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

Deliverable: D2.3 - Guidelines on best practices implementing the PaNOSC data policy framework

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:	D2.3
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	31 November 2020
Actual Delivery Date:	31 May 2021
EC project Officer:	Simona Misiti

Document Control Sheet

Document	Title: Guidelines for implementing a Research Data Policy
	Version: DRAFT 31 May 2021
	Available at: https://github.com/panosc-eu/panosc
	Files: 1
Date	31 May 2021
Authorship	Written by: A.Götz (ESRF), J.Taylor (ESS), R.Dimper (ESRF), J-F.Perrin (ILL+ESRF), F.Gliksohn (ELI-DC), D.Roccella (CERIC-ERIC), K.Wrona (EuXFEL)
	Contributors: T. Ivănoaica (ELI-DC), J.Malka (EuXFEL)
	Reviewed by: C.Collins (DLS)
	Approved: J.Bodera Sempere (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

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

Table of Content

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

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

Guidelines for implementing a Data Policy at Photon and Neutron Research Infrastructures

1. Purpose

The purpose of this document is to provide guidelines on how to implement a research data policy at Photon and Neutron (PaN) sources and at Research Infrastructures (RI) in general. The document is organised as a set of annotated guidelines based on the experience of six PaN institutes who have either already implemented a data policy or are in the process of implementing a data policy. The 6 institutes are - ESRF, ILL, EuXFEL, ESS, CERIC-ERIC and ELI. They are all members of the PaNOSC project.

2. Scope

The scope of the guidelines are to cover all steps a typical RI has to go through when implementing a Research Data Policy (RDP). It covers the steps of how to motivate adopting an RDP, how to write an RDP, how to adopt it, how to implement an RDP, and how to get it accepted.

Note: *This is a live document which will be updated regularly throughout the course of the PaNOSC project. The first version (the one submitted to the EC) covers mainly the period before the publication of the PaNOSC FAIR Data Policy Framework (DPF) [1]. The updating of existing data policies is still under discussion at most sites, therefore they are not fully covered in the first version of the guidelines. The next version of the document will cover in more detail the period following the publication of the PaNOSC FAIR DPF and how it has been applied to update data policies already in place.*

3. Guidelines

The guidelines are based on the feedback to the questions below. For each question a guideline is presented followed by a case study from each of the PaNOSC partners on the guideline.

3.1 Who are the main drivers within an Organization to adopt a Research Data Policy?

3.1.1 Guideline

Adopting a data policy is a management decision because the data policy will be part of the governance documents of the RI. The main drivers should include top management. They will need to be supported by IT experts, scientists and data managers.

3.1.2 Case Studies

ESRF: The main drivers were the Directors and Heads of Division. They were assisted by the two IT experts

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

who acted as data policy experts as they had participated in the writing of the PaNdata data policy framework. Together with other IT experts they estimated the costs of implementing the data policy and provided background information on what other sites (ILL and ISIS) had done to implement their RDP. The proposed DP was presented to the ESRF scientists before being presented and discussed by the SAC (Science Advisory Committee) and finally endorsed by the Council. Some ESRF scientists were openly in favour of the RDP. Most of them had no strong opinion for or against while only one or two expressed strong doubts about the Open Data policy either because they thought it not technically / financially feasible to store all raw data, or they doubted the usefulness of the RDP. A working group made up of beamline scientists, data managers, user office and IT specialists was set up in 2021 to discuss issues on data management. The working group discussed the updating of the ESRF data policy based on the PaNOSC DPF in May 2021.

ILL: The Data Policy project was initiated by the IT in the scope of the PaNData-Europe project in 2008 and was largely supported by the Directors during the 3 years of internal discussion that were necessary to obtain an acceptable consensus for all parties. It was officially adopted in Sept 2011 after validation by our Scientific Council. A working group (composed of instrument control, user office, scientific computing, IT and representatives of scientific groups) was then set up to discuss, steer and monitor its implementation. This working group has also the responsibility to propose policy modifications. The first policy revision, driven by this group, took place in July 2017.

EuXFEL: The introduction of Scientific Data Policy was driven by the IT and Data Management group with the strong support of the Scientific Director in their division. The draft of the policy was presented to the Management Board where main directions were endorsed. Then it was sent for comments to the leading Instrument Scientists and technical group leaders and also legal aspects were checked. After corrections were implemented the final draft was accepted by the Management Board, then presented to the Scientific Advisory Committee and finally accepted by the European XFEL Council on 30 June 2017 just before entering into the operation phase of the facility.

ESS: The main drivers for ESS to develop and formally agree a policy for scientific data early in the construction phase were the head of the Data Management and Software Centre, The Head of the Data Management Group, and the Director for science. The process to develop the policy and gain council approval was undertaken in 2017. The process took approximately 12 months. There was considerable support from the ESS Director General.

CERIC-ERIC: The Executive Director of CERIC-ERIC.

