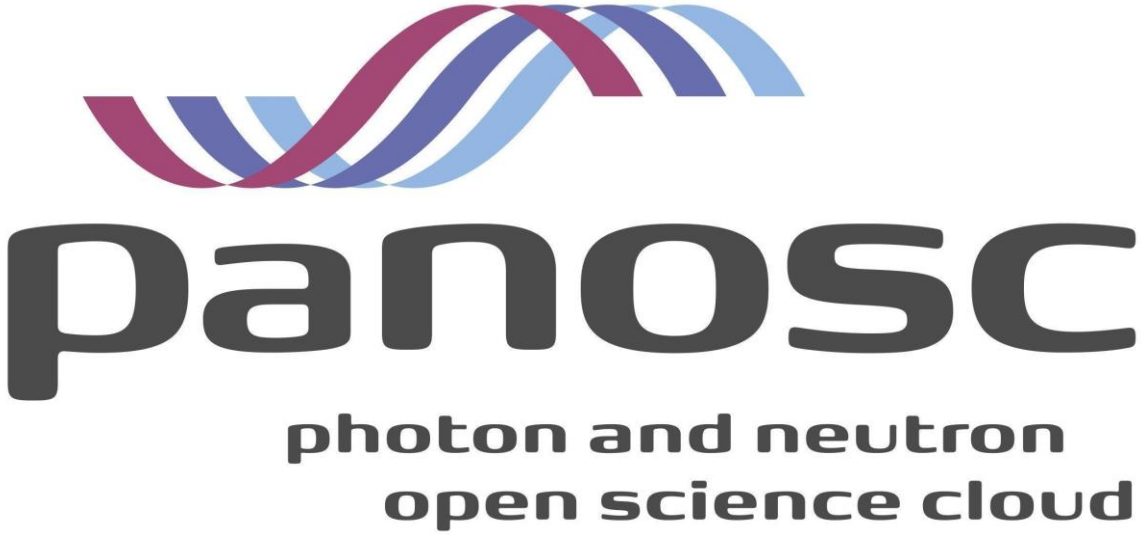
PaNOSC
Photon and Neutron Open Science Cloud
H2020-INFRAEOSC-04-2018
Grant Agreement Number: 823852



Deliverable: D2.2 - DMP Template for facility users

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Project Deliverable Information Sheet

Project Reference No.	823852
Project acronym:	PaNOSC
Project full name:	Photon and Neutron Open Science Cloud
H2020 Call:	INFRAEOSC-04-2018
Project Coordinator	Andy Götz (andy.gotz@esrf.fr)
Coordinating Organization:	ESRF
Project Website:	www.panosc.eu
Deliverable No:	D2.2
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	31 November 2021
Actual Delivery Date:	25/11/2021
EC project Officer:	Flavius Pana

Document Control Sheet

Document	Title: DMP Template for facility users
	Version: 1
	Available at: https://github.com/panosc-eu/panosc
	Files: 1
Date	1 November 2021
Authorship	Written by: Fredrik Bolmsten (ESS), Jonathan Taylor (ESS), Carina Lobley (ESS)
	Contributors: Andy Götz (ESRF), Marjolaine Bodin (ESRF), Janusz Malka (EuXFEL), Krzysztof Wrona (EuXFEL), Alessandro Olivo (CERIC-ERIC), Roberto Pugliese (CERIC-ERIC), Teodor Ivanoaica (ELI ERIC), David Bouchenot (ILL)
	Reviewed by: Jayesh Wagh (ESRF)
	Approved: Jordi Bodera (ESRF)

List of participants

Participant No.	Participant organisation name	Country
1	European Synchrotron Radiation Facility (ESRF)	France
2	Institut Laue-Langevin (ILL)	France
3	European XFEL (XFEL.EU)	Germany
4	The European Spallation Source (ESS)	Sweden
5	Extreme Light Infrastructure Delivery Consortium (ELI-DC)	Belgium
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	The Netherlands

Table of Content

Purpose

With the increasing data volumes (up to 100 TB for new 4th generation photon sources) Data Management Plans (DMPs) are becoming more and more necessary in order to ensure that users are aware of the data volumes that will be produced and how to process them. Currently, none of the PaN RIs have DMPs in place. PaNOSC and ExPaNDS are collaborating on a solution for generating and managing DMPs for PaN RIs.

The data management plan (DMP) is a 'living' document that supports a researcher through all phases of a project from the planning at the start of the project, through the collection or generation of data and its analysis, to the publication and archiving of data at the end of the project. The DMP is often a required element of an application for research funding and there are a number of online tools to assist in the preparation of a DMP.

It is in the best interest of PaN facilities and their users to provide an integrated DMP tool to their proposal workflow. This is because the PaN facilities could then accurately estimate the data volume produced by an experiment and the compute resources required to process and analyse that data. Additionally, many of the required elements for a DMP are also required for facility applications, so many of the steps can be shared between the proposal and the DMP, saving the user time and improving the efficiency.

In the current RI workflow, a user can apply for beamtime without considering the data volume or processing requirements needed. This can result in users collecting a large volume of data and then retroactively planning how to process and store results.

