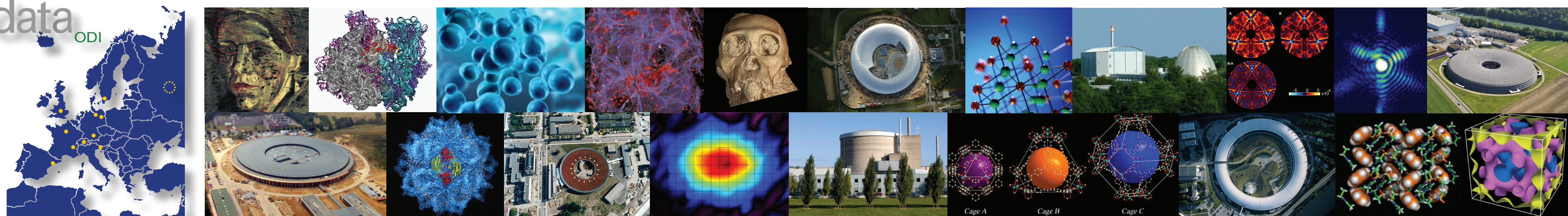


Open Data Infrastructure

Building blocks of the PaNdata

Open Data Infrastructure

J.Bicarregui (STFC), B.Abti (PSI), D.Dimper (ESRF), B.Gagey (SOLEIL), D.Herrendörfer (HZB), J.Klora (ALBA), T.Kracht (DESY), S.Longeville (LLB/CEA), B.Matthews (STFC), J.-F.Perrin (ILL), M.Prca (Elettra), R.Pugliese (Elettra), B.Pulford (DLS), F.Schlünzen (DESY), D.Spruce (MAXLab), M.van Daalen (PSI), H.-J.Weyer (PSI), J.Wuttke (JCN5/FRM-II)



Abstract

PaNdata is a long term co-operation of European Neutron and Photon Research Infrastructures. Within the FP7 framework PaNdata has launched two projects. PaNdata Europe, which was successfully completed in 2011, laid the basis for a common data management framework with the development of common policies on data lifecycles, user and software. PaNdata ODI has started end of 2011 and aims to provide a federated, sustainable data infrastructure across the participating facilities, to establish a common authentication and authorization system, to accelerate data analysis and to provide tools to link between experimental data, the analysis pathway and the scientific publications related to the experiment.

PaNdata

The PaNdata consortium involves thirteen major world class European Research Infrastructures. Most of these ERIs operate several lightsources (like e.g. DESY, ELETTRA) and/or a combination of Neutron and Photon Science facilities (like for example PSI, HZB or ILL and ESRF), providing hundreds of highly advanced scientific instruments, and ultimately serving several ten-thousands of users from a wide range of scientific disciplines. Neutron and Photon diffraction can exploit complementary aspects of physical and natural sciences. Although being quite different in various aspects, the basic principles are often quite similar. PaNdata intends to fully exploit the synergies arising from common approaches ranging from application development to user management. PaNdata aims to offer the user communities an infrastructure to fully exploit the complementarity within in unified environment, and at the same time facilitate user, data and resource management for the research infrastructures. PaNdata supports the movements towards Open Data by enabling users to rapidly analyse data, easily share information and results, and track and manage the process from the proposal to the scientific publication.

Science and Users

Neutron and x-ray scattering experiments for example can provide very different, complementary information of a protein structure, which has been recognized already 40 years ago. While x-ray diffraction provides a detailed view of the tertiary structure of a protein, it usually fails to locate hydrogen atoms. To fully understand the catalytic mechanism of an enzyme, the knowledge about the position and relocation of hydrogen atoms during catalysis is absolutely crucial. Neutron scattering can provide such information, however the joint refinement of neutron and x-ray data is often hampered by lack of standards, appropriate tools and accessibility of scientific data. In the long term PaNdata aims to create an open data infrastructure to provide access to all experimental data created at any of the participating facilities, thereby promoting collaborative research, distributed analysis, complementary use of data from different experiments and facilities. This will allow providing open access to unique, curated data, like the paleontological samples shown in (2). X-rays now make it possible for palaeontologists to study opaque amber, previously inaccessible using classical microscopy techniques. Scientists from the University of Rennes (France) and the ESRF found 356 animal inclusions, dating from 100 million years ago, in two kilograms of opaque amber from mid-Cretaceous sites of Charentes (France). Long term, open accessibility of such precious data is hence absolutely essential for comparative analysis to gain new insights into the evolution of species.

