D2.4: DMPs for Photon and Neutron RIs

Document Control Information

Settings	Value
Document Identifier:	D2.4
Project Title:	ExPaNDS
Work Package:	WP2
Work Package Lead	UKRI
Document Author(s):	Heike Görzig (HZB), Brian Matthews (UKRI), Abigail McBirnie (UKRI), Nicolas Soler (ALBA)
Document Contributor(s):	Fredrik Bolmsten (ESS), Emilio Centeno (ALBA), Andrey Vukolov (ELETTRA)
Document Reviewer(s):	Jonathan Taylor (ESS), Majid Ounsy (SOLEIL), Sophie Servan (DESY)
Doc. Version:	1.0
Dissemination level:	Public
Date:	30/11/2021

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

Abstract

In order to support the writing of Data Management Plans (DMPs), this report looks for sources to fill in the DMP (semi-)automatically. In particular, the report investigates from which sources, which roles, and when in a project or proposal lifecycle, the questions in a DMP can be answered. For this purpose, first general existing roles, workflows, and IT systems in projects in the literature and then their relevance to Photon and Neutron (PaN) Research Infrastructures (RIs) are examined. Using the DMP template developed by PaNOSC, roles, phases in the workflows and IT systems are then mapped to these DMP questions. The mapping shows that the users can be strongly supported by the facility staff when writing a DMP and that many of the DMP questions can be answered even before a proposal or project starts.

Document Log

Version	Date	Comment	Author/Partner
Complete Draft	04/11/2021	Internal review version	Heike Görzig (HZB)
1.0	30/11/2021	Final version for submission	Heike Görzig (HZB)

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

Abbreviations and acronyms

ADMP	Active Data Management Plan
API	Application Programming Interface
CCSDS	Consultative Committee for Space Data Systems
CERIC-ERIC	Central European Research Infrastructure Consortium
DCC	Digital Curation Centre
DESY	Deutsches Elektronen-Synchrotron
DLS	Diamond Light Source
DMP	Data Management Plan
DOI	Digital Object Identifier
EGI	European Grid Infrastructure Foundation
ELI	Extreme Light Infrastructure
EOSC	European Open Science Cloud
ESRF	European Synchrotron Radiation Facility
ESS	European Spallation Source
ExPaNDS	European Open Science Cloud (EOSC) Photon and Neutron Data Service
FAIR	Findable, Accessible, Interoperable, Reusable
FDO	FAIR Digital Object
FDOF	FAIR Digital Object Framework
HZB	Helmholtz-Zentrum Berlin
HZDR	Helmholtz-Zentrum Dresden-Rossendorf
IG	Interest Group
ILL	Institut Laue-Langevin
ISIS	ISIS Neutron and Muon Source
ISO	International Organization for Standardization
maDMP	machine-actionable Data Management Plan
OAIS	Open Archival Information System
PaN	Photon and Neutron
PaN RI	Photon and Neutron Research Infrastructure
PaN-data ODI	PaNdata Open Data Infrastructure
PaNOSC	Photon and Neutron Open Science Cloud
PID	Persistent Identifier
PMBOK	Project Management Body of Knowledge
PMI	Project Management Institute
PSI	Paul Scherrer Institute
RDA	Research Data Alliance

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

RDM	Research Data Management
RDMO	Research Data Management Organiser
RI	Research Infrastructure
STFC	Science and Technology Facilities Council
UKRI	UK Research and Innovation
URL	Uniform Resource Locator
VM	Virtual Machine
XFEL	X-Ray Free-Electron Laser

Additional abbreviations and acronyms specific to Figure 5

C3PO	Clever, Crafty, Content Profiling of Objects
CCEx	Curation Cost Exchange
CDE	Code, Data, and Environment packaging
DROID	Digital Record Object Identification
FITS	File Information Tool Set
LDAP	Lightweight Directory Access Protocol
OPM	Open Provenance Model
OSF	Open Science Framework
PREMIS	Preservation Metadata: Implementation Strategies
PROVO-O	PROV Ontology
SCAPE	Scalable Preservation Environments

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

Executive Summary

The goal of this report is to find sources (persons/roles and IT systems) to answer the questions in a Data Management Plan (DMP) and find out in which phase of a proposal or project they arise. This, in a later stage, should help to design an IT system that enables the creation of DMPs (semi-)automatically.

The analysis starts with a review of some selected DMP workflows and roles in combination with project management lifecycles from the literature. The relevance of these workflows and lifecycles are explored further for DMPs in the Photon and Neutron Research Infrastructure (PaN RI) context.

Drawing heavily on work presented in ExPaNDS deliverable *D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management* (Dec 2020), Chapter Three overviews a model of the PaN experimental lifecycle and the systems, roles, and workflows needed to support a metadata framework for Research Data Management (RDM) in PaN RIs. With this overview in mind, the section then considers the relationship of the PaN DMP to both this model of the experimental lifecycle and the metadata framework.

The following chapter presents ways to answer the questions in the PaNOSC DMP template (PaNOSC *D2.2: DMP Template for Facility Users*). Based on the project management lifecycle and phases of the PaN research lifecycle, DMP phases for PaN facilities are introduced. In these phases the DMP-relevant information either becomes more concrete or accumulates. In particular, the analysis looks at the sources of information and when information becomes available across the experimental lifecycle. IT systems and roles described in the previous chapter are mapped to the DMP questions and the phase of their emergence is noted.

Chapter Five concludes the report and sets out next steps.

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.

1. Introduction

1.1 Aims of the Deliverable and Related Documents

This ExPaNDS[1] deliverable addresses the requirements of the first part of the project's task on data management planning, which focuses on answering the questions in Data Management Plan (DMP) templates to help facility users and staff meet their DMP obligations. The second part of our data management planning task aims to make these steps more focused on the implementation. We will report next year on this work in a follow-on deliverable, ExPaNDS *D2.8: Active DMPs for Photon and Neutron RIs* (Nov 2022).

ExPaNDS deliverable D2.4 is the second in a series of three related documents:

1. PaNOSC[2] *D2.2: DMP Template for Facility Users* (Nov 2021)[3] proposes a discipline specific DMP template for Photon and Neutron Research Infrastructures (PaN RIs). The template sets out the questions to be answered to compile the DMP. D2.2 also analyses what information is required for what purpose in relation to the DMP.
2. ExPaNDS *D2.4: DMPs for Photon and Neutron RIs* (Nov 2021) (i.e. the present document) considers possible sources for the information needed to answer the questions proposed in the PaNOSC template. In some cases, the required information may be available from the Research Infrastructure (RI) (e.g. stored in the user office system, from the instrument scientist); in others, users may need to provide additional information themselves. D2.4 also examines when the relevant DMP information is available. Some information may be available straight away, at the proposal stage, but other information may not become available until much later in the experimental lifecycle.
3. ExPaNDS *D2.8: Active DMPs for Photon and Neutron RIs* (Nov 2022) will build on the work of the previous two DMP documents by demonstrating possible technical solutions for linking the DMP template with relevant knowledge sources (i.e. that hold the information needed to answer the questions that make up the DMP template).

[1] The ExPaNDS project "…aims to deliver standardised, interoperable, and integrated data sources and data analysis services for Photon and Neutron facilities". The project brings together 10 national PaN RIs: Deutsches Elektronen-Synchrotron (DESY), Paul Scherrer Institute (PSI), Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Diamond Light Source (DLS), MAX IV, Elettra, ALBA, SOLEIL, Helmholtz-Zentrum Berlin (HZB), and ISIS Neutron and Muon Source (ISIS). Additional ExPaNDS partners include: UK Research and Innovation (UKRI), Science and Technology Facilities Council (STFC), and European Grid Infrastructure Foundation (EGI). ExPaNDS (2020). ExPaNDS. https://expands.eu

[2] PaNOSC is a "…project for making FAIR data a reality in 6 European Research Infrastructures (RIs), developing and providing services for scientific data and connecting these to the European Open Science Cloud (EOSC)." The PaNOSC partners are: European Synchrotron Radiation Facility (ESRF), Central European Research Infrastructure Consortium (CERIC-ERIC), ELI Delivery Consortium, European Spallation Source (ESS), European Grid Infrastructure Foundation (EGI), European XFEL, and Institut Laue-Langevin (ILL). PaNOSC (2020). PaNOSC. https://www.panosc.eu

[3] Bolmsten, F., Lobley, C., and Taylor, J. (2021). D2.2: DMP Template for facility users. https://doi.org/10.5281/zenodo.5639428

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

1.2 Links between this Deliverable and Other Work in ExPaNDS

This ExPaNDS DMP deliverable links with other work around persistent identifiers (PIDs), metadata, and research data management (RDM) that is currently ongoing in the project. For example, in considering both the sources of information and when information becomes available (i.e. to answer the questions that make up the DMP template), we draw extensively on the experimental lifecycle model and metadata framework presented in ExPaNDS deliverable *D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management* (Dec 2020).[4] We also look ahead to other ExPaNDS deliverables, especially *D2.5: Advanced Infrastructure for PIDs in Photon and Neutron RIs* (Feb 2022) and *D2.7: Final Recommendations for FAIR Photon and Neutron Data Management* (May 2022), to ensure our DMP work relates to the emerging key aspects of these ongoing tasks.

