Cremlin Connecting Russian and European Measures for Large-scale Research Infrastructures

Document information

Deliverable no.	D4.4
Deliverable title	Recommendations on Common Data Standard Policy
Deliverable responsible	ILL
Related Work-Package/Task	T4.4
Type (e.g. Report; other)	Report
Author(s)	Jean-François Perrin
Dissemination level	Public
Submission date	16.08.2018
Download page	https://www.cremlin.eu/deliverables/

Project full title	Connecting Russian and European Measures for Large-scale Research Infrastructures
Project acronym	CREMLIN
Grant agreement no.	654166
Instrument	Coordination and Support Action (CSA)
Duration	01/09/2015 – 30/08/2018
Website	www.cremlin.eu

This project has received funding from the European Union‘s Horizon 2020 research and innovation programme under grant agreement No. 654166.

Recommendations on Common Data Standard Policy

1 INTRODUCTION.

This document has been prepared with the aim to ignite discussions on Data Policy for big data producers that represent scientific facilities. By experience, setting up a data policy standard within scientific infrastructures is a long process as results need to be understood and accepted by all the stakeholders, from policy makers to the end users. The important part is to not just have a data policy but full compliance with the policy from all parties. Setting up Open Data policies is known to be a disruptive process because it necessitates the adoption of an ever changing open culture.

2 CURRENT LANDSCAPE REGARDING DATA POLICIES IN ANALYTICAL FACILITIES.

As of today, most of the analytical facilities in operation have published a data policy. They propose very similar principles mostly originating from the collaborative work done during the European PaNData project (Grant Agreement Number: 261537).

Facility	Technic	DP document
ILL	Neutron scattering	https://www.ill.eu/DataPolicy
ESRF	X-Ray	http://www.esrf.eu/datapolicy
EU XFEL	FEL	https://www.xfel.eu/users/experiment_support/policies/scientific_data_policy/
PSI	Neutron, X-Ray and FEL	https://www.psi.ch/science/psi-data-policy
ISIS	Neutron scattering	https://www.isis.stfc.ac.uk/Pages/Data-Policy.aspx
ELETTRA	X-Ray and FEL	https://www.elettra.trieste.it/userarea/scientific-data-policy.html

We have used these texts to propose the key elements for reflection on working towards a standard policy.

3 WORKSHOP OUTCOME

A session was held on data management and Policies during the workshop on Big data Management (15-16 February 2017, NRC “Kurchatov Institute”, 1, Akademika Kurchatova pl., Moscow). We had to slightly extend the scope of the discussions to user communities, instruments, data reduction and analysis, in order to obtain a comprehensive overview of data management for neutron facilities.

The analytical facilities present a different range of techniques and instruments for the scientists. In order to conduct their research, scientists use the different tools, complementary techniques and travel between the facilities; this is demonstrated by the regular PaNData user surveys (http://pandata.eu/Users2014-Results). It is therefore extremely important to coordinate efforts in order to provide a consistent ecosystem for our users in terms of software, policies, infrastructures and services.

The current work to harmonize and support the development of the data treatment software (Mantid, MuhRec & KipTool, Born Again, SASView, NSXtool …) has been presented and discussed as well as the Data Management (i.e: Data and Metadata preservation, Policies, use and impact of DOIs for data). The urgent need for data analysis services has been highlighted by the evolution of experimental data in term of volume and complexity. The PaNDaaS collaboration, which tries to bring this data analysis service to the X-ray and neutron users' community as a whole, and the different approaches were explained.

4 KEY ELEMENTS OF A POSSIBLE DATA POLICY PROPOSAL.

4.1 WHAT TYPE OF DATA

In 2008, when the initial data policy Framework was created, it addressed mainly Raw data and meta-data. Today, with the increase in data volume and complexity, RIs tend to also provide analysis services where users can extract scientific knowledge by processing raw data. It is advisable to consider this processed data when drawing up modern data policies.

Typical glossary includes definition of raw data, metadata and processed data:

The term **raw data** pertains to data collected directly from instruments during experiments. This definition includes data that are created either automatically by instrument-control software (e.g. detector counts, angles, time stamps, etc) or manually by facility staff and/or the experimental team, but which have not yet been reduced or processed by any data-treatment software.

The term **metadata** describes contextual information that is complementary to raw data and possibly useful for subsequent data treatment. Metadata include (but are not limited to) part of the beam-time proposal, log files and parameter surveys generated by the instrument-control software, the instrument configuration (e.g. wavelength), the sample environment, the sample description and state points (e.g. temperature and pressure), the content of the instrument notebook, and other logistical information.

