

# PaN-data ODI

## Deliverable D5.2

### Report on the deployment of the specification of the three virtual laboratories

Grant Agreement Number	RI-283556
Project Title	PaN-data Open Data Infrastructure
Title of Deliverable	Deployment of specification of the three virtual laboratories
Deliverable Number	D5.2
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	31 Mar. 2013
Actual Delivery Date	7. June 2013

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

**Abstract**

The first deliverable of WP5, released in April 2012, was a list of specific requirements for the other work packages. This deliverable reviews the current status of the implementation and deployment of the elements from the other work packages from the viewpoint of WP5 and discusses how these should be taken forward to enable the three Virtual Laboratories.

**Keyword list**

PaN-data ODI, Virtual laboratories

**Document approval**

Approved for submission to EC by all partners on 7.06.2013

**Revision history**

Issue	Author(s)	Date	Description
1.0	Thorsten Kracht	8 Mar 2013	First version
2.0	Frank Schlünzen	10 Mar 2013	Additions
3.0	All partners	5 June 2013	ICAT deployment status

## Table of contents

	Page
<b>1 INTRODUCTION.....</b>	<b>4</b>
1.1 OVERVIEW OF THE DELIVERABLE .....	4
<b>2 THE COMMON AUTHENTICATION SYSTEM .....</b>	<b>6</b>
<b>3 THE METADATA CATALOGUE.....</b>	<b>10</b>
3.1 OVERVIEW OF THE ICAT DEPLOYMENT AT THE PAN-DATA PARTNERS .....	12
3.1.1 ALBA.....	12
3.1.2 DESY .....	12
3.1.3 DLS .....	14
3.1.4 ELETTRA.....	14
3.1.5 HZB.....	14
3.1.6 ILL .....	14
3.1.7 ISIS .....	15
3.1.8 SOLEIL.....	15
3.2 SUMMARY.....	15
<b>4 THE STANDARD DATA FORMAT .....</b>	<b>16</b>
4.1 THE NEXUS LIBRARY.....	16
4.2 NEXUS METADATA .....	17
4.3 THE NEXUS WRITER .....	18
4.4 THE COMPONENT DESIGNER .....	21
4.5 THE CDMA STATUS.....	21
4.6 SUMMARY.....	22
<b>5 SCALABILITY .....</b>	<b>23</b>
<b>6 SOFTWARE CATALOGUE .....</b>	<b>24</b>
<b>7 WORKFLOW MANAGEMENT SYSTEMS.....</b>	<b>26</b>
<b>8 INTEGRATION .....</b>	<b>29</b>
8.1 STANDARD DATA FORMAT .....	29
8.2 THE METADATA CATALOGUE.....	30
8.3 THE COMMON AUTHENTICATION SYSTEM .....	31
8.4 SUMMARY.....	31

# 1 Introduction

The PaN-data ODI project creates an Open Data Infrastructure among the participating institutes. The goal of work package 5 is to demonstrate the results of various Pan-data ODI activities by setting up three virtual laboratories. They are displayed in the following list together with the involved tasks:

- VL1: Structural ‘joint refinement’ against x-ray and neutron powder diffraction data
  - Raw data searched for by an authenticated user through the data catalogues.
  - Access is authorized and data downloaded from facility archives.
  - Relevant analysis software searched for in software database.
  - Software downloaded and run locally or at facility.
  - Analysis carried out.
  - Results (refined structure) and any relevant reduced data uploaded to facility archives.
- VL2: Simultaneous analysis of SAXS and SANS data for large scale structures
  - Raw data searched for by an authenticated user through the data catalogues.
  - Access is authorized and data downloaded from facility archives.
  - Relevant analysis software searched for in software database.
  - Software downloaded and run locally or at facility.
  - Analysis carried out.
  - Results (modeled structure) and any relevant reduced data uploaded to facility archives.
- VL3: Access to tomography data exemplified through paleontological samples
  - Setup a public access database for storing tomographic raw and processed data of paleontological data e.g. 2D tomographs and 3D processed images.
  - Provide authorized access from multiple institutes to store processed data in the database.
  - Enable public access to data in database.
  - Implement long term archiving of database.

