



# A Framework for Active DMPs in Photon and Neutron Science Large-Scale Facilities

PRACTICE PAPER

HEIKE GÖRZIG

ALEJANDRA N. GONZALEZ BELTRAN

FELIX ENGEL

BRIAN MATTHEWS

\*Author affiliations can be found in the back matter of this article

ubiquity press

## ABSTRACT

In this paper, a framework and a system architecture are presented to support researchers in DMP creation and execution, with a focus on the generation of FAIR data. Using the research data lifecycle within Photon and Neutron analytical facilities as a detailed exemplar of this approach in practice, it shows how combining the creation of the DMP with the project management framework PMBOK makes it easier to integrate DMP creation within the researchers' workflow and reuse pre-existing information within the research infrastructure and related project roles. The paper identifies requirements and introduces a lifecycle for pre-existing information that helps in automatic population of the DMP. This paper also discusses a data model for the reuse of pre-existing information. It shows possible approaches to support scientists through the (semi-)automation of the creation, execution, and use of a DMP and knowledge transfer. The approach is based on work within the PaNData ODI, ExPaNDS, and PaNOSC projects.

## CORRESPONDING AUTHOR:

**Heike Görzig**

Helmholtz-Zentrum Berlin  
für Materialien und Energie,  
Germany

[heike.goerzig@helmholtz-berlin.de](mailto:heike.goerzig@helmholtz-berlin.de)

## KEYWORDS:

DMP; active DMP;  
Synchrotrons; PaN sciences;  
Project management; Reuse;  
Ontology; Workflows

## TO CITE THIS ARTICLE:

Görzig, H, Gonzales Beltran, AN, Engel, F and Matthews, B. 2024. A Framework for Active DMPs in Photon and Neutron Science Large-Scale Facilities. *Data Science Journal*, 23: 4, pp. 1–18. DOI: <https://doi.org/10.5334/dsj-2024-004>

## 1. INTRODUCTION AND MOTIVATION

Photon and Neutron Science (PaN) facilities provide measurement and analysis infrastructure and services based on radiation sources for very diverse disciplines including: archaeology; material sciences; the natural sciences; and medicine. These support studies that require the detailed analysis of the structure of matter. The growing volume and complexity of the experimental data produced, as well as the required computing power for analyzing them, require a facility-wide infrastructural solution for data management, as they usually go beyond what an individual scientist can handle. Further, this data explosion drives the increasing need for on-site analysis infrastructure, rather than analysis on user's resources. Also, researchers need high quality data and reliable knowledge about previous experiments to form the basis of their work. Therefore, integrating the data into the existing infrastructure is required to ensure that data is used and reused to maximize the scientific return. It is now recognized that the use of Data Management Plans within PaN facilities is required to underpin sound data management processes and support the collection of FAIR data, as they ensure advanced consideration of how the data will be handled.

In many research projects, Data Management Plans (DMPs) are formally required, which is the case for most publicly funded research projects. Funding agencies very often request that research data generated in funded projects is made available for future re-use, and therefore are demanding the generated research data comply with the open access and FAIR data principles (Miksa et al. 2019). To meet these requirements, funding agencies are demanding the elaboration of DMPs at proposal time—or at least at research-fund contracting time. Within organizations that provide research infrastructure, DMPs assist in the planning of data curation, managing resource requirements for its storage and processing, and satisfying the requirements of funders and legal authorities. However, the creation of DMPs and FAIR data is an administrative burden, and different approaches exist to ease this burden through automation (Miksa et al. 2022). The creation of DMPs also asks for distinct knowledge of FAIR data in the community and of the infrastructure that creates research data in the facilities.

This paper has been written in the context of the EOSC projects PaNOSC (*The Photon and Neutron Open Science Cloud (PaNOSC) – Panosc, no date*) and ExPaNDS (*ExPaNDS is the European Open Science Cloud (EOSC) Photon and Neutron Data Service., no date*). Both projects together represent a significant portion of the PaN facilities in Europe. Both projects aim to provide a framework for PaN facilities to enable the production and use of FAIR data from experimental processes. The use of DMPs forms a significant part of that framework, which is reported in detail in Bolmsten et al. (2021), Görzig et al. (2021), and Görzig et al. (2022). This paper summarizes this work and places it within a wider project management framework.

Further, while we focus on PaN facilities, the approach, data model, and workflows should also be applicable to other infrastructures, where the research equipment providers or maintainers support visiting researchers using the equipment.

This paper will first introduce the Research Data Management (RDM) requirements of PaN facilities, introduce a data model that allows the (semi-)automation creation, execution, and use of DMPs, and conclude with the introduction of a DMP system. The following section will first introduce RDM and DMP requirements in PaN facilities, introduce a project management method, and show how to apply this project management method on information flows in PaN facilities.

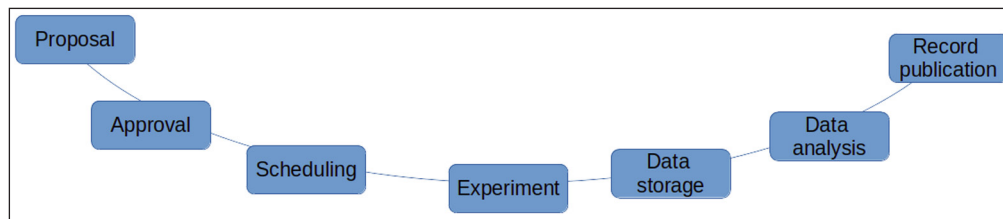
## 2. PaN RDM REQUIREMENTS AND PROJECT MANAGEMENT

### DMP INFORMATION REQUIREMENTS IN PaN FACILITIES

We first introduce the DMP requirements in PaN facilities by giving an overview of research workflows, involved roles, and infrastructures. In this section, the requirements of data handling in PaN facilities will be explained by listing the general requirements given by facility data policies and requirements resulting from PaN specific recommendations on FAIR data.

Inside the PaN facility, a typical workflow for facility experiments that will take place has been elaborated in Matthews et al. (2012), where the roles and workflows in different stages of the research lifecycle in PaN facilities were analyzed. Typically, the first contact of a researcher with

a PaN facility is the facility’s user office. The researcher must submit a proposal to request a slot for measuring time on a specific instrument provided by the facility. This is assessed for scientific merit and feasibility, and after the approval and scheduling of time on the instrument, the experiment will be executed. Then, the data will be generated, stored, and analyzed, and the results ultimately published (cf. [Figure 1](#)).

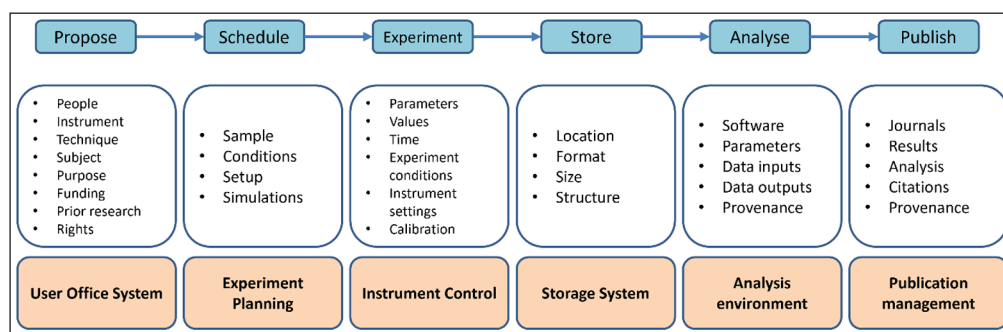


**Figure 1** An idealized facility research lifecycle, simplified (Matthews et al. 2012).

The major actors involved in the research lifecycle are listed below (Matthews et al. 2012):

- **The Experimental Team:** a group of largely external (e.g., university-based) researchers who propose and undertake the experiment.
- **The User Office:** a unit within the facility dedicated to managing external users of the facility. User Office staff and systems will typically register users and process their applications for beam-time, in addition to arranging their visit to the instrument.
- **The Instrument Scientist:** a member of the facility’s staff with specialist scientific knowledge of the capabilities of a particular instrument or beamline and its use for sample analysis.
- **Research Data Managers:** within the facility support the collection and curation of data and are becoming more prominent as this task gains importance.

