



PaN-data ODI

D2.4

Open Source software distribution procedure

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.

Table of contents

Description of D2.4	3
PaNdata Software catalogue as the strategic distribution platform	3
Access to the software catalogue	4
Deployment and version control	4
Debian Repository DebianPAN	4
Loose ends	5
Identity Management Tools	6
Data Management Tools	7
ICAT	7
Deployment status	7
Roadmap	8
Service verification	8
Software Releases	9
ICAT – Quick Facts	9
NeXus libraries	10
The NeXus PNI libraries	10
NexDaTaS - The NeXus data collector	11
HDF5 extensions	11
Data Processing Tools	12
DawnScience	12
Mantid	12

Description of D2.4

The task associated with D2.4 was the *provisioning of the open source software and appropriate documentation to potential partner bodies*. The description of work in the PaNdata ODI proposal does not outline a specific procedure or name particular software products to be deployed. This was a conscious choice since PaNdata ODIs open source developments are largely embedded into collaborations and contributions to (external) open source software products. The software distribution procedure therefore needs to comply with policies of a variety of different software packages and software deployment mechanisms.

This deliverable will describe the platforms to commonly document and summarize the software developments and briefly outline the particular procedure for a number of important tools.

PaNdata Software catalogue as the strategic distribution platform

PaNdata ODI has developed a software catalogue, which collects and aggregates information on all software products relevant for experiments at photon and neutron sources and the subsequent data process and analysis. The software catalogue is openly accessible under http://software.pandata.eu/.



The software catalogue has been populated at an early stage with the most important software products, which were identified through a survey at the facilities. Contributions were mostly coming from beamline scientists. The beamline scientists are usually intensively involved in data processing and analysis and could hence identify software packages essential for running an experiment as well as the subsequent data analysis.

Software packages can be identified through experimental techniques, instruments, associated facilities and other tags. The tags are not yet derived from a common, unified dictionary, which is currently still under development and hence need to be further refined at a later stage.

Access to the software catalogue

The software catalogue has recently been enhanced to support the UmbrellalD.org identity management system. UmbrellalD authentication is fully implemented and permits to do any task related to software products like additions, annotations, commenting or evaluation of products. The software catalogue is hence open for anyone interested.

Additions to the data catalogue are automatically announced through twitter, so that it becomes easy to keep track of new or updated software products. Queries and discovery of software products is also possible through web-services documented under https://software.pan-data.eu/help/web-services.

Further additions and enhancements to the catalogue are currently being investigated, like for example by registration of primary citations and discovery through DOIs.

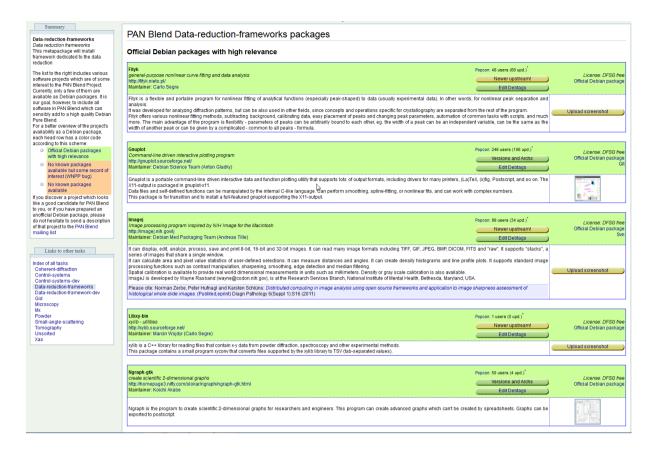
Due to the characteristics of the software catalogue it is the most suitable platform to register software packages developed by or of particular relevance for PaNdata. PaNdata related software packages will be tagged accordingly and become easily discoverable through a PaNdata category which will be implemented very soon.

