# PaNOSC

# Photon and Neutron Open Science Cloud

## H2020-INFRAEOSC-04-2018

## Grant Agreement Number: 823852

**Deliverable: D2.1 - PaNOSC data policy framework**

# 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: | D2.1 |
| Deliverable Type: | |
| Dissemination Level | |
| Contractual Delivery Date: | |
| Actual Delivery Date: | |
| EC project Officer: | |

## Document Control Sheet

| | |
|---|---|
| **Document** | Title: |
| | Version: |
| | Available at: |
| | Files: |
| **Authorship** | Written by: |
| | Contributors: |
| | Reviewed by: |
| | Approved: |

## 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

# 1. Introduction

The increasingly high profile of issues surrounding preservation and access of research data makes it important to work towards harmonisation of data policies across the research base. PaNOSC brings together a significant number of major world-class European research infrastructures to lay the foundation for a fully integrated, pan-European, information infrastructure supporting the complete scientific cycle from experiment definition to publication.
This document describes the common framework for scientific data management at photon and neutron facilities. The participating facilities are used by researchers in universities, publicly funded research entities, and industry. The general process includes the generation of raw data from each experiment, which is then analysed by the research team. The results of non-proprietary research are published in the standard periodic literature and publicly available. In case of proprietary research, beamtime has to be purchased by the experimental team and the results are kept confidential.
The data format recommended by PaNdata for the raw data is NEXUS/HDF5, which in addition to the detector data includes administrative metadata, instrument metadata and scientific metadata. The data framework presented in this document focuses on this "raw data" stage. It must be stressed that the full strength of this digital approach is only reached when all data from detector data to final publication are included, giving full advantage to the experimenting team and the scientific community.
The framework has been strongly influenced by the OECD "*Principles and guidelines for access to research data from public funding*". It strives for a careful balance between aspects of competition and collaboration in science.
Having an open access data policy with data in well-defined formats has many benefits:

- Raw data becomes open to scrutiny by other researchers, which helps to uncover cases of scientific fraud. Thus open access policies foster scientific integrity.
- It makes previously measured data available for further analysis without the necessity to repeat the example.
- It promotes interdisciplinary research.
- Scientists can mine data in previously unknown ways or reapply new methods to existing data.

There are in part controversial views concerning the ownership of raw data (e.g. universities or researchers as persons). These issues will come into play when researchers are changing affiliation. Solutions may depend on local and national rules.

Say something about fair.

# 2. Recommendation

The PaNOSC consortium strongly recommends to their respective facility managements to adopt and publish a defined data policy framework. Having an identical approach to the management of scientific data will ease the life of scientists using more than one facility and add to the overall transparency of the scientific process. It is suggested that the implementation at each facility should

be versioned and citable using a persistent identifier. Metadata should be preserved as long as possible after the deletion of the raw data.

Legal compliancy: If and to the extent scientific data and metadata include personal data, the data protection legislation of the [*country of facility*] and the European Union, respectively, will be applicable. This Scientific Data Policy will be governed by and constructed in accordance with the law of the [*country of facility*]. Exclusive place of jurisdiction is [*location + country of court*].

# 3. Generic data management policy

## 3.1. Definitions

For the purposes of this policy:

3.1.1. The term **facility** refers to one of the Photon and Neutron facilities participating in the PaNOSC initiative.

3.1.2. The term **raw data** pertains to data collected from experiments performed on facility instruments. This definition includes data that are created automatically or manually by facility specific software and/or facility staff expertise in order to facilitate subsequent analysis of the experimental data.

3.1.3. The term **metadata** describes information pertaining to data collected from experiments Instruments, including (but not limited to) the context of the experiment, the experimental team, experimental conditions, electronic logbooks generated during the experiment and other logistical information.

3.1.4. The term **principle investigator** (PI) pertains to the PI identified on the experiment proposal. For experiments outside of the facilities proposal system, the PI is the person initiating or performing the experiment.