During the different stages of the research lifecycle, these actors interact with various information systems supported in the facility. Depending on the different instruments, these information systems and roles will have to be considered when implementing a system that supports the creation of DMPs (cf. [Figure 2](#)).



**Figure 2** Metadata collected and information systems supporting the stages of the Experimental lifecycle.

Data curation, validation, and reporting have been identified as main goals to achieve with active DMPs, known as aDMPs (Bolmsten et al. 2021; Görzig et al. 2022). Validation comprises validation against the existing standards for the different PaN measurement techniques, but also validation against data policies. Reporting should be used to plan infrastructure usage in PaN facilities, as well as meeting funders’ requirements (Görzig et al. 2022). It is specifically stated that for data curation, information should be reused as much as possible, which can be done by also reusing pre-existing information available before a project proposal is even submitted.

Most PaN facilities in Europe have adopted data policies for their research data during the last decade. These data policies detail embargo conditions, open access requirements, access rules while under embargo, and the applicable licence for the research data (Wilson et al. 2011). Some facilities have updated their policies, integrated the usage of FAIR data, and included the obligation of DMP creation in their data policy (Götz et al. 2020; McBirnie et al. 2021).

Further, recommendations for minimal metadata within FAIR Photon and Neutron Data management have been elaborated in Salvat et al. (2020) and Soler et al. (2022). These recommendations are guided by the FAIR Data Maturity Model, or FDMM (RDA FAIR Data

Maturity Model Working Group 2020) to prioritize metadata as *essential, important, and useful*. These recommendations are meant to concretize the information required for re-usability in PaN, on top of specifications in the FDMM. We give the information considered as essential and important for experimental data in PaN as follows:

Essential:

- Visiting experimental team (user id)
- Experiment date
- Sample information
- Instrument information
- Calibration information

Important:

- Experimental planning
- Environmental parameters
- Laboratory notebook
- Instrument scientist

## PROJECT MANAGEMENT STRUCTURES AND PHASES

While the previous sub-section described requirements on DMPs and FAIR data in PaN facilities, this sub-section will introduce a project management method that will later be used to form a strategy for organizing information about RDM in PaN facilities.

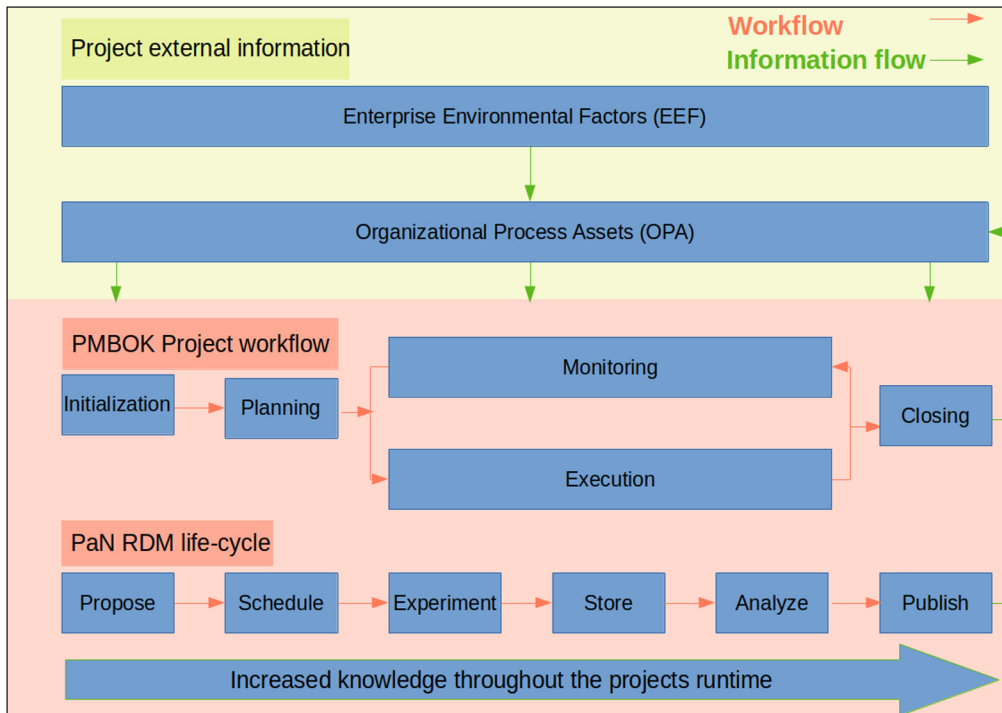
PMBOK® (Project Management Body of Knowledge Methodology) helped to structure and phrase the description of our FDM workflow; it is a project management standard and reference of the US-American Project Management Institute, PMI (PMI 2013). In PMBOK®, a project is divided into five process groups: initiation, planning, execution, controlling/monitoring, and closing—creating a basic projects' workflow. Here, we describe two key characteristics of the method used for the purpose of organizing information.

The first characteristic considers the availability of project knowledge, which increases throughout the project's runtime and can be divided into four phases. There is initial knowledge; planning knowledge, which can be updated during experiment execution; knowledge generated during the execution of a project; and knowledge generated after a project when it is finalized. With the knowledge generated in the final stage, the organizational knowledge might be updated (Muldoon 2014). How project knowledge about RDM increases during the project's runtime has been described in Giarretta (2015), cf. Figure 3.

While the first characteristic organizes the knowledge along the project's runtime, the second characteristic organizes the knowledge by its origin and changeability. In PMBOK®, knowledge to be managed can be divided into: user/project knowledge; organizational knowledge, in PMBOK® called Organizational Process Assets (OPA); and general knowledge concerning laws, policies, and standards produced outside the institution, in PMBOK® called Enterprise Environmental Factors (EEF). EEFs are independent of the organization and evolve over the time, but are quite stable during a project. This knowledge is generated before a project starts. Organizational knowledge is generated in the organization independently from the project, and general knowledge is incorporated into organizational knowledge. Organizational knowledge evolves as instruments, software, and institutional policies change, for example. Organizational knowledge exists before a project is initiated (Muldoon 2014).

## PaN RDM LIFECYCLE AND PROJECT MANAGEMENT STRUCTURES AND PHASES

In this sub-section, information sources and phases for DMP information are described as they were analysed in Görzig et al. (2021). Then, we will explore how the aforementioned project management phases and structures can be used to analyze an update of pre-existing information on research infrastructure and normally created data.



**Figure 3** Increasing knowledge throughout projects' runtime and information flow.

Information sources for DMP creation related to DMP questions have been investigated in Bolmsten et al. (2021) and Görzig et al. (2021). The first paper defined a DMP template suitable for facility users modifying the questionnaire from the Research Data Management Organiser, RDMO (Michaelis et al. 2021), to adapt it to the requirements in PaN RIs. In the latter paper, this template has been reviewed, and suitable information sources for the answers to the questions in the DMP were identified and assigned to different phases in the PaN research lifecycle (see Table 1). The review of information sources shows that a high percentage of the questions could be already filled by reusing pre-existing information available before a project proposal is even submitted.

