Active DMPs for Photon and Neutron RIs

Document Control Information

Settings	Value
Document Identifier:	D2.8 active DMPs for Photon and Neutron RIs
Project Title:	ExPaNDS
Work Package:	WP2
Document Author(s):	Heike Görzig (HZB), Vasily Bunakov (UKRI), Alejandra Gonzalez-Beltran (UKRI), Janusz Malka (EuXFEL), Brian Matthews (UKRI), Abigail McBirnie (UKRI), Nicolas Soler (ALBA), Noel Vizcaino (UKRI), Majid Ounsy (SOLEIL) Others: Marjolaine Bodin (ESRF), Fredrik Bolmsten (ESS), Andrei Vukolov (Elettra)
Document Reviewer(s):	Steve P. Collins (Diamond), Oliver Knodel (HZDR)
Responsible Partner:	HZB
Doc. Issue:	1.0
Dissemination level:	Public
Date:	20/12/2022

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

Abstract

This deliverable envisions an integrated aDMP system which supports the (semi-) automated generation of DMPs by maximising reuse of information generated during the research life-cycle and also aiming to support the scientists and other stakeholders by supporting information transfer. Therefore, a data model for aDMPs focusing on the integration of information that exists before a proposal is submitted and its life-cycle integration is introduced. It relates the data model also to the template created in PaNOSC.

Licence

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

Executive Summary

After having created a Data Management Plan (DMP) template for PaN sciences in (Bolmsten *et al.*, 2021) and analysed sources and DMP phases for collecting the answers to the questions in the DMP template, this deliverable aims to present a data model and an architecture for a possible active DMP implementation in PaN facilities.

This deliverable consists of four sections. The first section introduces the aims of this deliverable and relates it to other ExPaNDS and PaNOSC deliverables. Related documents in ExPaNDS and PaNOSC are especially mentioned where DMPs, Metadata, FAIR data, and data policies were discussed.

The second section presents aims and applications of active DMPs (aDMPs) in the sense of automation support for DMP creation as well as their integration in institutional IT infrastructure first. In the second part of this section aims for aDMPs in PaN facilities are presented as they have been discussed in the previous deliverable especially in the PaNOSC deliverable on the *DMP template for facility users* (Bolmsten *et al.*, 2021).

In the third section a data model for aDMPs focussing on the integration of information that exists before a proposal is submitted and its life-cycle integration is introduced. Relating the data model also to the template created in PaNOSC.

The fourth section aims to envision an integrated aDMP system which supports the (semi-) automated generation of DMPs by maximising reuse of information generated during the research life-cycle and also aiming to support the scientists and other stakeholders by supporting information transfer.

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

Table of Contents

Executive Summary **3**

1. Introduction **5**
  1.1 Aims of the Deliverable and Related Documents 5
  1.2 Aims of this Deliverable 5
  1.3 Links between this Deliverable and Other Work in ExPaNDS 5
  1.4 Overview of the Deliverable 7

2. Aims of “active” DMPs (aDMPs) **7**
  2.1 active DMPs State-of-the-Art 7
    2.1.1 Existing DMP tools 8
    2.1.2 Components and services for active DMPs 11
    2.1.3 Including active DMPs in infrastructure and workflows 13
  2.2 PaN Usage Scenarios 17

3. Data Model for aDMPs **18**
  3.1 Data Model for Workflows / DMP Execution 18
  3.2 Data Model for Pre-existing DMP Information 21

4. Suggested Architecture for active DMPs **26**
  4.1 Relevant IT Systems 27
    4.1.1 Proposal system 27
    4.1.2 Facility knowledge system 28
    4.1.3 Sample management system 31
    4.1.4 Instrument DAC system 31
    4.1.5 Data analysis environment system 33
    4.1.6 Facility storage system 34
    4.1.7 Facility librarian system 35
  4.2 Information Flow between the IT Systems 36

5. Concluding Remarks **37**

References **39**

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

Three deliverables on Data Management Plans (DMPs) are in the work plan of the EOSC projects PaNOSC and ExPaNDS, as already explained in the ExPaNDS deliverable D2.4. This document (D2.8) is the third and final document and the objectives of each document are summarised here:

1. PaNOSC *D2.2 - DMP Template for facility users* (Bolmsten *et al.*, 2021) proposes a PaN Data Management Plan (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) (Görzig *et al.*, 2021) 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 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 RI*s (Nov 2022) (i.e. the present document) builds 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). As well as integration with foreseen usage scenarios.

1.2 Aims of this Deliverable

The aim of this deliverable is to create a high level architecture for DMPs within facilities, including a view on IT systems and their data flow related to the creation, execution, and validation, as well as the creation of practical advantages of DMPs.

Based on this architecture, a future RDM system, its components and interfaces can be created and a roadmap for future possible implementations can be sketched.

1.3 Links between this Deliverable and Other Work in ExPaNDS

The previous DMP deliverables proposed a DMP template for PaN facility users, consisting of a knowledge base of more than 100 questions to be included within a DMP. Most of them are meant to be answered automatically via communication with the different systems present at facilities, with only a small core that a facility should select to put to users.

In parallel, ExPaNDS task 2.3 (deliverables D2.2 (Salvat *et al.*, 2020) and D2.7 (Soler *et al.*, 2022)) established a Common Metadata Framework for FAIR data in PaN facilities. This framework (see D2.7 P13) consists in a list of metadata fields ordered in a way that follows the traditional sequence of steps followed during a PaN experiment, i.e., from proposal to publication. The framework itself has been designed so as to confer a greater degree of FAIRness to raw and derived PaN data when they leave the facility. Some of the information requested in the framework can be provided at an early stage, and in fact could be directly extracted from a filled data management plan so as to directly feed the corresponding metadata records of the datasets acquired during the experiment. The compliance of these recommendations as described in the framework should be part of the achievements of data management planning. See below in the sections of "PaN usage scenarios".

It is also valuable to compare the DMP template for PaN facilities to a broadly-used DMP template used for scientific project funding such as the one developed for Horizon Europe (Horizon Europe, 2022). The Facility's DMP can then be used to populate the funder's DMP, reusing information thus reducing the burden on users and the potential for error. The nature of the questions posed in this DMP template, together with its organisation, are compared with the PaNOSC DMP template for PaN facility users (Bolmsten *et al.*, 2021) in the template table pp.15.

ExPaNDS, jointly with PaNOSC, has also undertaken work on developing a FAIR data policy framework for PaN RIs. The final version of the ExPaNDS data policy framework is published as part of deliverable *D2.3: Final data policy framework for Photon and Neutron RIs* (McBirnie *et al.*, 2021) (see also ExPaNDS deliverable D2.1 for the draft version of this framework (Matthews *et al.*, 2020)). In PaN RIs, the data policy describes the expectations and obligations on how research data in the facility should be handled.

As noted in ExPaNDS deliverable *D2.4: DMPs for Photon and Neutron RIs*, some of the DMP questions relate to information that would normally also feature in PaN data policies. License type and embargo period are examples of such information. Given this, PaN RI data policies form a framework on the 'static information' a facility holds that could be used repeatedly for the purposes of automatically populating DMP templates. This said, PaN facilities define their policies to differing levels of granularity, and they do not all include the same types of information. For example, the previous 2011 PaNdata data policy framework (Wilson *et al.*, 2011), which still forms the basis of some current facility data policies, addressed information such as licence, embargo period, and access rights. In turn, where facilities have drawn on more recent policy developments in PaNOSC and ExPaNDS, these facilities' data policies may be updated to include reference to FAIR and DMPs, although exactly how or if these concepts are incorporated directly and explicitly into the facility data policy itself (i.e. as opposed to supported through common practice and implementation at the facility) may differ across facilities.

Relevant Persistent Identifiers (PIDs) for PaN science and guidelines for facilities are described in deliverable D2.5 (Bunakov *et al.*, 2022). Section 5.14 of this deliverable introduced the PID graph as a method to contextualise facilities research if PIDs and their relationships are used more frequently. On a proposal or grant level the DMP is expected to be the first opportunity in the facility research lifecycle for creating such relationships. As here the proposal is first related to e.g., instruments, scientific techniques, and persons. The RDA DMP common standard (Miksa, Walk and Neish, 2020) discussed in Section 2 of this deliverable can be used as a source for building the PID graph, along with the record of funded research projects (grants), record of the facility organisation, facility instrument and sample, all bearing PIDs assigned to them. In return, a particular instance of a DMP can be enriched with the mentioned PIDs, including in an automated way through the content negotiation mechanism of PIDs resolution where possible.[1] One example of this would be populating DMPs with organisation-specific or researcher-specific metadata elements automatically retrieved via PIDs resolution. PIDs can be also used in photon and neutron sources community ontologies (Collins *et al.*, 2021), or exchangeable with ontology terms, to improve the quality of metadata for DMPs.

1.4 Overview of the Deliverable

This deliverable incorporates five main parts:

- The current section (Chapter One) introduces the aims of the deliverable and links it to other work in ExPaNDS.
- Chapter Two reviews the aims of “active” DMPs (aDMPs) in the literature and outlines some use cases related to facilities DMP systems.
- Based on the previous sections and the PaNOSC DMP template a data model for aDMPs is introduced.
- Chapter Four presents a high level architecture of a facility’s DMP system.
- Chapter Five concludes the report and sets out the next steps.

2. Aims of “active” DMPs (aDMPs)

2.1 active DMPs State-of-the-Art

While creating a DMP, researchers very often have their first contact with RDM concepts. A DMP is a tool to help researchers establish a plan of action to manage their data (Cardoso, Proença and Borbinha, 2020). Funding agencies are laying more and more emphasis on this tool and including it in the evaluation process of grant proposals. In parallel, many publications consider the bureaucratic burden of the creation of a DMP. In order to make the task of creating DMPs more beneficial to the researcher rather than a bureaucratic burden, many tools have been developed and the concept of the DMP has evolved.