ELI: The ELI Facilities started to enter into initial operations in 2019 and will gradually open up their capabilities to the user community in the coming years. They will be operated by ELI ERIC in the process of being established. A Data Policy will be adopted by the General Assembly of the ELI ERIC shortly after establishment. There is therefore an institutional push to adopt such a policy, even more so since the data policy is mentioned as one of the statutory data policies in the draft version of the ELI ERIC Statutes. There is also a strong expectation and request within the organization for such policy from scientists and staff in charge of data management as clear data policy principles are needed as guiding elements in the on-going communication with prospective users and design and implementation of the data infrastructure. PaNOSC from that standpoint adds to the pressure in a positive way, creating a favourable environment for ELI to develop the policy. It is foreseen that a working group, involving IT and data management staff, scientists, staff in charge of instruments) will be involved, as well as the Scientific Advisory Committee prior to approval by the ELI ERIC General Assembly.

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

3.2 Which are the main reasons/benefits for adopting a Research Data Policy?

3.2.1 Guideline

The reasons are many and range from the need to make science reproducible and replicable by adopting Open Science approach, following the recommendations of international bodies like the OECD, ISC, IUCr, implementing the FAIR principles to enable the re-use of data, providing scientists with new data services, archiving of important datasets, to improving the quality of scientific data.

3.2.2 Case Studies

ESRF: The reasons the ESRF adopted a RDP are multiple. They were motivated by the fact that the ESRF (like other RIs) produces huge quantities of data which need to be managed and curated in order to provide services for data and allow scientists to profit from them fully. Without a RDP many fundamental issues like ownership, embargo, sharing, archiving, open data etc. were ill defined. The RDP allowed us to define these. The RDP was motivated by the need to define and collect metadata and raw data in well defined formats. The changing scientific publishing landscape which requires data to be made available and citable was an additional motivation for the RDP. Without the RDP the burden is on the users to store and curate the data. Another motivation for the RDP was the growing volumes of data produced which makes it more and more difficult to export data and therefore requiring them to be kept on site. This would be difficult without a RDP because the ownership is otherwise undefined. A strong motivation of the RDP was to make data openly available after an embargo period to increase its usefulness.

ILL: The initial driver was the reuse of data: ILL was archiving experimental data since its first run in 1973, the IT was taking care to migrate the data files with every technology change (e.g. from IBM tapes to LTO), the cost of preservation as always been relatively important for a limited number of requests to access legacy data. As soon as we started the implementation, other excellent motivations for this work, such as the improvement of the service quality for our users, became obvious.

EuXFEL: The main reasons to introduce the Scientific Data Policy was to impose a coherent approach to data management across all instruments and to allow defining obligations and rights with respect to data for the facility and facility users who had to accept it upon the registration in the User Portal. The policy defined the basis for implementation of data management services and it turned out to be extremely useful.

ESS: The key reason to develop a policy for scientific data early in the ESS construction phase (some 5 years before beam on target) was to set the policy framework in place to assist future developments in scientific computing. As an example one can use the existence of a data policy as a lever for developing scientific computing in a way that is commensurate with empowering open data for the ESS scientific community. An important but certainly not the only reason for the ESS data policy was to maintain compliance with the core EU ambition for Open data.

CERIC-ERIC: The data policy is necessary to be compliant with H2020 funding. Also, CERIC committed to the ORDP for the data generated in the ACCELERATE project. CERIC believes that open data will benefit researchers and institutions, increasing the visibility, enhancing collaborations and allowing a better use of resources in general.

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

ELI: For an emerging infrastructure like ELI entering into operations, there is simply an expectation, both from users and funding agencies, that experimental data will be made available and comply with the FAIR principles. In other terms, not having a data policy is considered a failure. In this context, the data policy is a necessary framework, because it addresses a number of critical issues that organise the relation with visiting users and the user community in general when it comes to data (ownership, embargo, access to data, storage, curation, etc.). It is also an internal driver pushing ELI to look at data also in terms of services leveraging the value of experimental data. Naturally, this perspective has a direct impact on technological choices made around data management.

3.3 To write a Research Data Policy, should one use a template, a management platform or an existing policy? If yes, please specify.

3.3.1 Guideline

The obvious place to start for Photon and Neutron Research Infrastructures is with one of the existing Research Data Policy frameworks developed specifically for the Photon and Neutron RIs, namely the most recent PaNOSC [1] one (written in 2020) which specifically treats the FAIR principles and is an update of the original PaNdata [2] one (written in 2010).

3.3.2 Case Studies

ESRF: ESRF based their data policy on the PaNdata data policy framework. As ESRF had actively participated in the writing of the PaNdata DP (ESRF was WP leader for the deliverable) it had strong knowledge of the contents. The PaNdata DP was then modified based on the input from the ESRF management, discussions with scientists and SAC. Since 2021 ESRF is in the process of updating the first data policy based on the PaNOSC DPF to incorporate the FAIR concepts and take into account the experience gained.