The term **processed data** refers to raw data and/or metadata that have been processed or reduced by data-treatment software and then curated alongside to the raw data.

The term **data** without qualifiers refers to the ensemble of raw, meta- and processed data.

4.2 OWNERSHIP OF DATA

4.2.1 Proprietary research

Data that is obtained via proprietary research, conducted at a RI, is owned by the client who has purchased the access. Proprietary users must agree prior to their experiment with the facility management on how they wish their data to be managed. A blanket solution cannot be proposed due to the wide variety of industry needs and requirements, therefore agreements have be made alongside the contract.

4.2.2 Public Research

Regarding national regulations, there are very few legal mechanisms to define ownership of immaterial objects (e.g. Copyright owner in Art) but none of these seem to be applicable to scientific data.

Legal ownership is usually not defined in the current published data policies in this domain.

By default, the data belongs to the public domain and all data obtained via public research conducted at a public RI are destined for open access. However, the RI acting as curator can impose certain legitimate constraints on the access to these data. Current examples are the decision of an initial non-disclosure period or referencing of the original authors (including the RI) when publishing work based on these data.

4.3 Non-disclosure period

Facility users propose the experiment, prepare part of the work such as the samples and analyse the data. Furthermore, the experiment conducted at the RI is often a single step of much larger research process. Because each experiment differs greatly, only the user knows how much time is needed to treat the data and eventually publish their work.

The core idea is that all data used to achieve a specific publication should be made publicly available as soon as the corresponding scientific paper is published.

It is now common practice to allocate by default a non-disclosure period of 3 years where only the proposal or experimental team has exclusive access to the data. This practice should allow sufficient time for users to publish their work. Nevertheless, this can happen before the end of the non-disclosure period and it is therefore important that the user can trigger this release process at any time before the 3-year deadline. After the 3-year deadline, a user should be able to request an extension period of the non-disclosure agreement in justifying that his research demands more time to be publishable (e.g. ground breaking, fundamental research). This request should be formal written document validated the facility science directorate.

An embargo-type non-disclosure period seems to be the simplest and most general mechanism that we can put in place to protect the users from data theft. Putting in place this type of solution enforces a trusting relationship between users, funders and their research infrastructures. If a scientist feels that his or her work is secure and he would gain personal recognition for his work, he is more likely to endorse open data after his non-disclosure period. This adhesion of the scientist is of the utmost importance to permit high quality data and comprehensive open results.

However, other protective solutions are certainly possible, but none have emerged for the time being.

4.4 Licence

Public domain data means that there exist no residual rights of the original authors. This is usually not exactly the solution that scientists are looking for, as enforcement of good citation practise and referencing is extremely important.

In most countries a licence seems to be the only way to formalise the practise of using data from a fellow scientist. Even if they are not always legally binding depending on the legislations, it has the advantage to present clearly the will of the original author. A licence like the CC-BY who has become a standard in most domain, has an added advantage of being easily understood by anyone.

Recognition is vital for scientist and their science, therefore a licence agreement appears to be the best solution at this time.

4.5 Curation and access

It is common practice that facilities act as curators of data. This is increasingly the case, as the facilities understand the long-term value of the data and with the ever-growing volume of the datasets, which are becoming difficult to transfer for the users.

Data access services are proposed by user facilities as part of their experiment package. It also ensures that access of data is protected in line with the Data Policy. Other mechanisms could be foreseeable (e.g. central archive for the community) but nothing has been implemented so far for user facilities on a large scale.

4.6 Persistent identifiers

In order to provide proper referencing and citation of datasets, it is advisable to allocate a persistent identifier at the beginning of the curation process.

Different mechanisms exist as persistent identifiers (Handle, Archival Resource Key ARK, ...) but the standard has become the Digital Object Identifier (DOI) for scientific data, probably because it is already intensively used by the scientific publishing industry and support services of DataCite have simplified its application.

4.7 Data retention options for huge data sets (see Eu-XFEL)

Not all data can be stored on site due to its volume and the financial constraints of the facility. In case the facility is unable to store huge datasets for a long period it is advisable to provide clear information for the users, maybe in a separate documents to the data policy, about realistic retention period (like done by EU-XFEL in their “Data retention policy” document (https://www.xfel.eu/users/experiment_support/policies/scientific_data_policy/index_eng.html)

Conclusion

This document is aiming to serve as a basis for future discussion at a high level of the different Research Infrastructures. Collaborative work needs to continue to achieve a standard policy and to ensure the compliance by all parties involved. Work to be done will not necessarily be simple but it will be worthwhile as it is the key element for scientific open exchange.