conditions	P3 - U
Proposal	F	Experiment Description is stored	P1 - E
Proposal	F	Prior art are identified	P2 - I
Proposal	FA	Co-Investigators are declared	P1 - E
Proposal	F	Facility Information is provided	P1 - E
Proposal	F	Proposal unique identifier is declared	P1 - E
Approval	-	Approval Panel	P3 - U
Scheduling	FA	Allocated time on instrument	P2 - I
Scheduling	FA	Scheduled Visiting Experimental Team	P2 - I
Scheduling	-	Training	P3 - U
Scheduling	F	Details experimental planning	P2 - I
Scheduling	FR	Sample preparation	P2 - I

[69] RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. 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.

Scheduling	-	Sample reception	P3 - U
Experiment	FA	Visiting Experimental Team	P1 - E
Experiment	FR	Sample information	P1 - E
Experiment	FR	Instrument information	P1 - E
Experiment	FR	Experiment Planning	P2 - I
Experiment	FR	Environmental Parameters	P2 - I
Experiment	FR	Calibration information	P1 - E
Experiment	FR	Laboratory Notebook	P2 - I
Experiment	F	Instrument Scientist	P2 - I
Experiment	FA	Experiment date	P1 - E
Experiment	R	Experimental Report	P3 - U
Storage	F	Dataset information	P1 - E
Storage	AIR	File identifiers	P2 - I
Storage	FR	Instrument Parameters	P3 - U
Storage	AR	Preservation Description Information	P1 - E
Storage	IR	Representation Information	P3 - U
Storage	FA	Persistent Identifiers	P1 - E
Analysis	AIR	Analysis Team	P2 - I
Analysis	IR	Data Formats	P1 - E
Analysis	IR	Dataset Information	P2 - I
Analysis	AIR	File Identifiers	P1 - E
Analysis	IR	Instrument Parameters	P3 - U
Analysis	IR	Calibration Information	P3 - U
Analysis	IR	Software Package Information	P1 - E
Analysis	R	Dependence tracking and workflow	P2 - I
Analysis	IR	Original Data	P1 - E

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

Publication	FA	Authors/Co-Authors	P1 - E
Publication	FA	Proposal Information	P1 - E
Publication	F	Publication information	P1 - E
Publication	F	Supplementary data information	P3 - U
Processing	AIR	Processing team	P2 - I
Processing	IR	Data format	P1 - E
Processing	IR	Original data	P1 - E
Processing	AIR	Dataset information	P2 - I
Processing	R	Processing information	P1 - E
Processing	R	Software package information	P1 - E
Record	FI	Resource identity	P1 - E
Record	F	Related resource	P2 - I
Record	F	Creator	P1 - E
Record	F	Contributor	P2 - I
Record	F	Title	P1 - E
Record	FI	Publisher	P1 - E
Record	FI	Publication year	P1 - E
Record	IR	License	P1 - E
Record	IR	Release date	P1 - E

Table 2: Metadata FAIR prioritization summary

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

Annex 2: Note about Nexus Format

Although this document does not intend to be an implementation document, the need for a standard format implementation is clear.

Many RIs recommend using the Nexus file format to record metadata and data. The Nexus data format offers either base classes that can be used separately to adapt to specific use cases within the facility or application definition classes.[70]

According to the Nexus documentation, application definition is considered a *contract* between a data provider (such as the instrument control /data acquisition system) and a data consumer (such as a data analysis program for a scientific technique). It describes the information that is certain to be available in a data file.

The application definitions mainly focus on raw data files rather than on providing required metadata for processed data that could make data FAIR. Furthermore, there is a distinction between the metadata required for data processing and data analysis.

[70] Könnecke, M., Akeroyd, F. A., Bernstein, H. J. et al. (2015). The Nexus data format, *J. Appl. Cryst.* **48**. https://doi.org/10.1107/S1600576714027575

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