Deployment and version control

Most of developments like NeXus implementation are not mission critical and are hosted on git repositories, which permits full version control and easy deployment. The ICAT data catalogue software as well as the Umbrella developments cannot easily be updated without careful verification of service compliance. In these cases, mechanisms have been put in place to ensure continuous availability of services. The procedures are outlined below.

Debian Repository DebianPAN

The Software catalogue is complemented by the <u>DebianPAN</u> Debian Repository. DebianPAN is a Photon And Neutron dedicated Debian Pure Blend. The objective of this Pure Blend is to package software used to perform and analyse data collected in Synchrotron radiation or Neutron Facilities. This blend is not only reserved to peoples working on this kind of facilities, but also to smaller laboratories where diffraction experiments can be performed, or it can potentially interest also all users of these facilities when they need to analyze the collected data.



2 Snapshot of the DebianPAN repository

Loose ends

Any developments or materials that won't fit into a software catalogue (like for example ICAT service verification deployment packages) will be hosted on https://code.google.com/p/pandata/.

Identity Management Tools

PaNdata ODI is developing an identity management system (IdM) named Umbrella providing a unique and persistent identity to the photon and neutron user communities. The system got meanwhile in production under http://umbrellalD.org.

UmbrellaID.org is a fairly complex system of identity providers (IdP), service providers (SP) and directory services (DS) developed by PaNdata ODI. Certain aspects of the system like the development of bridging mechanisms to enable integration of other identities are developed in cooperation with <u>CRISP</u> and/or GN3+¹.

The IdP is replicated at several facilities. The user-database is implemented as an OpenDJ instance and automatically synchronized between IdP replicas. SPs and user requests are directed to the nearest IdP instance by a geodns location services. Availability of instances is monitored and the geodns service automatically configured accordingly. The configuration provides a highly distributed, redundant authentication system, and an easy integration into site-specific authorization infrastructures. However, it requires well-coordinated actions to keep the instances synchronized whenever updates of the software, the web-layout, the IdP or SP metadata are being deployed.

To warrant continuous availability of the service, any updates are being introduced by the umbrella development team, further tested and verified by the technical team and subsequently authorized by the management team or steering committee. The exact details and procedures are currently being described as part of the Umbrella MoU.

The IdP software itself, SP configurations, cookbooks for SP integration and documentations are available from github and the UmbrellaID.org web-site. A concise overview can be found under http://pan-data.eu/Umbrella.

Umbrella - Quick Facts

Web-Site (IdP): umbrellaID.org

Contact: contact@umbrellaid.org

Developed by: PaNdata ODI. Parts in co-operation with CRISP and GN3+

In Production: ESRF, ILL, PSI, (DESY)

Frontend repository: https://github.com/flowedback/UmbrellaFrontend

Documentation: https://umbrellaid.org/euu/

Umbrella cookbook: https://umbrellaid.org/euu/help/help.htm

Overview: http://pan-data.eu/Umbrella

License: open source

¹ An outline of the project can be found under https://portal.nordu.net/download/attachments/34571610/GN3plus+Intro.pdf. Short version of the Gn3+ announcement under https://pan-data.eu/node/72

Data Management Tools

PaNdata ODI aims to standardize data formats, metadata schemata and data catalogues as much as required to establish a highly efficient, federated data infrastructure. Correspondingly, there are three areas of developments

- HDF5 as the basis for fast data recording and analysis
- NeXus as a standard (meta-) data format building on top of HDF5
- ICAT as a standard data catalogue with a NeXus compliant metadata schema

ICAT

ICAT is an open source meta-data management system designed for large facilities. ICAT development is a collaboration involving ISIS, the Scientific Computing Department of STFC, ILL and the Diamond Light Source. PaNdata WP4 is aiming to provide use cases and requirements for the ICAT development to produce a product which meets the functionality of the proposed virtual labs. Important tasks of PaNdata WP4 include the deployment of ICAT and service verification; this work is progressing well. Currently, ICAT is in production at ILL, ISIS and DLS and several labs (e.g. ALBA, DESY, ELETTRA and ESRF) have deployed ICAT prototype instances, which have been tested during service