The scope of VL3 is currently being reconsidered. It is intended to widen it to include not only data from paleontological samples but also other biological specimen. This reorientation would increase the general acceptance of VL3 because the catalogue could serve educational purposes, provided that the data are well prepared and that a portal is created as a user interface.

## 1.1 Overview of the Deliverable

The first deliverable of WP5 was a list of specific requirements for the other work packages. It was released in April 2012. This deliverable describes the current status of the implementation and deployment of the elements from the other work packages enabling the virtual laboratories. Section 2 discusses the common authentication system which will al-

low users to access various related services, like digital user offices, metadata catalogues, portals and the software database, using a single-sign-on mechanism. The next section deals with the issues related to the metadata catalogue. Implementation details are exemplified for DESY, the institute that leads WP5. The situation at the other PaN-data partners is summarized. Section 4 elaborates on the standard data format by introducing the features of a newly created NeXus library and by describing the metadata treatment. This involves the configuration of discipline-specific metadata models and how metadata are captured during a measurement. To cope with the output rates of modern 2D detectors is an issue which is relevant also for the virtual laboratories. Section 5 explains how this issue is addressed by parallelizing I/O. The virtual laboratories process data with applications that are commonly available. A central software database, which is covered in section 6, facilitates the search for suitable applications. Experiments in the field of powder diffraction and small angle scattering generate data volumes that can only be processed by automated procedures. Workflow systems, as discussed in section 7, are suitable tools to be employed. This deliverable ends with a summary of the current status of the integration of the virtual laboratories. The summary discusses also the components that will be achieved during the remaining months of the project.

## 2 The Common Authentication System

The PaN-data partners decided to use Umbrella<sup>1</sup> as a common identity management (IdM) and authentication system. The development of the Umbrella was initiated at the PSI as an activity within the EuroFEL<sup>2</sup> project. PaN-data ODI (WP3) has taken up the initial developments and aims to implement and deploy it at the participating facilities. A series of tests with the involvement of the user communities have been successfully concluded.

The ESFRI cluster project CRISP<sup>3</sup> is currently developing enhancements of the Umbrella system to permit subsequent integration of other identity systems. This activity as well as related activities in projects like NMI3, CALIPSO and Biostruct-X are closely co-ordinated with the PaN-data-ODI WP3 efforts.

The development of a Memorandum of Understanding (MoU) between partners involved in PaN-data and CRISP has been started recently to guarantee a sustainable joint operation of the IdM system. It was agreed that PaN-data/WP3 concentrates on implementation issues and CRISP/WP16 focusses on further developments. This year Umbrella will be deployed at two groups of institutes. ESRF, ILL and PSI will be first, HZB, ALBA and DESY will follow. Since the authentication service is a critical component for the Digital User Offices, the roll-out will be divided into four phases. The functionality of the system will be gradually increased from phase to phase to guarantee the availability of the user offices.

Once Umbrella is generally deployed, users will be able to access resources like digital user offices, portals, data catalogues and software databases with a single-sign-on method based on Shibboleth/SAML2. Original prototypes and tests were based on a single identity provider (IdP) at [umbrella.psi.ch](http://umbrella.psi.ch) and various service providers (SPs) at different facilities. User information was provided by a single directory service.

The applicability of the Umbrella services was demonstrated during the ‘friendly users’ phase. DESY, Diamond (ICAT service, Moonshot), ESRF, and PSI participated in this test phase. As an example the DESY test will be described here: DESY participated with an Umbrella-compliant instance of its digital user office DOOR (Fig. 1). The single-sign-on functionality could be verified, i.e. users with valid Umbrella credentials were able to connect to DOOR without supplying username and password. A mechanism was setup to match Umbrella and DOOR accounts.

Integration of the Umbrella with the data catalogue has been demonstrated as part of the TopCAT development. TopCAT can act as a proxy storing credentials for several data catalogues and thereby supporting cross-site metadata searches. TopCAT will hence allow demonstrating the authentication based tasks even if the Umbrella is not fully productive yet. Similarly, analysis frameworks like DawnScience<sup>4</sup> or Mantid<sup>5</sup> are currently made Umbrella-aware to support direct data mining and analysis from distributed ICAT instances.