As well as these links with other ExPaNDS activity, as highlighted in section 1.1, D2.4 is tightly interwoven with parallel work in PaNOSC, especially the project's DMP template for PaN RIs.

1.3 Overview of the Deliverable

This deliverable incorporates five main parts:

- The current section (Chapter One) introduces the purpose and scope of D2.4 and relates the deliverable to other work done in ExPaNDS and PaNOSC.
- Chapter Two reviews some selected DMP workflows and roles from the literature and explores the relevance of these for DMPs in the Photon and Neutron (PaN) context.
- Drawing heavily on work presented in ExPaNDS deliverable *D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management* (Dec 2020),[5] Chapter Three gives an overview of a model of the PaN experimental lifecycle and the systems, roles, and workflows needed to support a metadata framework for research data management (RDM) in PaN RIs. With this overview in mind, the section then considers the relationship of the PaN DMP to both this model of the experimental lifecycle and the metadata framework.
- Chapter Four presents ways to answer the questions in the PaNOSC DMP template. In particular, the analysis looks at sources for information and when information becomes available across the experimental lifecycle.
- Chapter Five concludes the report and sets out next steps.

[4] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). D2.2: Draft recommendations for FAIR photon and neutron data management. https://doi.org/10.5281/zenodo.4312825
[5] Ibid.

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

1.4 The PaNOSC Template

The PaNOSC project undertook a requirements analysis on DMPs. As a result of this work, PaNOSC has developed a DMP template for use by PaN RIs.[6] The template relies heavily on the template developed by the Research Data Management Organiser (RDMO) project.[7] More specifically, the help text and question wordings have been adapted to fit the research carried out at PaN RIs. Further, questions that are of less relevance to PaN science have been removed.

The template consists of a knowledge base, a questionnaire, and views or filters. The questionnaire incorporates seven broad areas of focus, which are explained in Chapter Four (see section 4.2). The knowledge base is called 'domain' in the original RDMO questionnaire and provides the underlying structure for the template. Each question in the questionnaire maps to one (and only one) item of the domain or knowledge base. There can be different DMP questionnaires, but they will all map to the same underlying knowledge base. One or many items in the knowledge base can map to a view.

The questionnaire and template are meant to be extended when implemented by a facility or at an instrument or beamline. This then allows a question such as "Which tools, software, technologies or processes are used to generate or collect the raw data?" to be broken down or refined in a way that is relevant to that particular facility or beamline, e.g. to the specific data acquisition system that will be used in a measurement. More importantly, these new questions still have to map to the same items in the domain or knowledge base.

In this way, the knowledge base becomes a medium that makes the DMPs interoperable between the facilities.

2. DMP Workflows and Roles in the Literature

As described above, our approach to DMPs is to re-use information provided for other purposes to fill in the DMP as much as possible. Therefore, in this section, two approaches will be presented where first project workflows and then data sources are analysed by the information content they can offer. The goal is to identify phases and roles in projects or proposal workflow in which information relevant for the DMP arises and map these phases and roles to the DMP questions. In particular, the first approach was chosen as it will later on be applied to the PaN research lifecycle. In the next section, Chapter Three, sources and workflows specific to PaN RIs will be described.

2.1 DMPs within the Experimental Workflow

The requirement to automate DMPs is not a new topic. In 2017, 'Active Data Management Plan (ADMP) scenarios' were presented in a workshop at the International Digital Curation

[6] Bolmsten, F., Lobley, C., and Taylor, J. (2021). D2.2: DMP Template for facility users. https://doi.org/10.5281/zenodo.5639428
[7] RDMO (n.d.). RDMO Research Data Management Organiser. https://rdmorganiser.github.io/

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

Conference,[8] and a white paper about requirements on machine-actionable DMPs (maDMPs) was published.[9] The conclusions of this workshop were subsequently refined and published.[10] This presented 10 principles for maDMPs. We shall not go into these principles in detail, but we note that the first two are in line with our approach:

- **Principle 1: Integrate DMPs with the workflows of all stakeholders in the research data ecosystem.** We propose that the completion, *and importantly the use*, of DMPs should be integrated into the experimental workflow practise of the research facility, and be the responsibility of all the relevant stakeholders, including instrument scientists, user office and IT staff as well as the Principal Investigator (PI).

- **Principle 2: Allow automated systems to act on behalf of stakeholders.** By integrating the DMP into the workflow, we can integrate its completion and use into the existing process of collecting and communicating information about the experiment, exploiting existing information systems.

Further, at the beginning of a project, some information required for the DMP is still missing while other planning information is already quite clear. Conversely, at the end of the project, plans about data storage, data format, etc. become better defined while other information might not be available anymore. Therefore, DMPs are commonly regarded as living documents[11] that need to be updated during the project's runtime in order to be more than a bureaucratic burden.

The Research Data Alliance (RDA) Active Data Management Plans Interest Group (IG)[12] had initially proposed the division of the creation and application DMPs into four phases during a project[13] (see also Figure 1 below). This emphasises the use of DMPs as much as the collation, and the need to support the DMP process with tools and monitoring; however, while funders are important stakeholders for RIs' DMPs, they are not necessarily the prime beneficiary of the DMP. Rather, the value lies:

- in the smooth operation of the data management and compute process within the facility,
- in the benefit to the facility user in having well documented and accessible data,
- to the wider scientific community in the production of FAIR (Findable, Accessible, Interoperable, Reusable) data.

[8] DCC (2017). 12th International Digital Curation Conference: Workshops. https://www.dcc.ac.uk/events/idcc17/workshops#workshop1
[9] Simms, S., Jones, S., Mietchen, D. and Miksa, T. (2017). Machine-actionable data management plans (maDMPs). https://riojournal.com/articles.php?id=13086
[10] Miksa, T., Simms, S., Mietchen, D. and Jones, S. (2019). Ten principles for machine-actionable data management plans, *PLOS Computational Biology*, 15:3, e1006750. https://doi.org/10.1371/journal.pcbi.1006750
[11] Open AIRE (2017). What is a Data Management Plan and how do I create one. https://www.openaire.eu/what-isa-data-management-plan-and-how-do-i-create-one
[12] See https://www.rd-alliance.org/groups/active-data-management-plans.html .
[13] RDA Active Data Management Plans IG (2017). ADMP scenarios. https://rd-alliance.org/group/active-data-management-plans/wiki/admp-scenarios.html

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

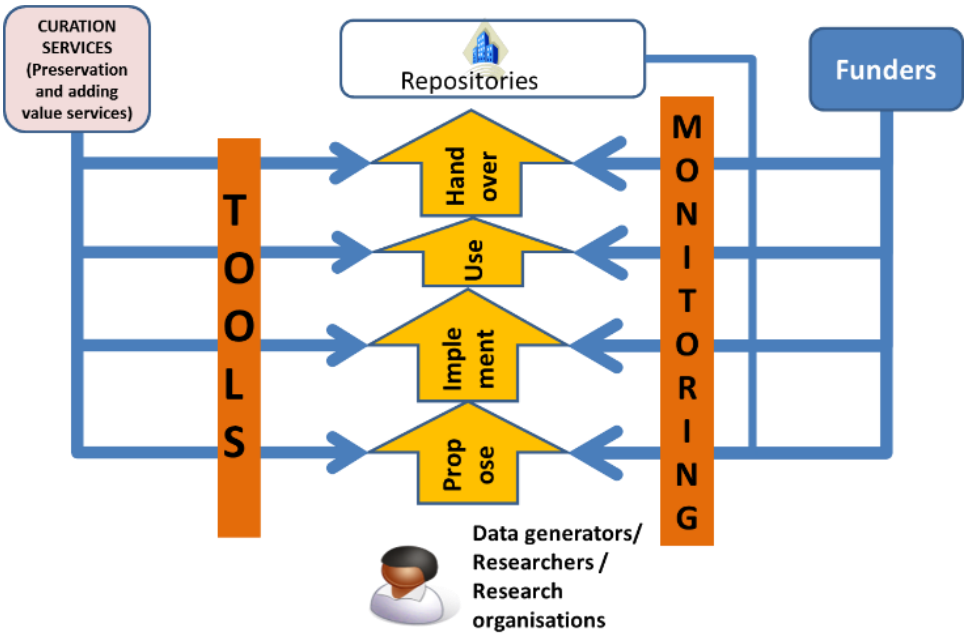

Figure 1: ADMP Phases, as proposed by the RDA Active DMP Interest Group (IG)[14]

In 2016, there was an International and Interdisciplinary workshop on Active Data Management Plans[15] organised by the RDA Active Data Management Plans IG. A project management method called Project Management Body of Knowledge (PMBOK)[16] was applied to DMPs. In PMBOK a project is divided into five process groups: **initiation**, **planning**, **execution**, **controlling/monitoring** and **closing** (see Figure 2 below).

[14] Ibid.
[15] RDA Active Data Management Plans IG (2016). International and interdisciplinary workshop on Active Data Management Plans. https://indico.cern.ch/event/520120/
[16] Project Management Institute (PMI) (2013). PMBOK® Guide – Fifth Edition.

