Addressing the challenges of implementing the ESRF Data Policy

Andy Götz, Alex de Maria, Armando Solé, Roberto Homs-Regojo, Bruno Lebayle, Joanne McCarthy, Jens Meyer, Dominique Porte and Rudolf Dimper

ESRF, 71 ave des Martyrs, 38000, Grenoble, FRANCE

E-mail: andy.gotz@esrf.fr

Abstract. The ESRF, the European Synchroton, has recently adopted a Data Policy which will archive all data collected at the ESRF for 10 years and be made freely available as Open Data after an initial embargo period of 3 years (can be extended on request). Currently the ESRF produces 2 PBs of raw data annually. This means archiving at least 70 PBs of data over the next 10 years if one assumes a linear growth of data production. The Data Policy introduces a number of new challenges for the ESRF. These challenges include persistent user identities, user rights, metadata definition and standardisation, automated collecting of metadata, metadata catalogue, data containers, long term archiving, and finding and re-using data. This paper will describe how these challenges are being solved. The paper describes how it is possible for a mature synchrotron to adopt and implement a modern Data Policy largely built on existing standards like the ICAT metadata catalogue and the HDF5/Nexus data format/convention. Archiving such large quantities of data is largely due to the availability of off-the-shelf tape technology which continues to evolve and improve.

1. Introduction

This paper describes the challenges faced by the ESRF in implementing an open data policy at an existing site which previously left the issues of data management to the users.

2. The Data Policy

The main points of the ESRF Data Policy [1] are :

- ESRF Council officially adopted a Data Policy (1/12/2015)
- ESRF is custodian of data and metadata
- ESRF collects high quality metadata to facilitate reuse of data
- ESRF will keep raw (or reduced) data for 10 years + metadata forever
- Data will be registered in a data catalogue (ICAT [2]) + published with a Digital Object Indentifier (DOI)
- The experimental team has exclusive access to data during the embargo period (3 years but can be extended on request)
- Data will be made public after the embargo period under CC-BY licence
- Data Policy will be implemented on all beamlines by 2020

3. The Challenges

This section will present the different challenges which need to be addressed to implement the data policy.

3.1. Defining and adopting a Data Policy

The first challenge in adopting a Data Policy is to convince the major stakeholders that they need one. Once this step has been achieved the next challenge is to propose or adapt an existing data policy which fits the needs of the stakeholders and the community being served.

The need for having a Data Policy even for a mature institute like the ESRF which has been producing data without an Open Data Policy since 1992 has been driven by the Open Science movement. An increasing number of scientific institutions, universities, governments, and inter-governmental organisation encourage, expect and in some cases require science to be more open and traceable. There are many examples of charters and manifests signed by these bodies which call for Open Science. Open Science for scientific data means Open Data. This has become a requirement for all projects financed by the European Commission. This change in the political landscape convinced the ESRF stakeholders to study the Data Policy question again.

Once it was decided that a Data Policy was necessary the next step was to define a Data Policy. The ESRF participated in the Photons and Neutrons Pandata-Europe project (funded by FP7) where it led the Work Package on Data Policy. The deliverable [3] was a generic Data Policy which took into account the Photon and Neutron community. The Pandata Data Policy was thus close to what the ESRF needs were too. Some modifications were made based on the specific needs of the ESRF e.g. industrial users. The proposal was presented to the Scientific Advisory Committee who requested some changes and then presented to the ESRF Council for final approval. The Council gave their official approval on the 1 December 2015.

The Data Policy was presented to the Users at the ESRF User Meeting in February 2016. Feedback was mostly positive with only a few critical comments concerning the duration of the embargo period. Users were informed they can ask for an extension of the embargo period.

The challenges concerning adopting the ESRF Data Policy were : to convince the ESRF management and scientists and the users that the Data Policy was (a) necessary and (b) feasible. The first challenge was addressed by the changing scientific and politicial landscape with the move towards Open Science and Open Data. The decision by the EC to require Open Data for all H2020 projects since 2015 is a good example of the changing landscape. The users were convinced by the improved metadata, the better logical structure of their data and the additional services like long term archiving and DOIs (see below).