There is a growing list of experimental modalities that can generate data volumes that exceed the capacity of individual users' infrastructure. In these cases, the user community relies upon the facility to provide an infrastructure for data storage and data processing / analysis. Efficient execution of the user programme (where experimental data is understood in as little time as possible) necessitates central provisioning of resources for storage and processing, a service now commonplace at PaN facilities. The challenge in this landscape is ensuring sufficient capacity in the centrally provided systems at any given time.

Facilities have a largely static view of the research data volume that is generated by the user programme. Provisioning of infrastructure for storage and compute is often based on a time averaged view of the user programme experimental modalities rather than a specific or dynamic evaluation of the IT infrastructure requirements on an experiment by experiment basis.
For facilities to move towards a more dynamic evaluation of the research data volume and required data services, there is a tacit understanding that some form of data management planning would be beneficial. From a facility management perspective, a forecast of the IT infrastructure required to support an upcoming experimental programme based on a more detailed understanding of users' need would be beneficial for the facility and user community.

The move to implement more effective data management for research infrastructures and the development of the European Open Science Cloud comes with a general expectation that data management planning is an essential aspect to the research data workflow.
The use of DMPs by the research community is commonly based around a static description of the data to be collected, made against a domain specific template that has been designed to codify essential parameters covering a variety of aspects which are pertinent to, but not necessarily derived from the FAIR concepts for data management. There are descriptions of best practice but essentially no standard for DMPs[1].

To be of real use, DMPs for PaN facilities should be aligned with the facility workflow for research (Fig 1) and be scalable to the size of the research community. The last point is very important, a plan by definition precedes execution, thus for users the planning stage is made before the experiment. The start of the DMP creation could be during submittion of a proposal and/or as part of activities preceding the experiment - such as sample declaration, visit planning, safety training, etc.

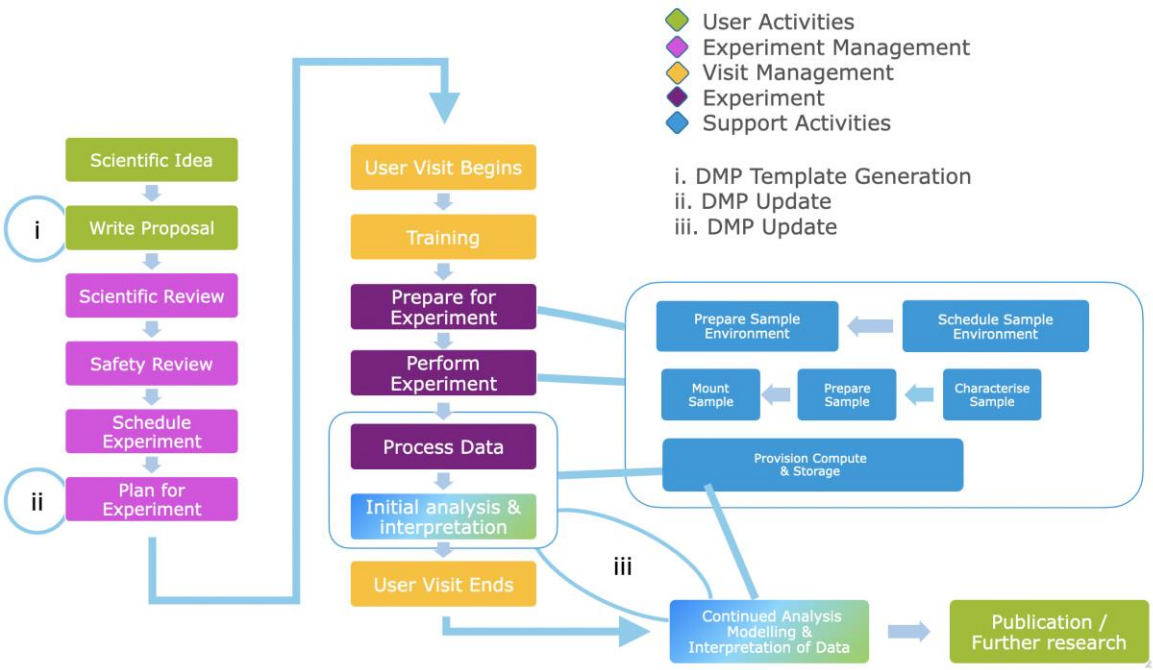


Fig 1. A generic facility workflow[2] for users and some support activities indicating breakpoints where Data Management Planning could be implemented.

[1] https://www.rd-alliance.org/groups/dmp-common-standards-wg
https://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf
[2] Image adapted from original created by A. Jackson ESS

Scope

This document describes the work undertaken to complete PaNOSC task 2.5:

"Task 2.5: Implement Data Management Plan template. Define and implement a template for data management plans for experiments performed at the PaNOSC research infrastructures. Implementations should support automatic filing of the template data based on existing information about the experiment."

This task links to the ExPaNDS project Task 2.2 that has the objective of presenting a DMP framework considering knowledge sources and related roles and activities for DMP-relevant information. This framework aims to integrate the DMP information into data lifecycle and metadata collections, and within the RDMO tool for policy enforcement and reporting.

