Deliverables

Deliverable Number	**D2.2**
Deliverable Title	**Data Management Plan (DMP) based on Data Policies of the different facilities**
Lead Beneficiary	**PSI**
Type	**ORDP: Open Research Data Pilot**
Dissemination Level	**Public**
Due date of delivery	**Month 6**

Table of Contents

FOR MORE INFO PLEASE CONTACT
Mirjam van Daalen (PSI) (author)
mirjam.vandaalen@psi.ch
0041/56/3105674

Premise

The present deliverable describes the Data Management plan for NA1, i.e. WP2 User tools for access and data management. The existing data policies of facilities (mostly based on the framework created within PaNdata Europe Deliverable D2.1; see ANNEX1) are the basis for this first version of the CALIPSOplus Data Management Plan.

The present Data Management Plan makes research data that are generated in the frame of the CALIPSOplus project findable, accessible, interoperable and reusable (FAIR). The DMP makes sure that the data are soundly managed.

The present DMP follows the guidelines on FAIR Data Management in Horizon 2020 given by the European Commission (Version 3.0; July 2016).

Description of Work

WP2: User tools for access and data management

1. Introduction

The present document presents the first version of a Data Management Plan for data generated within the frame work of the CALIPSOplus project. Data Management Plans are part of the Open Data Research Pilot of the European Commission.

2. SURVEY amongst CALIPSOplus members

In order to find out how the issue of data ownership, data curation and data archiving and open data access is handled at the different member facilities a survey was done amongst the members containing the following questions:

1. Which Data Policy is in place at you institute?
2. Do you have Data Management Plan for national research programmes?
3. Has your facility some e-infrastructures like metadata catalogue software in place?
4. How data curation is regulated at your facility?

Answers of beneficiaries CALIPSOplus:

Which Data Policy is in place at your institute?

	CALIPSOplus partners	**Country**	**Which Data Policy is in place?**
1	HELMHOLTZ-ZENTRUM DRESDEN-ROSSENDORF EV Germany	Germany	PaNdata*
2	ANKARA UNIVERSITESI	Turkey	No data policy
3	AARHUS UNIVERSITET	Denmark	No data policy
4	CONSORCIO PARA LA CONSTRUCCION EQUIPAMIENTO Y EXPLOTACION DEL LABORATORIO DE LUZ DE SINCROTRON	Spain	PaNdata*
5	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	France	No data policy
6	STIFTUNG DEUTSCHES ELEKTRONEN-SYNCHROTRON DESY	Germany	PaNdata*
7	DIAMOND LIGHT SOURCE LIMITED	United Kingdom	PaNdata* planned 2018
8	ELETTRA - SINCROTRONE TRIESTE SCPA	Italy	PaNdata*

9	EUROPEAN X-RAY FREE-ELECTRON LASER FACILITY GMBH	Germany	PaNdata*
10	HELMHOLTZ-ZENTRUM BERLIN FUR MATERIALIEN UND ENERGIE GMBH	Germany	PaNdata*
11	ISTITUTO NAZIONALE DI FISICA NUCLEARE	Italy	No data policy
12	INSTALLATION EUROPEENNE DE RAYONNEMENT SYNCHROTRON	France	PaNdata*
13	KARLSRUHER INSTITUT FUER TECHNOLOGIE	Germany	No data policy
14	LUNDS UNIVERSITET	Sweden	PaNdata*
15	PAUL SCHERRER INSTITUT	Switzerland	PaNdata*
16	STICHTING KATHOLIEKE UNIVERSITEIT	Netherlands	PaNdata* planned early 2018
17	SYNCHROTRON-LIGHT FOR EXPERIMENTAL SCIENCE AND APPLICATIONS IN THE MIDDLE EAST	Jordan	No user operation yet
18	Société Civile Synchrotron SOLEIL	France	PaNdata* planned early 2018
19	UNIWERSYTET JAGIELLONSKI	Poland	PaNdata* planned 2018

Table 1. Data Policies at different facilities

Table 1 shows that 13 out of 19 partners of CALIPSOplus have a data policy in place, or planned, based on the PaNdata Data policy. The rest of the facilities are in the process of planning to have the same policy. The PaNdata framework policy is a very good basis to handle the data that are generated in the frame of the CALIPSOplus project findable, accessible, interoperable and reusable (FAIR).