---

<sup>1</sup> [Umbrella.psi.ch/euu](http://Umbrella.psi.ch/euu)

<sup>2</sup> [www.iruvx.eu](http://www.iruvx.eu)

<sup>3</sup> [www.crisp-fp7.eu](http://www.crisp-fp7.eu)

<sup>4</sup> [www.dawnsci.org/](http://www.dawnsci.org/)

<sup>5</sup> [www.mantidproject.org/Main\\_Page](http://www.mantidproject.org/Main_Page)

**DOOR** DESY ONLINE OFFICE FOR RESEARCH WITH PHOTONS  
**Umbrella prototype** HASYLAB

HASYLAB | DORIS III | PETRA III | FLASH | DESY

**DOOR Home**  
 \* Contact  
 \* New User  
 \* Lost Password  
 \* Registered User

Next Deadlines	
Proposal Submission PETRA III	06-Sep-2012
Proposal Submission FLASH	05-Sep-2012
Beamtime Application PETRA III	31-Oct-2013
Beamtime Application FLASH	28-Feb-2013
Annual Report	01-Mar-2013
News	Was fuer ein schoener Tag! Ja, und morgen wird noch schoener!

Welcome to DOOR which is based on DUO from PSI.  
**This is a DOOR test system for the EAA/Umbrella project. It operates on a test database.**  
 Please note that you need to register in order to use this system to submit research proposals and apply for beamtime. Please do not hesitate to [contact us](#) in case you have further questions.

**Registered DOOR user**  
 Log in using your DOOR user name and password or your Umbrella (EAA) credentials.

**Forgotten password**  
 If you do not remember your DOOR user name and/or password, your login information will be sent to your previously registered e-mail address.

**New DOOR user**  
 To obtain a DOOR user name and password, please register [here](#).  
 Users with an **existing Umbrella (EAA) account** might first log on at Umrella [here](#).  
 Or you might **set up an Umbrella (EAA) account** before registering [here](#).

**Imprint**  
 DESY Imprint

Please [contact us](#) if you encounter problems using DOOR.

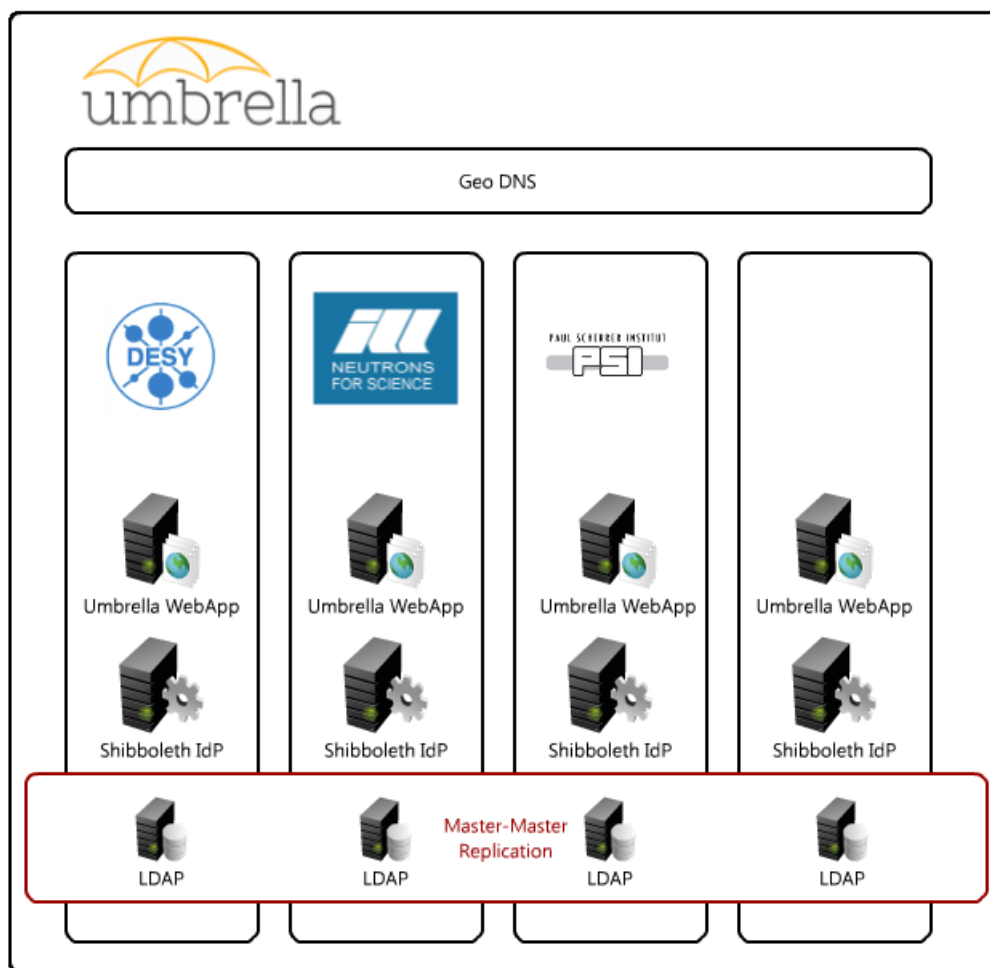
e-infrastructure | pan-data | HELMHOLTZ GEMEINSCHAFT | National Institute of Standards and Technology | WiscGrid