The second challenge about the feasability concerns mainly the cost. This was discussed in two stages - the cost of storing (a) metadata, and (b) raw data. An estimate showed that the cost of storing metadata indefinitely was neglible compared to the cost of available disk storage. It was therefore feasible to have (at least) a metadata policy. The second issue was the cost of long term archiving raw data. The projected storage needs for storing 10 years of raw data are 70 Petabytes assuming current data volume growth. The ESRF has the advantage that it has two high capacity tape libraries on site for backups. With new tapes of much higher density it turns out that 70 Petabytes can be stored with a modest increase in the budget of 100 000 € per year (over 10 years). This is much lower than the cost to produce the raw data.

3.2. Defining Metadata

One of the main goals and challenges of the Data Policy is to collect high quality metadata. The quality of the metadata will determine whether data can be re-analysed or understood by someone who was not part of the team who did the experiment which produced the data. High quality metadata also eases the task of the experimental team to (re)analyse their data. Metadata is often described as data about data or even as data not absolutely necessary for

the analysis. In the case of the ESRF Data Policy metadata are defined as those data which describe the experiment.

The challenge of defining metadata is three-fold - what metadata to define, what standards to follow, and finding engineers and/or scientists who are interested in metadata. At the ESRF the decision has been taken to follow the Nexus conventions for defining metadata. Where definitions are lacking new local definitions have been adopted.

Only metadata which is related to data produced at the ESRF is stored. The Data Policy does not attempt to cover metadata produced by processes upstream (or downstream) related to the sample preparation or other processes which impact the experiment. Using the Nexus conventions and maintaining a common set of metadata for all beamlines which use the same technique makes the challenge of defining metadata manageable.

3.3. Collecting Metadata from Experiments

Once the metadata is defined the next challenge is how to collect it automatically on the beamlines. Due to the diversity of beamlines there are a large number of different types of experiments. This makes the challenge more difficult because there is no single solution for all the beamlines and all the experiments.

The solution adopted is to have a generic Metadata server which can be configured via one or more setups in the database to collect the appropiate metadata per experiment and store it in a metadata master file and send it to a metadata catalogue. The metadata server is written in Python as a Tango device server and is configured via properties stored in the Tango database. Tango [4] is the ESRF distributed control system framework. The only remaining difficulty is to introduce the concepts of sample and dataset in the data acquisition sequence. A sample refers to the user samples which are put in the beam. A dataset refers to that group of data files which were acquired on the same sample and can be reduced and/or analysed independently. This needs to be done in the generic data acquisition macros as well as the data acquisition macros which are often beamline specific. Where this is difficult to implement the user is given the possibility to manually decide when to change sample and when to create a new dataset in the data acquisition sequence.

The current solution (configuring the metadata and modifying macros) needs roughly 1 week to get the minimum parameters and up to 4 weeks for all parameters per beamline. This fits in with the timeline of the implementation.

3.4. Choosing a Metadata Catalogue

Another challenge in implementing a Data Policy is the choice of a metadata catalogue. The metadata catalogue is the database and modules which store and allow users to find the metadata and download the raw data. There are a number of catalogues for managing metadata and data. The CRISP FP7 project which the ESRF participated in included a work package 17 [5] for evaluating metadata catalogues. A number of catalogues were compared e.g. ICAT, Dspace, Fedora, Ckan, Invenio, Tardis, ISPyB, iRODS, SRB-MCAT, MS. Zentity. Two of the catalogues, ICAT and ISPyB, were developed in the photon and neutron communities. ICAT was developed at STFC (see talk [6] at this conference) based on a generic data model and ISPyB at the ESRF for Macromolecular Crystallography and BIOSAXS. The other metadata catalogues come from different domains and consequently their data model are quite different.

The challenge of implementing a data policy for all experiments at the ESRF requires a very generic data model which implements the common features shared by all experiments. ICAT with its scientific data model describes the common features of all experiments.