Do you have a DMP for national research programmes?

	CALIPSOplus partners	**Country**	**Data Management Plan for national Research Programmes**
1	HELMHOLTZ-ZENTRUM DRESDEN-ROSSENDORF EV Germany	Germany	No
2	ANKARA UNIVERSITESI	Turkey	No
3	AARHUS UNIVERSITET	Denmark	No
4	CONSORCIO PARA LA CONSTRUCCION EQUIPAMIENTO Y EXPLOTACION DEL LABORATORIO DE LUZ DE SINCROTRON	Spain	No
5	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	France	No
6	STIFTUNG DEUTSCHES ELEKTRONEN-SYNCHROTRON DESY	Germany	No
7	DIAMOND LIGHT SOURCE LIMITED	United Kingdom	No
8	ELETTRA - SINCROTRONE TRIESTE SCPA	Italy	No
9	EUROPEAN X-RAY FREE-ELECTRON LASER FACILITY GMBH	Germany	No
10	HELMHOLTZ-ZENTRUM BERLIN FUR MATERIALIEN UND ENERGIE GMBH	Germany	No

11	ISTITUTO NAZIONALE DI FISICA NUCLEARE	Italy	No
12	INSTALLATION EUROPEENNE DE RAYONNEMENT SYNCHROTRON	France	No
13	KARLSRUHER INSTITUT FUER TECHNOLOGIE	Germany	No
14	LUNDS UNIVERSITET	Sweden	No
15	PAUL SCHERRER INSTITUT	Switzerland	Yes, there are drafts for Swiss National Foundation DMP's
16	STICHTING KATHOLIEKE UNIVERSITEIT	Netherlands	Yes, there are drafts for data managements plans available for national research programmes at the Faculty of Science and the Institute for Molecules and Materials at the Radboud University.
17	SYNCHROTRON-LIGHT FOR EXPERIMENTAL SCIENCE AND APPLICATIONS IN THE MIDDLE EAST	Jordan	No user operation yet
18	Société Civile Synchrotron SOLEIL	France	No
19	UNIWERSYTET JAGIELLONSKI	Poland	No

Table 2. Data management plans in place for national funding programs

Table 2 shows that only one of the 2 out of 19 partners of CALIPSOplus have a Data Management Plan in place for national research programs.

The implementation of a Data Policy can start only with the availability of a metadata catalogue software. Such a software manages the metadata of raw and derived data taken at experiment facilities (i.e. partners of CALIPSOplus). In a metadata catalogue different types of metadata are saved: 1) *administrative metadata:* data management lifecycle, ownership, file catalog and 2) *scientific metadata:* describing the sample, the beamline and experiment as well as parameters relevant for data analysis.

A data catalogue software enables management of the data lifecycle, i.e. from data aquisition to data analysis and eventual deletion of the data. The data can be linked to proposals and samples, to publications (DOI, PID) and can be migrated to and from longterm storage on tape.

A metadata catalogue helps keeping track of the data provenance (i.e. the steps leading to the final results) and it allows to check scientific integrity (checksum of data). It allows to find data based on the metadata (i.e. the users own data and handles open access to data). In the long term: metadata catalogues will help to automate standardised analysis workflows and support the standardisation of data formats.

Has your facility some e-infrastructures like metadata catalogue software in place?

	CALIPSOplus partners	**Country**	**Has your facility some e-infrastructures like Metadata Catagogue Software in place?**
1	HELMHOLTZ-ZENTRUM DRESDEN-ROSSENDORF EV Germany	Germany	Not planned yet
2	ANKARA UNIVERSITESI	Turkey	Not planned yet
3	AARHUS UNIVERSITET	Denmark	Not planned yet

