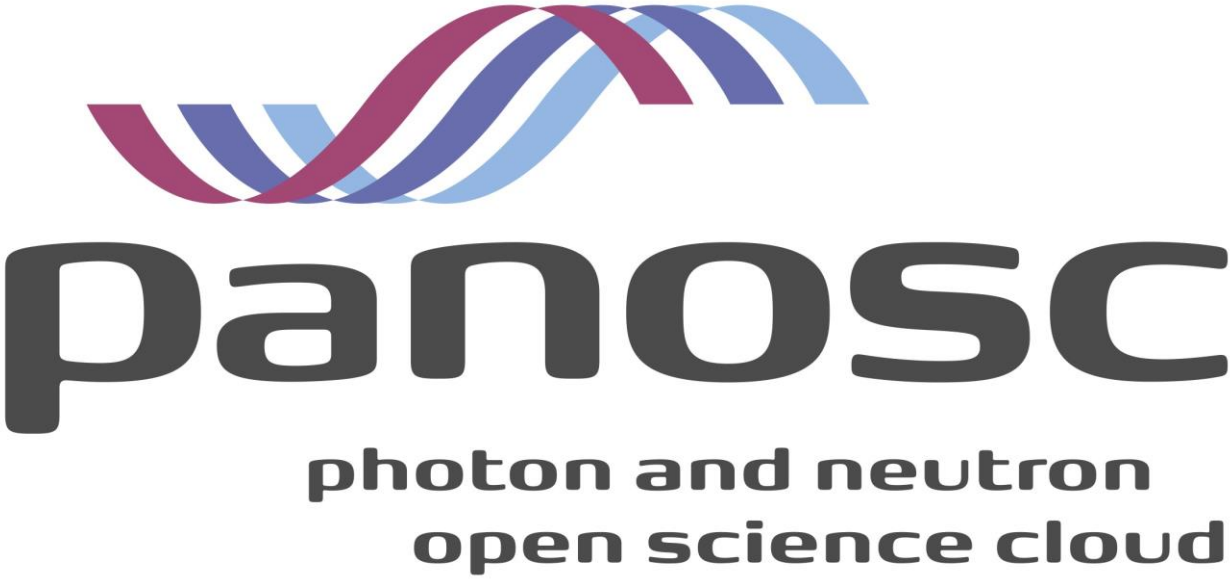
PaNOSC

Deliverable: D1.2 Data Management Plan

Version 2 – 17th September 2020



Photon and Neutron Open Science Cloud

H2020-INFRAEOSC-04-2018

Grant Agreement Number: 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:	D1.2 Data Management Plan
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	31/05/2019
Actual Delivery Date:	31/05/2019, resubmitted 17/09/2020
EC project Officer:	Geert Vancraeynest

Document Control Sheet

Document	Title: D1.2 Data Management Plan
	Version: 2
	Available at: https://github.com/panosc-eu/panosc
	Files: 1
	Change log: document renamed as D1.2 (instead of D1.4)
Authorship	Written by: Andy Götz
	Contributors: Jordi Bodera, Rudolf Dimper
	Reviewed by: Jordi Bodera Sempere
	Approved: Andy Götz

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 Contents

Introduction

This document is Deliverable 1.2 and concerns the **Data Management Plan** (**DMP**) for the **PaNOSC** project. The purpose of this deliverable is to support the data management life cycle for all data that will be collected, processed or generated by the project. It provides a description of the data types the project will generate and how the data will be collected and stored and made available for validation, exploitation and re-use by others.
The PaNOSC project is about providing support for scientific data management and linking scientific data to the EOSC. In this respect **PaNOSC does not produce any scientific data**. *PaNOSC produces services, software for managing scientific data and documents (deliverables) on scientific data management and services*. The DMP covers these data as well as recommendations for the scientific data which will be managed by PaNOSC.
The information in this DMP will be updated during the course of the project.

