PaN-data ODI

WP2

First year report on engagement and dissemination activities

Grant Agreement Number	RI-283556
Project Title	PaN-data Open Data Infrastructure
Lead Beneficiary	STFC
Dissemination Level	Public
Nature	Report

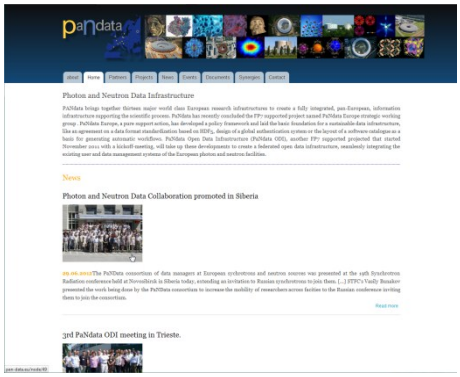
The PaN-data ODI project is partly funded by the European Commission under the 7th Framework Programme, Information Society Technologies, Research Infrastructures.

Work Package 2 (WP2) – Engagement and Dissemination

PaNdata ODI has undertaken a number of activities to engage with related projects and initiatives, has provided extensive input to strategic events and papers, organized a number of workshop as a platform to bring developers, communities and facilities together, and put essential efforts disseminating the results of the project. The activities are briefly outlined here.

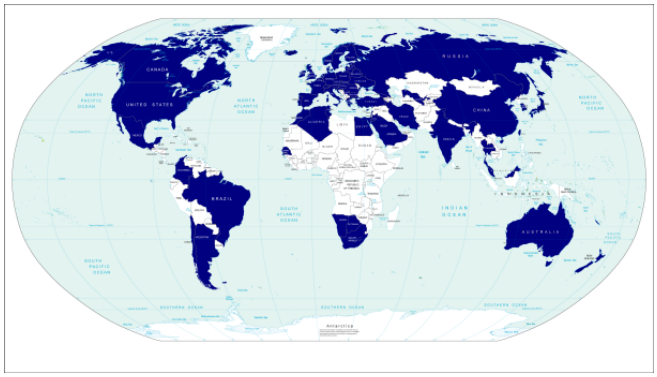
Website:

The first task was the creation of the PaNdata ODI website. The original PaNdata and PaNdata Europe web site was based on a media wiki hosting internal as well as published documents of all kind. With the new ODI project the wiki started to become difficult to maintain and separation of project specific content was poor. To make the web site better organized and more appealing, the PaNdata ODI project has been moved to a Drupal based web-site hosted by STFC. Only internal documents (like drafts of documents to be released after approval) are still posted on the wiki.



The site provides access to all published documents including the deliverables submitted to the Commission. Release of reports and deliverables is always immediate. The web-site is enriched by topical news and events, reporting participation in meetings and workshops as well as the PaNdata ODI organized events. The events are also registered in an event calendar.

Selected documents from PaNdata Europe, like the policies and user survey, with a clear continuation in PaNdata ODI are listed on the site as well. The user survey pursued in 2011 has meanwhile been continued in an enhanced fashion. In 2011 we were only seeking for information on common users or users performing complementary experiments at Neutron and Photon facilities. In the 2012 survey we extended the scope by collecting information on gender, countries of the users home institutions, age and number of visits to the facilities. The survey is still entirely anonymous to prevent any potential conflicts with data protection regulations. Preliminary result of the 2012 user survey can be found under http://www.pan-data.eu/Users2012.



Identity systems:

There are a number of PaNdata ODI goals, which require a high level of harmonization or standardization with projects or communities outside the PaNdata consortium. Dissemination of the core elements of the PaNdata Open Data Infrastructure are hence intensively discussed with and disseminated to the communities and projects. One of these core elements is the pan-European identity system.

PaNdata ODI aims to establish a common, unique identity for all its users. This identity system, the so called Umbrella, is intended to provide additional services to both users and facilities, like a harmonized proposal submission procedure, educational elements or a common facility database. The facility database is a major undertaking, since it requires establishing a common, multi-lingual schema and merging of literally ten thousand of entries. However, once established maintenance of

the database will become much easier and more efficient since the effort is divided through all partners.

Umbrella has been thoroughly tested not only by facility staff but also by a significant number of users providing valuable feedback to the developers. To promote implementation and deployment two teams have been organized, one concentrating more on management issue, the other on in-depth technical aspects. The teams hold regular telephone conferences and meetings. Since the topic is of high interest not only for the RIs organized in PaNdata, a number of additional RIs and projects are participating in the effort, in particular EMBL[1] as the Biostruct-X[2] project leader, members of the Calipso project, GSI/FAIR[3] and the European XFEL[4].