4	ALBA - CONSORCIO PARA LA CONSTRUCCION EQUIPAMIENTO Y EXPLOTACION DEL LABORATORIO DE LUZ DE SINCROTRON	Spain	Alba offers as well a portal for remote access to data from the experimental team authenticated with the proposal ID. Alba is working on the implementation of ICAT metadata catalogues for the beamlines. This will be progressively implemented as well as the whole data policy on the beamlines. The plan is to have the first prototype working on a beamline by 2018. In parallel ALBA is also working in other specific macromolecular metadata laboratory information management systems (ISPyB)
5	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	France	Not planned yet
6	STIFTUNG DEUTSCHES ELEKTRONEN-SYNCHROTRON DESY	Germany	ICAT metadata catalogue
7	DIAMOND LIGHT SOURCE LIMITED	United Kingdom	ICAT metadata catalogue
8	ELETTRA - SINCROTRONE TRIESTE SCPA	Italy	ICAT metadata catalogue planned • We collect different type of data. Each beamline has its own data acquisition system. • Most of the beamline acquisition systems are implemented using the Tango control system. Data acquired are saved in the storage system in an area called scratch. • Tango, Labview
9	EUXFEL - EUROPEAN X-RAY FREE-ELECTRON LASER FACILITY GMBH	Germany	Planned
10	HELMHOLTZ-ZENTRUM BERLIN FUR MATERIALIEN UND ENERGIE GMBH	Germany	We use a hierarchical storage management system for the central storage and long time archival of the data. The bulk of the data will be on tape. We use ICAT as the metadata catalogue and user portal for the access to the data.
11	ISTITUTO NAZIONALE DI FISICA NUCLEARE	Italy	Not yet in place
12	ESRF - INSTALLATION EUROPEENNE DE RAYONNEMENT SYNCHROTRON	France	ICAT metadata catalogue
13	KIT - KARLSRUHER INSTITUT FUER TECHNOLOGIE	Germany	Not planned yet
14	LUNDS UNIVERSITET	Sweden	MELANI
15	PSI - PAUL SCHERRER INSTITUT	Switzerland	MELANI
16	FELIX - STICHTING KATHOLIEKE UNIVERSITEIT	Netherlands	Not yet in place
17	SESAME - SYNCHROTRON-LIGHT FOR EXPERIMENTAL SCIENCE AND APPLICATIONS IN THE MIDDLE EAST	Jordan	Not in user operation yet
18	Société Civile Synchrotron SOLEIL	France	Not yet in place
19	SOLARIS - UNIWERSYTET JAGIELLONSKI	Poland	Not yet in place

Table 3. Implementation of metadata catalogue software

How data curation is regulated at your facility?

	CALIPSOplus partners	**Country**	**How data curation is regulated at your facility?**
1	HELMHOLTZ-ZENTRUM DRESDEN-ROSSENDORF EV Germany	Germany	-

2	ANKARA UNIVERSITESI	Turkey	-
3	AARHUS UNIVERSITET	Denmark	The curation of data is carried out by the beam line scientists. All of the original data files are kept here at the facility, but users are given copies of all their data for analysis and publication.
4	ALBA - CONSORCIO PARA LA CONSTRUCCION EQUIPAMIENTO Y EXPLOTACION DEL LABORATORIO DE LUZ DE SINCROTRON	Spain	Access to raw data and the associated metadata obtained from a public access experiment is restricted to the experimental team for 3 years. After this embargo period, the data can be made publicly available. These data are remotely accessible by the research group. Data preprocessing and analysis is depending on the Beamline partially done at the Alba premises and completed by the researchers at their home institutes. This means that the data repository contains always raw data and in some cases processed and curated data.
5	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	France	-
6	STIFTUNG DEUTSCHES ELEKTRONEN-SYNCHROTRON DESY	Germany	Long term archiving in tape library (up to 10 years)
7	DIAMOND LIGHT SOURCE LIMITED	United Kingdom	-
8	ELETTRA - SINCROTRONE TRIESTE SCPA	Italy	• The users decide if the data acquired are of good quality and if the case they transform them into datasets. Datasets are saved in an area of the storage system called online where they can be processed, accessed. In the long term the idea is to move the data from the online area to the offline data that in principle can be remote. • Preferred data format is HDF5.
9	EUXFEL - EUROPEAN X-RAY FREE-ELECTRON LASER FACILITY GMBH	Germany	Long term archiving in tape library (up to 10 years) • Preferred data format is HDF5.
10	HELMHOLTZ-ZENTRUM BERLIN FUR MATERIALIEN UND ENERGIE GMBH	Germany	We are currently implementing the Data Policy. Data curation is still work in progress.
11	ISTITUTO NAZIONALE DI FISICA NUCLEARE	Italy	We backup data on dedicated external Hard Drives
12	ESRF - INSTALLATION EUROPEENNE DE RAYONNEMENT SYNCHROTRON	France	The deadline for implementing the data policy on all ESRF beamlines is in 2020. At the moment 11 beamlines are connected to the metadata catalogue (6 are inprogress), data archiving is work in progress at 17 beamlines. This is long term archiving in tape library (data are being archived for 10 years in tape archive)
13	KIT - KARLSRUHER INSTITUT FUER TECHNOLOGIE	Germany	-
14	LUNDS UNIVERSITET	Sweden	Not yet in place
15	PSI - PAUL SCHERRER INSTITUT	Switzerland	• Data Analysis as a service DaaS project: focusses on offline data analysis and large offline disk storage. Finishes end of October 2017 • Petabyte Archive: focuses on enabling the data flows to and from a longterm data storage at HPC CSCS/Lugano. Finishes end of 2017 (data storage up to 10 years) • Data Curation Project, collaboration with ESS, focusses on data catalog and data analysis automation. Started 2017 and will last until end of 2019. Enabled to add dedicated manpower for data curation tasks.

