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

Document information

Deliverable no.	D5.4
Deliverable title	Recommendation on common data standard policy
Deliverable responsible	ESRF
Related Work-Package/Task	WP5/ Task 5.3
Type (e.g. Report; other)	Report
Author(s)	H. Reichert, ESRF
Dissemination level	Public
Submission date	31/08/2018
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 – 31/08/2018
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.

Recommendation on common data standard policy

CREMLIN Deliverable D5.4

CREMLIN WP5: “Science cooperation with the SSRS-4 synchrotron radiation source in the field of photon science”

Task 5.3 “Develop concepts on big data management in photon science”

Task Leader: ESRF

Author: Harald Reichert, ESRF

Preamble

Apart from a few scientific communities (high energy physics, astronomy) common data standards were not developed for most fields even though centralized large-scale facilities exist for a number of scientific disciplines. This has changed more recently with large data sets and large research collaborations becoming a standard. A strong push to develop (common) data policies comes also from the move towards open data / open science which will change the way science is being undertaken and perceived. Many existing large-scale facilities are therefore developing data policies with all the problems associated with implementing it in operating facilities. For a green-field facility in planning, such as the Russian ‘megascience’ facility SSRS-4, it is therefore prudent to consider such a data policy and its implementation from the outset.

Data Policies at existing large-scale facilities

In 2010/2011 the PaN-data Europe Strategic Working Group developed a general, standard data policy framework which it recommended for implementation in its partner organisations (FP7 Grant Agreement Number: 261537). This group was comprised of 11 European partners, mostly operators of large-scale facilities for the X-ray and neutron science communities. This was followed up by the PaNdata Open Data Infrastructure (PaNdata ODI) initiative to create a federated open data infrastructure, seamlessly integrating the existing user and data management systems of the European photon and neutron facilities.

PANdata - the Photon and Neutron data infrastructure initiative - created a fully integrated, pan-European, information infrastructure supporting the scientific process. Up to date, six of the original partners comprising 8 eight facilities have already implemented a data policy compliant with the standard data policy framework developed in 2011 (see Table below), while many other European facilities are still in the process of adopting such a data policy on this basis.

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

Experience shows that it is much harder to implement a data policy in facilities that are operating since many years. Changing the practices and routines of a large and diverse group of people takes considerable effort and some years of dedicated efforts. It has, however, become apparent that all European photon and neutron user facilities will adopt a data policy following the framework defined by the PANdata working group.
The main features of the data policy framework can be summarised as follows:

- The facility shall act as a "custodian" for the data.
- Ownership of the data left is left open and can be independently regulated for each facility (user owns data versus facility owns data).
- Data should become publicly accessible after an embargo period of 3 years.
- The embargo period can be extended on request.
- Analysis of openly accessible data must acknowledge the source of the data and cite its unique identifier and any publication linked to the same raw data.

The full text of the data policy framework is available at
http://pan-data.eu/sites/pan-data.eu/files/PaN-data-D2-1.pdf\

Actions within WP5

Common data standard policy for the SSRS-4 facility was the topic of a special session organised within the framework of the joint CREMLIN workshop on big data management (15-16 February 2017, NRC "Kurchatov Institute", 1, Akademika Kurchatova pl., Moscow). The results of this workshop are summarised and published within CREMLIN Deliverable 2.3. The discussions in the special sessions concerning large-scale X-ray facilities can be summarised as follows:

- The Photon Science community is a very large and heterogeneous user community exploiting a well-developed European network of accelerator-based photon sources (storage rings, XFELs). Most of the photon sources are set up as national sources, albeit with a major fraction of user time being allocated to users from across Europe (transnational access). The Photon Science community has been organising itself since 2006 in order to deal with the data challenge, mostly fostered by European projects. Basic points for an open data policy for all facilities have been established (PaNData-ODI project) and are in the process of being adopted and implemented at various facilities. Using the ESRF as an example, the challenges

of data management at the next generation of facilities, that are already in planning or under construction, were introduced with data production rates reaching the level of PBytes/week. Currently, the facilities are trying to organise a European network for the development of shared data analysis tools, which are mandatory to exploit the next generation of facilities. This will require support on the European level since (i) it goes beyond the level of support available from national funding agencies and (ii) the strong transnational use of the facilities fully justifies the use of common resources.

- For Photon Science at accelerator-based sources, there was very little interaction on issues of big data and data management. This will become more pertinent with the start of user operation of the European XFEL where Russia is a major partner and issues such as data transfer and data analysis capacities have to be dealt with. The construction of a 4th generation high energy photon source at the ESRF and in Russia also elevates the level of big data issues (large data sets, limited data analysis capacity individual user groups' home institution) for the Russian photon science community and is a docking point for common development efforts.

It is worth noting that many Russian users of accelerator-based photon sources are already frequent users of the European large-scale facilities. They are, therefore, already accustomed to adhering to data policies established within the general framework developed originally by the PANdata group.

Developments outside CREMLIN

LEAPS - the League of European Accelerator-based Photon Sources

LEAPS is a strategic consortium initiated by the Directors of the Synchrotron Radiation and Free Electron Laser (FEL) user facilities in Europe. Its primary goal is to actively and constructively ensure and promote the quality and impact of the fundamental, applied and industrial research carried out at their respective facility to the greater benefit of European science and society. LEAPS has been formally launched in Brussels, 13 November 2017 (https://www.leaps-initiative.eu/). In its goals and strategy LEAPS has made explicit reference to data policy and Open Science. It explicitely wishes to promote greater coherence in the developments of data-policy, -handling, -storage, -analysis, -access and the promotion of Open Science.

Once established, SSRS-4 could become a partner/associate to the LEAPS initiative.

Recommendations

Considering the context of SSRS-4, where European partner facilites are already involved in the preparation of the project, the WP5 CREMLIN partners recommend to adopt a data policy which is aligned with the data policy framework developed by the PANdata group. This concerns

- the defintion of data (raw data, metadata, processed data)
- the ownership of data (data created within public access, proprietary data)
- the curation of and access to data

- the definition of persistent identifiers
- the licensing of data (open data)
- the application of the FAIR principle (Findable, Accessible, Interoperable, and Re-usable)

Since SSRS-4 is a green-field facility, it will be much easier to implement this framework from the outset without having to change practices and habits.