ILL: The work done during the PaNData EU project with our colleagues from the other EU analytical facilities was the basis for the internal discussion. With the strong competition that exists between these user facilities to attract the best scientific team and the fear that existed, at this time, to lose some users because of such “Open Data” regulation, the fact that it was a common and de facto standard framework was extremely important.

EuXFEL: European XFEL based the scientific data policy on PaNdata recommendations and followed the majority of modifications made by ESRF and ILL.

ESS: The ESS policy for scientific data is based upon the proforma policy created by the PaNdata project. An initial comparison map was made of existing data policies from european research infrastructure This document and the PaNdata proforma was used to develop a policy for ESS.

CERIC-ERIC: Yes. We used the PaN-data data policy guidelines and incorporated elements of other existing policies (ALBA synchrotron, Elettra Sincrotrone Trieste, EuXFEL, ESS, ESRF, ILL).

ELI: ELI will use the PaN-data guidelines as initial reference, analyse the data policies collected within the framework of PaNOSC’s WP2 and build on the work of this work package.

3.4 Who should be consulted/involved when implementing the policy?

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

3.4.1 Guideline

The important groups of people to consult are the beamline scientists, User Office, legal office + management who will be confronted with the consequences of implementing the DRP. In addition the control engineers, data managers and IT engineers need to be involved in the implementation. Users have to agree to the policy when applying for beamtime. New data consumers (who do not have access to state of the art RIs) should also be consulted. The latter group is represented by community organisation (e.g. IUCr) and forums (e.g. RDA and GO-FAIR)

3.4.2 Case Studies

ESRF: The effort of implementing the DP at the ESRF involved the beamline scientists, the beamline control system staff, the data analysis scientists, and the IT staff. Two permanent staff positions are dedicated to data management to implement the data catalogue and manage the data curation.

ILL: A large part of the organisation was involved in the discussions, and more specially the Directorate, the instrument responsibles, the User office, IT, the legal office and the scientific council. The internal discussions took three years for reaching an agreement.

EuXFEL: Various scientific and technical groups were consulted after the main directions were accepted by Management Board. Substantial support was given by the legal office.

ESS: Implementation of the data policy from a technical perspective falls within the remit of the data management and software centre. DMSC was the driver for the development of the policy details DMSC staff were involved in development from the initial stage. The DMSC scientific and technical advisory panel were consulted for advice. For broader stakeholder engagement the policy was presented to the ESS scientific advisory council for discussion before being presented to the ESS Council for approval

CERIC-ERIC: CERIC-ERIC is a consortium offering access to 9 facilities in Europe. Our Partner Facilities were consulted and the final word was given in June 2019 by the General Assembly. Users were not consulted so far, this may happen at a later stage.

ELI: Policies, within the context of the future management system of ELI ERIC, are short high-level documents. It is expected that, as such, the data policy will be complemented by more detailed regulatory documents describing the practicalities of the principles enshrined in the policy document. The policy document will be approved by the General Assembly and will contain the core principles and strategic objectives of ELI in terms of data management and access to data. It is expected that the Scientific Advisory Committee of ELI ERIC will be invited to comment on the proposal. Complementary management and regulatory documents will be developed with the likely involvement of researchers and operators (scientific directorate, beamline scientists), of the staff involved in control systems and IT management and, possibly, of user representatives.

3.5 Before the adoption of a data policy, what compliance with legal and regulatory aspects should be assessed?

3.5.1 Guideline

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

The RDP should be reviewed by the legal counsel of the Research Infrastructure to ensure it complies with the legal statutes of the institute. The RDP should be reviewed by the Data Protection Officer to ensure it complies with GDPR for scientific data.

3.5.2 Case Studies

ESRF: The ESRF DP was submitted to the ESRF legal counsel for checking. She did not make any changes. This was before GDPR.

ILL: The initial version of the ILL data policy and especially the question of the protection of data was discussed with a layer specialist for IPR related questions and like for any policy was checked by our legal office.

EuXFEL: Yes. Amendments were introduced based on the Legal Office advise, especially personal data protection and liability aspects.

ESS: The ESS legal team were involved in the development of the policy text to ensure compliance with legislation and latterly GDPR.

CERIC-ERIC: No, it wasn't checked by a lawyer so far.

ELI: Yes, it is planned that a legal assessment will be performed.

3.6 Which data produced and related metadata are covered by the Research Data Policy? Which kind of data should be excluded (personal data, sensitive data, etc.)?

3.6.1 Guideline

The RDP covers scientific research data and metadata. Data can be raw data, processed data, auxiliary data or results (refer to the PaNOSC data policy framework [1] for a definition of the different types of data). It is highly recommended to exclude data from clinical trials or other data where the samples refer to identifiable humans as these are considered sensitive data. Paleontological human samples are not considered sensitive data. Proprietary research (resulting from commercial beamtime) is usually not covered by the RDP.

3.6.2 Case Studies

ESRF: The ESRF DP only excludes data produced by proprietary (commercial) research. All data from public research, including the CRG beamlines, are covered by the ESRF DP. Processed data are currently (May 2021) being included in the updated DP (based on the PaNOSC DPF).