DMP PHASE	ACTOR PROVIDING INFORMATION FOR THE DMP
0 Before proposal submission	Typically, knowledge of instrument from scientist or RDM team (static parameter).
1 Proposal submission	Typically, knowledge of the researcher, with support from the facility administration and RDM team.
2 Accepted experiment planning	Typically, knowledge of the researcher, 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 1** DMP Phases.

In the following, the lifecycle of pre-existing information with its rated roles, IT systems, and phases will be elaborated. Applying the knowledge sources and phases from Görzig et al. (2021) to the PMBOK®'s phases results in Table 2. Table 2 structures the information required for RDM in the PMBOK®'s phases with the additional phase 'Before project' and related roles. During the project phases *initiation*, *planning*, *execution*, and *finalisation*, the pre-existing RDM information is mostly used by project personnel to create or execute the DMP, and project specific information is generated, while in the *before and after* phases, facility personnel are evaluating information generated by the project and updating the pre-existing information collected previously.

The knowledge held by the data manager is mostly about standards, metadata schemata, and policies. Part of the information contributing to the knowledge is independent from the facility and evolves inside the scientific community, with funding regulations, or laws. This knowledge could be shared with other facilities and be classified as EEF. Examples are:

ROLE	BEFORE PROJECT/ OPA/EEF	PROJECT INITIATION	PROJECT PLANNING	PROJECT EXECUTION	PROJECT FINALISATION	AFTER PROJECT/ PA
Instrument scientist	Instrument/software description, selection of applicable metadata standards, general dataset description		Required project specific software, used instrumentations and their configurations, and standards			Adding/ actualisation of instruments and software information
Data manager	Controlled vocabularies and standards administration, mapping metadata to standards, general data policies, policy execution				Automatic metadata extraction and validation	Open access of research data, validation of policy execution; actualization of standards and policies
User office		Proposal information, instrument to be used, (co-) proposers				
Experimental team (research)			Concrete dataset description, references to additional information, metadata schema selection, estimated amount of datasets produced, dataset usage, special (own) software infrastructure requirements	Experiment execution: parameter and configurations	Dataset selection, metadata completion, and validation	
Experimental team (administration)		Specific policies and DMP requirements for project, funding, participating researchers	DMP actualization	DMP actualization after experiments	DMP actualization	

**Table 2** PMBOK®'s phases and roles in RDM information collection.

- Funder Policies including FAIR data requirements and Open Access requirements
- Laws like EU General Data Protection Regulation (GDPR)
- Scientific technique specific requirements on (FAIR) data
- General validation schemata for metadata
- Software for dataset usage

The data manager has an overview of e.g., how a specific standard is applied inside the facility and about facilities policies. The latter can evolve as facility's internal discussions lead to additions and changes, and the knowledge can be classified as OPA. A detailed overview can be found in the next section (cf. [Table 3](#)).

The other part of an organization's internal OPA is the information held by the instrument scientist. Here, the information evolves as changes to the instrument or software are made. Whilst the data managers have an overview about the standards and metadata used in the facility, the instrument scientists know which standards apply to the data measured at their specific instruments. The instrument scientist has knowledge about the typical characteristics of the datasets created by the instruments. A detailed overview can be found in the next section (cf. [Tables 4, 5](#)).

For some use cases, especially when automation and reproducibility are related, this information might be very detailed and underlie frequent changes. Therefore, this information often requires updates. How this can be supported by automation will be discussed in section 4.

The following section will focus on data modelling to store the pre-existing information that exists before a project starts and can be reused throughout various projects. The data manager and the instrument scientists are the main actors, holding information that exists before a proposal is submitted. Normally, the facility data policy applies to all projects and can be seen as a source of almost static, pre-existing information. Additionally, to the data policy requirements, other facility-wide guidelines and workflows can contribute to the pre-existing information held by the data managers. The instrument scientists have more concrete information about the instruments and the data they produce. The pre-existing organizational knowledge of data managers and instrument scientists will have to be connected and merged with information that arises during the proposal approval and execution of the experiment.

### 3. DATA MODEL FOR PRE-EXISTING INFORMATION AND AUTOMATION

In the previous section, the requirements for DMPs and the sources of information related to phases and origins have been introduced. To fulfill the envisioned tasks of data curation, validation, and reporting, in this section we describe the data model required to hold the information to comply with these tasks and support the creation and execution of DMPs. The first part will concentrate on the information and its meaning by giving some background; the second part will introduce an ontology describing the relations between the entities discussed before. After the pre-existing information about the research infrastructure and normally created data is modelled, the last sub-section will describe what actions in which phases of a projects/proposal's lifecycle need to be executed on the data model to use and enrich the information.

#### DATA MODELLING

In this sub-section, data models for facility, dataset, experimental techniques, projects, and policies will be introduced, taking as a starting point the questionnaire discussed in Bolmsten et al. (2021) and Michaelis et al. (2021). In the questionnaire, entries from DMP phase 0 containing information existing before proposal submission were combined with entries where only the RDM team as a supporting source were selected. The following tables hold the most basic entities and their description.

#### Facility information

These entries were the base for the data model for the facility information. The sources of this information are mostly facility data policies and other conventions in the facilities (cf. Table 3):

FACILITY INFORMATION	
repository	The repository information comprises the name and access URL where the data is made accessible.
licence	The license usually applied to the data in the repository.
security	Information about e.g., backups and replicas of the data and other special security information.
pid_system	The default PID system applied in the repository e.g., handles or Digital Object Identifiers (DOIs).
personal_data	In case personal information in research data is treated on a facility level, e.g., no personal information is allowed in research data more than required for provenance.
min_storage_period	The minimum period research data has to be available for good scientific practice.
archive	Data archive used. If it is the same as the repository, then no URL needs be provided, as the access procedures have to be described.
certificate	If the repository is certified and with which certificate e.g., CoreTrustSeal.

Table 3 Facility wide metadata.

(Contd.)

## FACILITY INFORMATION

arrangements	For the data produced in a research project, an arrangement with the data repository that will receive the research data has to be made. In case a proposal in PaN facilities is approved, it normally includes the usage of the repository.
embargo_period	In the data policy of the PaN facilities, there is also an embargo period defined.
access_control	How the access to the data repository is controlled.
costs	In case a proposal in PaN facilities is approved, it also normally includes the costs for research data management.

## Scientific technique information

Knowledge about the scientific techniques used in the experiment, with associated metadata standards and formats, are another area supervised by the data manager. This information tends to be stable for a set of experiments, although it evolves over time and grows as more standards are defined for particular techniques and others refined. Most of the information is independent from the facility and defined by the scientific community expert in those techniques.

The scientific technique very much dictates the structure and the semantics of a dataset. As seen below (cf. [Table 4](#)), the data model for scientific techniques, metadata standards, and formats does not have many entries, but its importance when organizing the data models is apparent, as the scientific technique prescribes the usage of metadata and formats. This will become clear in the next section about automation.

technique	Describes a scientific technique.
name	A name or label of the technique.
PID	E.g., the IRI in the PaNET Ontology ( <a href="#">Collins et al. 2021</a> ).
metadata schema	Related to the technique are the requirements on metadata. There can be more than one metadata schema and format used in practice.
structure	
format	
tools	
reading	Software for possible usage.
writing	Possible software for writing the format.
validation	Tools for validating the data against e.g., NeXus application definitions. <sup>1</sup>
validation schema	Schema complies with metadata schema above, used for validation with a tool.