1) Difference neutron scattering (green) and electron density (red) maps. Source: epm-campus.eu

2) Examples of 3D reconstructions of organisms embedded in opaque amber. Credits: M. Laki, P. Tafforeau, D. Néraudeau (ESRF Grenoble and UMR CNRS Rennes).

The majority of the research groups using neutron or x-ray sources visit more than just one facility, with varying group members and projects. In addition, users are commonly involved in small or large and volatile collaborations. Collaborating scientists might need access to data as well, though they have never been involved in an experiment and are hence not a priori known to the facilities. On the other hand, the process from proposal submission to radiation protection regulations require unique and persistent identification of a user/person. Collaborations and user are spread over all 5 continents. A solution for authentication/ authorization needs to be easily accessible from everyone and simple enough to be usable on a very infrequent basis.

Geographical location of the user communities' home institutions

Swiss Users of ESRF collaborate with scientists from many other countries worldwide

Infrastructure Building Blocks

Virtual labs integrate the different building blocks into a virtual research environment. DawnScience is for example an application integrating NeXus capabilities with advanced workflows and interfaces to explore the ICAT data catalogue.

Tracking of the analysis pathway, implementation of workflows into virtual labs and mining data repositories greatly benefit from the common software data catalogue.

Software

Photon and Neutron Software Catalogue

Please note: This website is currently in a BETA state. It is in the process of heavy development. Certain aspects of the website are likely to change. Furthermore, certain functionalities may not work.

Photon and Neutron Software Catalogue

Pathfinder is a database of software used mainly for data analysis of neutron and photon experiments. PaNdata is one element of a larger project, PaNdata, which aims to provide a complete, shared data infrastructure for neutron and photon laboratories.

This database can be freely consulted. It gives an overview of software available for neutron and photon experiments and their use with respect to instruments at experimental facilities.

By registering and logging in new software can be entered and it will appear in the database after moderation. Similarly, feedback can be given on the software presented herein and more generally via the forum hosted here.

RECENT SOFTWARE

- NAMID: A parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems.
- MOLREP: Processing, Molecular Replacement.

SEARCH FOR SOFTWARE

Looking for a piece of software? Use the search box below to find it.

Umbrella project architecture diagram showing the integration of various services and databases.

The whole process from proposal submission, running an experiment, accessing data and publication requires unique and persistent identification of the scientist. The Umbrella project provides tools to support the process in a unique way across the facilities.

Virtual Laboratories

Scalability

Federation

Data Continuum

Rapid data analysis and visualization is crucial to tune experiments in real time. In co-operation with projects like PNI-HDRI, GPGPU and multi-core acceleration of tomographic reconstructions is a particular successful development. Integration into virtual labs can tremendously enhance the scientific outcome.

Image Loader

Fetch slices for processing

Pool of Sinograms (host memory)

Pool of CPU and GPU processing threads

Store results

Pool of Vertical Slices (host memory)

GPU thread

1st Stage

2nd Stage

Double buffering

Filtering

Texture

Double buffering

PCIE Data Transfer

PCIE Data Transfer

ICAT Metadata Repository

Proposal

Approval

Scheduling

Experiment

Data analysis

Record Publication

Subsequent publication registered with facility

Facility committee approves application

Facility registers, trains, and schedules scientist's visit

Scientists visits, facility run's experiment

Raw data filtered and cleansed

Tools for processing made available

PanSoft

Science & Technology Facilities Council

The data catalogue is the core of the data infrastructure. ICAT provides the tools and services supporting the whole process from proposal submission to publication; data analysis and workflows; data ingestion, curation and provenance.

Federation of resources and services is a crucial element to provide scientists with a panEuropean research environment, including a common AAI infrastructure, the federation of data catalogues as well as affiliation and publication databases.

Contact: <http://www.pandata.eu>
Dr Juan Bicarregui
STFC e-Science

This work is partially supported by the European Commission under the 7th Framework Programme Grant Agreement 283745 (CRISP) and 283556 (PaNdata ODI)