ILL: The first revision of the data policy, published in July 2017, also addresses reduced data and more. Generally all scientific data resulting from the analysis of the raw data are stored by the ILL IT infrastructure. This revision also takes into account data generated from CRG instruments (Collaborating Research Groups instruments are instruments managed on ILL beamlines by third party organisation https://www.ill.eu/fr/users-en/instruments/crgs/). Only data resulting from proprietary research are excluded from the data policy.

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

EuXFEL: The Scientific Data Policy excludes data produced by proprietary research. It applies to all scientific data generated at European XFEL instruments including those contributed by third party organizations and User Consortia.

ESS: The ESS policy specifically excludes data from proprietary use of ESS beamlines / instruments. Metadata that constitutes sensitive data is not explicitly included or excluded. From the ESS perspective this aspect falls within other policies set by the organisation.

CERIC-ERIC: Our policy can be applied to all the data produced and relative metadata. Personal or sensitive data will not be disclosed.

ELI: Similar to other PaNOSC partner organisations, it is planned that our policy will apply to all data generated by ELI instruments and related metadata. It will address data from proprietary research and sensitive data that will not be disclosed.

3.7 Which personnel of your organization should be trained on how to apply the Data Policy?

3.7.1 Guideline

The implementation of an RDP requires dedicated personnel mainly in the form of data managers but also controls engineers, data scientists and IT personnel.

3.7.2 Case Studies

ESRF: ESRF dedicates two positions to data management. The DP was presented to staff and users are expected to use the data portal. It is planned to provide online training to users in the future.

ILL: The personnel was not trained, but support exists (data@ill.eu) to reply to data management related questions.

EuXFEL: The Scientific Data Policy is the first point in the data acquisition and data management training provided to instrument scientists.

ESS: DMSC has specific positions for data management. It is the intention of ESS to train users and staff in certain aspects of data management and aspects that directly pertain to our data policy (such as the SciCat data catalogue)

CERIC-ERIC: Not yet. We have just agreed on a final version that still needs to be approved by the General Assembly. Training will be necessary during the implementation.

ELI: No such training has yet been planned, but including data policy aspects in the compulsory user training is being considered.

3.8. Should the policy include a review cycle?

3.8.1 Guideline

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

It is necessary to review the RDP at regular intervals to take into account the evolving norms for research data (e.g. introduction of the FAIR principles in 2016) and experience gained in implementing the RDP. The data management landscape is evolving with the increased adoption of the FAIR principles and Open Science methodology thanks to the efforts of scientific communities and support from scientific bodies and governments and last but not least the EOSC. The RDP needs to be regularly reviewed to consider new guidelines like FAIR and be adapted if the new guidelines improve scientific data management. The review process should be foreseen and minor changes should be possible without going through the full approval process.

3.8.2 Case Studies

ESRF: Not yet. This is one thing we would like to introduce as part of the PaNOSC RDP framework. The review process began in 2021.

ILL: The current policy does not include a formal review process or cycle**.** Nevertheless it has already been reviewed when it has clearly appeared to the stakeholders that it was necessary. This review was simplified by the fact that the rationales were well understood by all parties, this will not necessarily be the case with a defined time scale review. The Data policy was reviewed in 2017 (see previous sections), the main drivers were to : (1) Handle the reduced data and other derived data in preparation of the set up of data analysis services. (2) Take into account the CRG instruments. (3) Create a more accessible text for the users (not enough people had a general understanding of the policy). For instance, the term PID was replaced by DOI.

EuXFEL: The Policy does not define review cycles. However, it allows within certain limits for modifications of storage periods on different levels of storage systems according to the experience and available resources. There is an ongoing attempt to redefine retention periods of different data types.

ESS: No specific timescale is included in the policy for review of the policy of the current ESS DP. A review process and frequency will be defined in the future ESS DP. It is the intention of the ESS to include a defined review process and frequency for the future DP policy update.

CERIC-ERIC: The CERIC DP is considered being a living document. Reviews may take place when necessary and in case major changes are required, the maximum time for a deliberation is up to 6 months.

ELI: Regular policy assessment and review are considered good management practice. A review cycle will therefore be proposed as part of the data policy submitted for approval to the ELI ERIC General Assembly.

3.9 If you used a template or model, do some standard definitions need to be changed?

3.9.1 Guideline

It is standard practice to adapt the definitions of certain terms in the template to the local vocabulary. If a definition needs to be altered significantly then it is better to introduce a new term.

3.9.2 Case Studies

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

ESRF: The definition of proprietary data was added to the PaNdata DP framework. We needed to add the definition of a session to the DP for DOIs.

ILL: There were no major changes but some definitions had to be adapted to the "language" of the ILL and its users, as an example we use the wording "main proposer" instead of PI. The revision also adopted a less formal and more practical approach in order to be more easily understood by users and personnel. It was mainly rewritten by a scientist whereas the initial one was mainly written by managers.