**Table 4** Metadata for scientific techniques related to metadata schema and file format.

## Dataset information

The information about the dataset generated in the experiment that can be contributed by the instrument scientist is less stable. It comprises basically the description of the dataset, and the infrastructure required for its creation, usage, and some DMP workflow-related information. The most basic components are typically hierarchically structured; the dataset contains one or more file-collections. A file-collection comprises files that are the output of one software that creates the files.

[Table 5](#) shows the minimum information required for a dataset to meet DMP and FAIR data requirements. This includes related software, hardware, and instruments, for example.

<sup>1</sup> NeXus application definitions are used to define the structure and semantics of a file for a certain application. They normally correspond to a measurement technique: <https://manual.nexusformat.org/classes/applications/index.html>.



DATASET	
name	a default name
description	a project independent description of the dataset that can be adapted in projects
contributor	contributing persons, typically identified via ORCIDs
reproducible	if the dataset is reproducible and under which efforts
interested_community	might be derived from disciplines
usage	there might be some default usage scenarios, like calibration; otherwise the data sets' intended usage in the project
archival	DMP question; moment, selection_criteria, and long_term_archival_reason might be used for automated execution and validation
data_security	measurements and responsible person
techniques	scientific techniques used to create the dataset
filecollections	a collection of files created by one software instance; a dataset can contain more than one filecollection
name	a default name
resource	instrument or laboratory used to create the filecollection; preferable identified by PID
storage	location and access to experimental storage
backup	location and access to experimental storage
quality_assurance	description and pointers of e.g., validation workflows
hardware	description of hardware components used to create the dataset; used for data curation
writing_software	description of software and its components used to create the dataset; used for data curation
files	files
name	can be a regular expression definition of default file names
format	the format of the file (could be related to a format registry and relates to the technique table above)
metadata_schema	the metadata schema applied in the file (could be related to a metadata schema registry and relates to the technique table above)
size	expected minimum and maximum size of the file; average size
amount	quantity of files; can be used together with size for estimating overall size and validation
processing_requirements	hardware and software requirements for processing the data
hardware_requirements	
type	type of hardware requirements like storage or processors of a certain type of computer; manufacturer and model are required
reading_software	possible software to use the data, including access and documentation, as well as required plugins
name	
PID	
type	
documentation	
URL	
plugins	
name	
type	
URL	

**Table 5** Metadata for datasets.

## Policy information

The DMP themes of the Digital Curation Centre (DCC), the University of California Curation Center (UC3), and the RDA Practical Policy WG break down RDM activities into a limited set of policies and resulting activities, although these may originate from a multiplicity of policy documents (Moore et al. 2015; DCC and UC3 2016).

The RDA WG Practical Policy has analyzed policies implemented in 30 different data management systems and identified 11 generic policies that were of interest (Moore et al. 2015): Contextual metadata extraction, Data access control, Data backup, Data format control, Data retention, Disposition, Integrity (including replication), Notification, Restricted searching, Storage cost reports, and Use agreements. Of this list, the *Contextual metadata extraction* and the *Data format control policies* are of special interest, as the need to validate against FAIR data requirements and against specific metadata schemata and PIDs schemas are part of the DMP requirements in Görzig et al. (2022). Other policies, such as data backup and use agreements, are inherent in the data management system or managed outside the data repository or management system, and thereby are out of control of the data producers.

For each of the listed policies, templates have been provided containing (Moore et al. 2015):

- Policy name;
- Example constraints that control application of the policy;
- State information that is needed to evaluate the constraint;
- Example operations that are performed by the policy;
- State information that is needed to execute the operations.

Table 6 below shows the minimum information required to find and execute a policy in an abstract way as described above.

POLICY	
name	The name of the policy.
constraint	When the operation should be executed.
type	Event trigger or scheduled.
value	Triggering event e.g. onCreation or date and time.
parameters	Array of required parameters like path to a file or metadata schema for validation. The parameters are divided in input and output parameters.
operation	Policy related operation or workflow (referencing an executable workflow).
categories	For finding the operation, e.g., validation, integrity, format, extraction, interoperability.
description	A textual description about what the operation does.

Table 6 Metadata for operations.

## Project information

The information mentioned in the previous sub-sections exists independently of a project and is rather stable. Therefore, it can potentially be reused throughout various projects. Apart from the experiment data, the project information is the least stable information. It changes for each project, but within a project, it can still be reused during its runtime. For modelling the project information again, the RDMO questions (Michaelis et al. 2021) and also the applied data structure for proposals in the ESS and the RDA DMP-common-standard (Miksa et al. 2020) have been reviewed. The table (cf. Table 7) below shows the minimum information required to find and describe a project and start a workflow for applicable policy and dataset retrieval.

## RELATED ONTOLOGY: JOINING THE INFORMATION

In the next step, the entities described above will be related to each other by creating an ontology describing the DMP information concepts and their relationships (Görzig 2022). Apart from describing these relationships, the ontology will allow future integration with other

PROJECT:	
name	name of the project
description	project description
funding reference	the usage of a PID is advisable for later curation and integration into a graph model
members	here as well the usage of ORCIDs is advisable
start_date	start of the project
end_date	end of the project
disciplines/keywords	to retrieve RDM requirements and improve findability of the data
jurisdictions	to retrieve policy requirements; jurisdictions can be funders, national, institutional, or a laboratory/instrument
resource	instrument or laboratory used to create data; used to retrieve possible dataset types created by the resource

Table 7 Metadata for projects.

ontologies and metadata schemata and facilitate the retrieval of information. Therefore, it will still need to be related to terms in standard ontologies. First, the dataset will be described, then the dataset will be listed, and then the project and policies.

### Dataset

The basic components of a *Dataset* are *Filecollections* that consist of one or more *Files*. *Dataset*, *Filecollection* and *File* are sub-classes of a *Digital object*. A *Dataset* is created in the environment of an instrument or a laboratory, in the following generalised to *Resource*. A *Resource* and its components are normally managing a stack of *Hardware* and *Software*. A *Filecollection* is created by one *Software* instance running on a *Hardware*. The *Filecollection* is normally also the collection of *Files* that can be processed by a *Software*. The number and sizes of *Files* forming a *Filecollection* can be predefined, as well as the *Format* and *Metadata schema* of the *Files*. Additionally, a basic general independent description of the *Dataset* can be provided. This description can later be used as a template for a *Project* specific description. The relations of the *Dataset* class are summarised in Figure 4.