3.1.5. The term **experimental team** includes the PI and any other person to whom the PI designates the right to access resultant raw data and associated metadata.

3.1.6. The term **public research** refers to research done through peer review and leading to publication(s).

3.1.7. The term **proprietary research** refers to research done through purchased (commercial) access to the research facility.

3.1.8. The term **on-line catalogue** pertains to a computer database of metadata containing links to raw data files, that can be accessed by a variety of methods, including (but not limited to) web-based browsers.

3.1.9. The term **results** pertain to data, intellectual property, and outcomes arising from the analysis of raw data. This does not include publications.

3.1.10. The term **open access** means belonging to the community at large, unprotected by copyright or patent and subject to appropriation by anyone.

3.1.11. The term **DMP** pertains to a Data Management Plan. A defined strategy that can include data types, volumes, metadata requirements archival duration processing and analysis requirements.

3.1.12. The term **auxiliary data** pertains to data that provides contextual information regarding the experiment and its dataset. Such as sample provenance information.

3.1.13. The term **processed data** pertains to the resultant data from processing of raw data.

## 3.2. General principles

TODO: refer to FAIR[1]

3.2.1. This data management policy pertains to the curation of and access to scientific data and metadata collected and/or stored at the facility.

3.2.2. Acceptance of this policy is a condition of the award of beamtime.

3.2.3. Users shall not attempt to access, exploit or distribute raw data or metadata unless they are entitled to do so under the terms of this policy.

3.2.4. Deliberate infringements of the policy may lead to denial of access to raw data or metadata and/or denial of future beamtime requests at the facility.

3.2.5. All data and metadata will be subject to the data protection legislation of the country in which the data and metadata are stored.

3.2.6. The facility should provision a method that allows updates to this policy

3.2.7. Users shall ensure raw data and processed data are collected with accurate meta data such that raw and processed data are FAIR. The facility will define a minimum subset of metadata as an appendix to this policy.

3.2.8. Users shall endeavour to include auxiliary data to augment the experimental data.

3.2.9. Users are encouraged to generate a DMP and if required will receive assistance from the facility to do so.

## 3.3. Persistent identifiers

3.3.1. Persistent identifiers shall be generated for raw data and metadata

3.3.2. Persistent identifiers shall be generated for processed data that is generated by facility maintained automated systems.

3.3.3. Persistent identifiers can be generated for custom data sets of raw, process and metadata.

3.3.4. The experiment team shall be able to create a DOI for a custom data set that has been / is to be made open access.

3.3.5. Users shall cite the data persistent identifier in any publication that refers to the data (or a subset).

3.3.6. The experiment team may include/append a persistent identifier to curated

auxiliary data.

3.3.7. Personally identifiable information that is included as part of a persistent identifier generated by the facility shall be generated in compliance with the privacy policy of the facility.

## 3.4. License

3.4.1. The Research Infrastructure will release open data under an appropriate license.

## 3.5. Raw data and associated metadata

3.5.1. All raw data and the associated metadata obtained as a result of publicly funded access to the research facilities are open access, with the research facility acting as the custodian.

3.5.2. All raw data and the associated metadata obtained as a result of proprietary research will be owned exclusively by the client who purchased the access. Proprietary users must agree with the facility management how they wish their raw data and metadata to be managed before the start of any experiment. [2]

3.5.3. It is the responsibility of the PI to ensure that the meta data collected meets the minimum required by the facility.

3.5.4. Curation of raw data and associated metadata:

3.5.4.1. All raw data will be curated in well-defined formats, for which the means of reading the data will be made available by the facility.

3.5.4.2. Metadata that is automatically captured by instruments will be curated either within the raw data files, within an associated on-line catalogue, or within both.