3.5. Defining a Data Format

A challenge for the Data Policy is to get experiments to adopt a common format. This has been a challenge for many sites and ESRF is not an exception. The Data Policy states that data will be made available in a well known format which can be read and interpreted for the next 10 years. In the past the formats were either very simple (ASCII based) or single image files in tiff or simple binary format. With the increasing complexity of experiments, huge data volumes and files and multiple techniques used on the same sample the simple approach is not adapted anymore. For example with this approach some experiments generate millions of files. This is very inefficient when transferring, backing up and/or exporting data. A data format which can store huge volumes of heterogeneous data in a few files is required. Without going into a long discussion the de facto standard for storing this type of data is HDF5 [8]. HDF5 is a binary format developed and maintained by the HDF Group and adopted by the majority of scientific applications which need to store large and diverse data. It allows any kind of data to be stored much like on a file system. Once the data format was chosen the next step is to define or adopt a convention for how to organise the data in the HDF5 file(s). Fortunately the Nexus convention already exists and defines a number of the standards needed. Where standards are missing (especially for beamline specific data and for new techniques) the ESRF has setup a metadata working group locally to endorse new definitions which are defined locally as they are needed.

3.6. Long Term Archiving

An important aspect of the ESRF Data Policy is to archive raw data for 10 years. This opens up a number of possibilities like remote data analysis, persistent identifiers (PIDs), and providing an Open Data policy. Due to the large volume of data (currently a few Petabytes are produced a year) this is quite a big challenge. The challenge is in finding a cost and energy efficient solution for large volumes of data. Tape storage is the most cost and energy efficient solution today. The ESRF has two high capacity tape libraries on site for doing backups. Recent advances in tape density make it possible to offer a solution based on tape only. This has led to upgrading the tape drives and tapes to the latest generation. The tape libraries are now equipped with T10000 tape drives and T2 tapes which have a capacity of 8.5 TB and an average read/write speed of 300 MB/s for large files. This brings the capacity of each tape library to 72 PBs. Data are duplicated in both tape libraries to ensure that two physical copies are kept in physically separate areas.

Tape archives are managed by a tape library and backup software (Time Navigator). In order to be efficient this imposes storing large (> few GBs) files and limiting the number of entries in the backup database. Some basic arithmetic showed that in order to stay within the limits of the backup database over a period of 10 years we need to store on average < 1000 files per shift per experiment (assuming the ESRF continues to operate a similar number of hours as currently and maintains around 40 beamlines). As said above a large number of experiments (roughly 50 %) are above this limit today. Some are much higher ($> 10^7$ files) and we will need to move to using HDF5 before we can archive the data. This can be done in one of two places - at the data acquisition or when the data is archived. Although both solutions are envisaged there is a preference for the first.

3.7. Publishing PIDs

An important added value of the ESRF Data Policy is the generation of Persistent Identifiers (PIDs). A persistent identifier (PID) is a long-lasting reference to a document, file, web page, or other object. Persistent identifiers look like web addresses but with the difference that they are guaranteed to be there in 10 years time. A number of PID schemes exist (ARK, DOI, PURL, URN and XRI). The CRISP project did a survey and proposed to use the DOI system. DOI's

consist of a landing page (web page) which contains high level metadata and a pointer to the data. The DOI are particularly useful for publications to be able to cite the data. It is possible to do the reverse and link DOIs to publications thereby tracking number of publications which cite ESRF data.

The main challenge with DOIs is not in implementing them but to decide at what level they should be generated i.e. per Investigation, per Sample or per Dataset. The cost of a DOI is low (cents) however it is important to make DOIs which can be generated with a minimum of human intervention. The solution currently proposed is to automatically generate one DOI per investigation and per sample. In the future users will be able to generate DOIs for a collection of datasets of their choice.

3.8. User Authentication + Authorisation

The archived data need to be kept under embargo for up to 3 years (or longer in special cases) with only members of the experimental team having access. This means all members have to be identifiable with individual ID's over a long period. This challenge is what is known as AAI - Authentification and Authorization Infrastructure. At the ESRF this is being solved with individual user accounts in the local LDAP server with the possibility of users being able to use their Umbrella credentials to authenticate. In the future other authentification mechanisms will be added such as eduGAIN.

