PaN-data ODI

Deliverable D2.5

Report on the 2nd Open Workshop

Grant Agreement Number	RI-283556
Project Title	PaN-data Open Data Infrastructure
Title of Deliverable	Report on the implementation of the three virtual laboratories
Deliverable Number	D2.5
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	30 Sep. 2014
Actual Delivery Date	15 Oct. 2014

The PaN-data ODI project is partly funded by the European Commission under the 7th Framework Programme, Information Society Technologies, Research Infrastructures.

Abstract

This report briefly describes the 2nd PaNdata ODI Open Workshop co-located with the RDA 4th plenary I Amsterdam September 2014.

Keyword list

PaN-data ODI, dissemination, engagement, workshop

Document approval

Approved for submission to EC by all partners on 15.10.2014

Revision history

Issue	Author(s)	Date	Description
1.0	Frank Schlünzen	Oct 2014	Final Version

The 2nd PaNdata ODI Open Workshop took place in Amsterdam on the 25th of September, co-located with the 4th plenary of the Research Data Alliance (RDA). The workshop was a one-day event, complemented by three RDA sessions and concentrating on three major themes:

- Standard data formats
- Evolution of scientific instruments and storage infrastructures
- Evolution of storage concepts

1.1 The workshop in brief

The workshop was a full day event and co-located with the RDA 4th plenary. Among the contributors were in particular a number of technology providers, namely HDFgroup, IBM and Dectris. The workshop was based on the PaNdata ODI developments with a focus on future challenges and developments, in particular the intertwined evolution of scientific instruments, data transfer technologies, file formats and storage infrastructures, hence covering most of the systems essential for off-line and on-line data recording, analysis and management.

The PaNdata workshop was complemented by three RDA sessions, two for the Photon and Neutron Science Interest Group (PaNSIG); one for Scientific File Formats (SFF) and HDF5 external filters. The PaNSIG sessions served mainly for networking with Photon and Neutron RIs outside the European domain and with related RDA activities like the Federated Identity IG, the Material Science IG, the Structural Biology IG and others. The BoF session on scientific file formats aims to support sustainable formats. The importance of the issue was emphasized by presentations of high-level members of de facto standardization bodies, namely HDFgroup for HDF5, UNIDATA for netCDF and NIAC for NeXus.

The workshop and RDA session were joined by more than 50 participants, more than half of it from outside the PaNdata collaboration.

All slides of all sessions and other materials are available from pan-data.eu.

1.2 Selected workshop topics

One emerging challenge - not only for lightsources like Free Electron Lasers - is the development of detectors with extremely high repetition rates in combination with high spatial resolution. The conventional approach considers a single image as the minimal entity to be stored sequentially in a container or on a physical storage device. Data recording is then expected to be concurrent with real-time analysis and fast visual inspection in fairly complex experimental environments. This approach is already posing severe problems, which can only be partially resolved with high-end storage devices and recent HDF5 developments. Clearly the evolution of experimental instruments needs to be accompanied by an evolution of the data recording and processing strategies. Therefore, several of the presentations were focusing on potential strategies and implementations. The strategy favoured by most of the RIs appears to be based on distributed on-line messaging frameworks were data streams rather than images are simultaneously but asynchronously broadcasted

and processed by dedicated components particularly designed to fulfil specific tasks. The approach appears as a common baseline among most of the RIs asking for cooperative developments. The workshop has certainly contributed to identifying common approaches and problems, and promoted joint solutions beyond the PaNdata project and beyond the PaNdata consortium.

This paradigm shift touches several aspects of the data lifecycle, including recording of meta-data (e-logbook), the aggregation of streams in HDF5/NeXus files, the analysis workflows, the storage devices and the scientific instruments (detectors). The presentations were covering most of these aspects from various angles revealing an amazing diversity of methods and implementations. Although the diversity might appear as a drawback to arrive at common solutions it on the other hand provides a rich playground to evaluate a wide spectrum of concepts and modules which can greatly facilitate future developments.

1.3 Standard Data Formats

Standard data formats are one of the core elements of the PaNdata project, and substantial efforts have been invested to develop HDF5-based standards; achieve interoperability with community standards; develop high-throughput capable APIs. The importance of standards and file formats as a crucial element for data storage, analysis, mining or visualization has lead us to propose two new RDA activities, namely an Interest Group for Scientific File Formats and a Working Group for HDF5 external filters.

1.3.1 The Scientific File Format IG

Despite the fact of the central role of file formats for any data related activity, RDA had until now no activities dealing with the file formats in general or scientific file formats in particular (other than the associated meta-data). Sustainability of data infrastructures requires however significant efforts to maintain file formats, preferably relying on open, portable standards. Preservation of file formats is at least to some extend a pre-requisite for preservation of data analysis tool chains, and the lack of sustainable workflows and file formats will obscure attempts to collect and archive related provenance data.

We hence initiated a Scientific File Format IG, which aims to support sustainability of open standards and to promote the development and evolution of file formats. Since the focus is clearly on open standards, we primarily considered HDF5, netCDF, NeXus and community standards like CIF as the main targets.

The major standards were represented by Elena Pourmal, Director of Development at HDFgroup, the non-profit company maintaining the HDF standards; Mohan Ramamurthy, Director of unidata, the maintainer of the netCDF standard and reference installations of major tools for the geosciences community; Eugen Wintersberger as a member of the NeXus International Advisory Committee (NIAC), the de facto maintainer of the NeXus standard.

The presentations were accompanied by fruitful discussion and explicit expressions of interest on the topic. A workplan has been agreed on, outlining a number of specific topics. Details will be pub-

lished on the RDA site. The presentations are available from the PaN-data.eu site (and hosted on the DESY conferencing site).

1.3.2 The HDF5 external filters WG

Recently, HDFgroup has introduced a new mechanism for pluggable extensions to the HDF5 core library. The extensions have a severe impact on the readability of HDF5 data encoded with external filters. Outsourcing some functionality to pluggable extension leaves the backward compatibility and portability of the HDF5 core intact and minimizing modifications on the application side, but that comes at a price. The non-availability of a specific module might make it impossible to access the data in a HDF5 container. To maintain a high degree of usability, a registry of extensions, rules and best practices as well as support for a number of HDFgroup supported platforms seem inevitable. It's not a major but non-neglible long-term support effort, which should best be provided by the community rather than HDFgroup, as was commonly agreed. Aim of the workgroup is hence to establish the support platform and extension registry as well as the necessary documentation on best practices and coding standards.

The working group meeting was supposed to take place as part of the Scientific File Format BoF session. Time was however too short to cover both topics without hampering the very interesting discussions. The HDF5 external filters and corresponding working group activities were hence partially covered in the PaNdata workshop and partially postponed to post-RDA video meetings. The slides and workplan for the HDF5 WG will be posted on the RDA-site.

1.3.3 Future HDF5 developments

The recent HDF5 developments have greatly enhanced he file formats capabilities to the benefit of scientific communities as well as developers of scientific instruments or software frameworks. One aim of the RDA and PaNdata activities is to strengthen the support for future sustainable developments. The efforts will be continued and further developed at the RDA 5th plenary in 2015 and beyond.