Umbrella has been presented at several meetings and conferences, in particular Biostruct-X and Calipso[5] meetings, the IUCR[6], the series of workshops on Federated Identity Management (FIM) for Research Collaborations[7] initiated by EiroForum[8] and more general events of EUDAT[9] or e-IRG[10].

A prototype implementation and essential documentation is available under https://umbrella.psi.ch (see screenshot above). Recent versions of the ICAT data catalogue are umbrella enabled and tests between some of the facilities have been successfully pursued.

Dissemination and Engagement with related projects and initiatives

The main goal of the PaNdata ODI project is to establish a common, federated, open data infrastructure. This involves a number of issue ranging from federated identity management to meta-data standards or persistent identifier for data, instruments and related publications. Quite a number

[1] European Molecular Biology Laboratory: http://www.embl-hamburg.de
[2] http://www.biostruct-x.eu/
[3] Facility for Antiproton and Ion Research: http://www.fair-center.de/
[4] European X-Ray Free Electron Laser: http://www.xfel.eu
[5] Calipso: Coordinated Access to Lightsources to Promote Standards and Optimization
[6] International Union of Crystallography: http://www.iucr.org/
[7] https://cdsweb.cern.ch/record/1442597
[8] http://www.eiroforum.org/
[9] http://www.eudat.eu/
[10] e-Infrastructure Reflection Group: http://www.e-irg.eu/

of projects and initiatives are working in closely related topics. Seeking for synergies, establishing co-operations and exchanging knowledge is therefore particularly important to arrive at solutions facilitating integration into a global data infrastructure.

Projects like Calipso or Biostruct-X are aim for example to ease the access to the facilities through transnational access support or harmonized proposal submission systems for protein crystallography (PX) beamlines. Both projects consequently need to tackle identity management issues, and the Umbrella also intends to incorporate certain parts of the proposal submission, but not restricted to PX instrument. To arrive at common solution satisfying the needs of the user communities, intense consultations between these projects and PaNdata ODI have been established and Biostruct-X is meanwhile in the Umbrella management team to exploit synergies in this area of common interest.

Federated identity management is also a crucial topic heavily discussed in the FIM workshops. Beyond these activities FIM has been intensely discussed with projects like Eur.XFEL, CRISP[11] and PNI-HDRI[12] or LSDMA[13] and the Umbrella developments are regularly tested and investigated by users from the scientific communities.

Project moonshot[14] is developing tools which permit integration of the Umbrella in non-web based analysis workflows and could provide access to compute resources through the Umbrella credentials. PaNdata ODI is currently investigating the set of tools with Project Moonshot.

Identity management and AAI is also an important topic of the EUDAT project.[15] PaNdata has provided feedback on the requirements of the Photon and Neutron user communities to EUDAT through participation in interviews, workshops and conferences. Likewise, PaNdata has submitted very detailed comments on the e-IRG bluepaper[16] on data infrastructures to e-IRG reflecting the particular view of our user communities.

One of the crucial topics for any data infrastructure is the persistent identification (PID) of digital objects and the citability of the corresponding PIDs. Some PaNdata partners are actively working on this topic in co-operation or at least based on knowledge exchange with projects and initiatives like EUDAT, OpenAire+[17] or DataCite[18]. For example, ILL is building up an infrastructure to assign DOIs to datasets and register them through DataCite.

The newly formed Research Data Alliance (RDA)[19] also aims to establish a number of working groups dealing with PIDs and related meta-data. PaNdata partners are participating in three of the RDA working groups, in particular on PID Information Types and Type registries.

[11] Cluster of Research Infrastructures for Synergies in Physics: http://www.crisp-fp7.eu
[12] Photon and Neutron Infrastructures - High Data Rate Initiative:http://www.pni-hdri.de
[13] Large Scale Datamanagement and Analysis: http://www.helmholtz-lsdma.de
[14] http://www.project-moonshot.org/
[15] http://www.eudat.eu/
[16] http://www.e-irg.eu/publications/blue-papers.html
[17] http://www.openaire.eu/
[18] http://datacite.org/
[19] http://rd-alliance.org/

Standards:

The standardization efforts in PaNdata focus mainly on the data formats, metadata schemes and defined vocabularies. HDF5 has been proposed by the EC to serve as the ISO standard for all binary data. PaNdata has adopted this proposal and selected NeXus as their community wide standard. NeXus is fully HDF5 compliant, but with a standardized meta-data scheme and controlled vocabularies. So NeXus is not really a new data format, but more of a convention how to organize data and meta-data in a HDF5 container, which greatly facilitates exchange and concurrent use of data from different scientific fields or experiments. The NeXus structure enables the automatic processing of the meta-data and ingestion of the data into data catalogues. For example, a NeXus ingestor has been developed for ICAT[20] and the ICAT scheme, which is fully compliant with the Dublin Core standard[21], is currently being adopted to support community requirements and the NeXus data model.