Originally DMPs were static documents that needed to be delivered to the funding agencies. The initial DMP tools, for example, as e.g., DMPonline and (Sallans and Donnelly, 2012), were focussing on guidance rather than automation, reuse, or interoperability of the gathered information as well as integration in research workflows. However, it became clear very quickly that DMPs are living documents where the content changes and becomes clearer throughout the runtime of a project (see our previous deliverable “DMPs for Photon and Neutron RIs“ (Görzig *et al.*, 2021)). This new generation of tools, concepts, and workflows can be referred to as active DMPs, also referred to as machine-actionable or machine-readable, or dynamic DMPs (Simms *et al.*, 2017).

[1] DOI content negotiation can be employed, as per https://citation.crosscite.org/docs.html

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

An important step was the development of a machine-readable representation of a DMP. The DMP-Common-Standard (DCS) has been developed by the RDA DMP Common Standards Working Group (Miksa, Walk and Neish, 2020) and includes a JSON and ontology representation. Many tools for creating DMPs have integrated or are planning to integrate import and export functionalities for this standard, thereby making the DMP exchangeable between tools and organisations. Another usage of the DCS ontology is the machine-readability of the DMP, linking datasets to projects and eventually the machine-resolvable URL of their location. The standard itself is more and more integrated on different topics. In 2021, there was a hackathon on “Interconnecting systems using machine-actionable data management plans” based on the DCS (Cardoso, Castro and Miksa, 2021). The topics covered were:

- Format serialisation and Integration of DMP Tools: the participants included developers of the tools Data Stewardship Wizard, EasyDMP, RDMO and Argos/OpenDMP.
- Further integration: participating tools were “Converis CRIS/RIMS with DMPRoadmap, and establishing maDMP export/import from Figshare”, and InvenioRDM integration with DCS, and other integration in researchers’ workflows
- Funder template mapping considering various funder templates.

2.1.1 Existing DMP tools

In the following sub-section, some DMP tools will be listed. The table below has been extracted from (Jones *et al.*, 2020):

Tool name	Developing organisations
DMPonline	Digital Curation Centre (DCC)
DMPTool	California Digital Library (CDL)
EasyDMP	EUDAT & UNINETT Sigma2
Data Stewardship Wizard (DSW)	Dutch Techcentre for Life Sciences (DTL, ELIXIR NL) & Czech Technical University in Prague (CTU, ELIXIR CZ)
Research Data Management Organiser (RDMO)	Operated by many institutions – self-deploy model. Creators are Leibniz-Institut für Astrophysik Potsdam (AIP), Potsdam University of Applied Sciences (FHP) & Karlsruhe Institute of Technology (KIT)
Research Data Manager (UQRDM)	University of Queensland
DataWiz	Leibniz Institute for Psychology Information and Documentation (ZPID)

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

ezDMP	Interdisciplinary Earth Data Alliance (IEDA)
OpenDMP	OpenAIRE & EUDAT

Table 1: DMP Tools

The DSW and the RDMO have been analysed further in the context of ExPaNDS and PaNOSC. Both have a central knowledge base that provides mapping to questionnaires for users on one hand and views for stakeholders with information requirements on the other. The questionnaire of the RDMO has been the basis for the questionnaire developed in the PaNOSC *DMP template for facility users* (Bolmsten *et al.*, 2021).

In the following we describe the most appropriate tools for our concept in more detail.

Data Stewardship Wizard (DSW)

DSW claims in its documentation to have an emphasis on FAIR metrics and templating capabilities, over and above other DMP tools capabilities. Internally, it uses a PostgreSQL (database) and MinIO (object storage) centric backend.[2] It should be noted that a docker-compose version is provided for development/testing only (standalone self-deployment). This is a common occurrence in widespread institutional software packages (e.g. OKF CKAN) as many software deployments are catching up with using docker-compose based solutions. One of the DSW's strengths is a strong API supporting the integration with other services such as facilities proposal systems.

Research Data Management Organiser (RDMO)

RDMO allows the user to create custom DMPs based on projects and profiles. The tool is specialised in supporting the use of DMP in the full research lifecycle. A deployment is achieved via a series of docker containers[3] implementing the different networked services.
One of the RDMO's strengths is a strong community around further developing the central knowledge base (domain), and mapping RDMO records to specific funder templates.

ARGOS

ARGOS is one of the OpenAIRE Services in the EOSC (Papadopoulou, 2022). ARGOS connects repositories to help populate sections of the ARGOS DMPs, and therefore facilitating the work of users (via automated filling of the DMP). The standalone service uses the OpenDMP software.

A major feature of ARGOS is the implementation of the RDA-DMP common standard (Papadopoulou, Kakaletris and Tziotzios, 2020). A further useful feature is the support of Horizon Europe research programme templates for publishing datasets into the Zenodo repository.

[2] *DSW Deployment Example*. Data Stewardship Wizard, 2022. Accessed: Oct. 31, 2022. [Online]. Available: https://github.com/ds-wizard/dsw-deployment-example

[3] 'GitHub - rdmorganiser/rdmo-docker-compose: RDMO running in different docker images held together by docker compose'. https://github.com/rdmorganiser/rdmo-docker-compose (accessed Oct. 31, 2022).

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

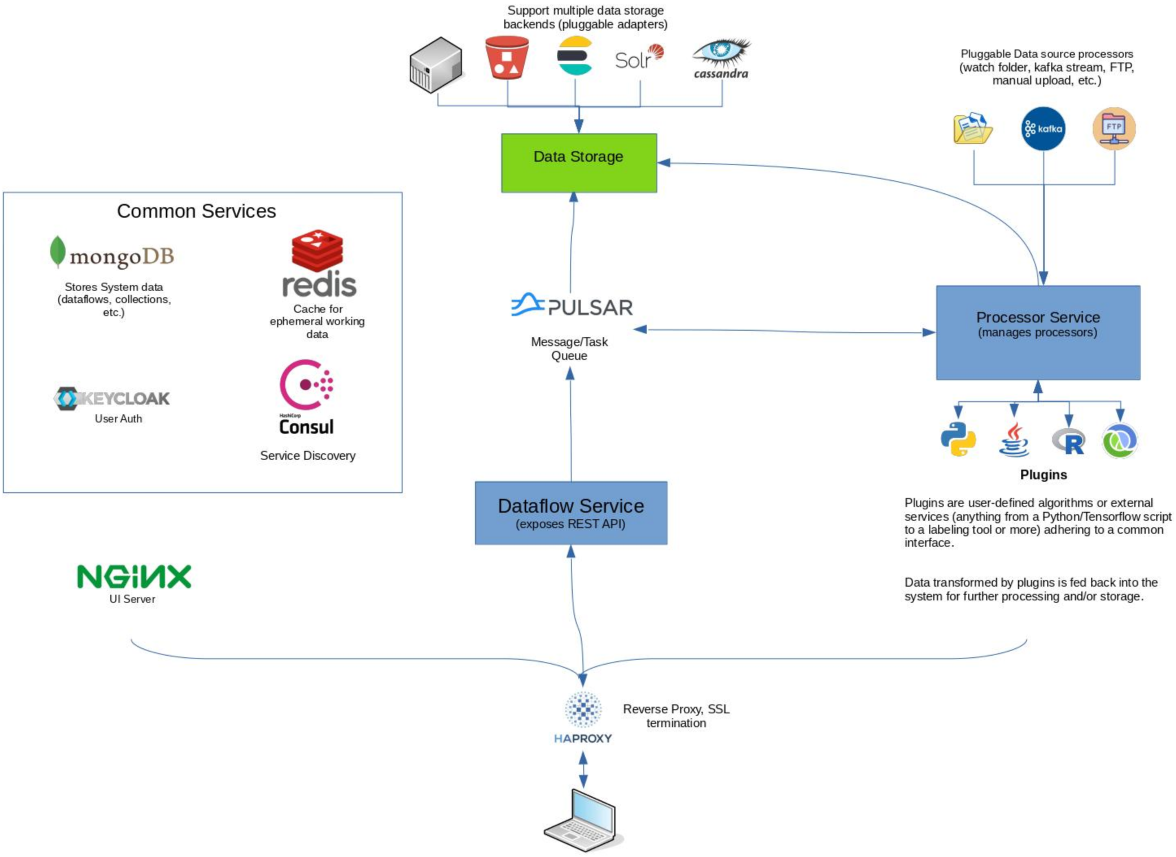

Support multiple data storage backends (pluggable adapters)
Solr
cassandra
Data Storage
Pluggable Data source processors (watch folder, kafka stream, FTP, manual upload, etc.)
kafka
FTP
Common Services
mongoDB
Stores System data (dataflows, collections, etc.)
redis
Cache for ephemeral working data
KEYCLOAK
User Auth
HashiCorp Consul
Service Discovery
PULSAR
Message/Task Queue
Processor Service (manages processors)
Plugins
Plugins are user-defined algorithms or external services (anything from a Python/Tensorflow script to a labeling tool or more) adhering to a common interface.
Data transformed by plugins is fed back into the system for further processing and/or storage.
Dataflow Service (exposes REST API)
NGINX
UI Server
HAPROXY
Reverse Proxy, SSL termination

Figure 1: OpenDMP Architecture [4]

The underlying technology of ARGOS is based on the OpenDMP[5] software for the **standalone deployment**. Features specified in the ARGOS docker-compose configuration file include the following, illustrated in the architecture diagram given in Figure 1:

- A mongoDB database backend.
- Keycloak for authentication, to provide single-sign-on infrastructure
- Redis for caching (ephemeral working metadata/data that needs to be readily available).
- Consul for service discovery.
- MinIO as Object Storage (data storage). Safe and high throughput data transfer over the network.
- Apache Pulsar for message queuing.
- Exposes a REST API.