Data Summary

PaNOSC is a project focused on providing services for people who generate scientific data. PaNOSC will generate documents and source code and services for data reduction, curation and simulation. Documents generated by PaNOSC will be in text, markdown, MS-Word, MS-Powerpoint, pdf, hdf5 (simulated data) and json (jupyter notebooks) formats. The following table summarises types of data and documents, expected data voumes and where they will be stored generated by PaNOSC:

Table 1 Summary of documents and data produced in PaNOSC

Data type	Contents	Volume	Archive
Text files	Meeting minutes and notes	10s Megabytes	Github.com
Word documents	Working versions of deliverables and milestones	100s Megabytes	Googledoc, github, CERIC file server
PDF documents	Final version of deliverables and milestones	100s Megabytes	CERIC file server, EU portal
Source Code	Source code for implementing data services like data catalogues, federated data search, simulation, Data reduction	10s Megabytes	Github.com
Simulated data	Data produced by simulation codes in WP5 for reference purposes	10s Gigabytes	Zenodo.org Github.com
Jupyter Notebooks	Recipes for data reduction and processing	10s Megabytes	Github.com
Training materials	Videos, tutorials integrated in moodle platform	100s Megabytes	Pan-training.org

Source code will be in text. Source code will be developed from scratch or extended from previous projects. PaNOSC has the following repositories for documents and source code:

- https://github.com/panosc-eu/panosc - the main repository for information, documentation and common issues for PaNOSC.
- https://github.com/panosc-eu/wp3 - for tracking data catalogues in WP3
- https://github.com/panosc-eu/fair-data-api - common API for exposing and accessing open data being developed as part of WP3
- https://github.com/panosc-eu/wp4 - repository for issues related to WP4
- https://github.com/PaNOSC-ViNYL - organisation for all WP5 repositories for simulation
- https://github.com/oasys-kit - organisation for the OASYS repositories for ray tracing of beamline optics (OASYS is part of WP5)

The size of the generated data is expected to be in the Megabytes range for documents and source code. On the other hand, the PaNOSC project will develop and provide data catalogues to manage petabytes of scientific data generated by scientists who come to the Research Institues (RIs) who are partners of PaNOSC.

The documents and source code will be useful to all PaN ESFRIs and the national Research Infrastructures (RIs) represented by the ExPaNDS project. The scientific data managed by PaNOSC will be useful for a wide range of scientists in many domains.

FAIR data

Making data findable, including provisions for metadata

PaNOSC will provide DOIs per site for scientific data as a service. If DOIs for documents are required they will be issued through a public service like zenod.org. Software will get a DOI on Zenodo. All the critical software for Photon and Neutron science are registered in the Photon and Neutron software catalog (https://software.pan-data.eu/) and are therefore easy to find. The software catalog will be further developed and enhanced with services as part of PaNOSC.

PaNOSC will use an internal naming convention for documents based on "PaNOSC - Dx.y_YYYYMMDD_version" e.g. "PaNOSC - D1.1_20190212_FINAL.docx". PaNOSC has chosen the NeXus metadata standard for scientific data - https://www.nexusformat.org. PaNOSC will extend the NeXus standard with new keywords and definitions for techniques not yet covered by NeXus.

Documents generated by PaNOSC can be searched for via the website or using the search feature on Github. They are organized in a logical way to make the easy to browse. The software catalog has categories of software to help users find software relevant to the technique they use. In Nexus and the data catalogue we have standard keywords which will help users find data. Github has a search facility for software stored there. The common API being developed as part of WP3 will allow searching for scientific data.

Version numbers for documents will be based on their dates. Software has a unique commit hash key and a tag with a version number for releases.

PaNOSC will use the Nexus metadata standard. For simulation an additional metadata standard called openMD is being developed as part of WP5. PaNOSC will explore integrating openMD into Nexus.