16	FELIX - STICHTING KATHOLIEKE UNIVERSITEIT	Netherlands	The experimental data are stored on a facility server; maintainenance and back-up is organized centrally by the Radboud University IT department. We are currently exploring and evaluating the possibilities to use local (Radboud) and a national (e.g. DANS) repositories.
17	SESAME - SYNCHROTRON-LIGHT FOR EXPERIMENTAL SCIENCE AND APPLICATIONS IN THE MIDDLE EAST	Jordan	Not in user operation yet
18	Société Civile Synchrotron SOLEIL	France	Not yet in place
19	SOLARIS - UNIWERSYTET JAGIELLONSKI	Poland	Not yet in place

Table 4. Data curation at different facilities.

Implementation of Data Management Plan for CALIPSOplus

As is clear from the survey 11 of 19 partners, by the end of 2018 will have a data policy in place based on the PaNdata data policy (Deliverable D2.1. of PaNdata Europe FP7 project in 2011). The PaNData data policy frame work defines (long term) goals concerning data storage, life cycle management, data access and ownership. Implementation of PSI data policy needs a metadata catalogue. This implies that the implementation of the data policies can only start with the availability of metadata catalogue software. Some facilities use the iCAT software (see Table 2) that was developed within the PaNdata ODI FP7 project, others are currently developing a new metadata catalogue software called MELANI (PSI, ESS, MAXIV) that is developed within the Data Analysis as A Service (DAAS) project of PSI. Implementation of the Data Policies is done step by step (i.e. role out from beamline to beamline) at all facilities. This stepwise process implies that a DMP will only be complete once these processes at the different facilities have been finished.

FAIR Data management at a glance: DMP components to be covered

1. **Data summary:** Within this project **data** are **collected** during the experiments at the facilities. The data are collected by users that received transnational access money from the CALIPSOplus project. Most data are collected in the HDF5 or Nexus **format**. Data are open access after the embargo period of 3-5 years and can be re-used by third parties. Data origin is from the experimental station of the CALIPSoplus partner large scale facilities.

2. The two types of metadata catalogues planned to be used by the CALIPSOplus partners (ICAT and MELANI; see Table 3) make data findable, accessable, interoparable and reusable **(FAIR data),** with standards for metadata creation.

2.2.The timeline in which **data** become **accessible** is defined by the embargo period (3-5 years) in the specific facility data policy (see example of data policy in ANNEX 1).

2.3 The **interoperability of the data** is guaranteed by the metadata catalogue softwares in place.

3. The implementation of the data policies at the different facilities and with this the putting in place of metadata catalogue software is a just started and an ongoing process. Therefore the **allocation of recources**, i.e. the cost of making our data FAIR and the costs for long term preservation of data will be descibed in detail in the next update of our DMP which is due in month 19 of the project.

Conclusion

The stepwise implementation of the Data Policies (i.e. role out from beamline to beamline) at all facilities implies that a DMP will only be complete once these processes at the different facilities have been finished. The present version of the DMP reflects the status as it is now for the different partners of CALIPSOplus. In the next update of the DMP in month 19, a clear progress in role out of this process will be visible.