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

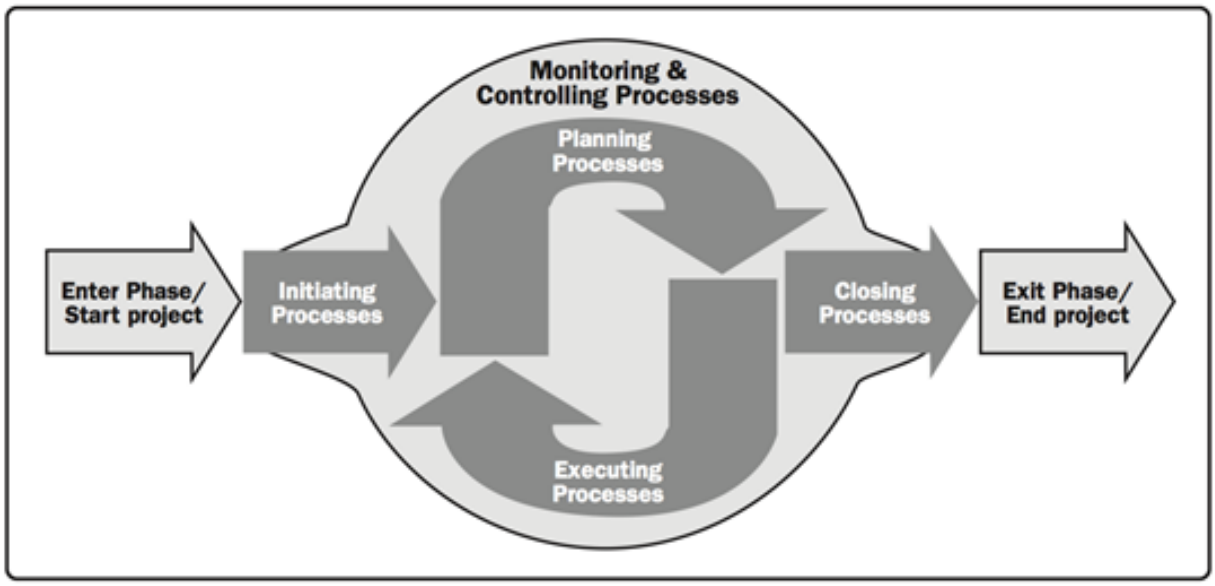

Figure 2: PMBOK Process Groups, as proposed in the PMBOK Guide[17]

2.2 OAIS and FAIR Digital Objects in Data Management Planning

The project phases of the PMBOK have been applied to the Open Archival Information System (OAIS) Information Model. The OAIS Information Model,[18] which can be seen as an earlier conceptualisation of a FAIR Digital Object (FDO), is a model which defines information that is required to make data findable, (re-)usable, and intelligible over a long period of time. Figure 3 below gives an overview of the OAIS Information Model. In the next section, DMP-relevant information in this model and its relation to project phases is shown.

[17] Ibid.
[18] Consultative Committee for Space Data Systems (CCSDS) (2012). Reference model for an Open Archival Information System. https://public.ccsds.org/pubs/650x0m2.pdf

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

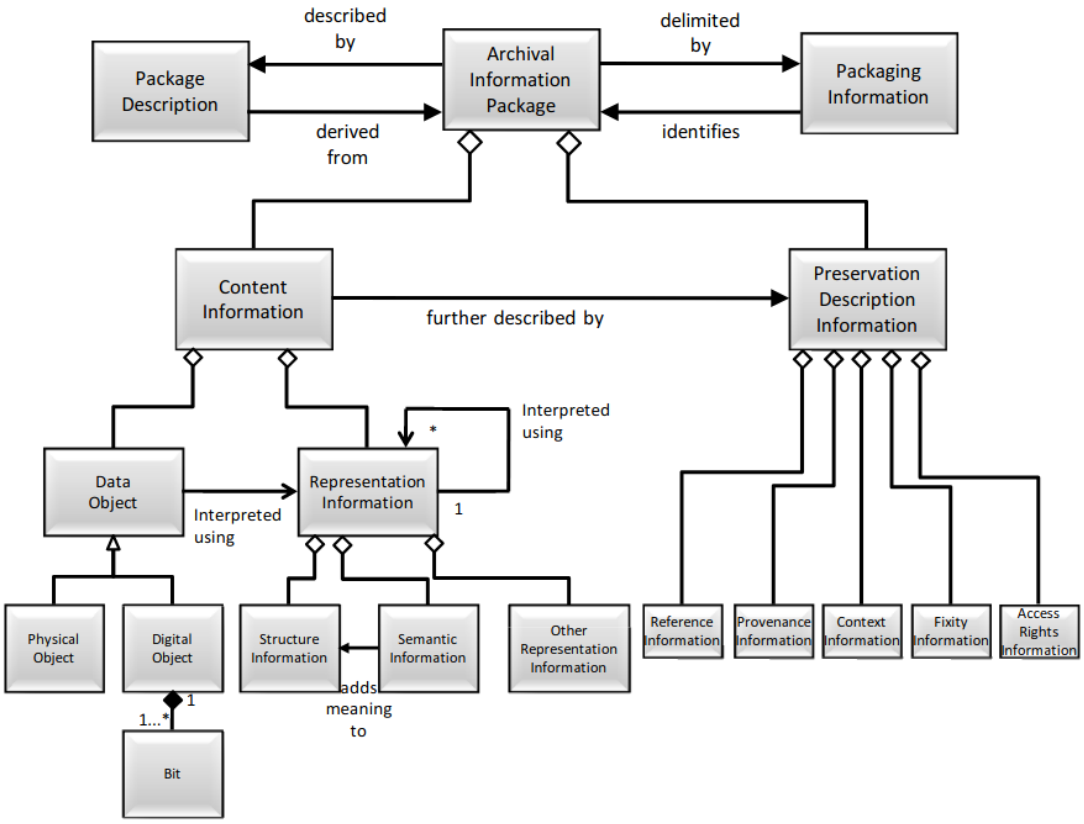

Figure 3: OAIS Information Model[19]

The scope of the OAIS Information Model (see Figure 3) and the scope of DMPs have considerable overlap. Typical DMP themes have been analysed.[20] Out of the scope of the OAIS Information Model are questions, e.g. relating to schedules, communication and costs, or if there is existing data to reuse.

The OAIS Information Model describes the packaging of data together with sufficient contextual information to support its long term preservation and reuse, potentially indefinitely into the future. The main components of the model include:

- **Data Object**: the data object represents the dataset without any additional information allowing interpretation of what the dataset is about and where it comes from. In the context of DMPs, it can be related to data volumes (expected and actual), data types, quality control, preservation and backup.
- **Representation Information**: the representation information includes all information that makes the data usable without giving the context of data production. In the representation

[19] Ibid.
[20] DCC (2016). Proposed revised set of themes for Data Management Plans. https://www.dcc.ac.uk/sites/default/files/documents/publications/DMP-themes-revised-Sept16.pdf

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

information, the structure of the data object, like formats and semantics as well as applicable metadata standards, are registered. Here also information of required soft- or hardware to use the data is described. Also original or other related data objects can be referred to here.

- **Reference Information**: the reference information is mainly about the usage of persistent identifiers.
- **Provenance Information**: the provenance information is about how the data object was created (hardware/software), and responsible persons (creation and each other processing step), which processes have been run on them and also about calibrations executed prior to the data object creation.
- **Access Rights Information**: the access rights information is about who has access to the data objects.
- **Context Information**: the context information can e.g. refer to the overlaying research project and the context the data object has been created in and the reason a data object has been created.

Thus, the OAIS Information model can be compared to the FDO, as described in FAIR Digital Object Framework (FDOF) Documentation,[21] and reproduced in Figure 4 below.

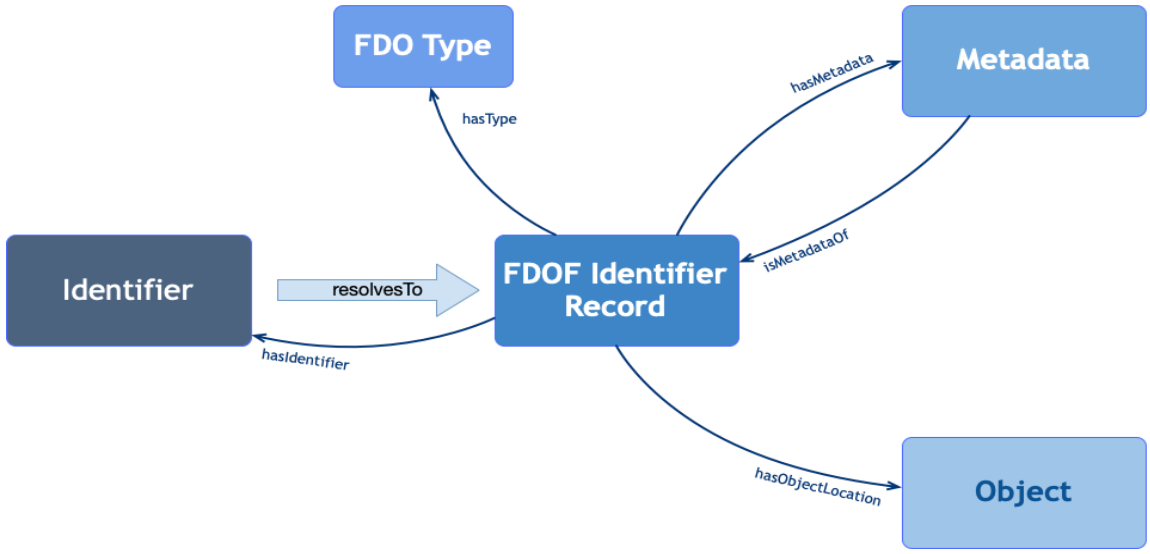

Figure 4: FAIR Digital Object (FDO), as proposed in the FAIR Digital Object Framework (FDOF)[22]

[21] Bonino da Silva Santos, L. O. (2021). FAIR Digital Object Framework Documentation. https://fairdigitalobjectframework.org/
[22] Ibid.