[4] J. Adam, *Open Data Management Platform*. 2022. Accessed: Oct. 31, 2022. [Online]. Available: https://github.com/rhinoman/odmp

[5] https://opendmp.eu › splash

It should be noted that various storage options are actually available for storage and also as pluggable sources (e.g., Kafka streams) via processor services.

2.1.2 Components and services for active DMPs

We now consider some general recommendations for implementing and using active DMPs. Miksa et al[6] identified **ten principles for active DMPs** that should be satisfied by custom solutions and DMP policies and processes or by turn-key DMP software to enable machine-actionable DMP. These are high level, but provide a guide to the desirable use and functionality of DMPs

These ten principles are:

1. Integrate DMPs with the workflows of all stakeholders in the research data ecosystem.
2. Allow automated systems to act on behalf of stakeholders.
3. Make policies (also) for machines, not just for people.
4. Describe, for both machines and humans, the components of the data management ecosystem.
5. Use PIDs and controlled vocabularies.
6. Follow a common data model for active DMPs.
7. Make DMPs available for human and machine consumption.
8. Support data management evaluation and monitoring.
9. Make DMPs updatable, living, versioned documents.
10. Make DMPs publicly available.

More concrete recommendations and good practices

Following the analysis from (Simms *et al.*, 2017) where they enumerate some characteristics desirable in active DMPs, for I/O for data and metadata, including its content model and serialisation, and for interfaces to other systems (APIs) have been compiled. Different literature sources would focus on different aspects but frequently overlap with well known good practices for data representation from standard setting institutions like e.g. W3C, RDA, ISO. The resulting metadata and its interfaces are expected to adhere to FAIR principles.

The Summary table below gives the internal and external I/O (Input/Output) of the different components or services as related to the good practices and recommendations previously mentioned. The state-of-the-art active DMPs tools from Section 2 already reflect these concerns with varying but high levels of maturity.

[6] T. Miksa, S. Simms, D. Mietchen, and S. Jones, ‘Ten principles for machine-actionable data management plans’, *PLOS Computational Biology*, vol. 15, no. 3, p. e1006750, Mar. 2019, doi: 10.1371/journal.pcbi.1006750.

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

Component I/O		
Content		Serialisation
Separation of data and metadata.		
Data	**Metadata**	• JSON is recommended as the **minimum** for interoperability, particularly with legacy systems. • JSON should be complemented by a *semantic context* for higher interop and e.g. for database ingestion. • Serialisation composability favours extendability (by addition). • W3C JSON-LD[7] is simultaneously the most advanced RDF document yet 100% backward compatible with **JSON**. • JSON-LD serialisation combined with JSON-LD framing[8] enables powerful solutions like e.g. for templating. • Multilingual support should be provided for international projects.
• (Large) binary objects of any kind. • Media type (IANA). • Media type (scientific, custom vocab). • Software may require the association of file extensions to media types. • Any readily harvestable metadata should be added to the metadata database.	• Metadata should be Human and machine readable. • It must make use of existing vocabularies and ontologies. • It must be extendible. • Metadata must be integrated in the same database. • It should cache what is immediately required. • Metadata enrichment is expected (and highly desirable) from local sources and remote sources.	
Enabling technologies		**Facilitates the following**
• Software-defined storage solutions (some networked). • It should be noted that large binary object storage may have **several modes of operation**: object storage	• Metadata needs to be highly structured yet flexible (triplestore/graph database) to leverage querying power on the Knowledge Graph. • The underlying graph should be	• DMP tools like DSW implement the **good practice** of making several RDF serialisations available as different **data views** to the users from a common data source. • The metadata in JSON/JSON-LD can be reused as a common format for the different

[7] 'JSON-LD 1.1'. https://www.w3.org/TR/json-ld11/ (accessed Mar. 09, 2022).

[8] 'JSON-LD 1.1 Framing'. https://www.w3.org/TR/json-ld11-framing/ (accessed Oct. 28, 2022).

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

(flat array), filesystem, etc ● Object storage (e.g. MinIO, Ceph), ● Mapping between the PID to the object and e.g. the temporary signed URL (allowing for safe direct access from external web clients), ● Distributed file systems (e.g. IPFS, Apache Hadoop).	as complete as possible. ● **Incremental** metadata enrichment from many sources is a must. ● Leveraging the database features may be essential to easily implement some services. ● Kafka and Redis as a fast cache. Albeit , they may have several other roles. ● PostgreSQL now has many “noSQL” functionalities (e.g. we can store graph nodes as **JSONB**), and even more specialist functionality via extensions.	services to interoperate internally acting as the system wide representation.

Table 2: Component I/O

2.1.3 Including active DMPs in infrastructure and workflows

A feature of active DMPs is to pre-fill them with existing information as much as practical. Current systems take different approaches to this. Prefilling DMPs or semi-automated DMPs is handled in the OpenAIRE context with ARGOS (Papadopoulou, 2022), in the Radbound University, Nijmegen, based on their CRIS system, and also in the University of Vienna, integrates various IT systems to support RDM (Jetten, Simons and Rijnders, 2019). These three examples will be introduced below. As a fourth example a discipline specific usage of DMPs will be presented.

The ARGOS system prefills DMPs with repository information about projects’ datasets e.g., from Zenodo. It is possible to create dataset templates for different types of datasets that allows some standardisation among dataset types. The workflow is facilitated by using conditional questions and multiple-choice answers. Thereby a researcher can be guided through filling in a DMP. It can connect to a collection of APIs to give standardised answers to e.g. formats and metadata standards extracted from other systems (Papadopoulou, 2022).

Research Information Systems (CRIS) have been extended to include various other aspects of the research life-cycle. One aspect has been the inclusion of RDM support, including Data Management Plan (DMP) support and data curation. The CRIS system could offer a seamless interface allowing dataset and publication registration. The approach here is that DMP information is recorded in the CRIS system as it holds key information about the management and preservation of the dataset. The functionalities of the DMPonline DMP tool have been used as an example for modelling the DMP functionalities of the CRIS system (Jetten, Simons and Rijnders, 2019).

In (Miksa, Oblasser and Rauber, 2022) an analysis of tasks done when creating DMPs has been given to create a high level workflow to implement a system supporting aDMPs or machine-actionable DMPs (Figure 2).

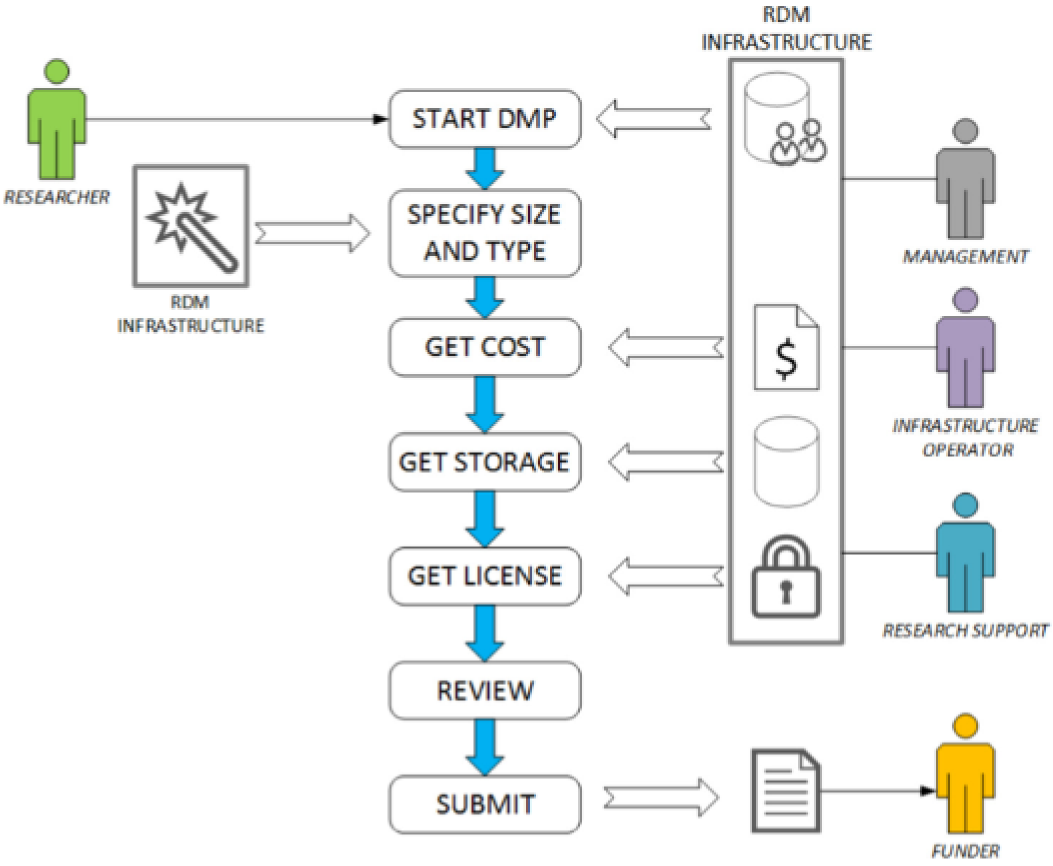


Figure 2: High-level Workflow of Creating an Initial DMP
(Miksa, Oblasser and Rauber, 2022)

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

This high level workflow is broken down in smaller workflows and tasks in order to define the required services. Figure 3 illustrates an example for “specifying size and type of research data”:

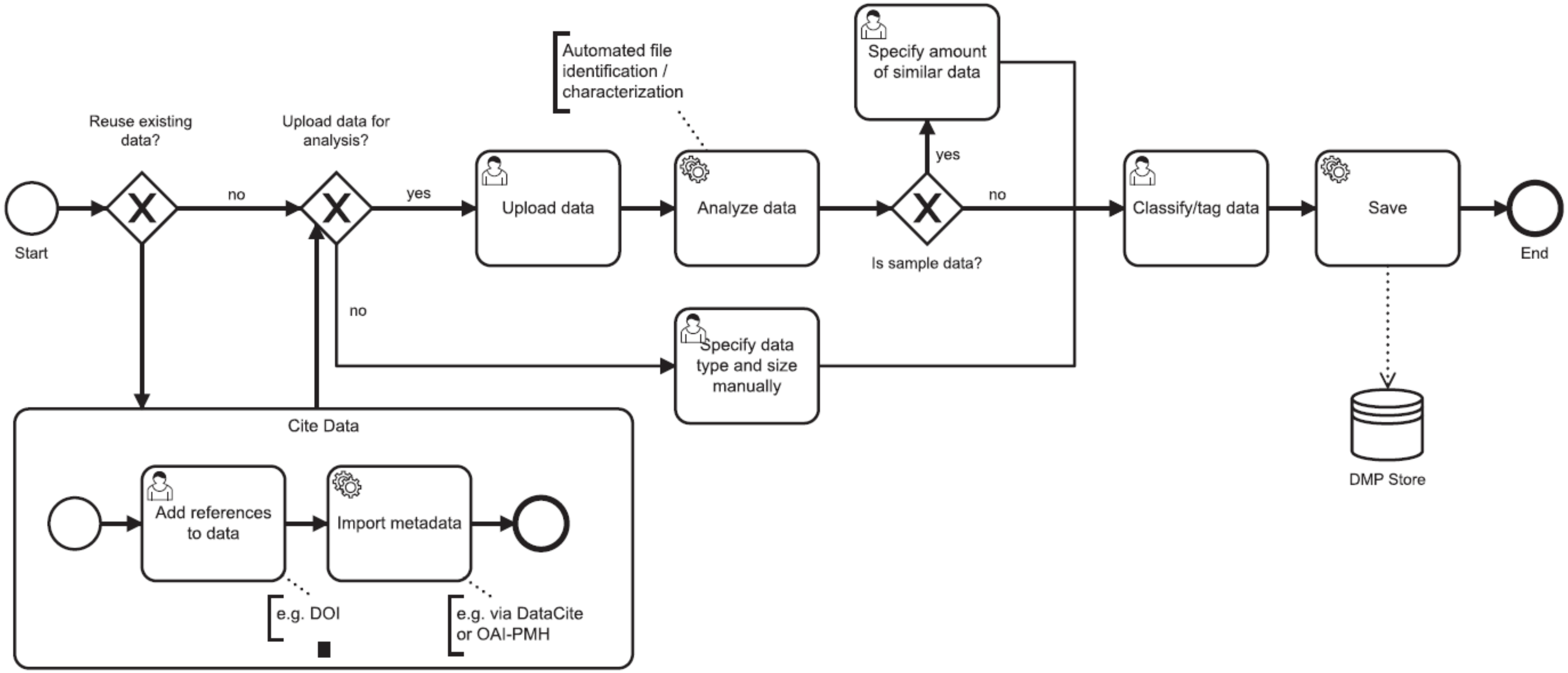
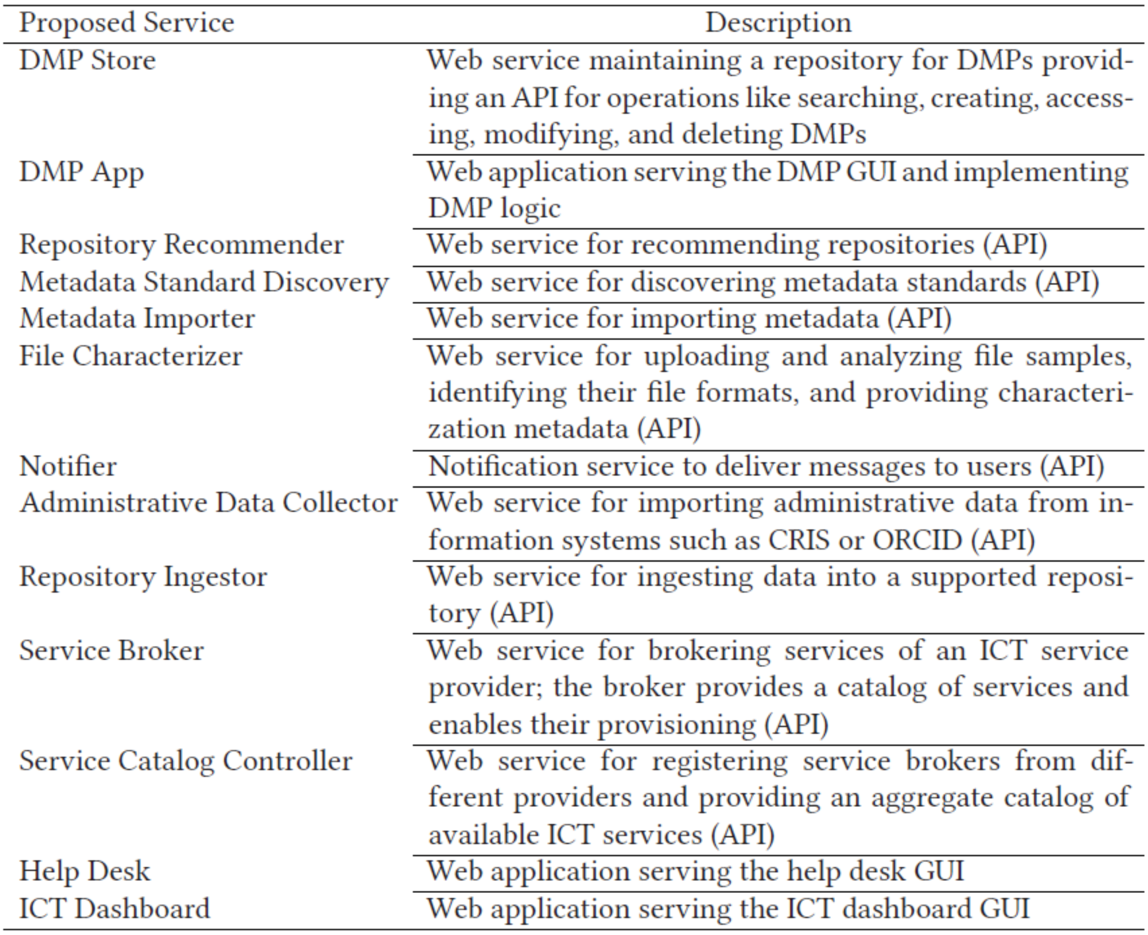

Start
Reuse existing data?
no
Upload data for analysis?
yes
no
Upload data
Automated file identification / characterization
Analyze data
Specify amount of similar data
yes
no
Is sample data?
Specify data type and size manually
Classify/tag data
Save
End
DMP Store
Cite Data
Add references to data
Import metadata
e.g. DOI
e.g. via DataCite or OAI-PMH

Figure 3: Specify Size and Type of Research Data (Miksa, Oblasser and Rauber, 2022)

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

Most of the tasks could be semi-automated. Steps like assigning tasks and roles or selecting datasets will always be manual, selection tasks can be supported by option lists and are thereby semi-automated, others as like calculating storage size is fully automated. This task was used as a demonstration of how the active DMPs services could be integrated with the IT infrastructure by using service brokers.

Finally, (Miksa, Oblasser and Rauber, 2022) identify these services as required to implement active DMPs:



Proposed Service	Description
DMP Store	Web service maintaining a repository for DMPs providing an API for operations like searching, creating, accessing, modifying, and deleting DMPs
DMP App	Web application serving the DMP GUI and implementing DMP logic
Repository Recommender	Web service for recommending repositories (API)
Metadata Standard Discovery	Web service for discovering metadata standards (API)
Metadata Importer	Web service for importing metadata (API)
File Characterizer	Web service for uploading and analyzing file samples, identifying their file formats, and providing characterization metadata (API)
Notifier	Notification service to deliver messages to users (API)
Administrative Data Collector	Web service for importing administrative data from information systems such as CRIS or ORCID (API)
Repository Ingestor	Web service for ingesting data into a supported repository (API)
Service Broker	Web service for brokering services of an ICT service provider; the broker provides a catalog of services and enables their provisioning (API)
Service Catalog Controller	Web service for registering service brokers from different providers and providing an aggregate catalog of available ICT services (API)
Help Desk	Web application serving the help desk GUI
ICT Dashboard	Web application serving the ICT dashboard GUI

Figure 4: Proposed Services for aDMPs (Miksa, Oblasser and Rauber, 2022)

A very discipline-specific approach has been described in (Romanos *et al.*, 2019) for materials design. In this paper, the DMP provides information about the data that has been created, collected, and processed. More specifically, the DMP was divided into sections and enhanced as follows.

Six main categories were created within the DMP, namely:

- General,
- Sample,
- Method,

- Raw data,
- Data analysis, and
- FAIR data

A concept of data documentation in materials characterization (CHAracterization DAta - CHADA), has been developed encompassing sample, method, raw data, and data analysis, In the sample section, features such as materials, dimensions, and quantity of the sample are described and also what the sample is used for e.g., calibration of the equipment.

The Method section states first whether the study in question is observational or experimental and then which methodologies were used to produce the data. The raw data section states the data type 'text', 'numeric' or 'audio visual', if it simulated or collected data, which code has been used, and whether the data has been reused. In the data analysis section, the data coming from the analysis of raw data is described. Naming method, description, purpose, and software used for the data analysis.The general section refers to relevant policies and giving general information e.g., the quantity of data and storage required. The FAIR section explains how to make data FAIR.

2.2 PaN Usage Scenarios

We propose three scenarios within the P&N Facilities experimental lifecycle where DMPs may be used. We note that in the PaNOSC deliverable 2.2 "DMP Template for facility users" (Bolmsten *et al.*, 2021), the section "High level requirements for DMPs at PaN ESFRI facilities" describes requirements for DMPs from each facility. The granularity of the information required to comply with the scenarios described in this deliverable is very diverse. The focus of these requirements has been on data curation, re-use of information, reports, and notifications. These are covered in the usage scenarios 1 and 3 below. The additional usage scenario 2 derives from the requirement of quality control and also for validation of the FAIR digital object.