ANNEX 1 Example of Data Policy based on PaNdata Europe D2.1 ESRF and PSI in PDF Format

30 November 2015

The ESRF Data Policy

The ESRF aims to implement a Data Policy starting as soon as possible in 2016. The main elements of this policy comprise:

- **Data ownership**
- **Data curation**
- **Data archiving**
- **Open access to data**

This policy follows largely the recommendations of the PaN-data Europe Strategic Working Group laying out a common framework for scientific data management at photon and neutron facilities (Deliverable D2.1, PaN-data Europe, co-funded by the European Commission under the 7th Framework Programme)

1. General Principles

1.1. The present data management policy pertains to the ownership of, the curation of and access to experimental data and metadata collected and/or stored at the ESRF.

1.2. Acceptance of this policy is a condition for the award of beam time.

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

1.4. Deliberate infringements of the policy may lead to denial of access to raw data or metadata and/or denial of future beam time requests at the ESRF, as well as actions of the ESRF in the court of law.

1.5. All data and metadata will be subject to the data protection legislation of France.

2. Definitions

For the purposes of this policy:

2.1. The term **raw data** pertains to data collected from peer-reviewed and in-house experiments performed on ESRF's instruments and includes data collected from peer-reviewed experiments performed on CRG beamlines. 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.

2.2. The term **metadata** describes information pertaining to data collected from ESRF instruments, including (but not limited to) the context of the experiment, the experimental team, experimental conditions and other logistical information.

2.3. The term **principle investigator** (PI) pertains to the PI identified on the Experiment Proposal (for peer-reviewed experiments) or the Safety Approval Form (for in-house experiments).

2.4. 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.

2.5. The term **public research** refers to research done through peer review or access via in-house research beam time.

2.6. The term **proprietary research** refers to research done through purchased (commercial) access.

2.7. 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.

2.8. The term **result** pertains to data, intellectual property, and outcomes arising from the analysis of raw data. This does not include publications.

2.9 The term **custodian** refers to the Institute storing, curating and providing access to raw data, metadata and results.

2.10. The term **long-term** means a minimum of 5 years and the ESRF will strive for 10 years. This will depend on the type and volume of data concerned and the economical consequences associated with long-term data storage. Thus the ESRF reserves the right to restrict the storage periods or data sets in consultation with the respective communities of high data rate instruments.

2.11. The term **open access** means belonging to the community at large, unprotected by copyright or patent and subject to appropriation by anyone. The ESRF data archive will be made available under CC-BY

(Creative Commons BY, http://creativecommons.org/licenses/by/4.0/legalcode).

3. Raw data and associated metadata

3.1 Access to raw data and associated metadata

3.1.1. All raw data and the associated metadata obtained as a result of peer reviewed access to the ESRF, in-house research and use of Management Contingency beamtime excluding proprietary research will be open access after an initial embargo period during which access is restricted to the experimental team, represented by the PI.

3.1.2 ESRF is the custodian of the raw data and associated metadata.

3.1.3. 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 and is not covered by the ESRF data policy. Data from proprietary research will be removed after the experiment from ESRF disk storage, unless otherwise agreed with ESRF management before the start of the experiment.

3.2 Curation of raw data and associated metadata

3.2.1. All raw data and metadata will be curated in well-defined formats, for which the means of reading the data will be made available by the ESRF.

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

3.2.3. Only data with metadata generated by ESRF software will be archived.

3.2.4. Raw data and metadata will be read-only for the duration of their life time.

3.2.5. Raw data and metadata will be migrated or copied to archival facilities for long-term curation.

3.2.6. It is planned that each experiment and data set will have a unique persistent identifier. Anybody publishing results based on open access data must quote the same identifier (and related publications if available & required).

3.2.7 High level metadata such as Title, Authors, Abstract, Beamline will be made public as soon as the experiment has been carried out. This information will be available via the persistent identifier landing page on the web.

3.3 Access to raw data and metadata

3.3.1. Access to raw data and metadata is foreseen to be via a searchable on-line catalogue.

3.3.2. Access to the on-line catalogue of the ESRF will be restricted to registered users of the on-line catalogue. The ESRF sets up the on-line procedure to become a registered user of the on-line catalogue.