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

Thus, the Identifier corresponds to the OAIS Information Model reference information; the FDOF Identifier Record, the Content Information; the Metadata and FDO Type, the Representation Information; and the Object, i.e. the OAIS Information Model Data Object itself. The OAIS model provides additional information on the preservation environment and actions necessary for long-term preservation.

Thus, if within a data management planning process, we seek to collect sufficient information to build an Archival Information package, we should be able to provide a data package which is FAIR and preservable for the long term.

2.3 DMP Scope in PMBOK Project Phases

In this section, the DMP scope of the OAIS Information Model is applied to the PMBOK project phases using data arising from space science (see Tables 1 – 5 below).[23]

For each component of the OAIS model, the typical information collected to support each item is outlined along with the level of detail which can be collected at each stage of the PMBOK process. Thus, for example, for Content Information, one of the items collected is Data volume. This may be only roughly estimated initially, with better estimates being formed during the experimental planning phase. The actual data volume can then only be given during the Execution phase; however, it should be noted that the planning estimate from the DMP is already of value as it allows storage allocation to be carried in readiness for the experiment.

Additional Information Topic	Detailed area	Initiating	Planning	Executing	Closing
Content Data	Inventory of data produced/ expected	Rough idea	Increasingly detailed	Becoming complete	Complete
	Types of data (raw, processed, etc.) which should be preserved?	Rough idea	Increasingly detailed	Becoming complete	Complete
	Type of data e.g. images, tables – which generic interfaces?	Rough idea	Increasingly detailed	Becoming complete	Complete
	Volume that would require preservation	Rough idea	Increasingly detailed	Becoming complete	Complete
	Quality constraints	Rough idea	Increasingly detailed	Becoming complete	Complete
	Quality checks which may be performed on the data by non-experts	Rough idea	Increasingly detailed	Increasingly detailed	Complete

Table 1: PMBOK and Content Data[24]

[23] Giaretta, D., Glaves, H.M., and Shiers, J. (2016). Active Data Management in Space. https://indico.cern.ch/event/520120/contributions/2171073/attachments/1299260/1938730/Active_Data_Management_in_Space.pdf

[24] Ibid.

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

Additional Information Topic	Detailed area	Initiating	Planning	Executing	Closing
Representation Information	Choice of data format	Rough idea	Increasingly detailed	Becoming complete	Complete
	Format definitions and formal descriptions	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
	Semantics of the data elements	Rough idea	Increasingly detailed	Becoming complete	Almost complete
	Data dictionaries and other semantics	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
	Information Model	Rough idea	Increasingly detailed	Becoming complete	Complete
	Other Data Documentation	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
	Applicable standards	Rough idea	Increasingly detailed	Becoming complete	Complete
	Hardware and Software Dependencies	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
	Other software which may be used on the data		Increasingly detailed	Increasingly detailed	Growing
	Calibration and system test tools and system test data that will be delivered.	Rough idea	Increasingly detailed	Becoming complete	Up to date and accumulating
	Relationships between data items	Rough idea	Increasingly detailed	Complete	Complete

Table 2: PMBOK and Representation Information[25]

Additional Information Topic	Detailed area	Initiating	Planning	Executing	Closing
Reference Information	DOI or other unique identifiers	Rough idea	Becoming complete	Up to date and accumulating	Up to date and accumulating; New methods could be introduced
	Rules, methods, tools for referencing data	Rough idea	Becoming complete	Up to date and accumulating	Up to date and accumulating; New methods could be introduced
	What standards will be used to ~~format,~~ identify and reference the data and metadata	Rough idea	Becoming complete	Up to date and accumulating	Up to date and accumulating; New methods could be introduced
	What may be used in future to identify the Information	Fairly firm	Increasingly detailed	Increasingly detailed	Evolving

Table 3: PMBOK and Reference Information[26]

[25] Ibid.
[26] Ibid.

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

Additional Information Topic	Detailed area	Initiating	Planning	Executing	Closing
Provenance Information	Record of origins of the project e.g. in a CRIS system	Fairly firm	Complete	Completed	Complete
	Documentation about the hardware and software used to create the data, including a history of the changes in these over time		Rough Idea then Increasingly detailed	Becoming complete	Up to date and accumulating
	Processing workflow	Rough idea	Increasingly detailed	Becoming complete	Complete
	Processing inputs		Rough Idea then Increasingly detailed	Becoming complete	Complete
	Processing parameters	Rough idea	Increasingly detailed	Becoming complete	Complete
	Who was responsible for each stage of processing		Increasingly detailed	Becoming complete	Complete
	When each stage was performed		Increasingly detailed	Becoming complete	Complete
	Record of any special hardware needed	Rough idea	Increasingly detailed	Becoming complete	Complete
	Calibration	Rough idea	Becoming complete	Complete	Complete
	System Testing	Rough idea	Becoming complete	Up to date and accumulating	Up to date and accumulating; New methods could be introduced
	Resident Archives			Rough idea	Becoming complete
	Who was responsible for each stage of processing (Fixity)		Up to date and accumulating	Up to date and accumulating	Up to date and accumulating

Table 4: PMBOK and Provenance Information[27]

Additional Information Topic	Detailed area	Initiating	Planning	Executing	Closing
Issues Outside the Information Model	Schedule of deliveries	Fairly firm	Increasingly detailed	Complete	
	Cost	Fairly firm	Increasingly detailed	Complete, but may Evolve	Complete, but may Evolve
	Pointers to the components to be transferred to the archive		Fairly firm	Complete	Complete, but may Evolve
	Potential preservation aims of the archive	Rough idea	Increasingly detailed	Increasingly detailed	Complete
	Potential risks to preservation and exploitation of the data	Fairly firm	Increasingly detailed	Complete, but may Evolve	Complete, but may Evolve
	The target archives and designated community for the solicitation.	Fairly firm	Complete	Complete, but may Evolve	Complete, but may Evolve
	The budget for archiving.	Fairly firm	Complete	Complete, but may Evolve	Complete, but may Evolve
	The schedule for major project milestones and deliveries to the archive.	Fairly firm	Complete	Complete, but may Evolve	Complete, but may Evolve
	Change Management		Complete	Complete, but may Evolve	Complete, but may Evolve
	The mechanism for communication between project and archive.	Fairly firm	Complete	Complete, but may Evolve	Complete, but may Evolve

Table 5: PMBOK and Information outside the OAIS Information Model[28]

A similar analysis of the information collected and used in a facility's experimental lifecycle can demonstrate the stages that it can be estimated, refined, completed, and used.

[27] Ibid.
[28] Ibid.

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

2.4 Sources for DMP Information

Another analysis carried out by Miksa et al. (2017) considers the questions used to develop DMPs and possible sources for their answers.[29] Tools and models in the experimental lifecycle that can provide information related to the DMP questions have been mapped to the Digital Curation Centre (DCC) DMP Checklist v.4.0,[30] which provides guidance on the information that a DMP should contain, sub-divided into a number of information categories. In Figure 5 below (from Miksa et al. 2017), the middle columns contain questions and keywords that are based on the guidance to the questions in the DCC DMP Checklist.

[29] Miksa, T., Rauber, A., Ganguly, T. and Budroni, P. (2017). Information integration for machine actionable Data Management Plans, *International Journal of Digital Curation*, 12:1, 22 – 35. https://doi.org/10.2218/ijdc.v12i1.529

[30] DCC (2013). Checklist for a Data Management Plan v. 4.0. https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

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

Table 1. Mapping of tools and models to sections of DCC DMP Checklist v4.