An important consideration in a meta-data catalogue is the meta-data schema. To satisfy PaNdata's needs, a number of modifications have been proposed and implemented in ICAT. The schema is closely related to the standardized meta-data schema implemented in NeXus. NeXus provides a complete description of the experimental conditions during an experiment, and covers essentially all of the data required for meaningful searches in a meta-data catalogue. Ingestion of NeXus into ICAT is straightforward; the ICATs in production at ILL, ISIS and Diamond ingest thousands of NeXus files every

ICAT has many attractive features, such as the programmable web service interface, the integration of provenance information and the registration of Data Object Identifiers. The developments at the moment include the ICAT Data Service (IDS) and a rich collection of authentication mechanisms.

ICAT is an open source metadata management system designed for large facilities. ICAT is developed as a collaborative project between ISIS, the STFC eScience department, ILL, Diamond Lightsource as well as the CLF. PaNdata WP4 is in particular aiming to provide use cases and requirements for the ICAT development to derive at a product perfectly suitable to fulfill the functionality of the proposed virtual labs. An important task of PaNdata WP4 is the service verification and deployment of ICAT, which is progressing very well. Currently, ICAT is in production at ILL, ISIS and DLS and several labs (e.g. ALBA, DESY, ELETTRA, ESRF) are operating ICAT prototype instances to verify services and deployment.

Deployment status

Currently, nine of the participating facilities operate an ICAT instance. A quick overview can be found on the <u>latest ICAT service verification</u>. Currently, ICAT is in production at ISIS, ILL and DLS (at some sites not involved in PaNData ODI). An instance at DESY will be in production in fall 2013. Prototype instances of ICAT are operational at ELETTRA, ESRF, HZB, JCNS and SOLEIL. STFC e-science in addition operates a number of virtual ICAT instances to demonstrate the federation of ICATs via TopCat and the federated search across catalogues.

Roadmap

A release manager has been assigned for PaNdata ODI to

- write a plan for the release management;
- be the main contact between the work packages and the test and the meta-data service work package;
- define and deliver a schedule of test events involving most of the partners of Pandata;
- coordinate the test events;
- ensure that the materials provided for test are in a testable state;
- provide feedback to the work packages on their materials;
- define the contents of a release;
- ensure that software releases are coherent and appear in a timely and reasonable form.

A detailed account of these efforts can be found here.

Service verification

The schedule for the service verification is in the following table. A later draft of this plan will contain further refinement of the contents of the service verification. Each service verification has detailed instructions and the report and can be found below. The service status will continuously be monitored on http://pandata.org/sls/.

SV	Date	Major topics	
4	10 May	Use of the bigv federation to verify the download manager. Use of the bigv authentication service.	
5	21 June	Initial test of the development federation. Verification of an download service. Verification of an ingest service.	
6	October	Initial test of the production service. Integration of materials from WP 5, 6, 7, 8.	
7	November	Verification of the Umbrella authentication service. Integration of materials from WP 5, 6, 7, 8.	
8	December	Verification of the production service.	
9	February	To be defined.	
10	March	To be defined.	

Note: These releases will not be packaged as a single package. The release will be a stack of components from a variety of sources. We will provide information in the service verification report of the versions and their performance.

Software Releases

Release	Date	Contents	Verification
0.0	May	ICAT, authn_db, authn_ldap, DM, Topcat. Low grade authentication service using Facebook.	<u>SV[4]</u>
0.1	June	As 0.0, plus development federation	<u>SV[5]</u>
0.2	October	As 0.1, plus materials from WP 5, 7 and 8	SV[6]
0.3	November	As 0.2, plus integration with Umbrella	SV[7]
0.4	December	As 0.3; A technology preview release	SV[8]
0.5	February	Release candidate	SV[9]
1.0	March	Production release	