3.3.3. Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for an embargo period of 3 years after the end of the experiment. Thereafter, the data will become openly accessible. Any PI that wishes data to retain *restricted access* for a period longer than three years will have this possibility by submitting a written request, specifying the reasons for the proposed prolongation, to the ESRF Directors of Research who decide on the request. In exceptional circumstances, data can be made openly accessible earlier than 3 years if the PI or the ESRF Directors of Research inform the ESRF to do so.

3.3.4. It is the responsibility of the PI to ensure that the experiment number is correctly entered into the metadata for each raw data set.

3.3.5. Authorized ESRF staff (e.g. instrument scientists, computing group members) have access to any curated data or metadata for facility related purposes. ESRF will undertake that confidentiality of such data is preserved during the embargo period.

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

3.3.7. The PI has the possibility to transfer parts or the totality of her/his rights during the embargo period to another registered person.

3.3.8. The PI has the possibility to create and distribute copies of the raw data.

4. Results

4.1 Ownership of results

4.1.1. Ownership of all results (intellectual property) derived from the analysis of the raw data is determined by the contractual obligations of the person(s) performing the analysis.

4.2 Curation of results

4.2.1 The ESRF will provide curation of results on a best effort basis, and acts as custodian of results in the long term.

4.2.2. The ESRF cannot be made liable in case of unavailability or loss of data or results.

4.2.3. The ESRF cannot be made liable in case of unavailability or loss of data analysis software.

4.3 Access to results

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

5. Good practice for metadata capture and results storage

5.1. The experimental team is encouraged to ensure that experiments metadata are as complete as possible, as this will enhance the possibilities for everybody to search for, retrieve and interpret the data in the long term.

5.2. ESRF provides 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.

5.3. 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 her/him and suggest a collaboration if required. Researchers must acknowledge the source of the data and cite its unique identifier as well as any publications linked to the same raw data.

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

6. Publication information

6.1. Publications related to data from experiments carried out at ESRF must cite the persistent identifier of the experiment and data in their publication.

PAUL SCHERRER INSTITUT PSI	Registrierung AW-11-16-05
Data Policy for PSI research data	Ersetzt
Dokumenten-verantwortung: SwissFEL Science Officer	Seiten 8 Datum 11.08.2016
Klassifizierung: ☒ NICHT KLASSIFIZIERT / NOT CLASSIFIED ☐ INTERN / RESTRICTED ☐ VERTRAULICH / CONFIDENTIAL ☐ GEHEIM / TOP SECRET	

Introduction:

PSI's research facilities enable the production of new knowledge by scientific investigation. This new knowledge is extracted from data. The procedure on how to deal with this data in the short, medium and Long-term for access, curation, analysis and preservation is defined in this Data Policy. The need of defining a Data Policy has become increasingly urgent over the last decade in view of the constantly increasing amount of data produced in laboratories such as PSI, and of the ethical and jurisprudential implications linked to the production and preservation thereof. By defining a Data Policy the following two goals should be achieved:

- Improve scientific output by providing a basis for the life cycle management of scientific data.
- Support the recommendations concerning scientific integrity.

Related Documents

- This document adheres to the *Guidelines Research Integrity at PSI* (https://www.psi.ch/integrity/research-integrity). The scope of the present data policy is described under General Principles 1.1 in the data policy below.
- It is in alignment with the PSI information governance policy framework *Informationssicherheit am PSI AW 11-16-01.*
- This policy follows largely the recommendations of the PaN-data Europe Strategic Working Group laying out a common framework for scientific data management at photon and neutron facilities (Deliverable D2.1, PaN-data Europe, co-funded by the European Commission under the 7th Framework Programme[1]).