Fig. 1: The Umbrella-compliant Digital User Office at DESY

Another WP3 objective is a pan-European affiliation database. Scientists as well as the facilities will be able to maintain their affiliation via a federated service allowing digital user offices to keep their user information up-to-date. A prototype of the affiliation database will soon be deployed at the ESRF.

Since it is preferential to realize Umbrella as a geographically distributed solution, the Umbrella team started to implement and test additional features:

- A distributed system of identity providers will be installed with a master-master replication of the LDAP user information to keep the servers up-to-date (Fig. 2). The implementation uses a distributed OpenDJ system which allows a straight-forward realization of the master-master replication. First tests have been successfully performed between PSI, ILL and DESY.
- It is intended to use a Geo DNS service to locate the nearest identity provider. Currently the different Geo DNS service providers are being evaluated.



**Fig. 2: Umbrella with Distributed Identity Providers (B. Abt, PSI)**

In conclusion: the results of the Umbrella tests clearly indicate that this single-sign-on system is ready to support the collaborative work among the PaN-data partners. Umbrella will presumably become productive at PSI, ILL and ESRF within a few months. DESY, HZB, DLS and ALBA will follow in fall 2013. The comprehensive Umbrella website (Fig. 3 The Umbrella incorporates all necessary information for users and technical supporters.

[Help](#)[About us](#) | [About the Umbrella](#) | [Services](#) | [User Offices](#)**ACTIONS**[Create Account](#)[Contact](#)[Login](#)**INDEX**

Welcome to the Umbrella. Please proceed with a following action:

- [Create new Account](#)
- [Login](#)



**Fig. 3 The Umbrella Home Page**

### 3 The Metadata Catalogue

The PaN-data partners agreed on ICAT<sup>6</sup> as a common metadata catalogue. ICAT was created by DIAMOND and ISIS and is now supported by an international community (ORNL, ILL, ESRF, ELETTRA). The comprehensive project home page (Fig. 4) fosters the collaborative work. It hosts documentation, source code, discussion groups and other related information.

**ICAT icatproject**  
ICAT metadata catalogue

Project Home | Wiki | Issues | Source

Summary | People

**Project Information**

Recommend this on Google+

Starred by 10 users  
[Project feeds](#)

**Code license**  
[New BSD License](#)

**Labels**  
ICAT, STFC, Neutrons, ISIS, Muons, Nexus, X-rays, Diamond, SNS, ANSTO, ScientificData, APIs, Java

**Members**  
[tom.griff...@stfc.ac.uk](#),  
[crcami...@gmail.com](#),  
[noris.ny...@stfc.ac.uk](#)  
34 committers

**Featured**

**Wiki pages**  
[Contribindex](#),  
[IcatRequirements](#),  
[MeetingActionsRecord](#),  
[Sweden2013](#),  
[Show all »](#)