EuXFEL: The definition of various data types was introduced. A separate paragraph on warranty and liability regarding scientific data was introduced.

ESS: The template was used to develop the overall concept of the data policy rather than a direct copy paste of text. Specific changes in definition of terms have been made to match other ESS user facing policies.

CERIC-ERIC: We needed to include the definition of 'Partner Facility', due to the particular nature of CERIC-ERIC, see table "PaNOSC definitions for data Policy, cell D5).

ELI: Not applicable.

3.10 Does one need to define one or more standard formats for the raw data? If yes, which one/s?

3.10.1 Guideline

The RDP should guarantee that all curated data can be read and understood by the custodians of the data i.e. the RI. Defining the data format in which (raw, processed, auxiliary and results) data will be curated ensures the data can be read. Standard metadata and/or using standard vocabularies are part of ensuring data can be understood by the community. The preferred data format and vocabulary should be mentioned in the RDP.

3.10.2 Case Studies

ESRF: HDF5 is the preferred data format with the Nexus conventions but currently not all data analysis programs can treat HDF5/Nexus.

ILL: The standard data format at ILL is NEXUS and it is in place for almost all instruments, exceptions only exist for instruments with an existing strong community standard (e.g. root format for the nuclear physics community). Nevertheless this standard is not defined formally in the Data Policy document.

EuXFEL: The policy document does not name any specific data format. In practice, the only format supported across the facility is HDF5.

ESS: The ESS DAQ system writes Nexus (HDF5) files as default. Nexus is the guideline data format that ESS prefers. Other file formats are not excluded as HDF files are not always readable by downstream data services.

CERIC-ERIC: Yes, it was decided that HDF5 may fit the needs of all the partners.

ELI: Not yet decided, but Nexus and HDF5 are preferred.

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

3.11 Which considerations should be taken into account in the choice of the embargo period?

3.11.1 Guideline

The two main common considerations to take into account are the length of a PhD which is commonly 3 years and the time needed between to analyse the data before publishing.

3.11.2 Case Studies

ESRF: The ESRF embargo period of 3 years is based on the length of a PhD.

ILL: The ILL formal embargo period of 3 years is based on the standard length of a PhD. There is also a possible extension of 2 years when no one is requesting access to the data, this period came from the discussion that took place in 2010 and was put in place to avoid having to face too many extension requests from the users. In practise, after 7 years of implementation only 1 request was received by the scientific director. It is difficult to know if the very limited number of extension requests is due to this mechanism, the feedback of the other facilities that did not implement it will be extremely interesting.

EuXFEL: Length of a PhD project and following the recommendation of PaNdata.

ESS: A 3 year embargo was chosen to match the average length of a PhD project and match the majority of facility embargo periods.

CERIC-ERIC: We chose 3 years of embargo period, which is the standard duration for a PhD degree, and it is a reasonable period in which all data should have led to a publication.

ELI: 3-year embargo currently being considered. It is considered to be a reasonable period of time and is the average duration of a PhD.

3.12 Should the embargo period be allowed to be extended and how to manage this?

3.12.1 Guideline

The embargo period is based on an average PhD and is a compromise for research projects that need more than 3 years. The RDP should foresee the extension of the embargo period for such projects and ensure the process is easy for researchers. It should not however encourage blanket extensions to the embargo period for research groups without good reasons.

3.12.2 Case Studies

ESRF: The ESRF DP allows the embargo period to be extended by the PI (Principal Investigator) on demand to the ESRF Scientific Directors. We have had one request so far but we have not defined the workflow for the implementation yet.

ILL: In the current policy, the non-disclosure period is extended to 5 years if no request has been received to

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

access the data.

EuXFEL: Any PI that wishes to extend the embargo period might submit a written request, specifying the reasons for the proposed prolongation, to the management board of European XFEL GmbH, which decides on the request at its own discretion. In exceptional circumstances, data can be made openly accessible during the embargo period if the PI informs the European XFEL GmbH to do so and subject to its own discretion.

ESS: The principle investigation can extend the period of embargo by application to ESS under a defined written procedure.

CERIC-ERIC: The PI can request an extension of the embargo period based on legitimate grounds defined by CERIC-ERIC.

ELI: An extension of the embargo period under the conditions and based on legitimate grounds defined by ELI ERIC.

3.13 What data services should be provided as part of the RDP?

3.13.1 Guideline

One of the main reasons for adopting an RDP is to improve the quality of scientific data and be able to provide data services to researchers. Adopting an RDP should go hand in hand with the proposal of new data services enabled by proper data management e.g. services like long-term archiving, download and data transfer services, data processing and analysis services, DOI services etc.

3.13.2 Case Studies

ESRF: The implementation of the DP has enabled the following data services:

- well defined metadata in HDF5 file and in metadata catalogue,
- DOI for sessions + on demand,
- long-term archiving of raw data,
- web portal (https://data.esrf.fr),
- search engine for metadata,
- download service,
- Jupyter notebooks.