Standardization requires naturally the involvement of many disjoint communities, like developers, vendors, users and IT-expert. PaNdata partners aim to continuously contribute and drive the process. One core platform to engage with the NeXus community is the Nexus International Advisory Committee (NIAC)[22]. Some PaNdata partners have been participating in the NIAC almost since it existence. The PaNdata ODI efforts have led to a stronger and broader representation in the NIAC. At the latest NIAC meeting, H.Bernstein representing imgCIF has joined the NIAC. CIF is another well-established standard widely used in the field of protein crystallography. To achieve interoperability between NeXus and CIF is hence an important issue tackled by the NIAC and the IUCr, and is well progressing[23]. More information can also be found on the IUCr forum on NeXus HDF5 CIF convergence[24].

High Speed data recording is becoming more and more important. Free Electron Laser facilities aiming to resolve processes at a femto-second scale strongly depend on the ability to record images at a MHz rate, but also synchrotrons, FAIR and even some Neutron facilities like ESS face the same problem with emerging new detector generations. To cope with these challenges PaNdata is co-operating with developers and vendors. There is for example an intense co-operation with Dectris[25], one of the leading detector companies, the NIAC and HDF5.org to enable detectors writing NeXus/HDF5 natively at the speed required. A workshop has been organized by PSI to support this process and DESY has developed in co-operation with the PNI-HDRI project an implementation of the NeXus API capable to deal with such data rates. Work on compression algorithms and methods to augment HDF5 with pluggable compression and image filtering modules is on-ongoing.

The recent NeXus developments have been presented at the NIAC and NeXus code camp and the NOBUGS conference in September 2012.

[20] http://www.icatproject.org/
[21] http://dublincore.org/
[22] http://wiki.nexusformat.org/NIAC
[23] http://wiki.nexusformat.org/NIAC2012#Meeting_Minutes
[24] http://forums.iucr.org/viewforum.php?f=31
[25] https://www.dectris.com/

Events:

PaNdata ODI at a whole as well as individual project partners have (co-)organized quite a number of events or participated in and contributed to events of related projects or initiatives. The PaNdata web site lists the majority of such activities, so we present below only a few, selected events with a particular outreach or impact.

In July 2012 the 11th International Conference on Synchrotron Radiation Instrumentation[26] took place in Lyon, by far the largest event of this kind. Rudolf Dimper (ESRF) and Philippe Martinez (Synchrotron SOLEIL) organized a round table on large data volume management[27].

The outcome of the round table discussion were presented[28] at the Workshop on Data Diffraction Deposition[29] (DDD) at ECM27[30], the European Crystallography Meeting in Bergen/Norway as a supplement to the talk by Heinz Weyer presenting PaNdata and related projects[31]. Erica Young and Brian Matthews were completing the PaNdata ODI presentations with a talk on *Linking raw experimental data with scientific workflow and software repository*[32]. The presentations revived the discussion on data deposition and open access to scientific data in the IUCr, and the PaNdata ODI policy framework and the implementation with ICAT at ISIS was featured in the workshop report as "*An exemplar of good practice demonstrating access to raw data is at the ISIS UK neutron source*"[33].

The workshop report lists some actions and recommendations. It suggests in particular "*to encourage and recommend to the IUCr Executive Committee that authors should provide a permanent and prominent link from an article to the raw data sets underpinning a journal publication with a view to making this a formal requirement on authors at such time as the community has adopted raw data deposition as a routine procedure*" and emphasized the "*urgent need to be clear about the metadata required for the various IUCr Commissions and their experimental raw data*".

This is an important step towards an open data infrastructure for one of our particularly prominent user communities. The deposition of scientific data, linked to persistent identifier cited in a

[26] SRI2012: http://www.sri2012.org
[27] Programme: http://www.lepublicsystemepco.com/files/modules/freezones/ProgrammeSRI2012-Web2.pdf
[28] http://www.iucr.org/__data/assets/pdf_file/0008/69479/Data-Mgt-RT.pdf
[29] http://www.iucr.org/resources/data/dddwg/bergen-workshop
[30] http://ecm27.ecanews.org/
[31] http://pan-data.eu/sites/pan-data.eu/files/03-BergenECM272012fb-Weyer.pdf
[32] http://pan-data.eu/sites/pan-data.eu/files/04-Linking-Raw-Data-with-Scientific-Workflow-PanData-ODI-Early-Experience.pdf
[33] http://forums.iucr.org/viewtopic.php?f=21&t=102

publication, standardization of meta-data with the idea to make data re-usable and accessible is exactly the type of infrastructure PaNdata ODI aims to provide.