ICAT - Quick Facts

Web-Site (IdP):
http://www.icatproject.org/

Developed by: ISIS/STFC, ILL, DLS

In Production: DLS, ILL, ISIS

Deployed at: ALBA, DESY, ELETTRA, ESRF, HZB, JCNS, PSI, SOLEIL, SLS

Source repository: http://code.google.com/p/icatproject/

PaNdata ODI WP4: http://pandatawp4.wordpress.com/

Documentation: for the → current release + release notes

ICAT installation, ICAT schema, ICAT java API

Demonstrators: WSDL API

glassfish online console

oracle online console

TopCat

Overview: http://pan-data.eu/ICAT

License: BSD 3

Note: The PaNdata specific parts like Demonstrators are currently being migrated to pandata.org to separate more clearly between PaNdataODI and the ICAT project.

NeXus libraries

NeXus is a common data format for neutron, x-ray and muon science. NeXus is developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data. NeXus itself builds on top of HDF5, which is by itself a widely adopted, standardized data format and has been proposed by the European Commission as an ISO standard for all binary data. Hence, any NeXus file is a fully valid HDF5 file, which can be read by a large number of applications without any further modification.

Two open source software packages have been developed in collaboration with the <u>PNI-HDRI project</u> and with support by the FP7/EU funding of PaNdata ODI, namely a complete re-rewrite of NeXus API in C++ and a NeXus data collector, which interfaces between instrument controls and the NeXus libraries, to automatically aggregate all relevant information into NeXus files.

The NeXus PNI libraries

The PNI libraries are a stack of related C++ libraries developed with the intention to simplify the development of scientific software in the field of Photon-, Neutron, and Ion-scattering. A utility library has been developed providing all kinds of data types and structures essential for scientific software developments. This library recently became the core of the PNI library stack.

Most of the code is written in C++. However, python-bindings exist to those C++ libraries. The following libraries are currently available

- **libpnicore** this is the central library providing data types and fundamental data structures like arrays and buffers
- **libpniio** provides functionality for data IO. This includes readers for several data types and the original Nexus API.
- **python-pniio** a Python binding to libpniio.

Documentation and open source code can be obtained from the project web under http://code.google.com/p/pni-libraries/wiki/Introduction. Debian packages can be found in the PNI-HDRI repository.

NeXus PNI libraries – Quick facts

Web-Site (IdP): http://code.google.com/p/pni-libraries/wiki

Developed by: PNI-HDRI in co-operation PaNdata ODI

Source repository: http://code.google.com/p/pni-libraries/downloads/list

Debain repository: http://repos.pni-hdri.de/apt/debian/

Documentation: http://code.google.com/p/pni-libraries/wiki

Overview: http://pan-data.eu/NeXus

License: GNU GPL v2

NexDaTaS - The NeXus data collector

All operations carried out on a beamline are orchestrated by the control client (CC), a software application operated by the beamline-scientist and/or a user.

With the appearance of complex data formats like Nexus the IO code for control clients became more complex. To cope with this complexity, NexDaTaS has been developed jointly by PNI-HDRI and PaNdata to provide an easy to use interface between the NeXus data integration and the control system. NexDaTaS is realized as a Tango server which allows storing NeXuS Data in H5 files. The server provides storing data from other Tango devices, various databases as well as passed by a user client via JSON strings. A detailed description can be found on the project page.

This NexDaTaS repository contains

- Nexus Data Writer implemented as a Tango server
- Configuration Tool written in PyQt , NDTS Component Designer
- Configuration Server implemented as a Tango server on a MySQL database
- Simple examples of configuration files and control clients

NexDaTaS - Quick Facts

Web-Site (IdP):
http://code.google.com/p/nexdatas/

Developed by: PaNdata ODI in co-operation with PNI-HDRI

Source repository: https://code.google.com/p/nexdatas/source/checkout

Documentation: http://code.google.com/p/nexdatas/w/list

Overview: http://pan-data.eu/NeXus

License: GNU GPL v2

HDF5 extensions

The NeXus developments and in particular the co-operation with detector manufacturers like Dectris have produced a set of additional requirements on HDF5 implementations. Two aspects were particularly important: the implementation of external filter and single-writer-multiple-reader (swrm) capabilities. External filter would for example permit to build trigger, veto or parallelized compression into the data recording flow, which would help tremendously to reduce data volumes and accelerate data recording, essential for next-generation x-ray detectors. Similarly, the swrm-feature would greatly enhance the throughput for real-time data processing and visualization.