In these scenarios, DMP information should be used for:

1. Curation: consistent propagation of reusable metadata across the information systems used in the Research lifecycle. This includes:
    a. Recording the availability of metadata and PIDs before dataset creation (re-use pre-existing information), also opportunities for the provision of more metadata, including PIDs, during the DMP creation
    b. Re-using repository information to populate the RDM database with pre-existing information
    c. Feeding information from DMPs (such as quality metadata) into other information systems such as data catalogues and files

2. Validation: validating the data and metadata generated with the experimental process, including:

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

   a. Validation against specific metadata standards, e.g., NeXus application definition
   b. Validation if related datasets like, e.g., calibration, sample, and alignment data are available
   c. Validation against FAIR recommendations (see ExPaNDS D2.7); general metadata requirements in PaN data and overall general requirements on FAIR data
   d. Validation against instrument specific metadata requirements; are metadata complete as defined
   e. Policy validation

3. Reports / DMP exposure: providing reports to inform decision making further along the experimental process, including:
   a. For infrastructure planning: to allocate and prepare resources, e.g,. storage, network, CPU/GPU to be made available to the experimental team.
   b. For funder: fulfil funder requirements on populating DMPs and assessing the quality of “FAIR data”
   c. For RDM team: generating overview on the provision of RDM (e.g., usage of formats and standards)
   d. Notifications about special requirements or large deviations from the normal expectations of data and computational needs.

3. Data Model for aDMPs

3.1 Data Model for Workflows / DMP Execution

In the previous ExPaNDS deliverable 2.4 (Görzig *et al.*, 2021) DMP sources, including IT systems, and the phases when DMP related information is first identified have been discussed. The following graphic shows the phases during the research life-cycle, related IT systems and information they hold (cf. Figure 5).

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

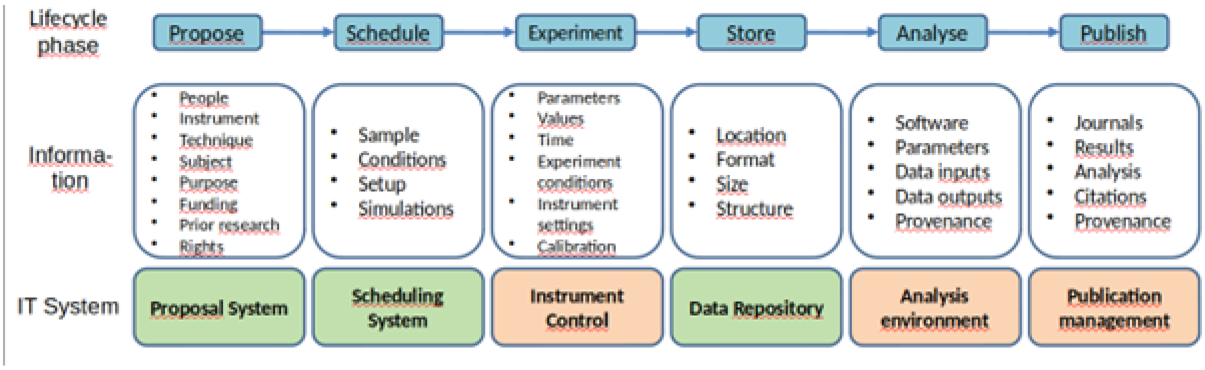


Figure 5: Research Life-cycle and IT Systems (Görzig et al., 2021)

In the same deliverable the phases of a DMP related to the research life-cycle have been proposed (cf. Figure 6).

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.

Figure 6: DMP Phases in Research Life-cycle (Görzig et al., 2021)

The DMP data flow below brings together the existing IT systems with the additional ones required for DMP as explained in the section PaN usage scenarios. The proposal system, the repository, and sometimes the sample DB are systems that already exist and correspond to usage scenarios 1.A and 1.B (re-use of information for filling in the DMP). The DBs pre-existing instrument information and static facility information also correspond to the same usage scenarios, but typically these systems are not as yet systematically supported within facilities. In the next subsection a data model for the expected information coming from these systems will be suggested.

The DMP, reporting, and notification tools correspond to usage scenarios in 3. Therefore, tools exist, but are not necessarily in use in the facilities. The curation and validation tools correspond to the usage scenarios in 1.3 and 2. Some of the required tools already exist in facilities, although they are not typically integrated with the DMP workflow.

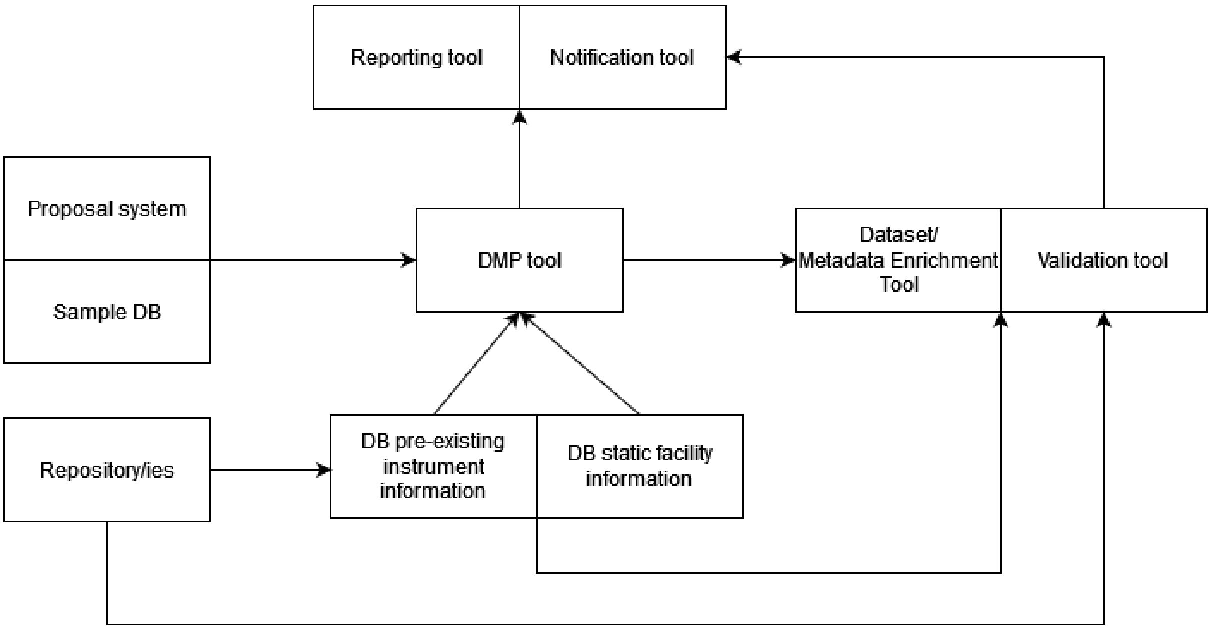

Reporting tool
Notification tool
Proposal system
Sample DB
DMP tool
Dataset/ Metadata Enrichment Tool
Validation tool
Repository/ies
DB pre-existing instrument information
DB static facility information

Figure 7: active DMP Components

Proposal system: DMP tool first extracts proposal specific information from the User Office IT system to e.g. instrument and scientific purpose information.

SampleDB: Sample database for first Sample registration. Specifies what the sample is and gives information about it.

DMP tool: the DMP tool could be any kind of existing DMP tool that can handle a knowledge base as described in the PaNOSC DMP deliverable. The DMP tool requests information from the static instrument information and facility information DBs.

Instrument information database:. It should hold information to describe the datasets produced by an instrument. This information should describe the resources to create and use the datasets, as well as to validate the datasets and describe the datasets themselves.

Facility information database: this system should hold information about data in the repository which is generally applicable to all data generated within the facility. This information is applicable to all instruments. .

Reporting tool: the reporting tool creates *human readable* reports about the data that will be or is produced during a proposal. The recipients can be e.g., funders, users, and/or facility staff and the reports should conform to the template required by each stakeholder. It also creates a *machine-readable* version of the DMP in the RDA DMP common standard.

Notification tool:. The Notification tool should send notifications mainly to facility staff. The notifications can be scheduled by e.g., due dates of reports or triggered by events such as conflicting demands on infrastructure by different proposals or strong deviations of default values in the DMP from registries in the *DB static instrument information*.

Dataset/Metadata Enrichment Tool: The Curation Tool should take machine readable information from the *DMP tool* and the *DB static instrument information* and other IT systems

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

and adds this information as metadata to the created datasets. It might also convert the measurement data to other formats.

Validation tool: It should validate created datasets and their files against validation schemes stored in the DB *static instrument information*. These validation schemes can be e.g., NeXus nxdl files or python recipes. The validation tools should also validate the datasets in a repository on compliance with the DMP information. Validation events can trigger notifications.

Repository: the data repository should support the maintenance of the *DB pre-existing instrument information*. By allowing export of characteristics of the data ingested in the repository by a specific instrument and inserting it into the Instrument DB.

3.2 Data Model for Pre-existing DMP Information

Pre-existing facility data:

We specify in Table 3 the information giving the context of the experiment which can be provided on a facility basis to all DMPs. The pre-existing facility data is closely related to the questions specified in the DMP template of the PaNOSC deliverable (Bolmsten *et al.*, 2021); we give the reference to the related questions in the template using the numbering from page 15 of that deliverable.

Fields	Question in template
facility	
repository	
name	107
URL	
licence	51
security	partly 47
pid_system	77
pid_responsible	
name	79
orcid	
personal_data	83
min_storage_period	105
archive	
name	107

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

Fields	Question in template
access	
URL	
certificate	108
arrangements	109
embargo_period	110b
access_control	111
costs	115

Table 3: Pre-existing facility data

In the facility data model, information described in the facility's data policy or other conventions used by the facility can be reused.