ILL: The services that were put in place for the implementation of the Data policy are:

- Better management of the data files and repository.
- Access to the data from any ILL computer directly available from the desktop of the users.
- Data portal, with search engine and user self management of the access (ACLs), including termination of the embargo period.
- Internet access to the data through SFTP service.
- Generation of DOIs.
- Elogbook available on the instruments.

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

EuXFEL: The following services deal with scientific data and are compliant with the policy:

- User Portal
- Metadata Catalogue service
- Automatic data acquisition service
- Data calibration service
- DOI generation
- Data processing on site
- Data archiving
- Data export services

ESS: Specific services are not discussed in the DP. Services developed at ESS should be compliant with the policy but not defined in the policy itself - this allows some flexibility in service provision and alteration thereof. That is to say one can change downstream policies, procedures and rules without changing the governing policy itself, which from our perspective is a more flexible approach.

CERIC-ERIC: Similar to ESS, specific services are not specified in the DP (and thus not approved by the General Assembly). The following compliant services are to be provided in the near-future:

- persistent identifier, for example DOI generation,
- data and metadata catalogues,
- access to and storage of raw, processed and auxiliary data,
- long-term data archiving service (10 years for data and indefinite for metadata).
- Automatic metadata ingestion (through e-logbook if available).

ELI: Similar to ESS, it is not anticipated that data management services will be discussed in the data policy, though data services will obviously have to be compliant with it. This being said, it is anticipated that the data policy proposal will be submitted to the General Assembly for approval together with some background information on the data services foreseen at ELI and an accompanying implementation roadmap.

4. Frequently Asked Questions

4.1 What granularity to apply for data DOIs?

The ideal granularity is to provide researchers with the possibility of minting DOIs for a bespoke set of datasets. To ensure all data including unpublished data are referenced by a DOI most sites offer a DOI which is minted automatically. The automatically minted DOIs granularity is usually at the level of the proposal or beamtime session.

4.2 What prefix to use for DOIs ?

The guideline given by the DOI Handbook is that the DOI prefixes should not be crafted for humans but more for machines. Some examples from PaN RIs are:

- https://dx.doi.org/10.5291/ILL-DATA.1-01-126 (for ILL proposal 1-01-126)

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

- https://doi.org/10.5286/ISIS.E.RB1820600 (for ISIS proposal)
- https://doi.org/10.15151/esrf-es-187197141 (ESRF data for an entire proposal)
- https://doi.org/10.16907/808de0df-a9d3-4698-8e9f-d6e091516650 (PSI dataset)

4.3 How long should data be archived for?

Data should be archived for as long as possible. Most of the PaN RIs implement archiving for 10 years but time is a compromise based on the costs of archiving and what was considered a good starting value. In the future this could be shortened or lengthened depending on available finances and the scientific interest in the data.

4.4 Should DOIs be generated for Instruments?

DOIs for Instruments allows them to be identified in a given configuration. Some sites are already generating DOIs for Instruments (see [10] in Further Reading) but only one (HZB) PaN RI is generating Instrument DOIs. This should evolve in the future to include more sites.

4.5 Which data catalogue to use?

Two Open Source solutions developed by the PaN community are used at multiple sites, namely ICAT (https://icatproject.org/)and SciCat (https://scicatproject.github.io/). In addition, there are local solutions used by individual sites. Other solutions from outside the PaN community more well-known abound e.g. DSpace. The table in the Appendix below has links to the catalogues for most photon and neutron sources in the world.

4.6 How to define raw data?

Raw data refers to the experiment data generated at the facility which are persisted and is implementation specific. Raw data do not necessarily only refer to the output generated directly by the detector but may refer to data produced further down the processing pipeline. Raw data represents the data closest to the ground truth to reproduce the results and which is stored for long-term archiving. This can refer to processed data which has been reduced in order to be archived.

4.7 Are Data Management Plans mandatory?

With the increasing data volumes Data Management Plans 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.

5. GDPR

Disclaimer: the recommendations below do not replace legal advice.

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

"*The General Data Protection Regulation 2016/679 is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area. It also addresses the transfer of personal data outside the EU and EEA areas.*"[1]

It is a legal requirement that the PaN RIs conform to the GDPR including for the data repositories governed by the RDP. A general guideline is to exclude any clinical trial data from the data repository. Seeing as most PaN RIs do not conduct clinical trials or only in very rare cases this recommendation is not difficult to follow.

A second guideline is to ensure users are informed of what metadata (name, affiliation, etc) will be made available as part of the DOI or data repository and that they accept this before applying for beamtime.

A common issue raised is that including the user's name in the DOI or dataset is potentially incompatible with the GDPR. This is not the case for the following reasons: (1) being able to identify the members of the experimental team for publicly funded research is part of the business process of a user facility, and (2) data repositories are archives of scientific data for the public good and therefore fall under the GDPR regulation Article 89 *Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes*. Similar to scientific publications (which are not anonymous) and public archives the data repositories of scientific data can store the minimum necessary information to ensure the validity and usefulness of the archive.

