PaNdata: A European Data Infrastructure for Neutron and Photon Sources

Juan Bicarregui, Simon Lambert, Brian Matthews, Michael Wilson
STFC e-Science, Rutherford Appleton Laboratory, Oxon, OX11 0QX

1 BACKGROUND

Since 2009 the PaNdata consortium of European photon and neutron laboratories has focused 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. These standards form the baseline for a new project, PaNdata Open Data Infrastructure (PaNdataODI) which will deliver common services that will integrate data across the consortium to create a fully integrated, pan-European, research data infrastructure supporting numerous scientific communities across Europe.

The participating facilities are major producers of scientific data which serve an expanding user community of about thirty thousand scientists across Europe. Historically the situation at many of the facilities, and in particular at the photon sources, has left data management largely to the individual users who, often literally, carried data away on portable media. These media are notoriously unsuitable to guarantee the longevity and availability of costly experimental data. Keeping track of experimental data is becoming increasingly important as the rate at which experiments can be performed and analysed is increasing and the experiments themselves are of increasing complexity, and increasingly performed in more than one laboratory by collaborations between international research groups. The resulting raw and processed data therefore need to be accessible over the Internet across facilities and user institutions enabling researchers from different organisations to seamlessly work together. It should remain on-line at least until the results are published, in many cases much longer to allow validation of analyses, and re-processing to test new theory. The increased resolution of modern electronic detectors and high-throughput automated experiments, will cause these facilities to produce even more enormous quantities of scientific data in the future, which makes it essential to establish an efficient and sustainable European infrastructure for data management and analysis.

The standardisation being done by the PaNdata consortium will tremendously simplify the landscape for multi-disciplinary exploitation of the instruments and lay the groundwork for common implementation of data management infrastructure across these participating facilities and beyond. This will, in turn, allow industry to utilise publicly available data, processing or reordering the data in such a way that it could be presented with added value to commercial market segments such as the life sciences, engineering or materials sciences.

To drive the development and evaluate the benefit of the services deployed, PaNdata ODI will implement three virtual laboratories which provide case studies in the use of the shared data infrastructure. These virtual laboratories will support the following techniques:

1. structural 'joint refinement' against X-ray & neutron powder diffraction data,
2. simultaneous analysis of SAXS (Small Angle X-ray Scattering) and SANS (Small-Angle Neutron Scattering) data for large-scale molecular structures,
3. tomography such as demonstrated in the rendering of palaeontology samples.

2 CONCEPTUAL DESIGN

Our vision is to standardise and integrate our research infrastructures in order to establish a common and traceable pipeline for the scientific process from scientists, through facilities to publications. At the heart of the vision is a series of federated catalogues which allow scientists to perform cross-facility, cross-discipline interaction with experimental and derived data, with near real-time access to the data. This will also deliver a common data management experience for scientists using the participating infrastructures particularly fostering the multi-disciplinary exploitation of the complementary experiments provided by neutron and photon sources.

Building on the unification of data management policies and adoption of common data standards by the PaNdata consortium, PaNdata ODI will develop and deploy the common technologies which will realise the benefits of standardisation. The aim is illustrated in Figure 1. Currently, each facility separately handles the full data management sequence, from generation of raw data to publication of results. In the future view, a single user experience will be enabled through the use of a common data management scheme at the different facilities. The common infrastructure will make data accessible, preserve the data, allow experiments to be carried out jointly in several laboratories and provide powerful tools for scientists to remotely interact with the data.

2.1.1 User catalogues and data catalogues

An integral component of the project is an authentication and authorization system that is normalised to include scientific users across the collaborating facilities and able to interoperate with similar systems across the ERA. The system delivered here is not to replace the local systems of the individual facilities, but rather to allow these systems to interoperate such that individual scientists can be identified on a pan-European level. One major benefit of this is that individuals will be able to seamlessly access all their resources at any of the facilities without having to authenticate themselves against the different systems in place at the participating partners. Other benefits include ease of maintenance which arises from the elimination of multiple entries for particular users and the ability to follow individual scientists as their careers progress through different roles, at different facilities, and across national boundaries. It therefore removes one significant obstacle to the coordination of research policy and practice across Europe.