Pre-existing instrument data:

In the case of the static data associated with an instrument, the mapping is not that straightforward as for the facility information. Here the perspective shifts from instrument as root to dataset as root. A dataset might be the result of a scan. A filecollection consists of the files written by one software during a scan, e.g. DAC (Data Acquisition and Controls) for the beamline, or a detector software.

In the repository the dataset and its filecollections will be catalogued together, or maybe at an earlier stage, e.g. by writing data into a common NeXus file. But by separating the output of a specific software it is possible to identify where they came from and what is possible to do with the files created by a specific software. The figure below is showing a hierarchical view on a dataset (Figure 8):

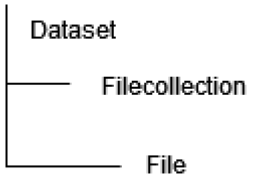


Figure 8: Dataset - Filecollection - File Structure

The dataset could be seen as the overall data produced in an experimental run using the instrument, the filecollection is a collection of files produced with one instance of software used in a scan, which is reflected in the hierarchical structure of the data. Table 4 below gives the fields that can be related to a dataset and its components that can be used to fill in the DMP or used for data curation, and again we give the reference to the question in the DMP template.

Fields	Comment	Question in template
Data-sets		
name	*Could be used for data curation*	
description		15
contributors[]	*Could be used for data curation*	17
reproducible		20
interested_community		19
usage	*e.g. sample_information, calibration, ELN, …*	36
archival		
moment	*point in time for archival*	112
selection_criteria		104
long_term_archival_reason		101
data_security		
measurements		47
responsible_person		
methods[]	*Information required to select standards (format and metadata)*	15/16
description	*could include controlled vocabulary terms, e.g., from PaNET*	
filecollections[]	*A file collection is created by one software instance*	
name		43
instrument		30
name		
pid	*e.g. Instrument DOI*	
storage		40/41

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

backup		45
location		
responsible_person		46
quality_assurance		57
interoperable		48
hardware	*metadata for data curation*	30
hardware_architecture[]		
type	*[chip, hardware peripheral, input/output device, processor, storage device, …]*	
manufacturer		
model		
hardware_peripheral[]		
type	*for example, electrometer, …*	
manufacturer		
model		
writing_sotfware[]	*metadata for data curation* *this is the software creating the file collection running on a hardware*	30
name		
type	*[ancillary, driver, operating system, plugin, renderer, software,…]*	
URL		
plugins[]		
name		
type	*[script, plugin, application, …]*	
URL		
files[]		

name	*Data curation*	43
format[]		29
metadata_schema		
size	*For estimating data volume*	26/27
amount_min		26/27
amount_max		26/27
processing_requirements		31
hardware_requirements[]		31
type	*[chip, hardware peripheral, input/output device, processor, storage device, …]*	
manufacturer		
model		
reading_softwares		31
name		
type	*[ancillary, driver, operating system, plugin, renderer, software,…]*	
documentation		32
URL		
plugins[]		
name		
type		
URL		

Table 4: Pre-existing Instrument Information

Pre-existing project data:

The pre-existing project data comprises information from the proposal systems, but also sample information. The available project data is very limited:

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

proposal

- title
- abstract
- principal investigator/proposer
- co-proposer

Sometimes:

- experimental team
- instrument

Other information as e.g., experimental technique, experimental team with ORCIDs or funding references could be useful in future. For data curation publications or beamtimes referenced in the proposal might be a plus.

4. Suggested Architecture for active DMPs

As has been well described in report outputs of the International Digital Curation Conference 2017 workshop on machine-actionable data management plan (Simms *et al.,* 2017), it is important to have in mind that for a DMP tool to be well shaped and easily scalable, it should be thought of as an IT service to be embedded in existing research life-cycle workflows where different research tools and systems are dynamically exchanging information of many different kinds.

The report then identifies that a well-designed DMP data model has to take care of interoperability between the DMP tool and the other systems, and this can be achieved using a common data model with a core set of elements, as we have defined it in the preceding chapters. It also proposed that the DMP tool should be highly customizable, which introduces the need for a final step of developing a common interface and default implementation in a variety of programming languages to enable a common way of accessing information in active DMPs. This way, all tools and systems involved in processing research data can be extended easily to be able to provide and access information to/from a DMP. For example, a workflow engine can add provenance information to the active DMP, or a repository system can automatically pick suitable content types for submission and later automatically identify applicable preservation strategies.

To pave the way to this final goal, we will address in this chapter all the relevant IT systems involved in a research life-cycle and describe their interaction with a DMP tool from a service oriented point of view which will help us to define which information they have to exchange in a standardised way.

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

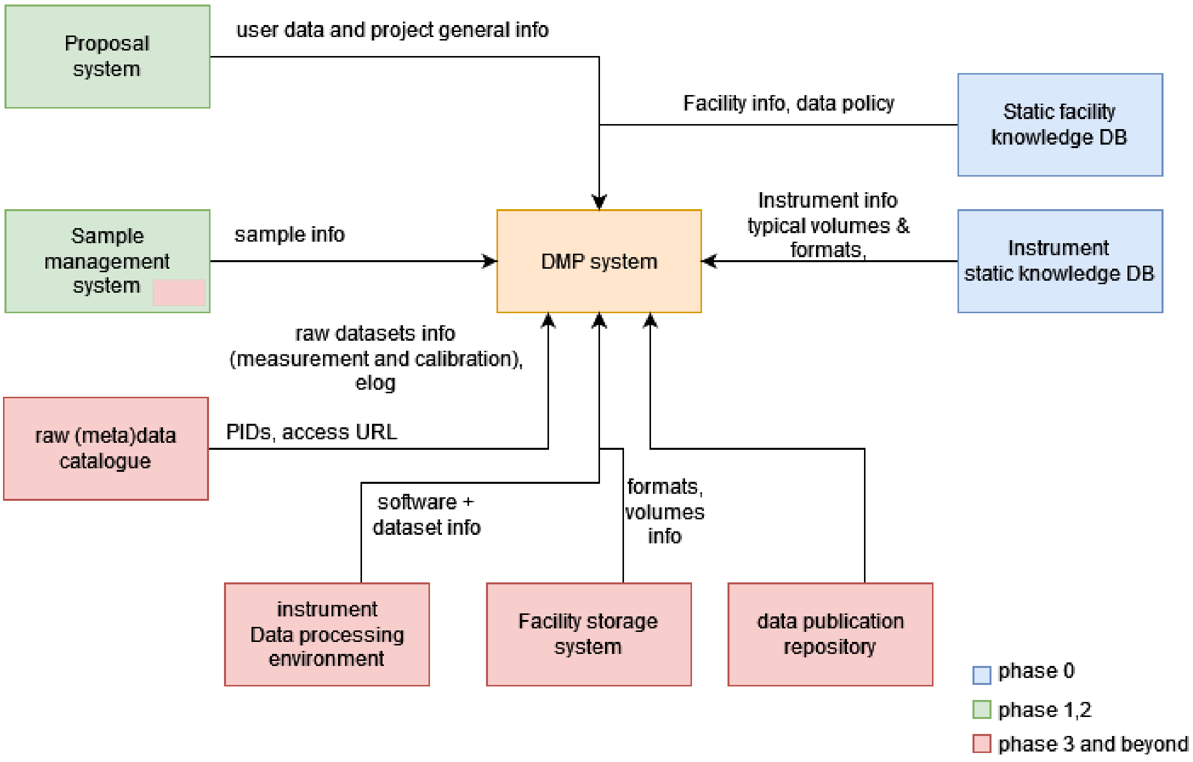

Proposal system
user data and project general info
Facility info, data policy
Static facility knowledge DB
Sample management system
sample info
DMP system
Instrument info typical volumes & formats,
Instrument static knowledge DB
raw datasets info (measurement and calibration), elog
raw (meta)data catalogue
PIDs, access URL
software + dataset info
formats, volumes info
instrument Data processing environment
Facility storage system
data publication repository
phase 0
phase 1,2
phase 3 and beyond

Figure 9: Sources for an active DMP System

4.1 Relevant IT Systems

The relevant IT systems will be described here as abstract services that act as intermediaries between the DMP service/system and the different overall facility information system building blocks. These intermediate layers are mapping (with slightly modified names) the ones sketched in Figure 4, with a supplementary layer representing the facility static information holders such as static facility knowledge DB or instrument static knowledge DB.

In the following sections, we will describe what information is required from each relevant IT (sub)system, focusing on one part on the elements that each component should output, that should enter the DMP. By the way, the dynamic aspect of an active DMP will be illustrated by the elements that may/should be updated by one component while they have been first introduced by another component. As a preparation for the stage where the active DMP will have to be concretely implemented through a given set of software modules exchanging data, the targeted information will be described as json structured data.

4.1.1 Proposal system

The required information from the proposal system is mainly administrative:

- administrative metadata (names, affiliation of the PI and the visiting team)
- start and end date
- instrument requested

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

- research area / discipline
- experiment description
- funding information

It can be exchanged through a json record similar to:

```
{
        “start_date”,
        “end_date””,
        “title”,
        “abstract”,
        “principal_investigator”:{
                "name",
                ”orcid”,
                “affiliation”
        },
        “visiting_team”:[
                {
                  “name”,
                  ”orcid”,
                  “affiliation”
                 },
                 {
                  “name”,
                  ”orcid”,
                  “affiliation”
                 },

                      ….
        ]
        “instrument”,
        “experiment_description”,
        ………
}
```

4.1.2 Facility knowledge system

The required information from this IT subsystem layer is related on one hand to the facility itself and on the other hand to the instrument that will be the support of the experiment declared in the proposal.