Before E-logbooks can be made public they need to be made GDPR compliant by excluding personal information. This can be done with a mixture of users agreeing to not add sensitive data to the E-logbook, by following best practices of data anonymisation, manual editing the logbook after the experiment and automated anonymisation.

6. Lessons Learned

An example of some lessons learned while implementing the research data policy at the **ESRF and ELI**:

1. Implementing a Data Policy is a long process, especially when it is being implemented on an existing installation where the implementation has to be retro-fitted to the running installation and habits of scientists and engineers need to be changed. At ESRF we started working on the Data Policy 10 years ago (2009) with the writing of the PaNdata data policy framework. Implementation on beamlines started in 2016.
   At the same time, for new RI's (such as ELI), it is equally important to engage the control systems, data acquisitions and data management divisions in the definition and review process of the data policy as key players in the implementation process and a reliable advocate raising the awareness of different internal stakeholders and promoting the Data Policy and all associated tools and services.

2. Support of upper and top management is essential to get the data policy accepted and implemented.

3. An initial hurdle was the feasibility and cost of storing all raw data for the ESRF (hundreds of petabytes

[1] https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

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

over 10 years) but we started off by discussing metadata and metadata policy. Because metadata is much less and there is no cost or feasibility issue it unblocked the discussion. It would be an option for all sites. Every site must have a metadata data policy. This also ensures that when there is a budget to store raw data the data are already well organised with metadata.

4. A top-down approach for deciding that we need a data policy reduced the discussion time. It was considered the right approach at the ESRF because policy is the prerogative of the management and not of the scientists or users.
   At the same time, in ELI's case, where we have to think about the standardization but also the integration of the ELI sites under ELI ERIC, the approach was top-down as the management supported the Data Policy. In parallel, a bottom-up approach was started by Control Systems and Data Management teams. The two teams are now trying to accelerate the standardization of tools, services and formats as well as the integration challenges raised by the fact that the facilities are joining together under a single ERIC. To gain momentum, we have used the timing advantage and drafted the Data Policy based on the PaNOSC community framework to address the challenges addressed by the engineering groups, thus making the Data Policy easier to adopt.

5. A majority of beamline scientists at the ESRF were not aware of or knew very little about data management concepts like PID i.e. DOI. They needed explaining and will need more training.
   For ELI, similar concepts like PID/DOI or Data Stewardship, even if they are easier to integrate since the facilities are now in the commissioning process, will require training and awareness campaigns presenting the Data Governance concepts and all the associated roles and services.

6. There were no ready to use solutions with all features we needed for the metadata and data catalogue. We found icat was the closest to what we needed because it had a data model which mimicked our proposal and scientific data flow. However, it suffered from lack of widespread adoption and an active user community. We invested in extending icat.
   For the metadata, ELI is evaluating three different file cataloguing solutions ICAT, SCICAT and InvenioRDM, aiming to select one that will be integrated to serve all ELI sites. The main challenge, since we are working in parallel with the experiments/lasers commissioning, is to actively engage our beam scientists to support and validate different scientific data management tools and services.

7. We identified the need for an electronic logbook in order for scientists who were not part of the original experiment to understand the experiment and data produced. We therefore developed one. We did not find any of the Open Source solutions which fitted our needs. One requirement was to have a modern web UI.
   The electronic logbook is a basic tool that evolves into one of the most interesting challenges driven mainly by the maturity of the Control Systems and Data Acquisition integration. For ELI, the electronic logbook challenge, presented in the last PaNOSC WP3 - Best Practices Workshop, will most likely require a custom development.

8. GDPR can be difficult to handle because there is no clear directive for scientific institutes coming from the standards bodies. In fact the standards bodies and EU projects have not been of much practical value so far. We hope this will evolve with the EOSC.

9. The PaNdata policy is implicitly FAIR but we have not mentioned FAIR explicitly in our data policy. We need to do this, but how is still an open question. We are working on this in PaNOSC.

10. We did not find any practical guidance on how to present landing pages for DOIs. We use a dynamic

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

web page built out of the datacite metadata. We are not sure this is the best solution as sometimes datacite is down. Dynamic pages are not indexed as well. Datacite search engines are not user friendly. Here is an example of an ESRF landing page: https://doi.esrf.fr/10.15151/ESRF-ES-135816585

11. It is essential to setup a contract with datacite for minting DOIs. This is done via the local Datacite representative in your country. In the case of France this is the CNRS institute INIST (https://www.inist.fr/).

For more information on the implementation of the ESRF data policy please refer to the article which was published in the journal SRN (see [11] in Further Reading).

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

Appendix – Survey of PaN data policies

The following table summarises the results from the survey of accessible data policies at Photon and neutron (PaN) facilities around the world (table prepared by **Jonathan Taylor** (ESS), private communication).