2.1.2 Provenance and preservation

Data catalogues provide a valuable capability to locate particular data sets seen as a 'snapshot' of the scientific process of which they are one of the outputs. The project aims to go a step further and, by representing the *provenance* of the data, link it into the context of the process—the lifecycle of scientific endeavour. This will have a number of impacts. It will become possible to validate published results, linking back through the analyses to the original raw data gathered from the instruments. This is not just a safeguard against errors or fraud, but enables the application of improved analysis techniques as they become available, without needing to repeat the entire experiment, by securely establishing dependencies and derivations from preceding data sets in the chain. Thus efficiency will be improved along with the reliability of the results of such subsequent analyses.

2.1.3 Scalability

Essentially all applications in photon and neutron science currently rely on a strictly sequential data access model. This poses a significant bottleneck for real time analysis and hinders efficient use of advanced computing technology. For example, it is not a major problem to dump a large number of individual files onto disk and do analysis almost simultaneously. However, this is a poor data management solution. A dataset logical combining digital objects should be compiled into a single, self-descriptive, structured data file (a Nexus-file), which however prevents simultaneous analysis. Development and implementation of a parallel pHDF5 capable Nexus API (pNexus) will overcome such limitations, which will not only accelerate the analysis workflow, but lays the foundation for efficient analysis of extreme data rates from highly advanced light source like x-ray free electron lasers.

3 SUMMARY

PaNdata-ODI will develop, deploy and operate an Open Data Infrastructure across the participating facilities with user and data services which support the tracing of provenance of data, preservation, and scalability through parallel access. It will be instantiated through three virtual laboratories supporting powder diffraction, small angle scattering and tomography.

4 ACKNOWLEDGEMENTS

The research reported in this paper was partially funded by a support action grant from the European Commission's 7th Framework Programme to the PaNdata collaboration. Formed in 2008, the PaNdata collaboration currently brings together eleven major world class European Research Infrastructures providers (See www.pandata.eu). The PaNdata Partners are: ISIS, the world's leading pulsed spallation neutron source; ILL, the most intense slow neutron source in the world; PSI, including the Swiss Light Source, SLS, and Neutron Spallation Source, SINQ, and is developing the SwissFEL Free Electron Laser; HZB operates the BER II research reactor the BESSY II synchrotron; CEA/LLB operates neutron scattering spectrometers from the Orphée fission reactor; ESRF is a third generation synchrotron light source jointly funded by 19 European countries; Diamond is new 3rd generation synchrotron funded by the UK and the Wellcome Trust; DESY operates two synchrotrons, Doris III and Petra III, and the FLASH free electron laser; Soleil is a 2.75 GeV synchrotron radiation facility in operation since 2007; ELETTRA operates a 2-2.4 GeV synchrotron and is building the FERMI Free Electron Laser; and ALBA, a new 3 GeV synchrotron facility due to become operational in 2011.

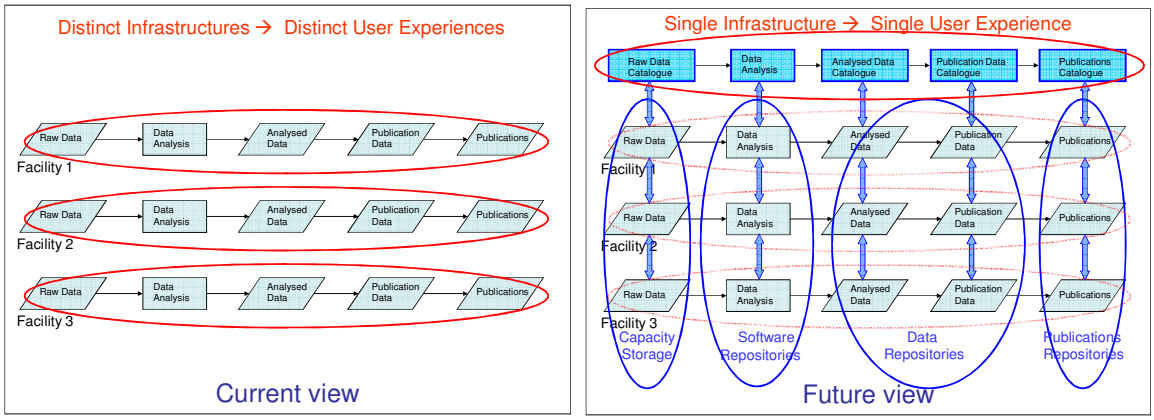

Figure 1: PaNdata Vision - Current and future views of data pipeline at facilities