The two projects will jointly develop the DMP functionality for the PaN facilities.

Overview of the deliverable

Chapter One: introduces the DMP, outlining the purpose of the DMP as well as the advantages to the user and PaN facilities. Additionally, we briefly review the current and desired future status of DMPs.

Chapter Two: (this chapter) provides a background to PaNOSC task 2.5 Implement Data Management Plan template and an outline of the remainder of the deliverable document.

Chapter Three: outlines the requirements for, and current status of, DMPs at the individual facilities participating in the PaNOSC.

Chapter Four: links the work of PaNOSC and ExPaNDS and describes the derivation of a DMP knowledge model. A detailed outline of the knowledge model template is provided.

Chapter Five: gives an overview of the investigated DMP tools and the reference implementation.

Chapter Six: gives a table showing all the questions in the knowledge model divided into the sections outlined in chapter 4. For each question, mapping to the original RDMO source is given and each question is mapped on to the Horizon 2020 DMP where applicable.

High level requirements for DMPs at PaN ESFRI facilities

A survey regarding high-level requirements and current DMP status was performed spanning the facilities in the PaNOSC project.

European Spallation Source

At ESS, we have a tiered set of requirements and ambitions for DMPs. We consider the DMP as an integral part of the experiment planning and feasibility assessment. At many current RIs the feasibility is performed in collaboration with the appropriate member of beamline staff, and for ESS we will endeavour to augment this with additional experiment planning aids to allow (at a far higher-level requirements) virtual test data generation for specific samples in certain beamline conditions.

At the lowest level of ambition, we see a clear need and benefit for evaluating (during part of the proposal submission process) the data volume that will be generated from a proposed investigation. The aim here is to evaluate the data volume and ergo provide requirements for provisioning data storage and data services (such as analysis VMs) for each proposed experiment.

The DMP should be online and should not present the user with a significant workload during the proposal submission process. Thus, a high level of automation should be considered. There are various ways that this automation could be envisaged, from a simple lookup table to a more elaborate inclusion of Monte Carlo ray tracing.

The data volume and data processing needs should be evaluated along with a high-level assessment of the data analysis software/ data analysis workflow. This would require a DMP that contains the early stages of the data lifecycle and in and off itself could define a draft FAIR digital object.

A more advanced DMP can be of use to the user beyond the important considerations of data volume and data workflows. One can imagine many scenarios where the planned experiment that is proposed is modified before the beamtime starts and then again during the beamtime. A DMP that has some determined level of integration with the data management workflow and data catalogue could provide a beneficial up-to-date picture of the data management needed for the ongoing investigation and the lifecycle of the data.

A dynamic DMP can be used as a defined point for allowing access to auxiliary data for the investigation in a streamlined way. Most neutron scattering investigations require auxiliary data as part of the data workflow. (Background runs, Vanadium runs etc). These data are not

necessarily owned by the principal investigator and their team. The provisioning of data processing and data analysis services against a data storage architecture with strong AI provides a real challenge for staff. The DMP in this context could be used as a tool to make provisioning more efficient. I.e. ensuring that the required auxiliary data are included as part of the raw experiment data workflow. There are tangible benefits for this in relation to reuse and interoperability of data.

ESRF

ESRF management and some large user groups are requesting DMPs for experiments which produce large to extremely large data volumes (100 GB to 100 TB). The cutoff value above which data are considered large is not yet defined. It is still undecided if the requirement for a DMP should not be extended to all proposals. The implementation has not started yet and we are in the requirements gathering process. The basic concept is that the user answers a minimum of questions needed for the DMP as part of the proposal submission process. This means it must be accessible via the ESRF Scientific Management System (SMIS) used for proposal submission. It should be an external tool integrated into the different web applications for users, e.g as an extra tab in the data portal. An example is DMPonline which provides an API for calling it from another application. Ideally, it should not be a development from scratch in order to profit from the existing implementations and know-how of other communities. The resulting DMP should machine actionable (cf. paper by Miksa et. al cited above). The DMP should be pre-filled in by the tool based on the beamline for which time is being requested. This raises the question of when to fill in the DMP. At the ESRF, the beamline is only known for sure once the beamtime is scheduled. One option is to base the DMP on the technique and fill it in based on the technique being requested. The DMP would need to be slightly modified to take the technique into account, e.g. data format is known so the question should be posed differently. The DMP should be an active document so it can be updated and verified throughout the experimental workflow e.g. the exact amount of data produced is known after the beamtime and can be filled in. The DMP can be used to communicate information to the users concerning data management issues e.g. data format, tools available, etc. The DMP must be stored with the data and be governed by the data policy. The ESRF can help by integrating and testing the DMP implementation. We need to involve the ESRF User Office and SMIS developers as well as the scientists.

EuXFEL

European XFEL has explored the advantage of introducing a data management plan (DMP) for all submitted proposals. We concluded that there should be a mechanism implemented with the current workflow to make the DMP accessible for users and the facility without additional overhead. The DMP should be easy to answer and light to fill up. Most of the questions present in the form should be answered in automatic or semi-automatic ways using existing data sources like the Scientific Data Policy, user portal or metadata catalog. It should help to automatize the communication between users and facility in the way that both sites will clearly profit out of it.