The json record that should be added to the DMP for facility description is the following (translated from the table **Pre-existing facility knowledge DB**):

```
{
   "facility": {
      "repository": {
```

```
                "name",
                ”URL”
        },
        “licence”,
        “security”,
        “pid_system”,
        “pid_resposible”:{
                "name",
                ”orcid”
        },
        “personal_data”,
        “min_storage_period”,
        “archive”:{
        “name”,
        “access”,
        “URL”
        },
        “certificate”,
        arrangements”,
        “embargo_period”,
        “access_control”,
        “costs”
}
```

Regarding Instrument static information, two kinds of information are required.

The first kind of information is more related to the description of the instrument as a set of physical devices such as sensors, detectors, etc. as well described in the document: “the Metadata Schema for the Persistent Identification of Instruments” (Krahl et al., 2022), published as a recommendation from the RDA Persistent Identification of Instrument Working Group.
The json record that should be added to the DMP for this kind of information may look like the following:

```
{
   “instrument":”instrument name (beamline?)”,
   “responsible",
   “configuration_date”,
    ……
    [
        {
          “Instrument_component”:”monochromator?,mirror?,diffractometer?, detector?...”
          “identifierType”,
          “SchemaVersion”,
          “LandingPage”,
          “Name”,
          ….
          “Manufacturer”,
          ….
          “Model”,
          ….
        },
        {
          “Instrument_component”:”monochromator?,mirror?,diffractometer?, detector?...”
          “identifierType”,
          “SchemaVersion”,
          “LandingPage”,
          “Name”,
          ….
          “Manufacturer”,
          ….
          “Model”,
          ….
        },
        ….

    ]
}
```

This json record structure reflects then that the instrument hosting the experiment is viewed as a set of components identified by metadata conforming to the PIDINST metadata schema adopted by the RDA.

The other kind of information is more related to the Instrument as a tool playing an essential role in creating research data and as stated in (Krahl et al., 2022), the ability to link a physical instrument to the data that it generated and contextual metadata such as when, where and how it was operated, is critical for the accurate interpretation of that data.

The json record that should be added to the DMP for this kind of information is the following (translated from the table **Pre-existing instrument data**):

```
{
“datasets”:{
        “name”,
        ….
        “archival”:{
                “moment”,
                “selection_criteria”,
                “long_term_archival_reason”
                }
        ….
        “filecollections”:[
                {
                  “name”,
                  “instrument”,
                  “storage”,
    “backup”,
    ….
    “files”:[
        {
          “name”,
          “format”,
          “metadata_schema”,
          “size”,
          “amount_min”,
          “amount_max”,
        },
        {
          “name”,
          “format”,
          “metadata_schema”,
          “size”,
          “amount_min”,
          “amount_max”,
        },
        ….
    ]
  },
  {
    “name”,
    “instrument”,
    “storage”,
    “backup”,
    ….
    “files”:[
        {
          “name”,
          “format”,
          “metadata_schema”,
          “size”,
          “amount_min”,
          “amount_max”,
        },
        {
          “name”,
          “format”,
          “metadata_schema”,
          “size”,
          “amount_min”,
          “amount_max”,
        },
        ….
    ]
```

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

```
                }
                ]
        }
}
```

4.1.3 Sample management system

No well adopted standard metadata schema for sample identification has emerged yet in the PaN community. The RDA "Physical Samples and Collections in the Research Data Ecosystem"[9] interest group is trying to address this topic and the work is still ongoing. The one widely used schema at this time is the "global persistent identifier for physical samples" IGSN: an alphanumeric code that is assigned to specimens and related sampling features to ensure their unique identification and unambiguous referencing (Klump et al., 2021), though this is not as yet widely used in the PaN community. In the PaN community the LEAPS-STARS[10] project started in 2021 is addressing this topic. It is anticipated that this project will propose the use of persistent sample identifiers and the development of a set of common standard information items about samples. In all cases, the information that should be registered in the DMP may consist at least of these PIDs.

4.1.4 Instrument DAC system

The information registered in the DMP from the Instrument DAC system will be used later (after the end of the experiment) to populate the "Pre-existing Instrument data". For consistency then the json record associated to this part must conform as much as possible to the structure described above. This record may look like:

```
{
"datasets":{
        "name",
        ….
        "filecollections":[
                {
                  "name",
                  "instrument",
                  "storage",
                  "backup",
                  ….
                  "files":[
                      {
                        "name",
                        "format",
                        "metadata_schema",
                        "size"
                      },
                      {
                        "name",
                        "format",
```

[9] https://rd-alliance.org/groups/physical-samples-and-collections-research-data-ecosystem-ig

[10] https://leaps-initiative.eu/wp-content/uploads/2021/10/DIGITAL-LEAPS-August-2021.pdf

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

```
                        “metadata_schema”,
                        “size”
                      },
                      ….
                  ]
                },
                {
                  “name”,
                  “instrument”,
                  “storage”,
                  “backup”,
                  ….
                  “files”:[
                      {
                        “name”,
                        “format”,
                        “metadata_schema”,
                        “size”
                      },
                      {
                        “name”,
                        “format”,
                        “metadata_schema”,
                        “size”,
                      },
                      ….
                  ]
                }
                ]
        },
        “cumulated_size”
}
```

4.1.5 Data analysis environment system

This part of the DMP is not that straightforward to describe in a systematic way due to the variety of usages in the scientific community regarding data analysis activities. Depending on the experimental technique some of the scientists use standard well known software tools commonly shared in their research field and references/identifiers for them can be easily found. Others develop home made python scripts or jupyter notebooks to analyse their data.

In all cases, if things are done following the recommendations stated about this topic in the D2.7 deliverable, there is room for valuable information to cite in the DMP as at least any software repository (git, github, gitlab…) where the software can be found, the other more detailed meta-data being injected in the data catalogue.

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

The other relevant information is related to the data analysis environment for which the actual size and format of analysed data should be filled in in the DMP. Here again we can use a json record similar to the one for Instrument DAC system:

```
{
“datasets”:{
        “name”,
        ….
        “filecollections”:[
                {
                  “name”,
                  “software”,
                  “storage”,
                  “backup”,
                  ….
                  “files”:[
                      {
                        “name”,
                        “format”,
                        “metadata_schema”,
                        “size”
                      },
                      {
                        “name”,
                        “format”,
                        “metadata_schema”,
                        “size”
                      },
                      ….
                  ]
                },
                {
                  “name”,
                  “software”,
                  “storage”,
                  “backup”,
                  ….
                  “files”:[
                      {
                        “name”,
                        “format”,
                        “metadata_schema”,
                        “size”
                      },
                      {
                        “name”,
                        “format”,
                        “metadata_schema”,
                        “size”,
                      },
                      ….
                  ]
                }
                ]
        },
        “cumulated_size”
}
```

4.1.6 Facility storage system

Information to be registered in the DMP has been well described in the D2.4 deliverable (Questions 101 to 115 of the DMP template).

The json record associated to this part of the DMP may look like the following:

```
{
        “proposal”,
        ….
        “rules_of_selection”,
        “responsible_for_selection”,
        “minimum_period_of_storage”,
        “repository_storage_after_end_of_project”,
        “repository_certification”,
        “embargo_period”,
        “responsible_for_data_access”,
        “date_of_storage”,
        “storage_costs_recovery”,
        ….
}
```

4.1.7 Facility librarian system

This part of the DMP overall system is in charge of the IT information system which deals with publications that close the whole experimental life cycle of a research project. It is generally taken in charge by the IT librarian team.

A json record associated to it should look like the following:

```
{
        “proposal”,
        “responsible_for_publication”,
        “curator”,
        [
                {
                  “title”,
                  “publisher”,
                  “authors”:[],
                  “date_of_publication”,
                  “DOI”
                },
                {
                  “title”,
                  “publisher”,
                  “authors”:[],
                  “date_of_publication”,
                  “DOI”
                }
                ….
        ]
}
```