3.5.4.3. The facility will allow user supplied equipment access to write meta data for curation.

3.5.4.4. Raw data and metadata will be curated by the facility for a minimum of 10 years.

3.5.4.5. Data will be read-only for the duration of its life-time.

3.5.4.6. Data will be migrated or copied to archival facilities for long-term curation.

3.5.4.7. It is planned that each data set will have a unique identifier. Anybody providing data with the same identifier must make sure that the copy is identical to the data in the facility database. Anybody publishing results based on open access data must quote the same identifier (and related publications if available & appropriate).

3.5.5. Access to raw data and metadata:

3.5.5.1. Access to raw data and metadata in the facility is foreseen to be via a searchable on-line catalogue.

3.5.5.2. Access to the on-line catalogue of the facility will be either open access or restricted to those who are registered users of the on-line catalogue.

*(Registration may be necessary for certain access to open access data due to potential bandwidth problems with large data sets. The underlying AAI (Authentication and Authorisation Infrastructure) is being worked on within PaNOSC and other EU funded projects.)*

3.5.5.3. Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of 3 years after the end of the experiment. Thereafter, it will become openly accessible. The PI can request an extension of the restricted access period by following the facility defined procedure for extension. Data can always be made openly accessible earlier on simple request of the PI.

3.5.5.4. Appropriate facility staff (e.g. instrument scientists, computing group members) has access to any facility curated data or metadata for facility related purposes. Every facility will undertake that they will preserve the confidentiality of such data.

3.5.5.5. The on-line catalogue will enable the linking of experimental data to experimental proposals. Access to proposals will only ever be provided to the experimental team and appropriate facility staff, unless otherwise authorized by the PI.

3.5.5.6. The PI has the right to transfer or grant parts or all of their rights to another registered person.

3.5.5.7. The PI has the right to create and distribute copies of their raw data.

## 3.6. Processed data and metadata

3.6.1. All processed data and meta data that is generated by facility maintained automated systems during publicly experiments will be made open access with the facility as custodian.

3.6.2. All processed data and meta data generated by facility maintained automated systems during proprietary research will be owned exclusively by the client who purchased the access. Proprietary users must agree with the facility management how they wish their processed data and meta data to be managed before the start of any experiment. [3]

3.6.3. All processed data and meta data generated by the instrument team during or after the experiment will not be made open access after the period of restricted access.

3.6.4. Users can make manually processed data open access if required.

3.6.5. Curation of processed data and metadata:

3.6.5.1. Processed data generated by facility maintained systems shall be curated in well-defined formats.

3.6.5.2. The facility does not guarantee readability for user generated processed

data.

3.6.5.3. Processed data and meta data will be curated by the facility for a minimum of 10 years.

3.6.6. Access to processed data and metadata:

3.6.6.1. Access to processed data and metadata in the facility will be primarily through a searchable on-line catalogue.

3.6.6.2. Access to the on-line catalogue of the facility will be either open access or restricted to those who are registered users of the on-line catalogue.

3.6.6.3. Access to processed data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of 3 years after the end of the experiment. Thereafter, it will become openly accessible.

3.6.6.4. The PI can request an extension of the restricted access period by following the facility defined procedure for extension.

3.6.6.5. Data can always be made openly accessible earlier on simple request of the PI.

## 3.7. Auxiliary data

3.7.1. Auxiliary data for publicly funded experiments should be made open access with the facility acting as custodian.

3.7.2. Auxiliary data for proprietary research will be owned exclusively by the client who purchased the access. Proprietary users must agree with the facility management how they wish their processed data and meta data to be managed before the start of any experiment. [4]

3.7.3. Curation of auxiliary data:

3.7.3.1. Auxiliary data shall be curated the original format.

3.7.3.2. The facility does not guarantee that readability of auxiliary data.

3.7.3.3. Auxiliary data and meta data will be curated by the facility for a minimum of 10 years.

3.7.3.4. The upload of auxiliary data may be subject to volume restrictions.

3.7.3.5. The facility cannot be made liable in case of unavailability or loss of data.

3.7.3.6. The facility cannot be made liable in case of unavailability or loss of data analysis software.

3.7.4. Access to auxiliary data:

3.7.4.1. Access to auxiliary data will be primarily through a searchable on-line catalogue.

3.7.4.2. Access to the on-line catalogue of the facility will be either open access or restricted to those who are registered users of the on-line catalogue.

