# Towards an open data infrastructure for photon and neutron facilities

*Juan Bicarregui and Brian Matthews*
*STFC Rutherford Appleton Laboratory*
*{Juan.Bicarregui, Brian.Matthews}@stfc.ac.uk*

## Introduction

Today's scientific research is conducted not just by single experiments but rather by sequences of related experiments or projects linked by a common theme that lead to a greater understanding of the structure, properties and behaviour of the physical world. This is particularly true of research carried out on large-scale facilities such as neutron and photon sources where there is a growing need for a comprehensive data infrastructure across these facilities to enhance the productivity of their science. In this short paper, we describe such an infrastructure under development by the PaNdata consortium.

Photon and neutron facilities support fields as varied as physics, chemistry, biology, material sciences, energy technology, environmental science, medical technology and cultural heritage. Applications are numerous: crystallography reveals the structures of viruses and proteins important for the development of new drugs; neutron scattering identifies stresses within engineering components such as turbine blades, and tomography can image microscopic details of the structure of the brain. Industrial applications include pharmaceuticals, petrochemicals and microelectronics. The experiments undertaken in these facilities are of growing complexity, they are increasingly done by international research groups and many of them will be done in more than one laboratory. As a result of the these trends and the increased capability of modern electronic detectors and high-throughput automated experiments, these facilities will soon produce a "data avalanche" which makes it essential that forces are joined to implement and deploy a framework for efficient and sustainable data management and analysis.

The PaNdata collaboration[1] brings together thirteen large multidisciplinary Research Infrastructures across Europe which operate hundreds of instruments used by over 30,000 scientists each year. Founded in 2008, PaNdata aims to construct and operate a shared data infrastructure which will enhance the research done in this community by providing powerful tools for scientists to interact with the data and allowing experiments to be carried out jointly in several laboratories .

A feature of these facilities is the extent to which their scientific community is shared. An estimate of users in common undertaken by PaNdata[2] showed that more than 20% of all users use more than one facility, with considerable numbers of users using both types of facility, and typically, 30-40% of the users of any of the photon or any of the neutron sources also use at least one other facility. This supports the case that there is likely to be considerable value if users have a similar user experience at each facility, and can easily share and combine their data and other resources as they move between facilities.

In the rest of this paper, we discuss the approach of the PaN-data consortium to providing a common open data infrastructure to support the users of facilities as they pass through and between different facilities.

---

1 Photon and Neutron Data Inftrastructure: http://www.pandata.eu
[2] Counting Users of the European Photon and Neutron Facilities: http://wiki.pandata.eu/CountingUsers

## The Facilities Scientific Lifecycle

While the Photon and Neutron scientific techniques and instruments vary, from a data infrastructure point of view, facility users typically go through a lifecycle as in Figure 1.

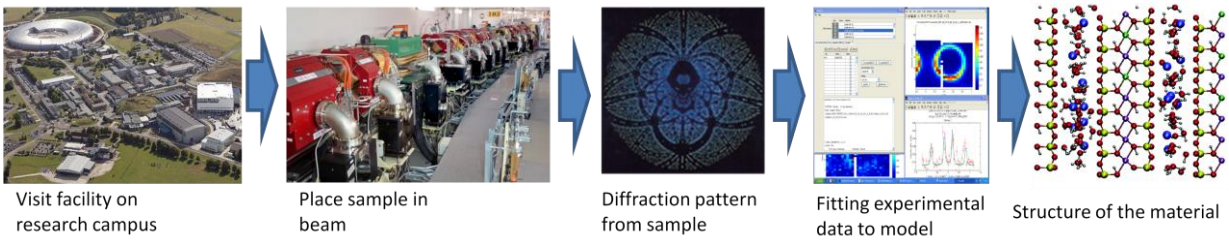


**Figure 1: Idealised lifecycle of an experiment within a facility**

These stages are as follows:

1. **Proposal :** The user submits a proposal applying to use a particular instrument on the facility for time to undertake experiments on particular material samples. This is lodged with the Facility and the application is judged on its scientific merits of the proposal, successful proposals being allocated a time period within an operating cycle of the instrument.
2. **Experiment:** During a visit to the facility, a set of samples are placed in the beam and a series of measurements are taken. Different instruments at the facilities have their own characteristics, but all have data acquisition software which will take data on the parameters of interest. Historically, this data is stored within the file systems associated with the instrument. However, as data volumes have grown, there has been a need to provide systematic support for this.
3. **Data Storage:** Data is aggregated into data sets associated with each experiment, stored in secure storage, within managed data stores in facility, and systematically cataloged.
4. **Data Analysis:** The scientist takes the results of the experiments (the "raw data"), and carries out futher analysis. The data from the instruments is typically in terms of counts of particles at particular frequencies or angles, and needs highly specialized interpretation to derive the required end result, typically a "picture" of a molecular structure, or a 3-D image of a nano-structure.
5. **Publication:** a suitable scientific result having been derived from the data collected, then the scientist will report the results within journal articles. The facility would usually like to be acknowledged and informed of its publication, so that it can track the impact of the science derived from the use of its facilities.

Thus there is a *Data Continuum* from proposal to publication. Added value can be provided to the user by providing integrated support for this continuum. Proposals registered with the user office can initiate the cataloguing of experiments and their datasets, and this can in turn be associated with analysed data and publications. This provides both consistent access to the experiment for the user, and efficient management and impact tracking for the facility. This lifecycle is similar at all facilities, so when a scientist uses more than one facility, the infrastructure and thus the user experience should also be similar. This will make the facility easier to use, and simpler to access and combine the results of experiments.

At the heart of the PaNdata vision is a series of federated catalogues which allow scientists to perform cross-facility, cross-discipline interactions with experimental and derived data, with near real-time access to the data. The architecture follows the data pipeline from data creation through to publication of analysed results which feed back into new research proposals. Thus the consortium is developing:

- A common user access system allowing users registered at one facility to access resources across the consortium using one identity.
- Standard formats so that data generated by one instrument can be readily combined with data from others and made accessible to common software tools.
- A federated data cataloguing system with a common metadata model, so that experiments and their associated data can be registered and published, so that users can access their data generated from different sources in a uniform way, and to allow searches of experiments across the facilities.
- A common registry of software which is suitable for use with each facility's data.
- Uniform access to publication systems so that the results of experiments can be tracked and accessed.

This is all within a common policy framework for data storage, access, publication and preservation, and also for software development.

Within the data continuum, there is currently limited support for systematic tracking of the provenance of data as it passes through the analysis phases of the lifecycle. This is an area which is difficult to manage as it tends to be within the control of the user scientist and require specialised techniques. Nevertheless, this is required for a verifiable record of the experiment, and extensions to the data and software catalogues to support provenance trails are under investigation.

## Current status

The first project involving the PaNdata partners, PaNdata Europe[3], has recently completed. It focussed on standardisation activities in the areas of data policy, user information exchange, scientific data formats, interoperation of data analysis software, and integration and cross-linking of research outputs. Essential elements of a scientific data policy were agreed, covering aspects such as storage, access, and acknowledgement of sources. A second project, PaNdata Open Data Infrastructure[4], is building on these results to put in place some key components of a common data infrastructure across the participating facilities, including a system for pan-European user identification across the participating facilities; a generic federated catalogue of scientific data across the participating facilities; and a metadata model, to record the analysis process, enabling the tracing of the derivation of analysed data outputs. Further, such a data infrastructure will need to be sustainable to gain the best value from data in the long-term; thus there is an investigation on how the capabilities of the infrastructure can be oriented towards long-term preservation

The infrastructure will be instantiated through three virtual laboratories supporting powder diffraction, small angle scattering and tomography. Deploying this common open data infrastructure across these major scientific facilities will open up new frontiers in data exploitation, with significant potential economic benefits, as the 'time to market' of the supported science is reduced.

[3] *PaN-data Europe project funded by the European Commission under the 7th Framework Programme, grant agreement RI-261537, June 2010-Nov 2011.*
[4] *PaN-data ODI project funded by the European Commission under the 7th Framework Programme, grant agreement RI-283556, Oct 2011-March 2014.*

Track:  Submission to the Digital Research Innovation Showcase track

Presentation: Oral