We understand that filling the entire DMP could be difficult at the proposal submission phase, however providing partial information will still be beneficial, therefore the updates of the DMP form should be possible before submission of the -arrival form and during the beam-time or even after. The good data management plan will help to understand users what facility offers to users to store and analyse data and facility will learn how users are going to process data and what are their needs to produce valuable scientific results.

The DMP will allow European XFEL to prepare the data resources more accurately for an experiment but also efficiently plan how the resources can be shared among experiments. Although the current scheme of the data storage and analysis infrastructure allocation works in most cases, we clearly see the benefits of having a DMP at the early stage of proposal submission as a tool to optimize this process. After the internal consultation with beam-line scientists and the user office, we think that the understanding of the following aspects in the DMP will be advantageous.

First, we should obtain information about the data sources, their configurations, and operation mode, this will allow us to better estimate a data volume. We would also like to ask users if there are other than the experiment data which are needed to perform an experiment or its data analysis, like laboratory data or external data and if those data need to be transferred and preserved in our facility.

In order to provide the best support for data analysis, we would like to understand what analysis method will be used. We need to know if users have technical knowledge and resources to analyse data or what will be the amount of the support that we, as the facility need to provide in that regard. It will be beneficial to know user's estimation for the duration of data analysis. This will help to allocate the human resources and computing infrastructure needed for the experiment.

To make all components available and usable on time, we would like to know which piece of software will be used for data analysis. Furthermore, if users are planning to deploy any home-grown software which they would like to integrate into our data analysis chain.

Finally, we would like to make enough hardware components like CPU, GPU available for users, considering needs and fair share between all users' groups which are performing experiments or analyze data simultaneously. If amount of hardware components required to perform data analysis is challenging, we might even take this as an argument while scheduling the beam-time.

Different storage and computing resources may be needed, if an experiment team plans to calibrate detectors on their own instead of using official facility calibration pipeline and if such approach is preferable how long they need access to the processed data on disk. Since the space on disk is limited, do the users plan to reduce data. Do they have a strategy on how to reduce the data volume, for example data pre-selection. Would they be able to assess the quality of data within a few months and select only partial data-sets for long-term storage?

Very often, users would like to have the data storage and computing infrastructure available before scheduled beam-time. Based on the answers in the DMP, we would like to understand how long in advance such an infrastructure should be provided to them and in what size.

Some users might prefer to take data to their home institute and analyse it in their home computing cluster. Considering the large data volumes, it makes the data export not trivial and sometimes not possible. It will be useful to know in advance which data, users would like to export or access.

European XFEL Scientific Data Policy defines the embargo period after which the data become open. There might be multiple reasons to change the time when data become available to the wider community. As an example, a publisher requires a subset of the data which was used for publication to be open at the moment the article is published. The DMP users might provide this information to the facility well in advance so that the facility can make the data public on time.

Discussing the open access and data reproducibility, it is extremely important which data are essential to reproduce the results of the analysis and if there are any additional data or metadata which should be preserved together with RAW data to make data reusable in the future.

European XFEL considers the DMP supplementary to the Scientific Data Policy. It will address the open questions and allows to take relevant action. It will increase the understanding of the data stewardship among the users and our facility. We think that the data management template provided below is usable and essential to achieve this goal.

ELI

The Extreme Light Infrastructure (ELI) consists of complementary facilities in the Czech Republic, Hungary and, in the future, Romania. The ELI facilities, built as individual construction projects, are now coming together as an integrated organisation, the ELI European Research Infrastructure Consortium (ELI ERIC). The ERIC will be in charge of the joint operations, facilitating the integration of the first two sites under a single ERIC and preparing them for sustainable users' experiments. ELI ERIC was established in April 2021, with the Czech Republic and Hungary as host members; Romania will join at a later stage when accession conditions have been agreed. Under ELI ERIC, ELI will operate as a multi-site organisation with a single management and single governance.

ELI Data Policy is the core policy governing the DMP and, together, they represent the cornerstone of the future ELI Scientific Data Management System. The first ELI FAIR Data Policy developed based on the PaNOSC FAIR Data Policy Framework and based on our particular setup and implementation, was presented to ELI Management, is soon going to be presented to the ISTAC (International and Scientific Advisory Board and will be later this year discussed at the level of the ELI General Assembly. We expect to have the data policy adopted by the end of the year.

Using the data management experience of our team, gathered during the experiment commissioning phase, and by allocating the extra resources needed for supporting the implementation of the Data Policy to facilitate de adoption of the ELI DMP, ELI is preparing for the implementation of the integrated Scientific Data Management System.

The developed DMP will be provided, together with the Data Policy, the necessary support for the users and, via the proposal submission system (User Office), will raise the awareness of the PIs and initiate the DMP data collection without adding a huge overload for the team submitting the proposal. Furthermore, during the proposal preparation, the applicants shall be guided to fill in the DMP, as it is equally important for the proposal evaluation process, informing technical teams about specific requirements of the experiments, but also for Computing and IT Teams that should prepare the support infrastructure to support Data and Data Management operations of the approved experiments.