**Links**

**External links**  
[ICAT Website](#),  
[TopCAT Code](#),  
[ICAT Data Server](#),  
[PaNdata](#),  
[PaNdataWP4](#)

**Groups**  
[ICAT Discussion Group](#),  
[ICAT Developers](#),  
[ICAT Announcements](#),  
[ICAT Support](#)

The project has a web site at: <http://www.icatproject.org>

If you wish to download a version of ICAT, then the materials are available from the project web site at <http://www.icatproject.org>. The material on the project web site has been compiled and packaged for deployment.

If you wish to obtain your own copy of the source code, then you should download it using SVN from this google code web site.

To browse the ICAT source code, visit the [Source](#) tab.

If you have isolated a problem or want a new feature to be included in ICAT, please [submit an issue](#).

Make sure to include all the relevant information when you submit the issue such as:

- ICAT version;
- One line of issue summary and a detailed description;
- And any workarounds if you have them.

The more information you provide, the quicker the issue can be verified and prioritized. A test case that demonstrates the problem is greatly preferred.

The ICAT code includes libraries under various compatible open-source license. A full list is available [here](#).

- The ICAT Development Team

Fig. 4: The ICAT Project Home Page

The data catalogue records meta-information about digital objects stored in some appliances, independent of the actual realization of the storage system. The primary purpose of the data catalogue is data discovery, typically providing all metadata and physical storage location of a set of digital object matching certain search criteria. For Photon and Neutron science applications a digital object would usually correspond to a dataset, but could also relate to an arbitrary collection of datasets or a single file. Utilizing NeXus as the primary storage format, a single file will in most cases be an entire dataset in a composite container. ICAT supports the ingestion of NeXus data through appropriate open source tools,

<sup>6</sup> [www.icatproject.org](http://www.icatproject.org)

which are part of the NeXus toolkit. ISIS has so far ingested more than 320.000 objects into their ICAT instance with the NeXus ingestor. Data discovery and potential re-use requires a careful and complete annotation of the digital object. ICAT permits to record all information about an experiment in a standardized and NeXus compliant schema, and can hence represent the entire process from a measurement to a subsequent publication. ICAT offers a Java API and a WSDL interface to communicate with the catalogue.

The client TopCAT is part of the distribution. It supports parallel file searches across facility borders.

ICAT is in regular operation at ISIS and ILL. About half of the other PaN-data partners installed ICAT instances by now. The ICAT team scheduled monthly dates for a verification of the service<sup>7</sup>. Fig. 5 displays the results for one of these dates. The operability of the client and server functionality is checked for each partner separately.

Client	Destination	Alba – Spain (cells)	DESY – Germany	DLS – UK	Elettra – Italy	ESRF – France	FRM – Germany	HZB – Germany	ILL – France	ISIS – UK	JCNS – Germany	LLB – France	MAX – Sweden	PSI – Switzerland	Soleil – France	STFC – UK (domains)	Sum	Notes
Alba – Spain		1	1	1					1	1				1	1		7	
DESY – Germany		1	1	1					1	1				1	1		7	
DLS – UK		1	1	1					1	1				1	1		7	
Elettra – Italy		1	1	1					1	1				1	1		7	
ESRF – France		1	1	1					1	1				1	1		7	
FRM – Germany																		
HZB – Germany																		
ILL – France		1	1	1					1	1				1	1		7	
ISIS – UK																		
JCNS – Germany																		
LLB – France		1								1				1	1		4	
MAX – Sweden																		
PSI – Switzerland		1	1	1					1	1				1	1		7	
Soleil – France		1	1	1					1	1				1	1		7	
STFC – UK		1	1	1					1	1				1	1		7	
SNS/ORNL – USA																		
ANS – Australia																		
Sum		10	9	9					9	10				10	10		Sum	67
Count		17	17	17	17	17	17	17	17	17	17	17	17	17	17		Count	255
Coverage		59%	53%	53%					53%	59%				59%	59%		Coverage	26%
									Failures	18%	12						Servers	7
									Success	82%	55						Clients	10

Fig. 5: ICAT Service Verification

Eventually catalogues will be federated to allow for file searches across facility borders. This functionality is most efficient, if all catalogue instances accept Umbrella credentials and if all PaN-data partners define a subset of mandatory metadata. Mandatory metadata are those which are always provided. Examples are:

- Investigation title,
- User id,
- Investigation start date,
- Instrument name,

<sup>7</sup> pandatawp4.wordpress.com

- Facility name,
- File name,
- File description,
- File format and
- File size.