Dokumenten-verantwortung:	Unterschrift:	Datum:	Dokumenten-freigabe:	Unterschrift:	Datum:
VM16	Mirjam v Daalen	2.11.16	MG95		3.11.2016

Verteiler (Zugriffsberechtigte)

Organisationseinheit oder Person
PSI

Änderungen

Revision	Datum	Änderung	Autor
0.0	25.07.2016	Final Draft	Mirjam van Daalen
0.1	11.08.2016	Formatierung als Weisung	Tanja Wagner, Gerd Mann
0.2	02.11.2016	Fertigstellung des Dokuments	Mirjam van Daalen

Table of Contents

Data policy for PSI research data (the "Policy")

The Policy defines the rules for the following topics:

- **Data ownership**
- **Data curation**
- **Data archiving**
- **Open access to data**

1 General Principles

1.1. The present Policy pertains to the ownership of, the curation of and access to experimental data and Metadata collected and/or stored by PSI Research Infrastructure.

1.2. Acceptance of this Policy is a condition for the award of access to Research Infrastructures and a binding part of each PSI employment agreement.

1.3. Users must not attempt to access, exploit or distribute data or Metadata unless they are entitled to do so under the terms of this Policy.

1.4. Deliberate infringements of the Policy may lead to actions concerning the employment of PSI employees to denial of access to data or Metadata and/or denial of future beam time requests at PSI, as well as to legal actions.

1.5. All data and Metadata will be subject to Swiss law.

2 Definitions

For the purposes of this Policy the term:

2.1 **Custodian** refers to the head of the institute divisions (or a delegated person) storing, curating and providing access to Raw Data, Metadata and Results.

2.2 **Experimental Team** includes the PI and any other person to whom the PI designates the right to access data and associated Metadata related to this proposal.

2.3 **Long-term** means a minimum of five (5) years and PSI will strive for ten (10) years, depending on the type and volume of data concerned and the economic consequences associated with Long-term data storage. Thus, PSI reserves the right to restrict the storage periods or data sets in consultation with the respective communities.

2.4 **Metadata** means information pertaining to data collected from Research Infrastructures at PSI, including (but not limited to) the scientific and administrative context of the experiment, the Experimental Team and the experimental conditions.

2.5 **On-line Catalogue** pertains to a computer database of Metadata containing links to data files, that can be accessed by a variety of methods.

2.6 **Open Access** means belonging to the public at large, unprotected by most copyrights or patents and subject to appropriation by anyone. Those data will be made available under CC-BY-SA (https://creativecommons.org/licenses/by-sa/4.0/).

2.7 **Principle Investigator** (PI) means the main proposer identified on the experiment proposal for peer-reviewed experiments or the leader of the Experimental Team for non-peer-reviewed experiments, both at the PSI large-scale facilities, or at other Research Infrastructures.

2.8 **Proprietary** research refers to research done through commercial access.

2.9 **Public Research** refers to research done through peer reviewed experiments and experiments done during in-house research or management contingency beam time.

2.10 **Research Infrastructure** includes but is not limited to PSI facilities SLS, SINQ, SμS, SwissFEL, and other PSI research infrastructures.

2.11 **Raw Data** means data collected from experiments performed at PSI. This 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.

2.12 **Result** pertains to data, intellectual property, and outcomes arising from the analysis of Raw Data (i.e. derived data).

Unless there is something inconsistent in the subject or context, words denoting the singular number include the plural and vice versa; words denoting one gender include the other gender and the neuter.

3 Raw Data and Metadata

3.1 Data resulting from Public Research and Proprietary Research

3.1.1. All Raw Data and Metadata obtained as a Result of Public Research will be Open Access after an initial embargo period during which access is restricted to the Experimental Team, represented by the PI.

3.1.2. PSI is the Custodian of the Raw Data and Metadata.

3.1.3. All Raw Data and Metadata obtained as a Result of Proprietary Research will be owned exclusively by the client who purchased the ac-

cess and is not covered by this Policy. Data from Proprietary Research will be removed after the experiment from PSI storage, unless otherwise agreed with the PSI management before the start of the experiment.

3.2 Curation of Raw Data and Metadata

3.2.1. All Raw Data and Metadata will be curated in well-defined formats, for which the means of reading the data will be made available by PSI.

3.2.2. Metadata that are automatically captured by instruments will be curated either within the raw data files, within an associated On-line Catalogue, or within both.

3.2.3. Raw Data and Metadata will be read-only for the duration of their life time.

3.2.4. Raw Data and Metadata will be migrated or copied to archival facilities for Long-term curation.

3.2.5. Each experiment and data set will have a unique persistent identifier. Anybody publishing Results based on open access data must quote the same identifier (and related publications if available & required).