CERIC-ERIC

CERIC-ERIC considers the preparation of a DMP a critical aspect of the infrastructure and will dedicate resources to design and implement what is needed as part of the quality management system of the research infrastructure. DMP involves users, beamline scientists, researchers, administration, legal, IT infrastructure. At the moment, we are designing surveys for the specific stakeholders in order to collect their respective view on the field. Then, when strategic decisions will be taken, we will be able to develop a DMP template for each instrument offered to the community. We will consider the available above mentioned DMP generation tools. An experiment involving different instruments will require a dynamic combination of the instrument specific DMP. The generated DMP will be presented to the PI as part of the proposal submission in order to make him/her aware. It is not yet clear at the moment if the PI will have a more active role, for example, modifying the DMP fields to suit specific needs.
The resulting DMP especially if machine readable and in aggregated form will be useful to plan the required IT infrastructure.

ILL

ILL management considers DMPs as a real improvement of the quality of data and science efficiency. The facility also sees a real asset regarding the better IT resource management brought by the early data gathered with DMPs.

ILL gives a thought to applying DMPs to all proposals since the process should be a benefit in all ways without generating an overload for the user.
At the moment DMPs are not implemented. Since the ILL proposal portal is about to get redesigned starting next year, the DMP's workflow will be implemented at that point.

A high level of automation will be present with pre-filled forms from existing data to not be redundant in any kind for the user.
As a dynamic document, the DMP will be updated during the experiment, automatically in the portal. In order to use DMPs in the most effective way, an important care will be made

about giving the users explanations about the DMP process at all stages, and how it will get updated along the process.
The DMP and all the workflow will be defined in the ILL Data Policy.

DMP knowledge model

The proposed DMP template will cover a wide range of requirements from PaN facilities, as well as funding agencies. It takes into consideration best practices developed from the RDMO and aims to help users produce a DMP in an easier fashion by populating the vast majority of fields in an automated manner.

Building from the work from ExPaNDS Task 2.2, and discussions within the PaNOSC and ExPaNDS projects, the objective is to develop a DMP knowledge model for PaN facilities. The DMP knowledge model is a set of questions that can map to a wide range of DMP templates. The reason for building a knowledge model instead of a specific DMP for PaN facilities is to service a growing set of requirements.

When surveying the DMP landscape, RDMO (https://rdmorganiser.github.io/en/) stood out due to the broad range of questions applicable to PaN facilities. While we kept all the sections of the RDMO knowledge model, we filtered the questions based on applicability for the facilities. Where necessary, questions were rephrased to suit facility users.

DMP knowledge model framework

Here, we discuss the broad areas of the DMP knowledge model. This framework provides a basis for a DMP which PaN facilities can adapt and change to meet their individual needs.

There are two overriding principles.

1. The questions should be lightweight for the user and beamline scientist(s), and where possible content should be automatically generated from the experiment proposal.
2. Completion of the template should be staged, with the staging aligned with the access mechanism of the research infrastructure and the data lifecycle

The knowledge model consists of 7 sections grouping related questions for ease of completion. The remainder of chapter 4 addresses each section in summary, while chapter 5 is a detailed description of the template, containing all the questions in the DMP knowledge model.

General / Topic

This section introduces the project, with questions surrounding the science to be conducted. We also identify the users and the source of funding for the research.

Content Classification / Datasets

This section asks questions about the individual datasets covered by the DMP.

Technical Classification / Data Collection

This section records the dates when data collection will take place, as well as when the data will be processed and analysed. The total volume of data and volume of data per year are recorded. The user is asked to provide information about the software required to work with the data. Finally, questions are asked to understand if and how the user will carry out versioning as they work with the data.

Data Usage / Usage Scenarios

This section establishes what PaN facility resources will be required in the lifetime of the data. It defines who handles the data and data backups as well as who can access the data and what provision is made for data security. The extent that data can be shared or form part of a collaboration is recorded. The user is asked to estimate the personnel and non-personnel costs associated with the data.

Metadata and Referencing / Metadata

Here, the user is asked to indicate what metadata are required to understand the data and to indicate whether this is collected automatically, semi-automatically or manually. Additionally, the use of persistent identifiers (PIDs) is requested. The user is asked to estimate the personnel and non-personnel costs associated with metadata and PIDs.

Legal and Ethics / General Legal Issues

This section establishes whether the data are under the jurisdiction of more than one country and whether they include personal or sensitive data. The user is asked to detail any recommendations the funding body has about data management.

Storage and Long-Term Preservation / Selection

Here, the user is asked to explain their criteria for archiving data as well as the duration and accessibility of such archived data.

DMP Tool

There are a number of services and open source developments available to generate online DMP's, for example:

- Data Stewardship Wizard (DSW)
- Open DMP
- DMPTool

Each Service provides a static DMP template for completion by users following a questionnaire style workflow. Services provide functionality to create specific templates, and in the case of services such as DMPTool and DSW provide standard templates developed by specific organisations. These tools provide a static DMP document that can be used as augmentation for a research proposal, where a DMP is a requirement of the funder the tools are adapted to be specific for the organisation's requirements.