The selection of the items which fall into this category and the precise semantics is currently being negotiated. It may be worth mentioning that apart from the mandatory metadata each facility has the freedom to store additional metadata in the catalogue according to the local requirements.

### **3.1 Overview of the ICAT Deployment at the PaN-data partners**

This section gives an overview of the ICAT deployment at various PaN-data institutes. The DESY part contains more details because it hosts the virtual laboratories.

#### **3.1.1 ALBA**

ALBA is a third generation Synchrotron Light Facility which is in user operation since May 2012. First experiments data are being produced, and the need of having a catalogue has grown over the last months. That is why ALBA is implementing ICAT and it is planned to use TopCAT as well. The objective is to have the tools as flexible as possible to provide our users with access to other institutions' data catalogues which are also using the ICAT service.

Today, the implementation of this service is in progress. Visits management is available in ICAT (synchronized with the ALBA User Office) and the data can be visualized with TopCAT. Metadata synchronization and the ICAT Data Server implementation, for the download feature, are in progress.

It is expected to have the implementation finished within the following 2 months, and the beta version should be available at the beginning of September 2013.

There is an IDS feature which will be out of scope: downloading data stored in tapes or external resources. It is planned to be implemented once the current phase is consolidated.

#### **3.1.2 DESY**

Two ICAT instances have been installed at DESY. One of these, Science3D, is used exclusively for VL3. It is worthwhile to have dedicated catalogue for VL3 because this virtual laboratory is less demanding than VL1 and VL2 as far as the data management is concerned. Science3D will store open-access data only and it can be operated with the generic ICAT interface to the local file system (DataServer).

The other ICAT instance, DesyICAT, is intended to become the general DESY metadata catalogue. It was successfully populated with HDF5 test files using the Python ingestion script which is provided by the ICAT project team. Files can be searched using metadata. The first real use case for DesyICAT will be a catalogue of open-access data. This way, an interface to the DESY authorization system is not needed. Still this catalogue can be used to study catalogue federation and the service providers will make valuable experiences with ICAT. If the ICAT authorization can be integrated into the DESY infrastructure, other files will be ingested. So far, the authentication does not support Umbrella credentials. It is possible to implement this feature in the near future.

The Gamma-Portal (Fig. 6) is an example for a service that gives users remote access to their data. The portal has been built using the APEX technology<sup>8</sup>. It supports keyword-based searches, file downloads and stage requests. Staging is necessary, because the Gamma-portal acts as an interface to the dCache<sup>9</sup> which serves for long term data storage. The dCache simulates a file system of unlimited size. It consists of a tape storage system with a disk frontend. Sophisticated caching and staging algorithms are applied to optimize the I/O performance of the system. The word staging denotes the process of copying a file from tape to disk, where it can be read from analysis applications.

Users are authenticated by their DOOR credentials. Read access is granted for those files that were created during a beam time of the authenticated user. In response to a keyword based search a list of files is displayed with the following metadata: the beam time id, the facility, the beamline, the proposal id, the creation date, the local contact and the file size. The user may select single files to view additional information or to download a file or to submit a stage request.

By now the Gamma-portal works with a DESY-specific metadata catalogue. At the time when the project was started this was the only way to integrate the local authentication and authorization procedures. The Gamma portal will switch to DesyICAT as soon as the local authorization scheme is implemented and a DataServer which is adapted to the DESY file servers has been developed.