Section	Question	Keywords	Tools and Models
Administrative Data	ID, Funder, Grant Reference Number, PI / Researcher ID, Contact, Date of Last Update	administrative	**Tools:** directory services (LDAP, Active Directory) and ORCID for user information; DOI for DMPs; **Models:** FOAF, Dublin Core;
Data Collection	What data will you collect or create?	type, format, size	**Tools:** content profiling (FITS, DROID, C3PO, exiftool),risk registries (PRONOM); **Models:** PREMIS;
	How will the data be collected or created?	provenance, process description, versioning, naming convention	**Tools:** execution monitoring (CDE, PMF, reproZip), workflow engines (Taverna, Pegasus), virtualisation and containers (Docker, VBox), code repositories (GitHub); collaboration platforms (OSF), virtual environments (Jupyter); **Models:** PROV-O, OPM, Dublin Core, Context Model;
Documentation and Metadata	What documentation and metadata will accompany the data?	metadata, documentation	**Tools:** wikis (redmine, confluence, OSF, GitHub), readmes, generated documentation (nanopublications, javadoc), Docker file for Docker images; **Models:** domain specific standards (biosharing.org), Dublin Core;
Ethics and Legal Compliance	How will you manage any ethical issues?	ethics, access control	DMPs are awareness tool, text description needed (DMP Roadmap); **Models:** see access control and security;
	How will you manage copyright and Intellectual Property Rights (IPR) issues?	licenses, policies	**Tools:** EUDAT tool, PERICLES Policy editor; **Models:** Creative Commons Ontology[25]
Storage and backup	How will the data be stored and backed up during the research?	storage, backup	**Tools:** Institutional storage and cloud services (ownCloud, data centres) **Models:** new developments needed

25 Creative Commons Ontology: https://www.w3.org/Submission/ccREL/

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

Section	Question	Keywords	Tools and Models
	How will you manage access and security?	access control, security	**Tools:** collaboration platforms, directory services; **Models:** Basic Access Control ontology[26]
Selection and Preservation	Which data should be retained, shared, and/or preserved?	preservation planning	**Tools and Models:** preservation planning (Plato and SCAPE project)
	What is the long-term preservation plan for the dataset?	repository, costs	**Tools:** repositories (re3data, Zenodo, Phaidra); costing tools (CCEx); **Models:** PREMIS
Data sharing	How will you share the data?	sharing, marketing, identifiers	**Tools:** repositories, data sharing platforms (GitHub, datahub, Zenodo), social media (twitter, Facebook), community portals (Researchgate, LinkedIn); **Models:** new developments needed;
	Are any restrictions on data sharing required?	embargo, legal regulations	**Tools:** DMPs are awareness tool, text description needed (DMP Roadmap); **Models:** PREMIS, Publishing Status Ontology[27];
Responsibilities and Resources	Who will be responsible for data management?	roles	**Tools:** directory services; **Models:** FOAF, Dublin Core;
	What resources will you require to deliver your plan?	resources	DMPs are awareness tool, text description needed (DMP Roadmap), applies mostly to initial DMPs.

Figure 5: Mapping of tools and models to sections of DCC DMP Checklist v4 (from Miksa et al. 2017)[31]

[31] Miksa, T., Rauber, A., Ganguly, T. and Budroni, P. (2017). Information integration for machine actionable Data Management Plans, *International Journal of Digital Curation*, 12:1, 22 – 35. https://doi.org/10.2218/ijdc.v12i1.529

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

2.5 Lessons to Be Drawn from the Literature

As we have seen, prior work on DMPs gives a detailed approach to developing DMPs which we can adapt to the PaN facilities context. Further, in developing an approach to DMPs for PaN RIs, we can draw some general lessons from the literature:

1. The information collected within the experimental DMP should be sufficient to support FAIR and preservable data. The OAIS model gives an abstract framework for the information which should be collected within the DMP; however, this should be interpreted and instantiated within the context of the PaN experimental lifecycle.
2. The DMP should be actively developed and used within the experimental lifecycle. Thus, it is not expected to be complete at proposal or planning phase, but rather is an active component of the data and computational environment in support of the experimental process. The information it records should thus be refined and added to, and also contribute to the configuration of the computational infrastructure at later phases of the experimental lifecycle.
3. The DMP is a collective responsibility of the stakeholders of the experiment, not just the PI.
4. The DMP should be integrated into the information environment of the facility, linked to the computational and information systems which support the experiment, automatically recording and sharing information recorded for one purpose down the lifecycle.

3. PaN RI Roles, Systems, and Workflows Relevant to the Creation of DMPs

In ExPaNDS deliverable *D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management* (2020)[32] and in the PaN-data Open Data Infrastructure (PaN-data ODI) Deliverable *D6.1: Model of the Data Continuum in Photon and Neutron Facilities* (2012)[33] relevant roles/actors, IT systems, and related workflows are described. The following sections give a short summary of these two deliverables where relevant for the creation of DMPs.

These two deliverables respectively define and refine the Data Continuum, which formalises the series of steps during which data and associated metadata are created. These steps, in their refined form are: **proposal**, **approval**, **scheduling**, **experiment**, **storage**, **data processing**, **data analysis, data record,** and, finally, **publication**. Note that these processes do not follow a strict sequential scheme since some of them can be performed in parallel (e.g. raw data storage and processing). Also note that the data processing and analysis steps generate new data by themselves, which should in turn undergo proper annotation and storage.

[32] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). D2.2: Draft recommendations for FAIR photon and neutron data management. https://doi.org/10.5281/zenodo.4312825
[33] 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.

As mentioned before, the DMP is not a static document but will rather evolve along with the completion of the different steps forming the Data Continuum, thereby helping the different stakeholders to keep a clear idea about how their scientific data should be disseminated. There are several use cases for the exploitation of these data outlined in Chapter Four of ExPaNDS D2.2. All of them can be grouped into the three categories listed below in section 3.1. We also review the IT systems and workflows associated with these categories in sections 3.2 and 3.3.

3.1 Roles

There are two main types of roles that interact with data through the Data Continuum. These roles are not always present in each stage of the lifecycle and will depend on the step we are actually analysing. In all the use cases that we previously analysed in ExPaNDS D2.2, we identified the following generic roles:

- **Users**
  - **Co-Investigators** are the primary members of the whole Experimental Team. Only one person is identified as a Principal Investigator.
  - The **Principal Investigator (PI)**, who is the main contact of the Experimental Team, works 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.
  - The **Experimental Team** finishes the project and prepares papers to be published in a journal.

- **Facility staff**
  - The **User Office** is responsible for formal review via the **Approval Panel**, which is usually made up of external scientific experts who provide relevant feedback and assessment on how the experiment shall be performed. At the end, the User Office will accept the Experimental Report and record the association of a paper with an experiment.
  - **Instrument Scientists** contribute as part of the feasibility review of the proposal (aka Technical Review) and help the experimental team during the experiment with the set-up, software, calibration, etc. They also arrange the data to take them off site. Instrument Scientists very often appear as co-authors of the papers to be published.
  - **Facility Operations** (e.g. Acquisition Systems Staff) who facilitate the collection of the data and metadata.
  - The **Data Infrastructure Team** manages the data storage, transfer, and publication process.
  - The **Library Service**, or **Library Team**, will lodge a metadata record and an appropriate copy of the publication. **Data Managers** validate the integrity of the data (e.g. curation,

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

custody). This refers to the facility's **Research Data Management (RDM) Team**, which ensures compliance with the FAIR principles and the data policy along the data lifecycle. The RDM team has therefore to make sure that the experimental data is properly annotated (i.e. with metadata) and that a consistent mechanism of persistent identification (e.g. PIDs) is in place for long lasting reference, archiving, and findability via external search APIs.

3.2 IT Systems

Today, information systems are crucial to achieving the objective of any data management endeavour. In this context, we identify the IT systems that map to the roles described in the previous section, keeping in mind that a particular system can map to more than one role.

In the context of IT systems, the data production role is necessarily assumed by the **data acquisition system** of the instrument under consideration. Furthermore, this system also acts as a metadata producer, like the **proposal and scheduling systems** (chronologically, being the first to be used, the proposal system acts as an ideal entry point for initiating the DMP).

Other systems like the **data processing and analysis pipelines** act both as data consumer and producer, whether they belong to the beamline software suite or constitute external programs in the context of later data reuse.

Last but not least, data management relies on a stable **storage system** and a **data catalogue** that contains all the metadata required for findability, preservation, and reuse, as well as licence and authorship information, which should be included in the DMP. Whenever processed data are made available through the catalogue, their provenance is crucial information that needs to be recorded from the corresponding **data processing and analysis pipelines**. Finally, findability is ensured by a **PID minting system** and related metadata schemas and various **external metadata discovery services** such as B2Find,[34] which, if used, should be mentioned in the DMP.

3.3 Workflows

In PaN-data ODI deliverable *D6.1: Model of the Data Continuum in Photon and Neutron Facilities* (2012),[35] a simplified and idealised view of the stages of the science lifecycle within a single facility is introduced as shown in Figure 6 below. The lifecycle starts with three planning stages: Proposal, Approval and Scheduling. During these early stages, information such as motivation for the experiment, planned techniques, intended instruments (along with information related to those instruments), and samples, as well as the people involved such as PI, co-proposer, and the experimental team are known. Getting closer to the experiment, information about the datasets to be produced becomes clearer in each step of the lifecycle.

During the experiment, the actual datasets are created. In this stage, information such as data volumes and applied configurations becomes clearer. After the creation of a dataset, normally the

[34] See http://b2find.eudat.eu/
[35] 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.

knowledge about its content is complete. In the data analysis step(s), the datasets created during the experiment are used and new ones are created. In order to do this, information about the created datasets, e.g. software (including versions) and hardware requirements, needs to be known.

Before the dataset is stored, aggregation, cleansing, and/or other processes such as data reduction are performed on the datasets. Then, the datasets are stored, preferably in the facility's data repository.