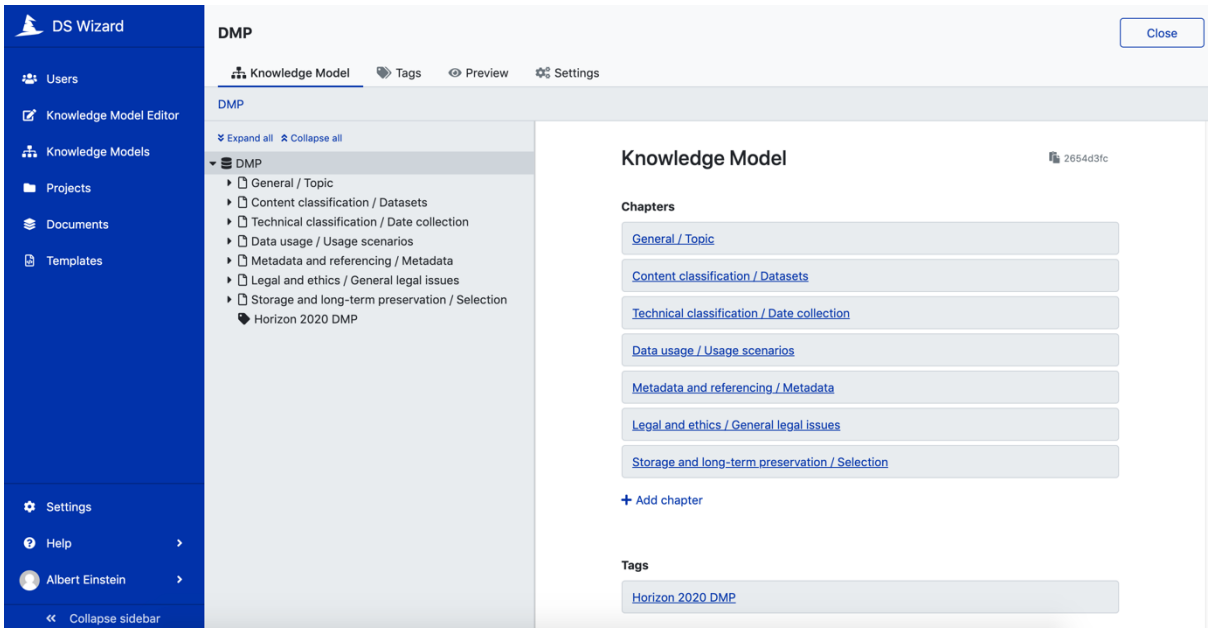


Fig 2. The DMP template proposed in chapter six implemented in the Data Stewardship Wizard tool.

After investigating the above mentioned DMP tools a reference implementation was made using the Data Stewardship Wizard. The template proposed in chapter six was implemented and a connection made to a mock-up facility infrastructure to automatically fill parts of the proposal. It was demonstrated that significant parts of the DMP template could be populated by the facility and therefore reducing the workload for the user.

Proposed template

It is envisioned that facilities will modify and extend the proposed template questions to fit their facility while keeping the underlying meaning, this will keep the mapping to the RDMO and also the funder templates.

RDMO NR.	Questions	Section	Help text	Funder template
1	What is the scientific motivation for the experiment?	General / Topic	Provide a summary of the experiment. Could be retrieved from the proposal abstract. This abstract might be later used in the data catalogue to describe the experiment.	
2	Please provide a minimum of two keywords describing the experiment	General / Topic	Provide Science area keywords as you would have in your publication. This should also include scientific methods used and keywords about the sample. Part of the keywords can be provided by the instrument scientist.	Horizon 2020
3	What is the primary research area?	General / Topic	Provide a scientific field descriptor as in your proposal.	

RDMO NR.	Questions	Section	Help text	Funder template
4	When does the project start?	General / Topic	This should be the start of the overall project, which includes experiment time	RDA-DMP-common-standard
5	When does the project end?	General / Topic	This should be the end of the overall project, which includes experiment time	RDA-DMP-common-standard
6	Who submitted the proposal and is responsible for the project coordination?	General / Topic	This normally is the proposer or a person nominated by the proposer.	
7	Who are the institutional experiment partners?	General / Topic	Provide a list of institutions for the proposal team as in the proposal.	RDA-DMP-common-standard
8	Provide institutional policy URL and relevant statements of the policy.	General / Topic	More and more universities and scientific institutions adopt research data management policies. These contain recommendations and / or demands concerning the handling of research data by researchers of the institution.	

RDMO NR.	Questions	Section	Help text	Funder template
9	Who is responsible for RDM in your institution?	General / Topic	This person is the experiments contact person to the facilities RDM team. Please give the name and an email address.	Horizon 2020,RDA-DMP-common-standard
10	Is the proposal supported by externally funded research?	General / Topic		RDA-DMP-common-standard
11	In which special funding programme is the project located?	General / Topic		RDA-DMP-common-standard
12	Does the funder have rules or recommendations for data management? If so, provide policy URL and relevant statements of the policy.	General / Topic	The facility needs to know about rules that precede the facilities' data policy, like embargo time and open access requirements.	