4.2 Information Flow between the IT Systems

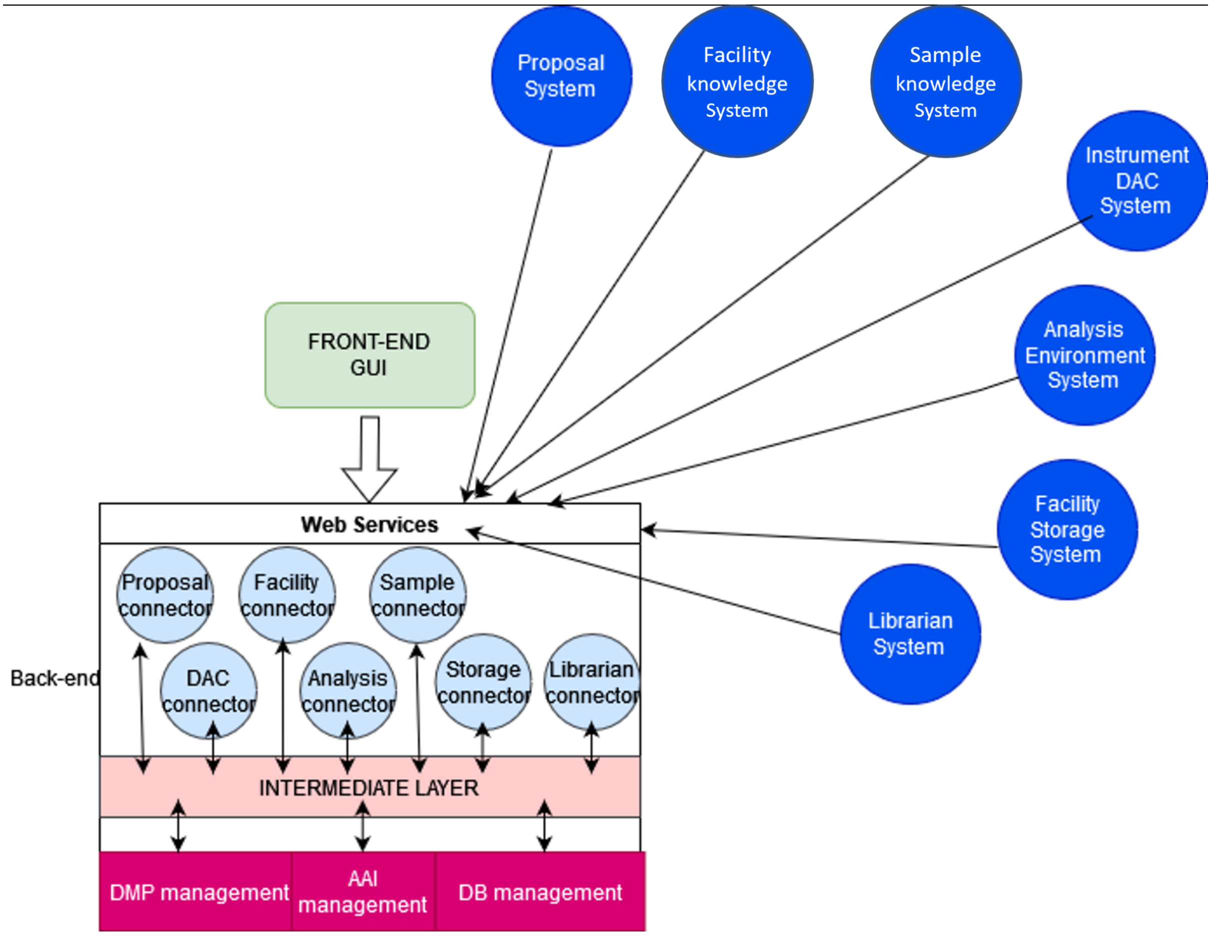


Figure 10: Possible implementation of the active DMP

One possible implementation of the active DMP is shown above. To use modern software paradigms, it is architectured as a web application made of two modules (that can be developed separately using different technologies): the front-end and the back-end.

The front-end is provided to the users to interact with the DMP giving access to features like:

- filling in answers to questions as those presented in the DMP template
- browsing sections of the DMP and manually updating them
- searching for DMP examples in the DMP database
- extracting DMP documents in different formats (PDF,...)

The back-end holds the machinery for core services (implemented as web application modules) such as:

- DMP lifecycle management with business processes
- AAI management to handle authorisations and access rights to the database depending on the user role
- DMP data base management
- Management for background jobs and Workflows

The front-end interacts with the back-end through well defined Web Services (published outside as a REST API to be designed and implemented.

These web services are also in charge of the communication between the DMP core system and the outside world and in particular with all the relevant IT systems described above. For each of these IT systems a dedicated connector to the DMP core is available handling its specific data exchange.

The DMP system connectors are not isolated. They have to interact between themselves and with the core services (a very evident example is for the proposal connector to update the DMP management module). Another type of interaction is for a DMP connector able to update some information injected early on in the process by a precedent DMP connector in the overall workflow of dynamic active DMP handling (The proposal connector will first introduce preliminary information about the sample, which will be later on updated by the sample connector). To this end there is no point to point connection, these interactions are triggered through the intermediate layer in order to assess consistency of the DMP data model (and also efficiency, depending on the system load this layer can implement load balancing using a fifo pool associated to asynchronous messaging).

The overall DMP handling workflow is also to be taken in charge by the intermediate layer.

5. Concluding Remarks

Based on the work in the previous deliverables and the analysis of Section 2 of this deliverable where, first integration of the creation of DMPs in infrastructures and the requirements of services is described and then the PaN usage scenarios are presented. This

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

work introduces in Section 3 a possible data model and in Section 4 ideas for an infrastructure for active DMPs in PaN sciences. The data model of pre-existing instrument and facility information shows that the creation of DMPs can be supported by holding this information in a dedicated database and what minimal information can be retrieved from the proposal system. In Section 4 this data model has been more concretized in relation to infrastructure used in the overall research life-cycle in PaN sciences as presented in figure 4 of this deliverable. The DAC system and data analysis environment have been integrated to concretize the pre-existing information retrieved from the facility knowledge system. Figure 10 in Section 4.2 gives an overview of a possible structure of an implementation of an active DMP system. For an active DMP implementation a web application consisting of a front-end and a back-end have been suggested. Where the front-end is created for the interaction with the users. The back-end consists of three layers: one layer is holding the connectors that are connecting to the infrastructure and preparing the retrieved information to be joined with other information in the intermediate layer where the data will be prepared for usage.

The data model and the possible overview of an implementation are still in the planning phase. For an implementation a more detailed elaboration will be required. In this deliverable aspects such as specific technologies or concrete metadata schemata have not been discussed, as the aim of this deliverable was to create a high level architecture for DMPs within facilities, including a view on IT systems and their data flow.

References

Bolmsten, F. *et al.* (2021) 'DMP Template for facility users'. Available at: https://doi.org/10.5281/ZENODO.5639428.

Bunakov, V. *et al.* (2022) 'Advanced infrastructure for PIDs in Photon and Neutron RIs'. Available at: https://doi.org/10.5281/ZENODO.5905351.

Cardoso, J., Castro, L.J. and Miksa, T. (2021) 'Interconnecting systems using machine-actionable data management plans – hackathon report', *Data Science Journal*, 20(1). Available at: https://doi.org/10.5334/DSJ-2021-035.

Cardoso, J., Proença, D. and Borbinha, J. (2020) 'Machine-actionable data management plans: A knowledge retrieval approach to automate the assessment of funders' requirements', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12036 LNCS, pp. 118–125. Available at: https://doi.org/10.1007/978-3-030-45442-5_15.

Collins, S.P. *et al.* (2021) 'ExPaNDS ontologies v1.0'. Available at: https://doi.org/10.5281/ZENODO.4806026.

Görzig, H. *et al.* (2021) 'DMPs for Photon and Neutron RIs'. Available at: https://doi.org/10.5281/ZENODO.5636096.

Horizon Europe (2022) 'Horizon Europe programme guide', in. Available at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf (Accessed: 7 November 2022).

Jetten, M., Simons, E. and Rijnders, J. (2019) 'The role of CRIS's in the research life cycle. A case study on implementing a FAIR RDM policy at Radboud University, the Netherlands', in *Procedia Computer Science*. Elsevier B.V., pp. 156–164. Available at: https://doi.org/10.1016/j.procs.2019.01.090.

Jones, S. *et al.* (2020) 'Data Management Planning: How Requirements and Solutions are Beginning to Converge', *Data Intelligence*, 2(1–2), pp. 208–219. Available at: https://doi.org/10.1162/dint_a_00043.

Klump, J. *et al.* (2021) 'Towards globally unique identification of physical samples: Governance and technical implementation of the igsn global sample number', *Data Science Journal*, 20(1), pp. 1–16. Available at: https://doi.org/10.5334/DSJ-2021-033/METRICS/.

Krahl, R. *et al.* (2022) 'Metadata Schema for the Persistent Identification of Instruments'. Available at: https://doi.org/10.15497/RDA00070.

Matthews, B. *et al.* (2020) 'Draft extended data policy framework for Photon and Neutron RIs'. Available at: https://doi.org/10.5281/ZENODO.4014811.

McBirnie, A. *et al.* (2021) 'Final data policy framework for Photon and Neutron RIs'. Available at: https://doi.org/10.5281/ZENODO.5205825.

Miksa, T., Oblasser, S. and Rauber, A. (2022) 'Automating Research Data Management Using Machine-Actionable Data Management Plans', *ACM Transactions on Management*

Information Systems, 13(2). Available at: https://doi.org/10.1145/3490396.

Miksa, T., Walk, P. and Neish, P. (2020) 'RDA DMP Common Standard for Machine-actionable Data Management Plans'. Available at: https://doi.org/10.15497/RDA00039.

Papadopoulou, E. (2021) 'Argos - How tools can facilitate our DMP discussions'. Available at: https://doi.org/10.5281/ZENODO.5549595.

Papadopoulou, E. (2022) 'Argos automates the writing of DMPs - OpenAIRE Blog'. Available at: https://www.openaire.eu/blogs/argos-automates-writing-of-dmps (Accessed: 2 November 2022).

Papadopoulou, E., Kakaletris, G. and Tziotzios, D. (2020) 'Implementing the DMP Common Standard on the Argos service for active DMPs'. Available at: https://doi.org/10.5281/ZENODO.4278006.

Romanos, N. *et al.* (2019) 'Innovative Data Management in advanced characterization: Implications for materials design', *Materials Today Communications*, 20, p. 100541. Available at: https://doi.org/10.1016/j.mtcomm.2019.100541.

Sallans, A. and Donnelly, M. (2012) 'DMP Online and DMPTool: Different Strategies Towards a Shared Goal', *International Journal of Digital Curation*, 7(2), pp. 123–129. Available at: https://doi.org/10.2218/ijdc.v7i2.235.

Salvat, D. *et al.* (2020) 'Draft Recommendations for FAIR Photon and Neutron Data Management', (857641), pp. 1–63. Available at: https://doi.org/10.5281/zenodo.4312825.

Simms, S. *et al.* (2017) 'Machine-actionable data management plans (maDMPs)', *Research Ideas and Outcomes*, 3, p. 10. Available at: https://doi.org/10.3897/rio.3.e13086.

Soler, N. *et al.* (2022) 'Final recommendations for FAIR Photon and Neutron Data Management'. Available at: https://doi.org/10.5281/ZENODO.6821676.

Stocker, M. et al. (2020) 'Persistent Identification of Instruments', Data Science Journal, 19(1), pp. 1–12. Available at: https://doi.org/10.5334/dsj-2020-018.

Wilson, M. *et al.* (2011) *Deliverable D2.1 Common policy framework on scientific data*. Available at: https://doi.org/10.5281/zenodo.3738498 .

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