Fig. 6: The Gamma Portal At DESY

<sup>8</sup> Apex.oracle.com

<sup>9</sup> www.dCache.org

### 3.1.3 DLS

DLS intends to make the DLS production ICAT available to PaN-data later in 2013. At the moment, ICAT 4.2 is in pre-production operation within DLS, but requires the Central Authentication Service for Diamond (CAS) before it can be made public. This will be ready by November 2013. CAS will be supported for TopCAT access. CAS may not be supported for scripted applications. CAS will support Umbrella, Eduroam and STFC FederalID. A future service verification will test that partners can obtain suitable credentials, access the DLS catalogue, find and download data.

### 3.1.4 ELETTRA

In the context of PaNdata-ODI, ELETTRA leads the Data Catalogue service (WP4). ELETTRA has established a set of evaluation criteria for data catalogue systems (D.4.1). ICAT emerged as the most suitable solution for the requirements of the PaN-data Consortium.

ELETTRA currently has three active ICAT installations. One of these is accessible even outside the facility and is used for PaN-data ODI integration tests and ICAT service verifications.

The second installation is dedicated to an onsite 4th generation light source facility, the FERMI@ELETTRA Free Electron Laser (FEL). The data of FERMI is stored mainly in HDF5 files. The data acquisition workflow has been designed contemplating the data cataloguing step.

The third ICAT installation is for the ELETTRA synchrotron radiation facility. Due to the use of legacy formats and limited HDF5 use, the data ingest step aims at both newly produced data but also past archived data that may require custom conversions.

As a web front-end, a TopCAT installation is under evaluation. The final decision on the front-end has not been made yet. It will be either TopCAT or a custom plug-in for ELETTRA's Virtual User Office (VUO).

### 3.1.5 HZB

The HZB is currently revising its IT-infrastructure for the experimental stations, including data acquisition and data storage. ICAT will most likely be incorporated into the future design. Currently a test implementation of the data catalogue is being build up. A first test of this service, on a selected experimental station should start until the end of 2012. Further timing will heavily depend on progress to switch data acquisition to NeXus based files. There is no decision yet if TopCAT will serve as a Web-Frontend to the catalogue.

### 3.1.6 ILL

As of May 2013, ICAT and the Web client TopCAT are installed on a production server at the ILL<sup>10</sup>. The server has been set up and the information concerning users and experimental proposals done under the new ILL Data Policy (in place since October 2012) have been imported. Nevertheless the data ingestion to the database is only in a preliminary phase. Data generated in the NeXus file format will be ingested by September 2013. Concerning the other instruments not yet producing NeXus files, the data ingestion will be done progressively during the next 2 years, in parallel to the

---

<sup>10</sup> <https://icat.ill.eu>

passage to the NeXus format foreseen for all the ILL instruments. Umbrella will allow authentication to both ICAT and TopCAT as soon as the SAML and SAML Delegation protocol are implemented in ICAT and TopCAT.

### 3.1.7 ISIS

ISIS intends to make the ISIS production ICAT available to Pandata later in 2013. At the moment, ICAT 4.2 is in pre-production operation within ISIS, but requires ICAT 4.3 before it can be made public. This will be ready by November 2013. ISIS will support TopCAT and script access to ICAT. They will support authentication with Umbrella, Eduroam and STFC FederalID. A future service verification will test that partners can obtain suitable credentials, access the ISIS catalogue, find and download data.

### 3.1.8 SOLEIL

ICAT and TopCAT instances have successfully been installed for tests. SOLEIL actively participates to the tests of ICAT service verification managed by the ICAT project manager. The test results are regularly sent to him. The ICAT instance has been populated during one of the service verification tests by means of a python program provided by the ICAT project manager and it has been possible to provide services to other partners so that they can search for files and download data from the SOLEIL test server.

For the future, SOLEIL has to provide for our server a certificate from a recognized certificate authority. It is foreseen to test the API injector which should be provided by ICAT team for the last version of ICAT. In the production, the SOLEIL Data Retrieval (SDR) tool linked to the User Office software is used to search for files and retrieve any kind of experimental data format.

But as soon as the API injector and link to the User office tool will be ready ICAT could be used as a parallel option for NeXus files.