RDMO NR.	Questions	Section	Help text	Funder template
13	Are there requirements regarding the data management from other parties (e.g. the scholarly/scientific community/publisher)?	General / Topic		Horizon 2020
14	Which are these additional requirements regarding data management?	General / Topic		Horizon 2020
15	Please provide a brief description of the dataset (eg.Diffraction data taken as a function of temperature)	Content Classification / Datasets	Description information that contains both context (experiment description), contents (raw data plus log book sample information and file and technique description. File and technique descriptions can be retrieved from the instrument scientist.	RDA-DMP-common-standard
16	Is the dataset experimental, processed, or analysed?	Content Classification / Datasets		Horizon 2020
17	If processed or analysed, who created the dataset?	Content Classification / Datasets		Horizon 2020

RDMO NR.	Questions	Section	Help text	Funder template
18	If processed or analysed, under which address, PID(s), or URL, can the original data be found?	Content Classification / Datasets		Horizon 2020
19	Which individuals, groups or institutions could be interested in re-using this dataset? What are plausible scenarios?	Content Classification / Datasets		Horizon 2020
20	Is the dataset reproducible in the sense that it could be created / collected anew in case it got lost?	Content Classification / Datasets	Some data can, technically, be created anew, as with scientific experiments or digitised versions of analog objects (as long as the originals are still there and in good shape). However, this can consume a considerable amount of time and cost. Regarding the long-term preservation, the effort of re-creation has to be weighed up against the effort of long-term preservation. Other data cannot be collected or created anew. The answer should be discussed by the user with the instrument scientist. Is all information available to repeat the experiment?	

RDMO NR.	Questions	Section	Help text	Funder template
21	When does data collection or creation start?	Technical Classification / Data Collection	This should be the start of your beamtime	
22	When does data collection or creation end?	Technical Classification / Data Collection	This should be the end of your beamtime	
23	When does data cleansing / data preparation end?	Technical Classification / Data Collection		
24	When does data analysis start? This creates another dataset and we need to know when the data needs to be available.	Technical Classification / Data Collection		
25	When is analysis infrastructure required?	Technical Classification / Data Collection	This question is only relevant if you require special resources.	
26	What is the expected size of the raw dataset?	Technical Classification / Data Collection		

RDMO NR.	Questions	Section	Help text	Funder template
27	Estimate the raw data volume generated from the proposal. Number of hours of instrument multiplied by data rate per hour	Technical Classification / Data Collection	Calculation or range of data volume based on proposed experiment	Horizon 2020, RDA-DMP-common-standard
x	Actual size of data collected during experiment	Technical Classification / Data Collection		
28	How many datasets were created?	Technical Classification / Data Collection		
29	Specify the file formats used for the raw data (default Nexus)	Technical Classification / Data Collection		Horizon 2020
30	Which instruments/hardware, tools, software, technologies or processes are used to generate or collect the data?	Technical Classification / Data Collection	This information is necessary to reconstruct the process by which the data was generated. It is also a prerequisite to judge the objectivity, reliability, and validity of the dataset. For reproducible data, it is also required to re-generate the data if need be. (Information to come from the instrument scientist or user.)	RDA-DMP-common-standard

RDMO NR.	Questions	Section	Help text	Funder template
31	Which software, processes or technologies are necessary to use the data?	Technical Classification / Data Collection	To re-use data (e.g. to replicate studies, for meta-analysis or to solve new research questions), along with the data, the software, equipment and knowledge about special methods to use the data are required. Just as with the formats, the recommendation is: the more standardised, open and established, the better for re-use. (Instrument scientist to give possible options for data usage.)	Horizon 2020
32	Is documentation about relevant software needed to use the data?	Technical Classification / Data Collection	Please provide a link to relevant software and documentation	Horizon 2020
36	How / for what purpose will this dataset be used during the project?	Data Usage / Usage Scenarios		Horizon 2020
38	Estimation of infrastructure needs select from list	Data Usage / Usage Scenarios	e.g. selection of cpu hours required (if needed)	

RDMO NR.	Questions	Section	Help text	Funder template
42	Are there internal project guidelines for a consistent organisation of the data? If so, where are they documented? Divide in instrument specific and project specific	Data Usage / Usage Scenarios		
43	Is there an internal project guideline for naming the data? If so, please briefly outline the naming conventions and, link to the documentation. Divide in instrument specific and project specific	Data Usage / Usage Scenarios		Horizon 2020
44	Who is allowed to access the dataset?	Data Usage / Usage Scenarios		Horizon 2020
45	How and how often will backups of the data be created?	Data Usage / Usage Scenarios		
46	Who is responsible for the backups?	Data Usage / Usage Scenarios		

RDMO NR.	Questions	Section	Help text	Funder template
47	Which measures or provisions are in place to ensure data security (e.g. protection against unauthorised access, data recovery, transfer of sensitive data)?	Data Usage / Usage Scenarios		Horizon 2020
48	Is this dataset interoperable, i.e. allowing data exchange and re-use for researchers, institutions, organisations, countries, etc.?	Data Usage / Usage Scenarios		Horizon 2020
49	Where will this dataset be published or shared?	Data Usage / Usage Scenarios		Horizon 2020
51	What license will be applied to the dataset?	Data Usage / Usage Scenarios		Horizon 2020
53	When will the data be published (if they are)?	Data Usage / Usage Scenarios		Horizon 2020