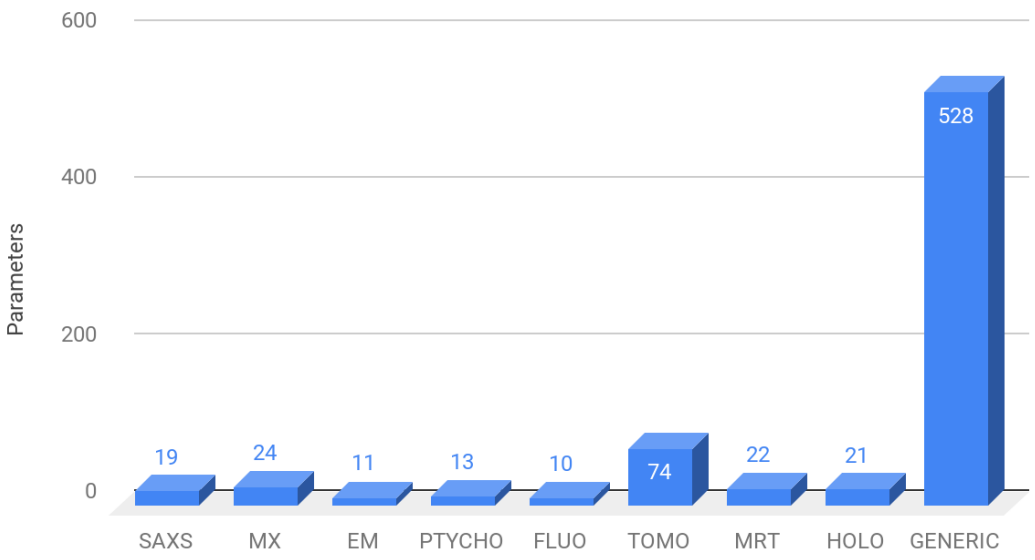


Figure 1 Number of Nexus metadata defined for different techniques at ESRF in 2019 (see [1] for more details)

Making data openly accessible

By default, documents are open access in PaNOSC (majority of deliverables are public). All minutes and discussions are on GitHub and publicly accessible. The default for source code developed by PaNOSC is Open Source and are available on GitHub. A few projects are extensions to closed source projects and therefore might not be made open source e.g. the XFEL metadata catalogue. Some intermediate working versions of documents might be public. The final versions will be.

PaNOSC Source code is by default Open Source however in the rare case where source code extends existing closed source projects they will not be made open source. PaNOSC will extend the data policy framework known as the PaN-data data policy. In this data policy scientific data is closed during the embargo period and then made available under as Open Data. Scientists can publish data as Open Data before the end of the embargo period. The typical embargo period is 3 years.

PaNOSC provides repositories for storing and curating scientific data. The goal of PaNSOC is not to generate the scientific data but to manage it. PaNOSC will extend and develop data catalogues for making scientific data easier to find, search and be accessible. Documents generated by PanOSC will be stored on GitHub and the H2020 portal. Source code is stored on GitHub. The PaNOSC web portal will store deliverables as well.

Regarding access and software tools required to access the data, text documents written in markdown require a simple text editor to read and modify. Deliverables will require a pdf reader to read and MS-Word to modify. Source code will be editable in any plain text editor. The software catalogue will provide services to launch software for reading and reducing data via a web browser. The software catalog provides links to documentation for the software packages.

The default for software developed as part of PaNOSC is Open Source. Therefore, software will be under a recognized Open Source license e.g. Apache2 or MIT. Data, associated metadata, documentation and code

will be in GitHub (https://github.com/) and will be under a Creative Commons license.

Software is managed in git repositories - GitHub and GitLab at the PaNOSC partner sites. Git is a distributed version control system. This means there are multiple copies of the repositories.

There are no restrictions on use for documents and source code. In the case of the scientific data in the data catalogues the principal investigator must decide if they can share a copy of the data while the data are under embargo. After that data are made available under as Open Data e.g. under a Creative Common license.

Open Source provides anonymous access for reading. Identity will be managed via GitHub authentication mechanisms for modifying source code and editing documents on Github. Catalogues for scientific data will use the AAI being integrated in EOSC as part of WP6 (UmbrellaID[1]) to register and identify users. Anonymous access will be possible for Open Data.

Making data interoperable

PaNOSC is all about enabling FAIR data policies and repositories. Even if PaNOSC does not produce any scientific data of its own it will be a major step towards making scientific data produced at the member institutes inter-operable. This will be ensured by adopting the community standard metadata standard Nexus and storing data in standard formats like HDF5. PaNOSC will extend the PaN software catalogue to ensure standard software applications are easily available (e.g. packaged as containers) for users to reduce and analyse scientific data.