The external filter capabilities have meanwhile been implemented by HDFgroup and are part of recent HDF5 distributions, for the benefit of the entire HDF5-user communities. The development has been funded by PaNdata partners. The swrm-development is currently under discussion and might be funded by PaNdata partners as well.

Data Processing Tools

Data processing is a crucial and very often he time-limiting step in the whole scientific process. To fully benefit from the PaNdata ODI data infrastructure, the seamless integration of ICAT and support for NeXus and HDF5 are important. The development of data analysis frameworks is not a PaNdata ODI task. There are however two software frameworks which play an important role in the Virtual Lab demonstration, namely DawnScience and Mantid. DawnScience is an open source framework predominantly developed by members of the PaNdata consortium. Mantid development is driven by SNS/ORNL in co-operation with ISIS/STFC and other partners mostly active in neutron sciences. The active participation of PaNdata partners in the development of these two data analysis frameworks guarantees the alignment of the basic requirements with the project goals. Consequently, both frameworks provide a strong integration with ICAT and NeXus/HDF5.

DawnScience

Data Analysis WorkbeNch (DAWN) is an eclipse based workbench for doing scientific data analysis. It implements sophisticated support for visualization of data in 1D, 2D and 3D; Python script development, debugging and execution; Workflows for analyzing scientific data calling Python and binary codes.

DAWN is not restricted to one scientific domain. It is developed by and for the synchrotron community foremost but has strong overlap with other communities like neutron scattering, photon science and any scientific communities with the above or similar needs.

DAWN is an open source software built on the Eclipse/RCP platform in order to scale to address a wide range of applications and to benefit from the workbench and advanced plugin system implemented in Eclipse. Due to standardization of the plugin mechanism of eclipse, the framework is highly modular and easily extendable. There exists already a rich set of plugins ranging from an ICAT interface to a TANGO-module covering substantial parts of the scientific workflow.

DAWN - Quick Facts

Web-Site: http://www.dawnsci.org/home

Developed by: DLS, ESRF, EMBL

Source repository: http://www.github.com/DawnScience

Documentation: http://www.dawnsci.org/documentation

Overview: http://pan-data.eu/Dawn

License: Eclipse Public License - v 1.0

Mantid

The Mantid project provides a framework that supports high-performance computing and visualisation of scientific data. Mantid has been created to manipulate and analyse Neutron and Muon scattering data, but could be applied to many other techniques. The Mantid framework provides support for visualization and data transformation. It allows the user to implement their own tailored analysis algorithms, and maintains full virtual instrument geometry. Various input and

output file formats are supported including NeXus files and loading direct from an data acquisition electronics.

Mantid can be easily extended using a plug-in mechanism to add further user-defined algorithms, data structures and file formats. It has been developed using object-oriented C++ for execution performance and portability and makes use of several publicly available libraries such as BOOST and GSL. Care has been taken to isolate as much complexity as possible from the areas that may be extended by users, while providing them with powerful objects to manipulate their data.

MantidPlot, a generic visualization and algorithm execution application, is the first graphical user interface that has been developed to use Mantid and also provides Python scripting capabilities. Mantid itself currently provides C++ and Python APIs.

Mantid - Quick Facts

Web-Site: http://www.mantidproject.org

Developed by: SNS/ORNL, ISIS/STFC, Tesella, ILL, Los Alamos, McStas

Source repository: https://github.com/mantidproject

Documentation: http://www.mantidproject.org/Documentation

Overview: http://pan-data.eu/Mantid

License: GNU GPL v3