# PaN-data Europe

## Deliverable D2.1

## Common policy framework on scientific data

| | |
|---|---|
| Grant Agreement Number | 261537 |
| Project Title | PaN-data Europe Strategic Working Group |
| Title of Deliverable | Common policy framework on scientific data |
| Deliverable Number | D2.1 |
| Lead Beneficiary | ESRF |
| Deliverable Dissemination Level | Public |
| Deliverable Nature | Report |
| Contractual Delivery Date | 1 Oct 2010 (Month 4) |
| Completed on | 14 December 2010 |
| Actual Delivery Date | 11 Feb 2011 |

*The PaN-data Europe project is partly funded by the European Commission under the 7th Framework Programme, Information Society Technologies, Research Infrastructures.*

**Abstract**

The increasingly high profile of issues surrounding preservation and access of research data makes it important to work towards harmonisation of data polices across the research base. PaN-data 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. The participating facilities are used by researchers in universities, publicly funded research entities, and industry. This document describes the common framework for scientific data management at photon and neutron facilities. This document and any future revisions is also available at www.pandata.eu.

**Keyword list**

Data, Metadata, Data policy, Data Access, Data catalogue, Preservation, Curation

**Document approval**

Approved for submission to EC by all partners on 11 February 2011.

**Revision history**

| Issue | Author(s) | Date | Description |
|---|---|---|---|
| V1.0 | M. Wilson and R. McGreevy | 19 June 2009 | |
| V1.1 | R. Dimper | 30 June 2009 | |
| V1.2 | H.-J. Weyer | 8 July 2009 | New introduction |
| V1.3 | F. Schlünzen and H.-J. Weyer | July 2009 | Comments by e-mail |
| V1.4 | | 5 October 2010 | Minor changes at Berlin PaN-data meeting |
| V1.5 | R. Dimper | 14 October 2010 | Complete rework, removed e-mails, integrated comments, removed facility specific parts |
| V1.6 | M. Könnecke, S. Egli | 21 October 20O0 | Additions from PSI draft hopefully useful for the other sites as well |
| V1.7 | R. Dimper | 14 November 2010 | Implemented various remarks from partners |
| V1.8 | R. Dimper | 14 December 2010 | Final version |

# Table of contents

Page

**1 INTRODUCTION ..... 4**

**2 GENERIC DATA MANAGEMENT POLICY ..... 5**

1 GENERAL PRINCIPLES ..... 5
2. DEFINITIONS ..... 5
3. RAW DATA AND ASSOCIATED METADATA ..... 6
4. RESULTS ..... 7
5. GOOD PRACTICE FOR METADATA CAPTURE AND RESULTS STORAGE ..... 8
6. PUBLICATION INFORMATION ..... 9

# 1 Introduction

The increasingly high profile of issues surrounding preservation and access of research data makes it important to work towards harmonisation of data polices across the research base. PaN-data 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 PaN-data 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’s “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 remeasure the same sample.
- 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.

**Recommendation**

The PaN-data consortium strongly recommends to their respective facility managements to adopt and publish the data 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.

# 2 Generic data management policy

**1 General principles**

1.1. This data management policy pertains to the ownership of, the curation of and access to experimental primary data and metadata collected and/or stored at the facility.

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

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 beamtime requests at the facility.

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

**2. Definitions**

For the purposes of this policy:

2.1. The term **facility** refers to one of the Photon and Neutron facilities participating in the PaN-Data initiative.

2.2. 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.

2.3. 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 and other logistical information.

2.4. 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.

2.5. 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.

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

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

2.8. 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.

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

2.10. The term **long-term** means a minimum of 5 years and facilities will thrive for 10 years. This may obviously depend on the type and volume of data concerned and the economical consequences associated to long-term data storage. Thus the facility reserves the right to restrict the storage periods in consultation with the respective communities for 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. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.

## 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 publically funded access to the research facilities are open access, with the research facility acting as the custodian.

3.1.2. 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.

### 3.2 Curation of raw data and associated metadata

3.2.1. 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.

3.2.2. 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.

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

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

3.2.5. 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 data-

base. Anybody publishing results based on open access data must quote the same identifier (and related publications if available & appropriate).

### 3.3 Access to raw data and metadata

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

3.3.2. 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 PaN-data and other EU funded projects.*)

3.3.3. 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. Any PI that wishes their data to remain *restricted access* for a longer period will be required to make a special case to the respective facility management. If data can only be stored at the facility for less than three years, then access is exclusive to the PI up to the end of the storage period. Data can always be made openly accessible earlier on simple request of the PI.

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, in order to correctly associate each data set with the PI. If this is not done, the experimental team will not be able to access the data via the on-line catalogue or other users may inadvertently be given access rights to the data.

3.3.5. 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.

3.3.6 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.

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

3.3.8. The PI has the right to create and distribute copies of his 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. 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.

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

4.2.3. 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.

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

4.2.5. The facility 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 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.

## 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 them to search for, retrieve and interpret their own data in the future.

5.2. 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.

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 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.

5.4. 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.

## 6. Publication information

6.1. References for publications related to experiments carried out at the facilities must be deposited in a publications database within 3 months of the publication date, or during any new application for beamtime, whichever is the earlier.