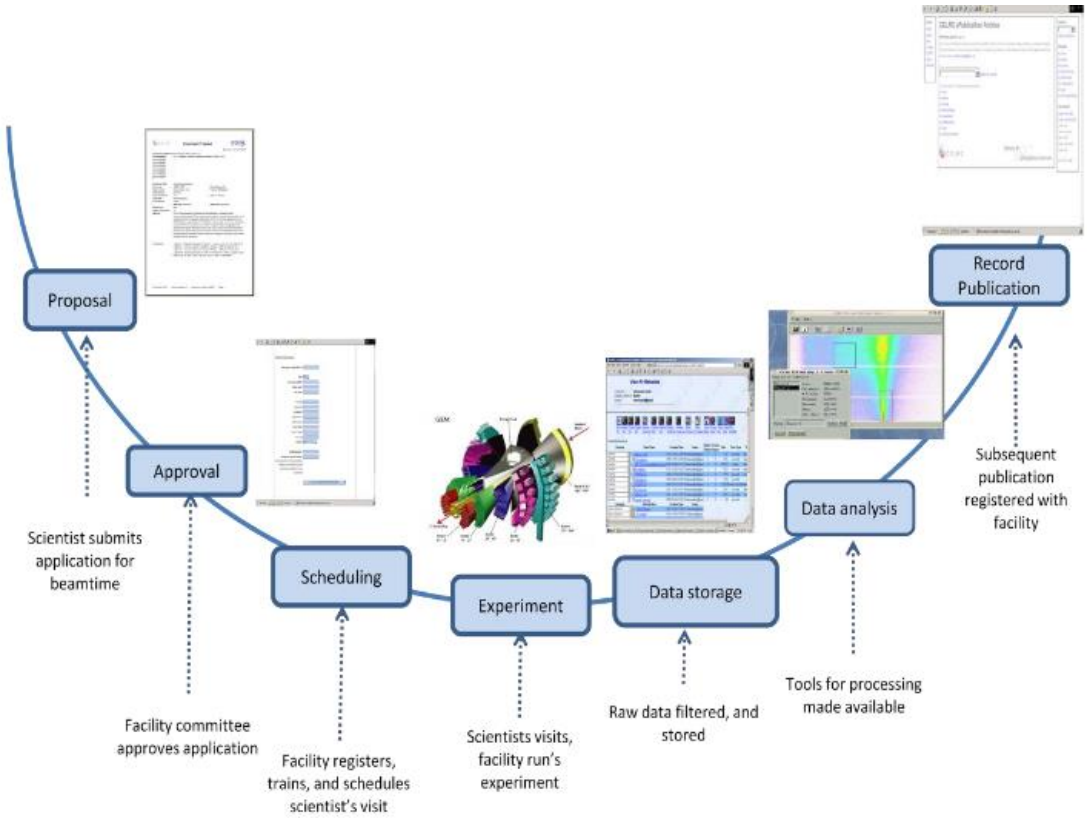

Figure 6: The idealised experimental lifecycle within facilities science (from PaN-data ODI D6.1)[36]

As discussed above, at each stage of the lifecycle, different information is collected and added to make a FAIR experimental dataset, and a set of IT systems are used as illustrated in Figure 7 below. In the figure, the relevant IT systems (green) are mapped to the phases in the research lifecycle and the upcoming information.

[36] Ibid.

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

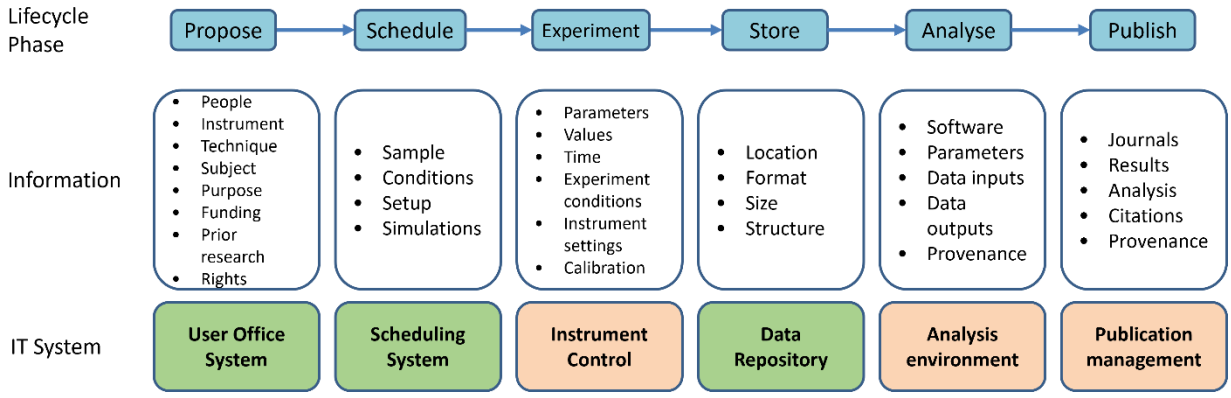

Figure 7: Metadata collected and information systems in the stages of the experimental lifecycle[37]

3.4 Relating DMP Knowledge to Roles in the Scientific Workflow

In section 3.1, facility roles as identified in ExPaNDS *D2.2: Draft Recommendations for FAIR Photon and Neutron Data Management* (2020)[38] were described. In this section, they are related to knowledge required in DMPs.

In a research project, the **PI** is the main person responsible for the creation of a DMP. She/He submits a proposal in the proposal system and, thereby, provides the first relevant information, such as the title and abstract. The PI also knows the overall research project that the proposal belongs to and likely knows the research domain, the measurement techniques required, and has an idea about the desired instrument and first sample information. The PI should also know about legal and contractual requirements of the research project.

The **Experimental Team** plan and execute the experiment, and they bring the sample to the facility. They can give a first estimate on how many datasets will be produced. They also know which software will be used. Together with the instrument scientist, they can give the most detailed description of the datasets to be produced. The experimental users also know if external storage and archival solutions are/will be used that are independent from the facility. The experimental team and the PI can be subsumed as **Users**.

The **Facility Admin**, and here most importantly, the user office, support the PI in submitting the proposal. The user office is in charge of the proposal system and the scheduling of the beam time.

[37] Matthews, B. (n.d.). Experimental Workflow.
[38] Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). D2.2: Draft recommendations for FAIR photon and neutron data management. https://doi.org/10.5281/zenodo.4312825

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

The **Instrument Scientists** can give information about data produced at the selected instrument. They know about data formats, volumes, creation methods, writing and reading software, and hardware requirements created at a certain instrument. The instrument scientists know about the organisation of files, the implemented quality control measurements, and they know what information is relevant to reuse the data and how the information is collected.

The **RDM Team** is a new and upcoming role in facilities. Its members are very often part of the IT team and work closely together with the libraries. The RDM Team has an overview about relevant standards and formats. They also know about data policies and RDM requirements in the wider PaN community. The RDM Team also knows about available RDM and some computing infrastructure at the facility. They can support the instrument scientists in compiling relevant information for creating the DMP, and finally, the FDO.

4. PaN DMP Template and Sources for Answers

After having described the roles, IT systems, and workflows in PaN RIs, the DMP questions are categorised by when the answer is known and by whom the answer is known.

4.1 The Four DMP Phases

According to the described workflows in the literature, four DMP phases can be defined (see Table 7 further below). Three of them, initiating, planning, and executing, correspond to the PMBOK Phases which can be mapped to some phases of the above described experimental lifecycle (see Table 6 directly below).

initiating	proposal submission
planning	accepted, experiment planning
executing	data collection / processing and analysis

Table 6: Mapping PMBOK to the experimental lifecycle

The first phase (Phase 0, i.e. see Table 7 below) starts before the project or proposal starts. It is outside of the project/proposals workflow and is determined by the facilities infrastructure and workflows/personnel. It belongs to the environment in which the datasets are produced such as instruments, hardware, and software. The DMP relevant information lies with the facility personnel responsible for the environment such as instrument scientists and, for RDM-related information, the RDM team. Very often the information known in Phase 0 prepares an option the user has to select during the research lifecycle, mainly in the proposal phase and in the planning before the experiment. In Table 8 (see section 4.3), supporting IT systems for this phase would be the RDM database and a vocabulary service. Both IT systems are normally not present in the facilities. The RDM database would hold all static RDM information about instruments and data produced by

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

instruments in a facility. The vocabulary service would be a registry of relevant metadata schemas and vocabularies. See, for example, the ontologies developed by ExPaNDS.[39]

When the proposal is submitted in Phase 1, the project/proposal is initiated, proposal information such as motivation of the research, requirements to execute the experiment, and administrative information are ingested, whereby the knowledge lies with the User submitting the proposal.

Phase 2 corresponds to the PMBOK planning phase. Here, the proposal has been accepted, the beamtime is scheduled, and the DMP information is improving as the instrument is now known and more concrete planning can be done.

During the experimental or execution phase in PMBOK Phase 3, information such as data volumes and more concrete dataset descriptions become clearer.

DMP phases	0 Before proposal submission	Typically knowledge of instrument scientist or RDM team (static parameter)
	1 Proposal submission	Typically knowledge of the user, with support by the facility administration and RDM team.
	2 Accepted experiment planning	Typically knowledge of the user, with support from the facility administration and instrument scientist.
	3 Data Collection / Data processing / analysis	Typically knowledge of the user, with support from the instrument scientist.