PaNOSC recommends to use the Nexus vocabulary for metadata. Nexus is a photon and neutron community standard documented at https://www.nexusformat.org/. Nexus is used for all standard scientific data. Where new experiments require new vocabularies they will be developed following the Nexus conventions and proposed as extensions to Nexus.

Regarding documentation generated to manage the project and its reports these will be in plain text, MS Word and PDF formats which are widely interoperable.

Where PaNOSC will extend or develop new vocabularies eg. openMD, they will be mapped to existing standards in the PaNOSC vocabulary (Nexus) and/or extensions will be proposed.

Increase data re-use (through clarifying licenses)

There are two kinds of data generated by PaNOSC:

1. Documents: these will be openly available in text, markdown, MS-Word and PDF formats under a Creative Commons (CC) license
2. Source Code: this will be open source and in plain-text format. Data and documents will be publicly available and source code open souce, thus ensuring the widest re-use possible.

Both code and reports will be available immediately after delivery with no embargo periods in place, remaining re-usable forever.

Quality assurance for PaNOSC-generated data and documents, reports and documentation is shared among the partners and reviewed by all, thus correcting and increasing the quality of these documents.

[1] https://www.umbrellaid.org/

Costs

Documents and source code are hosted in GitHub for free and public deliverables submitted and then hosted by the EU at no cost for the project.

Each work package leader is responsible for document and source code management regarding his/her work package, with the Project Management Committee and WP1 leader overseeing the fulfilment of the data management responsibilities.

Resources for long term preservation are discussed within PaNOSC. The scientific data catalogues will host petabytes of scientific data (generated by scientists who use the facilities at the PaNOSC partner sites) which will incur a significant cost. For this reason WP7 will study the cost of curating and storing petabytes and propose ways of making it sustainable.

Data Security

PaNOSC does not deal with sensitive data.

PaNOSC uses GitHub is its main repository with a file server (hosted at CERIC-ERIC) for documents which are considered private e.g. contracts etc.

GitHub is considered suitable for long term storage however it will be advisable to store a copy of the Gitlab repositories which are then backed in a long term storage system maintained by the partners.

Ethical aspects

PaNOSC does not have any ethical or legal issues impacting data sharing and does not deal with personal data, therefore consent is not required.

Other issues

PaNOSC does not use any other national/funder or otherwise procedures for data management.

References

[1] R. Dimper, A. Götz, A. de Maria, V.A. Solé, M. Chaillet & B. Lebayle (2019)"*ESRF Data Policy, Storage, and Services*", Synchrotron Radiation News, 32:3, 7-12, DOI:10.1080/08940886.2019.1608119

Appendix I: FAIR Data Management summary tables

DMP component	Issues to be addressed
1. Data summary	• State the purpose of the data collection/generation • Explain the relation to the objectives of the project • Specify the types and formats of data generated/collected • Specify if existing data is being re-used (if any) • Specify the origin of the data • State the expected size of the data (if known) • Outline the data utility: to whom will it be useful
2. FAIR Data 2.1. Making data findable, including provisions for metadata	• Outline the discoverability of data (metadata provision) • Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? • Outline naming conventions used • Outline the approach towards search keyword • Outline the approach for clear versioning • Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how
2.2 Making data openly accessible	• Specify which data will be made openly available? If some data is kept closed provide rationale for doing so • Specify how the data will be made available • Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? • Specify where the data and associated metadata, documentation and code are deposited • Specify how access will be provided in case there are any restrictions
2.3. Making data interoperable	• Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. • Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
2.4. Increase data re-use (through clarifying licenses)	• Specify how the data will be licensed to permit the widest reuse possible • Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed • Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why • Describe data quality assurance processes • Specify the length of time for which the data will remain re-usable
3. Allocation of resources	• Estimate the costs for making your data FAIR. Describe how you intend to cover these costs • Clearly identify responsibilities for data management in your project

	• Describe costs and potential value of long term preservation
4. Data security	• Address data recovery as well as secure storage and transfer of sensitive data
5. Ethical aspects	• To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former
6. Other	• Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)