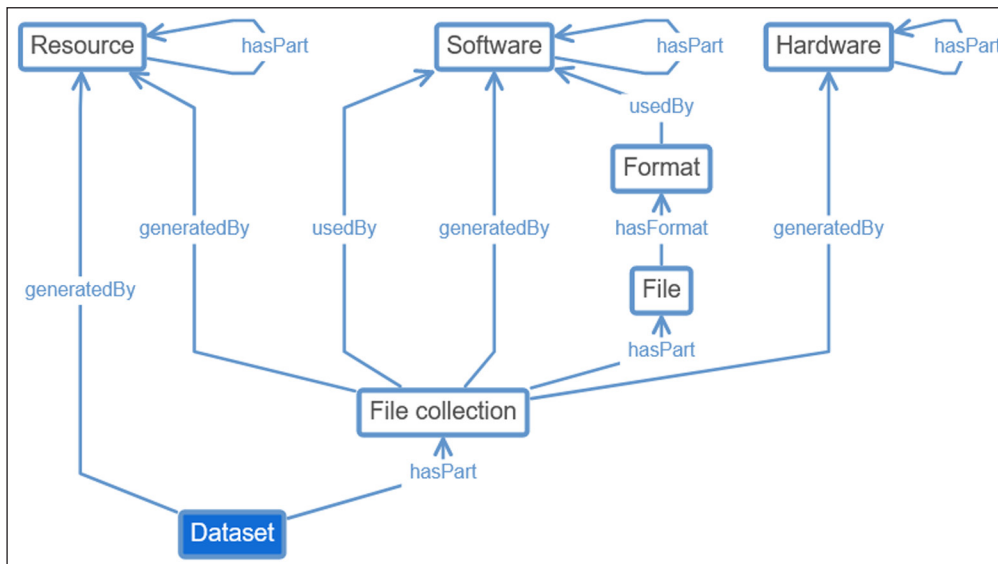
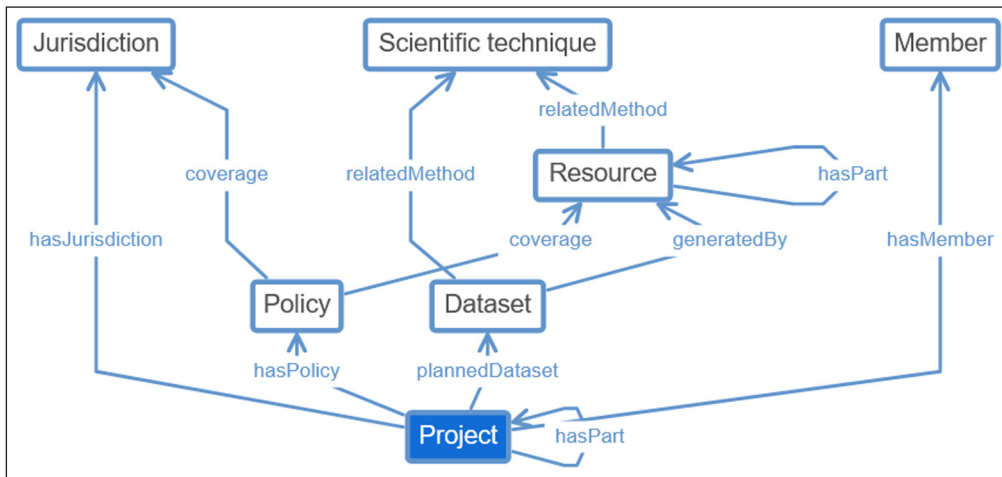


Figure 4 Relations of the Dataset class.

### Project – Datasets

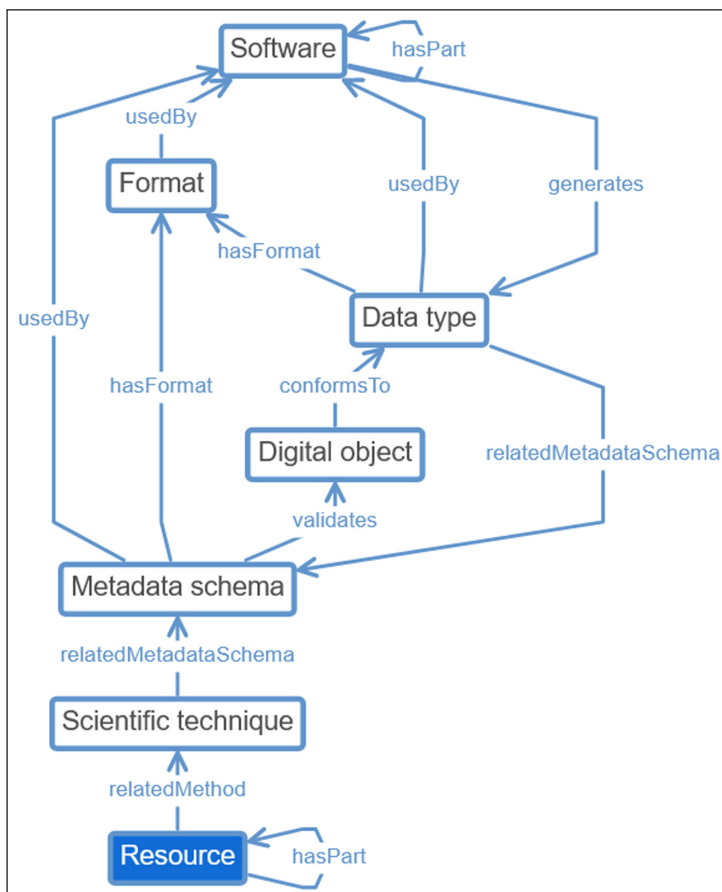
A *Project* or proposal as it is typically called in PaN facilities, is the top-level entity aggregating all related *Datasets*. *Projects* can be organised in a hierarchy representing for example, the work packages of an overlaying funded *Project*. When a *Project* is planned, there is information on which *Resource(s)* will be used for *dataset* creation. Thereby, knowledge about the *Datasets* that are planned will be available and can be used in the DMP. The relations of the *Project* class are summarized in Figure 5.



**Figure 5** Relations of the Project class.

### Scientific technique – Software

Tables 4 and 5 show that the *Scientific technique* is a central entity connecting experimental *Resources*, with *Metadata schemata*, *Formats*, and *Software*. A *Resource* can be used to conduct experiments with a certain *Scientific technique*. The *Resource* is chosen by a *Project* because of the *Scientific technique* it supports. The *Scientific technique* applied in an experiment defines what kind of metadata and data are produced, and which analysis methods can be applied to the data it generates. For the generated metadata and data, a *Metadata schema* can be applied when available. The semantics are described in a *Metadata schema* and depends on the *Scientific technique*. Therefore, the information a *File* or *Filecollection* should contain is prescribed via the *Scientific technique*. Nevertheless, the semantics of the *Scientific technique* can be expressed with different terminologies and in differing serializations, requiring different *Metadata schemata* and *Formats*; for example tomography measurements can result in a series of tiff images or images in a NeXus file, and the required information to process the data will stay the same. Thereby, the *Scientific technique* also defines which *Operations* can be run on a *File* or *File collection*. The relations of the *Resource* class and its associated techniques are summarized in Figure 6.



**Figure 6** Relationships associated with the Resource class and its associated technique.

In order to inform users and facility staff what *Operations* can potentially be executed by a *Software* on a *File* or *Filecollection*, and respective *Software* can be found, the *Data type* has been introduced. The *Data type* can be abstract or concrete, e.g., if a *Data type* ‘tomography images’ can be related to a specific set of *Operations* that can be executed potentially on all tomography data. To execute these *Operations* on tomography data, the concrete *Metadata schema* and *Format* of the images needs to be known. With this knowledge, a concrete sub-*Data type* of ‘tomography images’ can be created, e.g., ‘tomography images tiff beamline XY’. This concrete *Data type* can then be related to concrete *Commands* to execute the *Operation*. An *Operation* can be executed on a *Data type* using a concrete implementation with specific *Software* and *Commands*.

### Policies – Operations

*Policies* can be related to *Resources* and/or *Projects*. They result in executed *Commands* on *Datasets*. *Policies* are normally described in policy documents. *Operations* differ in their parameters among the documents e.g., the embargo period might be three years or five years. To execute a policy, a chain of *Commands* might have to be executed. The *Operation* itself has an input and an output scheme, which represent expected input and result. The input and output of the executed *Command* need to be validated against these schemes.