RDMO NR.	Questions	Section	Help text	Funder template
57	Which measures of quality assurance are taken for this dataset?	Data Usage / Usage Scenarios		
58	Is the integration between auxiliary information, measured and processed datasets ensured? If yes, by which means?	Data Usage / Usage Scenarios		
64	What is the amount of non-personnel costs associated with the storage of the data sets during the project?	Data Usage / Usage Scenarios		
65	Which information is necessary for other parties to understand the data (that is, to understand their collection or creation, analysis, and research results obtained on its basis) and to re-use it?	Metadata and Referencing		
66	Which standards, ontologies, classifications etc. are used to describe the data and context information?	Metadata and Referencing		Horizon 2020

RDMO NR.	Questions	Section	Help text	Funder template
67	Describe automatically generated metadata from experiment	Metadata and Referencing		Horizon 2020
x	Describe which metadata will be automatically collected during the experiment, which is essential for analysis of the data set. (Eg : Sample temperature)	Metadata and Referencing		
68	In case it is unavoidable that you use uncommon or generate instrument/project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?	Metadata and Referencing	Provide a pathway to convert non-standard formats onto Nexus, CBF or similar	Horizon 2020
69	Which metadata are collected semi-automatically?	Metadata and Referencing	For example, metadata collected using forms	
70	Describe which metadata will be manually added to the dataset.	Metadata and Referencing		

RDMO NR.	Questions	Section	Help text	Funder template
x	Provide description, link, or DOI relevant to the metadata for data processing and data analysis. Eg data processing script / input file	Metadata and Referencing		
x	Describe which auxiliary data will be added to the data set for processing and analysis.	Metadata and Referencing		
x	Which calibration data sets have been used?	Metadata and Referencing		
x	Which sample characterisation data will be added.	Metadata and Referencing		
71	Provide schema description and validation process	Metadata and Referencing		
72	Who is responsible for documenting the metadata and context information and for checking if they are correct and complete?	Metadata and Referencing	Normally the PI but can be delegated to other person	

RDMO NR.	Questions	Section	Help text	Funder template
75	What is the structure of the data? How are the individual components of the dataset related to each other? How is the dataset related to other datasets used in the project?	Metadata and Referencing		
76	Will persistent identifiers (PIDs) be used for this data set?	Metadata and Referencing		Horizon 2020
77	Which system of persistent identifiers shall be used?	Metadata and Referencing		Horizon 2020
79	Who is responsible for the maintenance of the PIDs and the object maintenance (i.e. who is responsible for notifying the PID-Service about object relocation and the new address)?	Metadata and Referencing		
83	Does this dataset contain personal data?	Legal and Ethics	Normally, there should not be any personal data. In case there are check how to proceed.	Horizon 2020

RDMO NR.	Questions	Section	Help text	Funder template
84	Does this dataset contain sensitive data other than personal data?	Legal and Ethics	Examples are data that contain business secrets or geoinformation on endangered species. Normally, there should not be any other sensitive data. In case there are check how to proceed.	
90	Has the project been approved by a research ethics committee?	Legal and Ethics		
95	Does the project use and/or produce data that is protected by intellectual or industrial property rights?	Legal and Ethics	Public funded projects normally create non-proprietary data.	
101	What are the criteria / rules for the selection of the data to be archived (after the project)?	Storage and long-term preservation		
102	Who selects the data to be archived?	Storage and long-term preservation		Horizon 2020

RDMO NR.	Questions	Section	Help text	Funder template
103	Does this dataset have to be preserved for the long-term?	Storage and long-term preservation		
104	What are the reasons this dataset has to be preserved for the long-term?	Storage and long-term preservation	Is there some special long-term value in the data that overrules the normal preservation policy?	
105	What is the minimum period that the data will be stored?	Storage and long-term preservation		Horizon 2020
107	Where will the data (including metadata, documentation and, if applicable code) be stored or archived after the project?	Storage and long-term preservation		Horizon 2020
108	Is the repository or data centre chosen certified (e.g. Data Seal of Approval, nestor Seal or ISO 16363)? (If the dataset is archived at several places, you may answer this question with yes, if this applies to at least one of these.)	Storage and long-term preservation		Horizon 2020

RDMO NR.	Questions	Section	Help text	Funder template
109	Have you explored appropriate arrangements with the identified repository?	Storage and long-term preservation		Horizon 2020
110a	Shall there be an embargo period before the data is open access?	Storage and long-term preservation		
110b	How long is the embargo period?	Storage and long-term preservation		
111	How will the identity of the person accessing the data be ascertained?	Storage and long-term preservation		
112	By when will the data be archived?	Storage and long-term preservation		
115	How will the data management costs of the project be covered?	Storage and long-term preservation		Horizon 2020

Where column RDMO NR. contains x this indicates a new question added to the template.