3.9. Searching + Finding Data

Open Data are supposed to be FAIR - Findable, Accessible, Interoperable and Reusable [7]. The latter two are related to having high quality metadata and a common data format. The first two are addressed by the metadata catalogue, publications and the Open Data repositories. This challenge is currently addressed by the search mechanisms implemented in ICAT based on the Lucene indexing algorithm ([9]). Currently it is possible to search on Investigation, Sample and Dataset. Searching on Parameters (metadata related to specific dataset) is very basic due to the large variety of parameters and needs extending in the future. The next step is to register the ESRF metadata catalogue with external open data repositories so ESRF data can be found more easily. This is an expanding field and the repositories depend on what are the most widely used ones at the time of registration. One solution is to implement a data harvesting protocol like OAI-PMH which can be used for data searching. The agency managing the DOIs (Datacite) provide some of these features but they could be offered by ICAT directly in the future.

3.10. Exporting Data

A challenge for users is how to get their data home, especially when they have a lot of it. One of the main services offered by ICAT is a data download service called IDS. The IDS currently supports multiple protocols - http, ftp, WebDav, gridftp. IDS retrieves data from the tape archives or disk (if online) to a staging area where the user(s) can download using it one of the above protocols. The ESRF Data Policy does not automatically offer compute resources for re-analysing the data. This service is considered part of the DAAS (Data Analysis as a Service) service which will profit from the implementation of the Data Policy but which will need a separate budget.

3.11. Cost

The expected additional investment cost of implementing the ESRF Data Policy is 100 000 € mainly for the long term storage. An extra 1.5 persons per year is required in human resources to implement the Data Policy. More resources are required to implement the metadata collection on beamlines - up to 1 month per beamline, and to convert data analysis programs to read

HDF5. ESRF has started on this work for specific applications and is developing the **silx** toolkit, a software library which provides builtin support for HDF5 and converts legacy formats (SPEC files) to HDF5 (see silx poster [10] at this conference). Additional work required for implementing single user accounts and authentication has already started and is foreseen to be completed by 2017.

4. Timescale

The goal is to implement the ESRF Data Policy on all beamlines by 2020. This corresponds to roughly 10 beamlines a year. The first beamline to implement the full Data Policy (including DOIs and long term archiving) is expected to be at the end of 2016.

5. Conclusion

The adoption of an ESRF Data Policy has put data management on the front of the ESRF agenda. The first group of people to benefit from this are expected to be the users themselves. The Data Policy is only the start of data management, it opens a number of new possibilities like Remote Data Analysis as a Service, data citing, data re-use to mention a few. These services will be possible to implement in the future largely due to the Data Policy.

5.1. Acknowledgments

Thanks to the ICAT development team @ STFC for providing a robust metadata catalogue. Thanks to the system admins at the ESRF for providing a powerful infrastructure. Thanks to those scientists who have helped to define and improve metadata.

6. References

[1] ESRF Data Policy `http://www.esrf.fr/files/live/sites/www/files/about/organisation/ESRFdatapolicy-web.pdf`
[2] ICAT project web site `http://icatproject.org`
[3] PaNData Europe Data Policy `http://wiki.pan-data.eu/imagesGHD/0/08/PaN-data-D2-1.pdf`
[4] TANGO Controls web site `http://tango-controls.org`
[5] CRISP WP 17 `http://www.crisp-fp7.eu/research-programme/work-packages-descriptions/wp17/`
[6] Fisher, S et. al. *Growth of the ICAT family*, NOBUGS 2016 conference (Copenhagen)
[7] Wilkinson, M. D. et al. *The FAIR Guiding Principles for scientific data management and stewardship.* Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).
[8] HDF5 home page `https://support.hdfgroup.org/HDF5/`
[9] Lucene project web site `https://lucene.apache.org/`
[10] Solé, V.A. et. al. *The silx toolkit*, NOBUGS 2016 conference (Copenhagen)