## DMP SUPPORT – PRE-EXISTING INFORMATION LIFECYCLE – KNOWLEDGE TRANSFER AND (SEMI-) AUTOMATION

The pre-existing information should be used to pre-fill DMPs. Pre-filling of DMPs or semi-automated DMP creation has already been implemented e.g., at the Radbound University, where the DMP creation has been integrated to the CRIS system (Jetten et al. 2019); and ARGOS (Romanos et al. 2019), which pre-fills DMPs using repository information; and at the University of Vienna, integrating the DMP creation with various IT systems (Cardoso et al. 2021). In these examples, CRIS and data repositories have been used to pre-fill the DMPs. In this paper, pre-existing information from laboratory or instrument infrastructure (previously named *Resources*) will be used to pre-fill the DMP. To hold the pre-existing information that has been introduced earlier, in this sub-section, the lifecycle of pre-existing information will be presented. Firstly, activities will be shown that are required for the maintenance, usage, and exposition of pre-existing information that is stored in a central knowledge base. Then, the filling and maintenance of the central knowledge base and consequently the usage for DMPs will be described.

The lifecycle of the pre-existing information includes the phases: Before project (OPA/EEF), its initiation, planning, execution, finalisation, and after project (OPA). The central knowledge base will be used for projects’ DMP creation, exposition, and execution. Before a project starts, the central knowledge base needs to be filled with the required information. This information might have to be updated during and after the end of the project. Table 8 lists the phases and related activities:

LIFECYCLE PHASE	ACTIVITIES
Before project (OPA/EEF)	Retrieve initial information for central knowledge base about Datasets of Resource from repository.  General update/insert knowledge base: <ul style="list-style-type: none"> <li>• Metadata standards</li> <li>• Formats</li> <li>• Policies</li> <li>• Mappings</li> </ul>
Project initiation	Insert new project  Relate to resource
Project planning	Specify projects datasets: <ul style="list-style-type: none"> <li>• Create project specific datasets</li> <li>• Relate to default datasets of resource</li> <li>• Update description</li> </ul>

**Table 8** Pre-existing information lifecycle phases and related activities.

LIFECYCLE PHASE	ACTIVITIES
	Create DMP: <ul style="list-style-type: none"> <li>Retrieve projects datasets and related information from central knowledge base and insert into a DMP tool</li> <li>Retrieve facility specific information</li> <li>Update information in DMP tool</li> </ul> Create concrete policy execution environment: <ul style="list-style-type: none"> <li>Retrieve to datasets related operations into execution environment</li> </ul>
Project execution	Update DMP <ul style="list-style-type: none"> <li>Retrieve projects datasets repository and update information in DMP tool</li> </ul> Execute operations on datasets
Project finalization	Update DMP Execute operations Select datasets for archival Update concrete dataset descriptions
After project (OPA)	Validate pre-existing information against projects datasets in the repository: <ul style="list-style-type: none"> <li>Update central knowledge base</li> </ul>

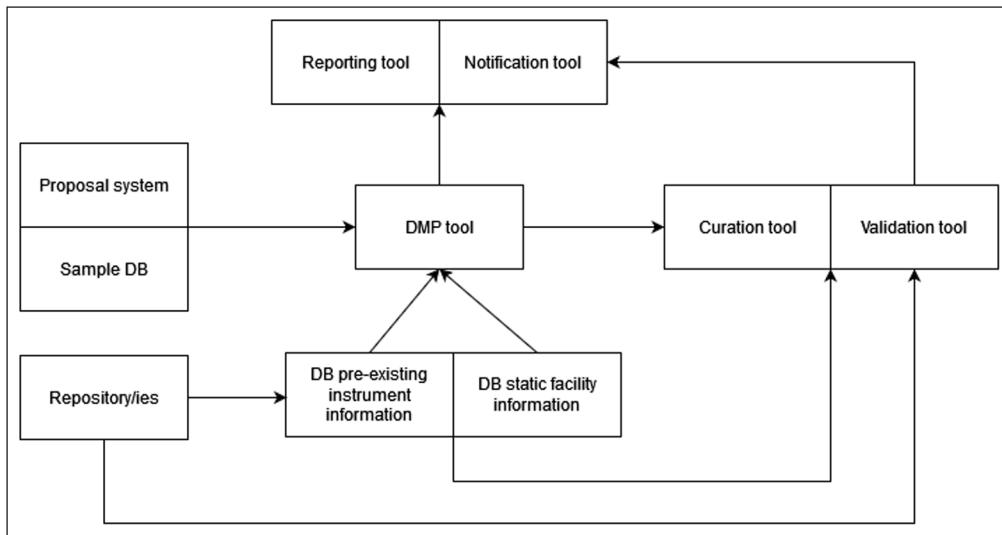
In order to start filling the central knowledge base for the first time, datasets ingested into the data repository need to be analyzed and relevant information extracted. Relevant information to be found in the repository is: *Data types* produced at a *Resource* and their *Formats*; in case a standard is used, this can be extracted. Also, average quantity and sizes of files can be retrieved. In case standardised formats are used, metadata from the files can be obtained and eventually be related to *Scientific techniques*, *File sizes* and quantities. When creating a DMP, this information is frequently reviewed and can be updated in the central knowledge base.

When a DMP is created, project specific information needs to be provided. Therefore, firstly information can be acquired from e.g., the facilities proposal or other administrative systems. In a second step, dataset information from the selected IT systems can be requested from the central knowledge base and be concretized manually. The concretization can be based on e.g., option lists provided by the central knowledge base. In the ExPaNDS deliverable 2.4 (Görzig et al. 2021), the IT systems and roles used as information sources to create a DMP are listed. For creating a DMP, normally a DMP tool such as the Data Stewardship Wizard, DSW (*Data Stewardship Wizard*), or the RDMO (*RDMO*) are used. These two DMP tools have been analyzed in the ExPaNDS deliverable 2.8 (Görzig et al. 2022) for having a central data structure. This central data structure is used to allow separate views for filling in the DMP questions and provide DMP information as required e.g., in the funder template. The data structure of the central knowledge base and the proposal or administrative system will need to be mapped to the central data structure of the DMP tool.

## 4. DMP SYSTEM

This section proposes a design of an integrated aDMP system that supports the (semi-) automated generation and execution of DMPs by maximising the reuse of information generated during the research lifecycle, while supporting the scientists and other stakeholders by transferring knowledge between stakeholders and IT systems.

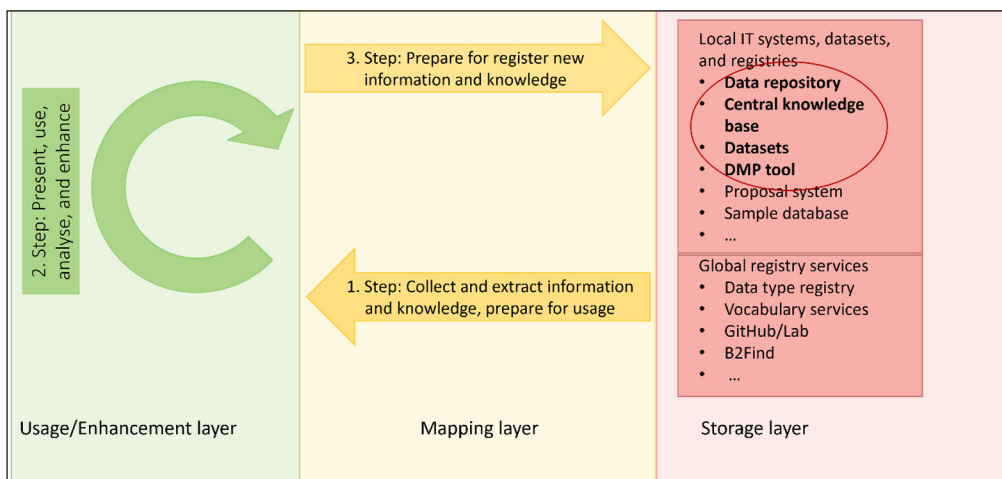
In Miksa, Oblasser and Rauber (2022), services required to automate DMPs are described. They analyzed tasks to be done when creating DMPs to design a high-level workflow to implement a system supporting machine-actionable DMPs. They have identified 13 application services with a varying degree of complexity, some institutional, and some shared with the outside world (Miksa et al. 2022). Existing IT systems that can contribute to the DMP creation have been analyzed in Görzig et al. (2021). In Görzig et al. (2022), components and their relations of a DMP system that are required to fulfil aDMP requirements as described in section 2 have been identified (Figure 7).