3.7.4.3. Access to processed data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of 3 years after the end of the experiment. Thereafter, it will become openly accessible.

3.7.4.4. The PI can request an extension of the restricted access period by following the facility defined procedure for extension.

3.7.4.5. The license of auxiliary data shall determine subsequent access rights.

3.7.4.6. Data can always be made openly accessible earlier on simple request of the PI.

## 3.8. Results

3.8.1. The intellectual property rights for results derived from the analysis of the raw data is determined by the contractual obligations of the person(s) performing the analysis.

3.8.2. Curation of results:

3.8.2.1. Each facility will provide a means for users to upload results and associated metadata to the facility and enable them to associate these results with raw data collected from the facility.

3.8.2.2. The upload of results and associated metadata may be subject to volume restrictions.

3.8.2.3. These results will be stored long-term by the originating facility. It will not be the responsibility of each facility to fully curate this data e.g. to ensure that software to read / manipulate this data is available.

3.8.2.4. The facility cannot be made liable in case of unavailability or loss of data.

3.8.2.5. The facility cannot be made liable in case of unavailability or loss of data analysis software.

3.8.3. Access to results:

3.8.3.1. Access to the results of analyses performed on raw data and metadata is restricted to the person or persons performing the analyses, unless otherwise requested by those persons. However, if the raw data being analysed is still restricted, access to the analysis results must be granted to the PI on request.

## 3.9. Good practice for metadata capture and results storage

2.9.1. The experimental team is encouraged to ensure that experiments metadata are as complete as possible, as this will enhance the possibilities for them to search for, retrieve and interpret their own data in the future.

2.9.2. Each facility undertakes to provide means for the capture of such metadata items that are not automatically captured by an instrument, in order to facilitate recording the

fullest possible description of the raw data.

2.9.3. Researchers who aim to carry out analyses of raw data and metadata which are openly accessible should, where possible, contact the original PI to inform them and suggest a collaboration if appropriate. Researchers must acknowledge the source of the data and cite its unique identifier and any publications linked to the same raw data.

2.9.4. PIs and researchers who carry out analyses of raw data and metadata are encouraged to link the results of these analyses with the raw data / metadata using the facilities provided by the on-line catalogue. Furthermore, they are encouraged to make such results openly accessible.

## 3.10. Warranty and liability regarding scientific data, metadata and results

3.10.1. The facility [*name*] will at its own discretion use reasonable efforts to ensure an accurate storing and curating as well as an uninterrupted access in accordance with the acknowledged IT standard. However, failures caused by technical or human mistakes cannot be ruled out regarding any data processing. The facility [name] cannot warrant an absolutely accurate storing and curating. Also, access might be temporarily limited or impossible, especially due to necessary maintenance or overhaul services or failure of third-party service providers.

3.10.2. The facility [*name*] shall not be liable in case of lost, inaccurate, or defective scientific data, metadata, or results as well as for access being limited or unavailable unless the facility [name], a representative, agent, or employee of the facility [name] acted in a grossly negligent manner or intentionally.

## 3.11. Termination of custodianship

3.11.1. If the facility [name] decides to not continue to act as custodian and/or to maintain and provide the metadata catalogue, the facility [name] will inform the PIs concerned in a timely manner and provide them with effective means to make a copy of the respective raw data, metadata, calibration data, alignment data, and results, provided facility [name] is aware of the effective email address of the PI at that time.

# APPENDIX - CHANGES

For changes to the original version ...

---

TODO: check DP against recommendations in the Making FAIR reality document

is this clause sufficient for proprietary access. should it still be included?