Organisation	Policy defined data retention period	Embargo Period preceding open access	Ref
ORNL	Dependent upon data volume	-	https://tinyurl.com/y9wrb463
Argonne APS	No guarantee for archival storage of data	-	https://tinyurl.com/3btw54p5
BNL NSLSII	1 year	-	https://tinyurl.com/eunup24e
NIST NCNR	Not specifically defined	None or 18m	https://tinyurl.com/3spkpza8
SLAC	Responsibilities of facility users	-	https://tinyurl.com/2yzzz487
SPring8	No Online Data policy information	-	
Sirius	No Online Data policy information	-	
SSRF	No Online Data policy information	-	
JPARC MLF	Not specifically defined	3 years	https://tinyurl.com/vj5u5rsm
ANSTO Australian Synchrotron	yes 12m or 36m	Public after 36m	https://tinyurl.com/3az3bk75
Diamond Light Source	yes 30 days & long-term archive	3 years	https://tinyurl.com/nc2uwdu6
ISIS neutron and Muon facility	no guarantee - long term archive	3 years	https://tinyurl.com/f3zhnpw3
ESRF	5 years minimum, 10 years expected	3 years	https://tinyurl.com/3rpe9vk6
ILL	5 years minimum, 10 years expected	5 years	https://tinyurl.com/2afuk755
Soleil	5-10 years	3 years	https://tinyurl.com/48vb9f73

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

DESY	Not specifically defined	Not Specifically defined	https://tinyurl.com/hrr4nzpb
FRMII	10 years	Not Specifically defined	https://tinyurl.com/tdkn67y9
HZDR	10 years	5 years	https://tinyurl.com/4brvdtuv
HZB	10 years	5 years	https://tinyurl.com/n62tnv62
EUXFEL	5 years minimum (separate policy)	3 years	https://tinyurl.com/zp6yjebh , https://tinyurl.com/2cmb8cjc
PSI	5 years minimum, 10 years expected	3 years	https://tinyurl.com/rmc4naj
MaxIV	3 months	Not Specifically defined	https://tinyurl.com/2bm53zc6
ESS	No Online Data policy information	-	
Elettra	5-10 years	3 years	https://tinyurl.com/3vp73tvr
Alba	5 years	3 years	https://tinyurl.com/usb59c9m
Sesame	Minimum 5 years	3 years	https://tinyurl.com/sm8fwa3z
PaNData Policy Framework	10 years	3 years	https://tinyurl.com/28rwdyjd
PaNOSC Data Policy framework	10 years	3 years	https://tinyurl.com/tw9hju5a

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

References

1. Gotz, A., Perrin, JF., Fangohr, H., Salvat, D., Gliksohn, F., Markvardsen, A., … Matthews, B., (2020), *PaNOSC FAIR Research Data Policy framework* (Version 1.1). Zenodo, https://doi.org/10.5281/zenodo.3826039
2. Dimper, R. (2011), *Common policy framework on scientific data*, Zenodo, https://doi.org/10.5281/zenodo.3738497

Further Reading

1. Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. and Goudie, S., 2020. *Developing a Research Data Policy Framework for All Journals and Publishers*. *Data Science Journal, 19*(1), p.5. http://doi.org/10.5334/dsj-2020-005
2. Mons, B, 2018, *Data Stewardship for Open Science*, Taylor and Francis, eBook ISBN 9781315380711, https://doi.org/10.1201/9781315380711
3. *CESSDA Data Management Expert Guide*, https://cessda.eu/DMEG (accessed 15/5/2021), https://doi.org/10.5281/zenodo.3820472
4. OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris, https://doi.org/10.1787/9789264034020-en-fr.
5. International Science Council (2021), *Open Science for the 21st Century*, https://council.science/publications/open-science-for-the-21st-century/ (accessed 15/5/2021)
6. International Science Council (2021). *Opening the record of science: making scholarly publishing work for science in the digital era*. Paris, France. International Science Council, https://doi.org/10.24948/2021.01
7. European Archives Group. "*Guidance on Data Protection for Archive Services: EAG guidelines on the implementation of the GDPR in the archive sector*". October 2018. https://ec.europa.eu/info/files/guidance-data-protection-archive-services_en (accessed 15/5/2021)
8. Matthews, B., McBirnie, A., Vukolov, A., Ashton, A., Collins, S., Da Graca Ramos, S., … Van Daalen, M., (2020), *Draft extended data policy framework for Photon and Neutron RIs*. Zenodo, https://doi.org/10.5281/zenodo.4014810
9. Salvat, D., Gonzalez-Beltran, A., Görzig, H., Matthews, B., McBirnie, A., Ounsy, M., … Vukolov, A. (2020), *Draft recommendations for FAIR Photon and Neutron Data Management*. Zenodo. https://doi.org/10.5281/zenodo.4312824
10. Stocker, M., Darroch, L., Krahl, R., Habermann, T., Devaraju, A., Schwardmann, U., D'Onofrio, C. and Häggström, I., 2020. Persistent Identification of Instruments. *Data Science Journal*, 19(1), p.18. DOI: http://doi.org/10.5334/dsj-2020-018
11. 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: 10.1080/08940886.2019.160811

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