Table 7: DMP Phases

4.2 The Seven Sections of the RDMO Questionnaire

The RDMO questionnaire is split up into seven sections. Almost all sections are divided into a part concerning the whole project/proposal and a part considering just a dataset:

1. The **General/Topic** section is about the project/proposal, its motivation, funding, and project partners.
2. **Content Classification** is about dataset contents, origin, reuse options, and reproducibility.
3. **Technical Classification** is about collection dates, formats, sizes, creation tools and methods, and versioning.
4. **Data Usage** is about data usage of a dataset in the project, dataset organisation, storage and security, interoperability, sharing of the dataset and its collaborative use, and quality assurance. On a project/proposal level, it is about integration of the datasets and data management and data production costs.

[39] Collins, S. P., da Graça Ramos, S., Iyayi, D. et al. (2021). ExPaNDS ontologies v1.0. https://doi.org/10.5281/zenodo.4806026

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

5. **Metadata and Referencing** is about metadata requirements and standards, structure of datasets and referencing of other datasets in the project, PIDs, and overall metadata and PID costs in the project.
6. **Legal and Ethics** is about data protection, personal and other sensitive data, approvals, property rights (software) and other legal issues, and related costs.
7. **Storage and Long Term Preservation** is about selection of datasets to be archived, repositories, embargos, and related costs.

4.3 The RDMO Questionnaire Adapted as a PaN DMP Template

Table 8 below shows the questions of the PaN DMP questionnaire grouped by primary knowledge source and supporting sources (human and IT system) and the phase when the knowledge becomes clear/available.

The RDM database should hold static information for each instrument that is relevant for RDM.

Please note:

- The 'RDMO NR.' column does not include all questions from the RDMO template. Some non-relevant questions have been left out; hence, their numbers are also missing.
- An 'x' in the 'RDMO NR.' column means that question is an additional question, i.e. added to meet specific demands/needs of the PaN context.
- If the user is in parentheses, i.e. (User), this indicates that the user does not directly provide the information but has to take the final decision.
- Where entries in the 'Phase when clear' column have two numbers separated by a forward slash, i.e. (x/y), this indicates that there are options prepared in an earlier phase that have to be selected in a later phase.
- Where entries in the 'Phase when clear' column have two or more numbers separated by a comma/commas, i.e. (x, y, z), this indicates that information is accumulated during the phases.

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
1	What is the scientific motivation for the experiment?	General / Topic	User		Proposal system	1
2	Please provide a minimum of two keywords describing the experiment	General / Topic	User	Instrument scientist (typical keywords for instrument)/ RDM team	Proposal system/ RDM database Vocabulary Service (Science area key words)	1
3	What is the primary research area?	General / Topic	User	Instrument scientist (typical research fields for instrument) / RDM team	Proposal system/ RDM database/ Vocabulary Service (Scientific field descriptor)	1
4	When does the project start?	General / Topic	User		Manual input	1
5	When does the project end?	General / Topic	User		Manual input	1
6	Who submitted the proposal and is responsible for the project coordination?	General / Topic	User		Proposal system	1

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
7	Who are the institutional experiment partners?	General / Topic	User		Proposal system?	1
8	Provide institutional policy URL and relevant statements of the policy.	General / Topic	User	(RDM team)	Manual input or RDM database	2
9	Who is responsible for RDM in your institution?	General / Topic	User		Manual input	2
10	Is the proposal supported by externally funded research?	General / Topic	User		Proposal system	1
11	In which special funding programme is the project located?	General / Topic	User		Proposal system	1
12	Does the funder have rules or recommendations for data management? If so, provide the policy URL and relevant statements of the policy.	General / Topic	User	RDM team	Manual input or RDM database	1
13	Are there requirements regarding the data management from other parties (e.g. the scholarly/scientific community/publisher)?	General / Topic	(User) only if there are no common requirements	RDM team	RDM database	1
14	Which are these additional requirements regarding data management?	General / Topic	(User)	RDM team	RDM database	1

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
15	Please provide a short description of the dataset (e.g. Diffraction data taken as a function of temperature)	Content Classification / Datasets	User	RDM team / Instrument scientist	RDM database	2
16	Is the dataset experimental, processed, or analysed?	Content Classification / Datasets	User			2
17	If processed or analysed, who created the dataset?	Content Classification / Datasets	User			2
18	If processed or analysed, under which address, PID(s), or URL can the original data be found?	Content Classification / Datasets	User			2
19	Which individuals, groups or institutions could be interested in re-using this dataset? What are possible scenarios?	Content Classification / Datasets	(User)	RDM team / Instrument scientist	Proposal system (research area) / RDM database	1
20	Is the dataset reproducible in the sense that it could be created / collected anew in case it got lost?	Content Classification / Datasets	User (Know-ledge about sample)	Instrument scientist	RDM database	1
21	When does data collection or creation start?	Technical Classification / Data Collection		Facility admin	Scheduling system (beamtime start)	2

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
22	When does data collection or creation end?	Technical Classification / Data Collection		Facility admin	Scheduling system (beamtime end)	2
23	When does data cleansing / data preparation end?	Technical Classification / Data Collection		Facility admin	Scheduling system	2
24	When does data analysis start? This creates another dataset and we need to know when the data needs to be available.	Technical Classification / Data Collection		Facility admin	Scheduling system	2
25	When is analysis infrastructure required?	Technical Classification / Data Collection		Facility admin	Scheduling system	2
26	What is the expected size of the raw dataset?	Technical Classification / Data Collection		Facility admin	Scheduling system	2
27a	Estimate the raw data volume generated from the proposal. Number of hours of instrument multiplied by data rate per hour	Technical Classification / Data Collection		Instrument scientist	RDM database	2
27b	Actual size of data collected during experiment	Technical Classification / Data Collection		RDM team	Data repository	3

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
28	How many datasets were created?	Technical Classification / Data Collection		Instrument scientist	RDM database + Scheduling system	2
29	Specify the file formats used for the raw data (default NeXus)	Technical Classification / Data Collection		Instrument scientist	RDM database / Data repository	2
30	Which instruments/hardware, tools, software, technologies or processes are used to generate or collect the data?	Technical Classification / Data Collection		Instrument scientist	RDM database	2
31	Which software, processes or technologies are necessary to use the data?	Technical Classification / Data Collection		Instrument scientist	RDM database	2
32	Is documentation about relevant software needed to use the data?	Technical Classification / Data Collection		Instrument scientist	RDM database	1
36	How / for what purpose will this dataset be used during the project?	Data Usage / Usage Scenarios	User			1
38	Estimation of infrastructure needs select from list	Data Usage / Usage Scenarios	User	Instrument scientist	RDM database	1

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
42	Are there internal project guidelines for a consistent organisation of the data? If so, where are they documented? Divide in instrument specific and project specific	Data Usage / Usage Scenarios	(User)	Instrument scientist	RDM database	1/0
43	Is there an internal project guideline for naming the data? If so, please briefly outline the naming conventions and, if necessary, link to the documentation. Divide in instrument specific and project specific	Data Usage / Usage Scenarios	(User)	Instrument scientist	RDM database	1/0
44	Who is allowed to access the dataset?	Data Usage / Usage Scenarios	User		Proposal system (Data policy)	0,1,2,3
45	How and how often will backups of the data be created?	Data Usage / Usage Scenarios		RDM team / Instrument scientist	RDM database	0
46	Who is responsible for the backups?	Data Usage / Usage Scenarios		RDM team / Instrument scientist	RDM database	0

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
47	Which measures or provisions are in place to ensure data security (e.g. protection against unauthorized access, data recovery, transfer of sensitive data)?	Data Usage / Usage Scenarios		RDM team / Instrument scientist	RDM database	0
48	Is this dataset interoperable, i.e. allowing data exchange and re-use between researchers, institutions, organisations, countries etc.?	Data Usage / Usage Scenarios		RDM team / Instrument scientist	RDM database	0
49	Where will this dataset be published or shared?	Data Usage / Usage Scenarios		RDM team / Instrument scientist	RDM database	0
51	What license will be applied to the dataset?	Data Usage / Usage Scenarios	User	RDM team	RDM database	1/0
53	When will the data be published (if they are)?	Data Usage / Usage Scenarios	User		Manually	1
57	Which measures of quality assurance are taken for this dataset?	Data Usage / Usage Scenarios	(User)	RDM team / Instrument scientist	RDM database	2/0

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
58	Is the integration between auxiliary information, measured and processed datasets ensured? If yes, by which means?	Data Usage / Usage Scenarios	(User)	RDM team / Instrument scientist	RDM database	2/0
64	What is the amount of non-personnel costs associated with the storage of the data sets during the project?	Data Usage / Usage Scenarios		RDM team / Instrument scientist	RDM database	2/0
65	Which information is necessary for other parties to understand the data (that is, to understand their collection or creation, analysis, and research results obtained on its basis) and to re-use it?	Metadata and Referencing	User	RDM team / Instrument scientist	RDM database	2/0
66	Which standards, ontologies, classifications etc. are used to describe the data and context information?	Metadata and Referencing	(User)	RDM team / Instrument scientist	RDM database	2/0
67	Describe automatically generated metadata from experiment	Metadata and Referencing		Instrument scientist	RDM database	0
x	Describe which metadata will be automatically collected during the experiment, which is essential for analysis of the data set. (e.g. Sample temperature)	Metadata and Referencing		Instrument scientist	RDM database	0