**Figure 7** Components for aDMP system (Görzig et al. 2022).

In most PaN facilities, the *proposal system* and the *data repositories*, and sometimes the *sample DB* are systems already in production. Most of them provide an API to access the data; *Curation* and *validation tools* may also exist for some data formats such as NeXus. The reporting and DMP tools are mostly integrated. PaNOSC implemented a DMP system, which integrated the proposal system with the DSW tool (Bodin et al. 2023). This represents a partial implementation of our proposed method, the missing components being the DataBase for holding pre-existing instrument information and the DataBase of static facility information, and their integration into the IT infrastructure.

The data repository with its catalogue and datasets, the DMP tool, and the proposal system additionally to the central knowledge base need to be considered when designing a dataflow between the systems that function as information sources and recipients. Beside them, the DMP tool, datasets, and the data repository, as well as the central knowledge base, are knowledge bases that are both source and recipients of further and enhanced information. Other IT systems that can contribute towards enhancing information may include external data and software repositories, data catalogues, vocabulary services, and data type registries, for example. The cycle of storing, enhancing, and using the information is shown in the image below (Figure 8).



**Figure 8** Pre-existing information enhance and use.

This cycle involves three layers in the architecture. In the storage layer, all the information is stored. This information needs to be extracted from this layer via APIs or metadata extractors and prepared for usage. This requires a mapping layer where mapping and joining information from different sources to the serializations and data formats required for usage and enhancement. Usage and enhancement happen in the usage layer, where for example, curation, validation, and notification tools, supporting GUIs from DMP tools, and metadata editors process the data either automatically or manually with human intervention. The newly generated information then needs to be fed back into the data layer.

The first part of this paper shows how DMP relevant information is connected to organizations and has its own lifecycle that is independent from a project where the DMP is required. This information can be collected and maintained by organizational or facilities staff, for example, such as data managers and instrument scientists to be used by the projects, which trigger an update of the existing information. PMBOK served here as a method to structure the information of the source of its origin and its emergence in time. In the second part, a data model is introduced that can hold this pre-existing information. This data model is divided in three main sections, the facility specific information, the information around a *Scientific technique*, and the *Dataset* information (linked to a *Resource*). To create a DMP, this has to be linked to a *Project*; and for DMP execution, *Policies* have been introduced. In order to have a machine-readable serialisation of the information to fill a DMP, an ontology has been introduced. The paper concludes with an outline of a DMP system that supports the usage of the pre-existing information for DMP creation, execution, and validation. It builds upon existing facility infrastructure with new DMP relevant components in its center. But it also outlines how an information base with pre-existing information can be enriched semi-automatically to support the maintenance of this information, and to reduce the additional input required from the user, thus reducing the significant barrier for use scientists may experience when asked to complete a DMP.

### NEXT STEPS

Future work is proposed in two areas: firstly, the integration of the internal inherent data models of the existing infrastructure with each other's inclusive the adoption of interfaces to access the information from external systems to present, use, analyze, and enhance the metadata; and secondly, the integration of the presented DMP ontology with standard ontologies and vocabularies.

The components of the DMP system presented here normally work with data models designed for their respective applications; we propose that they should be integrated with each other. Therefore, their internal data models will have to be mapped with the presented DMP ontology, and interfaces for their integration will have to be created. Initial work on this is presented in Görzig et al. (2022).

The proposed DMP ontology is still under development and needs to be integrated with existing data formats and standards and non-discipline specific ontologies. Appropriate concepts for the terms used in the presented DMP ontology can be found in ontologies such as DCAT (Gonzalez-Beltran and Winstanley 2022) and PREMIS (Di Iorio and Caron 2016). These terms will have to be mapped. For integration with DMPs, the RDA DMP common standard (Miksa et al. 2020) can be applied, and on a more PaN specific level, the data model described in Görzig et al. (2022) should be taken further. Other PaN relevant standards that need to be considered are: the the NeXus format standard for measurement files in PaN facilities (Könnecke et al. 2015); related scientific techniques such as for spectroscopy described in Hanson et al. (2022); and analytical chemistry in Rauh et al. (2022).

In the data repositories and their catalogues, different standards are used. For discovery purposes, DataCite metadata, an OAI-PMH interface, and a common search API are normally implemented (Richter et al. 2020). Further, several PaN facilities operate the ICAT as a data catalogue and repository, which uses a common central metadata schema (ICAT; Flannery et al. 2009; Matthews et al. 2009). In addition to the ICAT metadata schema, other standards-based metadata are under development e.g., PaN ontologies NeXus and PaNET (Collins et al. 2021). The data model in ICAT and also SciCat—another popular repository tool used in PaN sciences (SciCatProject · GitHub, no date)—should be reviewed to maximize the information required for DMPs. For execution and validation of the DMP, the operations in the DMP ontology have been defined and will need to be mapped to a workflow language such as CWL (Crusoe et al. 2021).

### ACKNOWLEDGEMENTS

We would like to thank all our colleagues in the European Open Science Cloud (EOSC) Photon and Neutron Data Service (ExPaNDS), and The Photon <https://expands.eu/> and Neutron Open Science Cloud (PaNOSC) projects. <https://www.panosc.eu>.



This work supported by the project European Open Science Cloud Photon and Neutron Data Services (ExPaNDS) <https://expands.eu/> received funding from the *European Union's Horizon 2020 research and innovation programme* under Grant Agreement Number: 857641, in collaboration with the project Photon and Neutron Open Science Cloud (PaNOSC), under Grant Agreement Number: 823852.

## COMPETING INTERESTS

The authors have no competing interests to declare.