3.2.6. High level Metadata such as title, authors, abstract, specific Research Infrastructure will be made public as soon as the experiment has been carried out. This information will be available via the persistent identifier landing page on the web.

3.3 Access to Raw Data and Metadata

3.3.1. Access to Raw Data and Metadata is foreseen to be via a searchable On-line Catalogue in addition to a direct access protected file based access within the PSI network.

3.3.2. Access to the On-line Catalogue of PSI will be restricted to registered users of the On-line Catalogue. PSI sets up the on-line procedure to become a registered user of the On-line Catalogue.

3.3.3. Access to Raw Data and Metadata obtained from an experiment is restricted to the Experimental Team for an embargo period of three (3) years after the end of the experiment. Thereafter, the data will become openly accessible. Any PI that wishes data to retain *restricted access* for a period longer than three (3) years will have this possibility on a yearly basis on a maximum prolongation of two (2) years. For longer extension a written request shall be submitted, specifying the reasons for the proposed prolongation, to the head of the corresponding PSI division who decides on the request, all at their sole discretion. Data can always be made openly accessible earlier on simple request of the PI. In exceptional circumstances the head of the corresponding PSI division can grant access to official committees at any time for the purpose of verifying data integrity.

3.3.4. Raw Data and Metadata explicitly used for peer-reviewed publication will become Open Access at the time of such publication.

3.3.5. It is the responsibility of the PI to ensure that the experiment number and requirements concerning the confidentiality, integrity and availability of the data is correctly entered into the Metadata for each Raw Data set.

3.3.6. Authorized PSI staff (including but not limited to facility management, instrument scientists, computing group members) have access to any curated data or Metadata for facility related purposes. PSI will undertake that confidentiality of such data is preserved during the embargo period.

3.3.7. The On-line Catalogue will enable linking experimental data to experimental proposals. Access to the full proposal text will only be provided to the Experimental Team and appropriate facility staff, unless otherwise authorized by the PI.

3.3.8. The PI has the possibility to transfer any or all of his rights during the embargo period to another person. This transfer has to be documented.

3.3.9. The PI may to create and distribute copies of the Raw Data, without violating against the rules that apply during the embargo period.

3.3.10. To the extent permitted by law, PSI cannot be held liable in the case of unavailability or loss of Raw Data.

4 Results

4.1 Ownership of Results

Ownership of all Results (intellectual property) derived from the analysis of the Raw Data is determined by the contractual obligations of the person(s) performing the analysis.

4.2 Curation of Results

4.2.1. PSI will provide curation of Results on a best effort basis, and acts as Custodian of Results in the long term.

4.2.2. PSI cannot be held liable in case of unavailability or loss of data or Results.

4.2.3. PSI cannot be held liable in case of unavailability or loss of data analysis software.

4.3 Access to Results

Access to the Results of analysis performed on Raw Data and Metadata, is restricted to the person performing the analysis, unless otherwise requested by that person. However, if the Raw Data being analysed is still restricted, access to the analysis results must be granted by the PI on request.

5 Leading practice for Metadata capture and Results storage

5.1. The Experimental Team shall ensure that experiments Metadata are as complete as possible, as this will enhance the possibilities for everybody to search for, retrieve and interpret the data in the long term.

5.2. PSI provides 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.

5.3. Researchers who carry out analyses of Raw Data and Metadata which are openly accessible shall, to the extent practicable, contact the original PI to inform him and suggest a collaboration if required. Researchers must acknowledge the source of the data and cite its unique identifier as well as the original publication linked to the same Raw Data.

5.4. PI and researchers who carry out analyses of Raw Data and Metadata shall link the Results of these analyses to the Raw Data / Metadata using the facilities provided by the On-line Catalogue. Furthermore, PI and researchers shall make such Results openly accessible.

6 Publication information

Publications related to data from experiments carried out at PSI must cite the persistent identifier of the experiment and data in their publication.

Implementation

The data policy for PSI research data will be implemented over time, depending foremost on our capacity to tag raw data with meaningful Metadata. A Metadata catalogue will be installed. A process document describes the details.

[1] http://pan-data.eu/PaNdataEurope