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
68	In case it is unavoidable that you use uncommon or generate instrument/project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?	Metadata and Referencing	(User)	RDM team / Instrument scientist	RDM database	(1)/0
69	Which metadata are collected semi-automatically?	Metadata and Referencing		RDM team / Instrument scientist	RDM database	0
70	Describe which metadata will be manually added to the dataset.	Metadata and Referencing	(User)	Instrument scientist	RDM database	1/0
x	Provide description, link or DOI relevant to the metadata for data processing and data analysis. E.g. data processing script / input file	Metadata and Referencing	User	Instrument scientist	RDM database	0,2,3
x	Describe which Auxiliary data will be added to the data set for processing and analysis.	Metadata and Referencing	(User)	Instrument scientist	RDM database	0,2,3
x	Which calibration data sets have been used?	Metadata and Referencing	User	Instrument scientist		3
x	Which sample characterisation data will be added?	Metadata and Referencing	User			2

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
71	Provide schema description and validation process	Metadata and Referencing		RDM team / Instrument scientist	RDM database	0
72	Who is responsible for documenting the metadata and context information and for checking if they are correct and complete?	Metadata and Referencing	User		Manually	
75	What is the structure of the data? How are the individual components of the dataset related to each other? How is the dataset related to other datasets used in the project?	Metadata and Referencing		RDM team / Instrument scientist	RDM database	0
76	Will persistent identifiers (PIDs) be used for this data set?	Metadata and Referencing		RDM team	RDM database	0
77	Which system of persistent identifiers shall be used?	Metadata and Referencing		RDM team	RDM database	0
79	Who is responsible for the maintenance of the PIDs and the object maintenance (i.e. who is responsible for notifying the PID-Service about object relocation and the new address)?	Metadata and Referencing		RDM team	RDM database	0
83	Does this dataset contain personal data?	Legal and Ethics	User		Manually / RDM database	1/0

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
84	Does this dataset contain sensitive data other than personal data?	Legal and Ethics	User		Manually	1
90	Has the project been approved by a research ethics committee?	Legal and Ethics	User			
95	Does the project use and/or produce data that is protected by intellectual or industrial property rights?	Legal and Ethics	User			
101	What are the criteria / rules for the selection of the data to be archived (after the end of the project)?	Storage and long-term preservation	(User)	Instrument scientist	RDM database	1/0
102	Who selects the data to be archived?	Storage and long-term preservation	(User)	Instrument scientist	RDM database	1/0
103	Does this dataset have to be preserved for the long-term?	Storage and long-term preservation	(User)	RDM team / Instrument scientist	RDM database	0
104	What are the reasons this dataset has to be preserved for the long-term?	Storage and long-term preservation	(User)	RDM team / Instrument scientist	RDM database	0
105	What is the minimum period that the data will be stored?	Storage and long-term preservation		RDM team	RDM database (data policy)	0

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
107	Where will the data (including metadata, documentation and, if applicable, relevant code) be stored or archived after the end of the project?	Storage and long-term preservation	(User)	RDM team	RDM database	0
108	Is the repository or data centre chosen certified (e.g. Data Seal of Approval, nestor Seal or ISO 16363)? (If the dataset is archived at several places, you may answer this question with yes, if this applies to at least one of these.)	Storage and long-term preservation		RDM team	RDM database	0
109	Have you explored appropriate arrangements with the identified repository?	Storage and long-term preservation		RDM team	RDM database	0
110a	Shall there be an embargo period before the data is open access?	Storage and long-term preservation		RDM team	RDM database	0
110b	How long is the embargo period?	Storage and long-term preservation				
111	How will the identity of the person accessing the data be ascertained?	Storage and long-term preservation		RDM team	RDM database	0

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

RDMO NR.	Questions	Section	Primary source	Supporting sources	IT sources	Phase when clear
112	By when will the data be archived?	Storage and long-term preservation		RDM team / Instrument scientist	RDM database	0
115	How will the data management costs of the project be covered?	Storage and long-term preservation		RDM team	RDM database	0

Table 8: DMP sources/phases mapping

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

The table above shows that, for many of the DMP related questions, answers can be given by the facility staff, either the RDM team, the instrument scientists, or, in some cases, the administrative team. Some information that relates to the planning phase (Phase 1, 2) of the project/proposal can be retrieved from the proposal/scheduling system. The majority of answers to the DMP questions are known in Phase 0, i.e. even before a proposal is submitted. The information for these answers is static and could be held in a database to support RDM. This information is sometimes about options given to the user, who has to choose (i.e. numbers separated by a forward slash). In a few cases, the information only arises in Phase 3, e.g. the actual amount of data collected, but also related to specific metadata and auxiliary data. Here, the information accumulates information previously added by the instrument scientist (i.e. numbers separated by commas).

5. Concluding Remarks

By relating DMP knowledge to roles in the scientific workflow and by using the PMBOK project phases to structure the information required in a DMP, the sources for answering a DMP can be identified. As Chapter Four makes clear, the information required to answer the questions in a DMP relies very much on information that is held in the organisational and technical infrastructure of a facility. A large amount of information is already available before Users submit proposals. This information is static and is available independently of a proposal/project. As such, the DMP could be largely prepared in advance by the facility staff, with the remainder of the DMP completed by the User or any other role that has the necessary knowledge. Thereby, the User could be significantly supported by the facility when creating a DMP.

In order to achieve this, the information generated in the various phases of the DMP lifecycle needs to be made available. At present, the required information resides primarily in the heads of facilities' staff or in facilities' documentation. In order to use the information to automatically fill in DMPs or in the RDM or research lifecycle, this information needs to be available in a machine-readable format.

In this deliverable, we have not investigated sources of information such as version control systems (e.g. Git), containers (e.g. Docker), VMs (e.g. VirtualBox), or notebooks (e.g. Jupyter) (i.e. as mentioned in Chapter Two). Also of note is that this deliverable refers to a RDM database, which has yet to be developed. Therefore, two major tasks can be foreseen as possible next steps. The first task would be a model of the RDM database, illustrating how it allows the capture and storage of the static information required for DMPs. The second task would be to model the integration of the various IT systems.


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

References

Bolmsten, F., Lobley, C., and Taylor, J. (2021). D2.2: DMP Template for facility users. https://doi.org/10.5281/zenodo.5639428

Bonino da Silva Santos, L. O. (2021). FAIR Digital Object Framework Documentation. https://fairdigitalobjectframework.org/

Collins, S. P., da Graça Ramos, S., Iyayi, D. et al. (2021). ExPaNDS ontologies v1.0. https://doi.org/10.5281/zenodo.4806026

Consultative Committee for Space Data Systems (CCSDS) (2012). Reference model for an Open Archival Information System. https://public.ccsds.org/pubs/650x0m2.pdf

DCC (2013). Checklist for a Data Management Plan v. 4.0. https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

DCC (2016). Proposed revised set of themes for Data Management Plans. https://www.dcc.ac.uk/sites/default/files/documents/publications/DMP-themes-revised-Sept16.pdf

DCC (2017). 12th International Digital Curation Conference: Workshops. https://www.dcc.ac.uk/events/idcc17/workshops#workshop1

Giaretta, D., Glaves, H.M., and Shiers, J. (2016). Active Data Management in Space. https://indico.cern.ch/event/520120/contributions/2171073/attachments/1299260/1938730/Active_Data_Management_in_Space.pdf

Matthews, B. (n.d.). Experimental Workflow.

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

Miksa, T., Rauber, A., Ganguly, T. and Budroni, P. (2017). Information integration for machine actionable Data Management Plans, *International Journal of Digital Curation*, 12:1, 22 – 35. https://doi.org/10.2218/ijdc.v12i1.529

Miksa, T., Simms, S., Mietchen, D. and Jones, S. (2019). Ten principles for machine-actionable data management plans, *PLOS Computational Biology*, 15:3, e1006750. https://doi.org/10.1371/journal.pcbi.1006750

Open AIRE (2017). What is a Data Management Plan and how do I create one. https://www.openaire.eu/what-isa-data-management-plan-and-how-do-i-create-one

Project Management Institute (PMI) (2013). PMBOK® Guide – Fifth Edition.

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

RDA Active Data Management Plans IG (2016). International and interdisciplinary workshop on Active Data Management Plans. https://indico.cern.ch/event/520120/

RDA Active Data Management Plans IG (2017). ADMP scenarios. https://rd-alliance.org/group/active-data-management-plans/wiki/admp-scenarios.html

RDMO (n.d.). RDMO Research Data Management Organiser. https://rdmorganiser.github.io/

Salvat, D., Gonzalez-Beltran, A., Görzig, H. et al. (2020). D2.2: Draft recommendations for FAIR photon and neutron data management. https://doi.org/10.5281/zenodo.4312825

Simms, S., Jones, S., Mietchen, D. and Miksa, T. (2017). Machine-actionable data management plans (maDMPs). https://riojournal.com/articles.php?id=13086

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