Another event with a significant impact was the joint PaNdata and PNI-HDRI workshop in Hamburg[34]. The workshop was visited by 52 participants from 14 different facilities and was focused on various technical and non-technical topics related to the data flow from the detector to the archive.

Several of the talks were presenting current developments of workflows for data processing and visualization. The developments were obviously all targeting the problems but in very different way, using different frameworks, programming languages and APIs. The workshop has raised awareness of these developments and led to a better communication about the on-going developments. Consequently, we have recently submitted a topic proposal on integrated analysis frameworks to harmonize the approaches and make frameworks easily interoperable through standardized APIs and modular structure of the integrated framework. The topic proposal also includes facilities like Eur.XFEL, ESS and ANKA, which are currently not participating in PaNdata but recognized the importance of a close co-operation on these kinds of topics.

Another important topic was the archival of data, data curation and standardization of data formats. Standardization of data and meta-data is progressing reasonably well. The implementation of data policies was however intensely discussed. Common concerns from users were the efforts to provide meta-data and the potential misuse of data without giving credit to the original producer. Indeed, citing data and the ability to track citations of data is still an open question under investigation with related projects, user communities and initiatives like RDA. However, some users and beamline providers like EMBL were extremely supportive to the PaNdata data policy and the new database for coherent x-ray imaging[35] is a beautiful example of users promoting open access to scientific data.

Finally, particularly fruitful were presentations on high performance data analysis and hardware developments to enhance data throughput and accessibility of detectors, which initiated more intense collaborations between PaNdata partners and non-PaNdata facilities. A more detailed description and links to slides and supplemental materials can be found in the workshop summary[36].

[34] https://indico.desy.de/conferenceDisplay.py?confId=5517
[35] http://cxidb.org/
[36] http://pan-data.eu/sites/pan-data.eu/files/hdri-pandata.workshop-2012.summary.pdf

The third important event was NOBUGS 2012[37] which was organized by STFC and DLS and took place at RAL Sept. 2012. The NOBUGS conference was accompanied by a number of PaNdata satellites, like Umbrella and ICAT workshops as well as a NeXus Code camp and NIAC meeting. PaNdata work was presented in several of the NOBUGS talk like the data management reports from DESY[38] and ALBA[39], a status report for NeXus[40], the data catalogue requirements analysis[41] and the practice of data citation at ISIS[42]. Several of the talks were again focusing on data analysis frameworks and workflows further emphasizing the need for an integrated data analysis framework. More information about the satellite workshop can be found on the PaNdata web and the NOBUS2012 site.

Co-operations with industry

As mentioned above we are seeking co-operations or exchange of technical knowledge with industrial partners. The focus is here the promotion of the standard data format and the improvement of system for high-speed data collection and recording.

The co-operation with a number of PaNdata partners, in particular DESY, STFC and PSI, and Dectris and the HDFgroup has already led to some results. Dectris is ready to enable their next-generation detectors to natively write NeXus/HDF5. To achieve the desired performance/bandwidth some adjustment to NeXus/HDF5 are required. The adjustments of the NeXus header is discussed and supported by the NIAC. Further acceleration of the HDF5 API is an on-going process between PaNdata partners and the HDFgroup. The HDFgroup will presumably receive funds to implement the desired accelerators, which would be beneficial not only for the PaNdata collaboration and detector vendors, but also for the entire HDF5 user community.

Another topic are low level co-operations with some storage system providers. DESY for example is testing new platforms from Netapp and IBM. The API, in particular NFS 4.1 are thoroughly tested against the current needs of the Photon and Neutron facilities, while keeping an idea on future requirements originating for example from Eur.XFEL, ESS[43] or SKA[44] [45]. Similar activities are pursued at other facilities also testing products like dCache[46], pvfs[47], Lustre[48], fhgfs[49] or hadoop[50]. Although

[37] http://nobugs2012.org/
[38] T Kracht , T Nunez, A Rothkirch, E Wintersberger: Experiment Control and Data Acquisition at PETRA III, DESY
[39] C Pascual-Izarra et al: Data management for the Beamlines at Alba
[40] M Koennecke et al: The State of NeXus
[41] M Prica: Requirements for data catalogues within facilities
[42] MD Wilson, BM Matthews, S Nagella, AJ Wilson: Using DataCite DOIs for ISIS Neutron Source Data
[43] http://ess-scandinavia.eu/
[44] http://www.ska.ac.za/
[45] http://www.ska.gov.au/
[46] http://www.dcache.org/
[47] http://www.pvfs.org/
[48] http://wiki.lustre.org/index.php/Main_Page
[49] http://www.fhgfs.com/cms/

these activities are mostly bilateral, experiences and knowledge are exchanged between all partners and the vendors hopefully accelerating the process to arrive at stable high performance systems tailored for the needs of current and future data infrastructures.

[50] http://hadoop.apache.org/