e-Infrastructure across Photon and Neutron Sources

Brian Matthews, STFC

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.

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 Research carried out at neutron and synchrotron facilities is rapidly growing in complexity. Experiments are also increasingly being carried out by international research groups and in more than one laboratory. Combined with the increased capability of modern detectors and high-throughput automation, these facilities are producing an avalanche of data that is pushing the limits of IT infrastructures.

In addition, there is a push from policymakers and some scientific communities to make data "open" in order to encourage transparency and sharing between scientists. It therefore makes sense to build a common data infrastructure across different photon and neutron facilities that makes data management and analysis more efficient and sustainable and maximises the science throughput of their user communities.

Established in 2008 by the ESRF, ILL, ISIS and Diamond, the PaNdata consortium now brings together 13 large European research infrastructures that each operate hundreds of instruments used by some 33,000 scientists each year. Its aim is to provide tools for scientists to interact with data and to carry out experiments jointly in several laboratories. Research undertaken by PaNdata partners show that more than 20% of all European synchrotron and neutron users make use of more than one facility[1]. It is therefore of considerable value to offer scientists a similar user experience at each facility, and to allow them to share and combine their data easily as they move between facilities.

At the heart of PaNdata are shared data catalogues that allow scientists to perform cross-facility and cross-disciplinary research with experimental and derived data almost in real time. The catalogues follow the data from its creation and analysis to the final publication of results, which feed back into new research proposals.

[1] See the PaNdata Counting Users Survey http://pan-data.eu/Users2012-Results

The first stage of the project, PaNdata Europe, which ran from 2000–2011, focussed on data policy, user information exchange, scientific data formats and the interoperation of data analysis software. Essential elements of a scientific data policy were agreed, covering aspects such as storage, access and the acknowledgement of sources.

A second project, PaNdata Open Data Infrastructure, which will conclude in 2014, includes a common user authentication system to allow users registered at one facility to access resources across the consortium using one identity; promoting standard formats so that data generated by one instrument can be readily combined with data from others; and building a federated data cataloguing system with a common metadata standard, allowing users to access data generated from different sources in a uniform way. The project is also extending data management across the data continuum, into analysis processes so that users will able to trace how data are used once they have been collected. Since a data infrastructure must be sustainable, the consortium is investigating which processes and tools need to be changed to allow a facility to move towards long-term preservation, and also considering approaches to scaling data management as data acquisition rates continue to grow.

PaNdata intends to continue working together, and also extend their collaboration. The recently established Research Data Alliance (RDA) will allow other data managers and scientists throughout the world to collaborate. The RDA is a new organisation backed by the EC, NSF in the US and the Australian National Data Service and includes several working groups in areas such as metadata, data publishing, digital preservation, and policy enactment. PaNdata is proposing an RDA Interest Group to develop best practise for data management in the photon and neutron community across the world. The ultimate vision of PaNdata is to allow users to move within and between facilities without having to learn and use new computing systems. By allowing data to be moved, shared and mixed together simply across the complete lifecycle of an experiment, scientists can concentrate on getting the best science from the facility.