## AUTHOR AFFILIATIONS

**Heike Görzig**  [orcid.org/0000-0001-9121-8643](https://orcid.org/0000-0001-9121-8643)

Helmholtz-Zentrum Berlin für Materialien und Energie, Germany

**Alejandra N. Gonzalez Beltran**  [orcid.org/0000-0003-3499-8262](https://orcid.org/0000-0003-3499-8262)

Science and Technology Facilities Council, UK

**Felix Engel**  [orcid.org/0000-0002-3060-7052](https://orcid.org/0000-0002-3060-7052)

University of Hagen, Germany

**Brian Matthews**  [orcid.org/0000-0002-3342-3160](https://orcid.org/0000-0002-3342-3160)

Science and Technology Facilities Council, UK

## REFERENCES

- Bodin, M**, et al. 2023. Data management plans for the Photon and Neutron Communities. *CODATA Data Science Journal*, 22(1): 30. DOI: <https://doi.org/10.5334/dsj-2023-030>
- Bolmsten, F**, et al. 2021. DMP Template for facility users. DOI: <https://doi.org/10.5281/ZENODO.5639428>
- Cardoso, J, Castro, LJ and Miksa, T**. 2021. Interconnecting systems using machine-actionable data management plans – Hackathon report. *Data Science Journal*, 20(1). DOI: <https://doi.org/10.5334/dsj-2021-035>
- Collins, SP**, et al. 2021. ExPaNDS ontologies v1.0. DOI: <https://doi.org/10.5281/ZENODO.4806026>
- Crusoe, MR, Abeln, S, Iosup, A, Amstutz, P, Chilton, J, Tijanić, N, Ménager, H, Soiland-Reyes, S, GavriloVIC, B and Goble, C**. 2021. Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language. *Communications of the ACM*, 65: 54–63. DOI: <https://doi.org/10.1145/3486897>
- Data Stewardship Wizard*. Available at <https://ds-wizard.org/> [Last accessed 14 February 2022].
- DCC** and **UC3**. 2016. Proposed revised set of themes for Data Management Plans. URL: <https://www.dcc.ac.uk/sites/default/files/documents/tools/dmpOnline/DMP-themes-FINAL-Dec2016.pdf>.
- Di Torio, A** and **Caron, B**. 2016. PREMIS 3.0 Ontology: Improving semantic interoperability of preservation metadata. *iPres*. Available at [https://www.loc.gov/standards/premis/pif/2016/iPresDiTorio\\_Caron\\_iPRES\\_2016\\_paper\\_150.pdf](https://www.loc.gov/standards/premis/pif/2016/iPresDiTorio_Caron_iPRES_2016_paper_150.pdf) [Last accessed 19 April 2017].
- ExPaNDS is the European Open Science Cloud (EOSC) Photon and Neutron Data Service*. Available at <https://expands.eu/> [Last accessed 17 December 2019].
- Flannery, D**, et al. 2009 ICAT: Integrating data infrastructure for facilities based science. *5th International Digital Curation Conference (IDCC 2009)*, 201–207. DOI: <https://doi.org/10.1109/e-Science.2009.36>
- Giaretta, D**. 2015. *Related work on data lifecycle phases from formulation to exploitation* | RDA. Available at <https://rd-alliance.org/group/active-data-management-plans-ig/wiki/related-work-data-lifecycle-phases-formulation> [Last accessed 4 November 2022].
- Gonzalez-Beltran, A** and **Winstanley, P**. 2022. The data catalog vocabulary (DCAT). DOI: <https://doi.org/10.5281/ZENODO.6142906>
- Görzig, H**. 2022. DMP Pre-existing Information Ontology. DOI: <https://doi.org/10.5281/ZENODO.7437982>
- Görzig, H**, et al. 2021. DMPs for Photon and Neutron RIs. DOI: <https://doi.org/10.5281/ZENODO.5636096>
- Görzig, H**, et al. 2022. Active DMPs for Photon and Neutron RIs. DOI: <https://doi.org/10.5281/zenodo.7223438>
- Götz, A**, et al. 2020. PaNOSC data policy framework. DOI: <https://doi.org/10.5281/ZENODO.3862701>
- Hanson, RM**, et al. 2022. IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles. *Pure and Applied Chemistry*, 94(6): 623–636. DOI: <https://doi.org/10.1515/pac-2021-2009>
- ICAT**. *ICAT* metadata, data and processing. Available at <https://icatproject.org/> [Last 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. *Procedia Computer Science*, 146: 156–165. DOI: <https://doi.org/10.1016/j.procs.2019.01.090>
- Könnecke, M,** et al. 2015. The NeXus data format. *Journal of Applied Crystallography*, 48(1): 301–305. DOI: <https://doi.org/10.1107/S1600576714027575>
- Matthews, B,** et al. 2009. Using a core scientific metadata model in large-scale facilities. In: *5th International Digital Curation Conference*. London: 106–118. DOI: <https://doi.org/10.2218/ijdc.v5i1.146>
- Matthews, B,** et al. 2012. Model of the data continuum in Photon and Neutron Facilities. DOI: <https://doi.org/10.5281/ZENODO.3897190>
- McBirnle, A,** et al. 2021. Final data policy framework for Photon and Neutron RIs. DOI: <https://doi.org/10.5281/ZENODO.5205825>
- Michaelis, O,** et al. 2021. rdmorganiser/rdmo-catalog: 1.1.0-rdmo-1.6.0. DOI: <https://doi.org/10.5281/ZENODO.5747651>
- 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): 1–22. DOI: <https://doi.org/10.1145/3490396>
- Miksa, T, Walk, P and Neish, P.** 2020. RDA DMP Common standard for machine-actionable data management plans. DOI: <https://doi.org/10.15497/rda00039>
- Miksa, T,** et al. 2019. Ten principles for machine-actionable data management plans. *PLoS Computational Biology*, 15(3): e1006750. DOI: <https://doi.org/10.1371/journal.pcbi.1006750>
- Moore, R,** et al. 2015. Practical policy. *Research Data Alliance*. DOI: <https://doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC>
- Muldoon, JP.** 2014. *PMBOK® Summarized*. Available at <http://johnmuldoon.ie/wp-content/uploads/2014/08/PMBOK-Summarized.pdf> [Last accessed 1 July 2016].
- PMI.** 2013. *PMBOK® Guide—Fifth Edition*. 5th edn.
- Rauh, D,** et al. 2022. Data format standards in analytical chemistry. *Pure and Applied Chemistry*, 94(6): 725–736. DOI: <https://doi.org/10.1515/pac-2021-3101>
- RDA FAIR Data Maturity Model Working Group.** 2020. FAIR Data Maturity Model: Specification and guidelines. *Research Data Alliance*. DOI: <https://doi.org/10.15497/rda00050>
- RDMO*. Available at <https://rdmorganiser.github.io/> [Last accessed 5 August 2018].
- Richter, T, Murphy, G and Bolmsten, F.** 2020. *API Definition (common search API)*. Available at [https://www.panosc.eu/wp-content/uploads/2020/12/D3.1\\_API-definition.pdf](https://www.panosc.eu/wp-content/uploads/2020/12/D3.1_API-definition.pdf) [Last accessed 8 November 2022].
- Romanos, N,** et al. 2019. Innovative Data Management in advanced characterization: Implications for materials design. *Materials Today Communications*, 20: 100541. DOI: <https://doi.org/10.1016/j.mtcomm.2019.100541>
- Salvat, D,** et al. 2020. Draft recommendations for FAIR Photon and Neutron data management. *Zenodo*, 857641: 1–63. DOI: <https://doi.org/10.5281/zenodo.4312825>
- SciCatProject*. *GitHub*. Available at <https://github.com/SciCatProject> [Last accessed 14 November 2018].
- Soler, N,** et al. 2022. Final recommendations for FAIR Photon and Neutron data management. DOI: <https://doi.org/10.5281/ZENODO.6821676>
- The Photon and Neutron Open Science Cloud (PaNOSC) – Panosc*. Available at <https://www.panosc.eu/> [Last accessed 20 February 2020].
- Wilson, M,** et al. 2011. Deliverable D2.1 Common policy framework on scientific data. DOI: <https://doi.org/10.5281/zenodo.3738497>

#### TO CITE THIS ARTICLE:

Görzig, H, Gonzales Beltran, AN, Engel, F and Matthews, B. 2024. A Framework for Active DMPs in Photon and Neutron Science Large-Scale Facilities. *Data Science Journal*, 23: 4, pp. 1–18. DOI: <https://doi.org/10.5334/dsj-2024-004>

**Submitted:** 15 December 2022

**Accepted:** 01 December 2023

**Published:** 22 January 2024

#### COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.