## 3.2 **Summary**

The ICAT installations meet most of the requirements listed in D5.1. The following list discusses the missing features:

- So far the authentication does not support Umbrella credentials. It is possible to implement this feature in the near future.
- The interface to the local authorization systems has to be implemented for most of the PaN-data partners.
- All metadata should be taken from the common dictionary which is currently being developed in WP6.
- All PaN-data partners have to agree on a common subset of mandatory metadata. Discussions on this issue are currently being carried out.
- Unique identifiers which allow for citations are implemented so far at ILL and ISIS. ESRF will follow soon. The PaN-data partners agreed on using DOIs for this purpose.
- The requirement that the API should support mounting the catalogue as file shares or via NFS has been implemented at DLS.

## 4 The Standard Data Format

The PaN-data community agreed on using NeXus/HDF5 as the standard data format. The contents and the internal structure of the NeXus files are closely related to the implementation of various PaN-data services. This section elaborates on the file production within the virtual laboratories. It addresses these issues:

- An efficient NeXus library with a proper interface has to be available.
- A metadata model has to be designed for every experimental technique.
- The NeXus file production has to be integrated into the data acquisition process.

The implementation details are exemplified for the experiments at PETRA III, because this is the place where the virtual laboratories will be setup.

### 4.1 The NeXus Library

The NeXus deployment at the Petra III is based on a program package<sup>11</sup> which was developed within the PNI-HDRI<sup>12</sup> project. The primary design goal for this package is to facilitate the creation of NeXus files. This is accomplished by creating a strictly object-oriented interface for application programs. It supports those data structures which are typically used by x-ray and neutron scattering experiments. The interface is type-safe in the sense that data structures have to be fully qualified when transferred between an application and the library. This way, misinterpretations are avoided and the usage of the libraries becomes very reliable. The NeXus package has a C++ API and the Python bindings. The underlying data format is HDF5 exclusively. The performance loss of the NeXus library compared to raw HDF5 is negligible ( $< 1\%$ ). The code is available on a public repository. Partners are encouraged to participate in the development. A full documentation is currently being prepared. The PNI-HDRI NeXus activities are coordinated with the NeXus International Advisory Committee.

---

<sup>11</sup> [Code.google.com/p/pni-libraries](http://Code.google.com/p/pni-libraries)

<sup>12</sup> [www.pni-hdri.de](http://www.pni-hdri.de)

The screenshot shows the PNI library stack website. The header includes the PNI logo and the text 'PNI library stack'. There is a search bar and a 'Search projects' button. The navigation menu includes 'Project Home', 'Downloads', 'Wiki', 'Issues', and 'Source'. The 'Wiki' section is active, showing a search bar and a 'Search' button. The left sidebar contains a list of links: 'Introduction', 'Installation', 'User Guides', 'Examples', and 'API documentation'. The main content area is titled 'Introduction' and features a 'Featured' section. The 'Introduction and History' section describes the PNI libraries as a stack of related C++ libraries developed to simplify the development of scientific software. It mentions the High Data Rate Initiative (HDRI) at DESY and the Nexus file format. The 'Overview' section includes a diagram showing the relationship between Python/Fortran/C applications, native C++ apps, bindings, and the PNI library stack (libpniio, libpniGUI, libpniAlgorithm, libpniCore).

**Introduction**

Featured

Updated Feb 12, 2013 by [eugen.wintersberger@gmail.com](mailto:eugen.wintersberger@gmail.com)

## Introduction and History

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. The development started within the High Data Rate Initiative (HDRI) at DESY. Originally only a strictly object oriented API for the [Nexus file format](#) should have been developed. Due to high performance requirements the API should have been implemented in C++. However, shortly after the development begun it turned out that the major problem with C++ was not the Nexus API but rather the fact that C++ provides no data structures required for writing scientific software. In particular the missing array types for numerical calculations turned out to become a serious problem. Additionally working with raw pointers would make the resulting code error prone and susceptible to all kinds of memory issues (in particular memory leaks). Thus a utility library was developed providing all kinds of missing data types and structures. This library recently became the core of the PNI library stack.

## Overview

The following graphics gives an overview how the libraries are related to each other and to application programs using them.

```

graph TD
    A[Python/Fortran/C applications] --> B[native C++ apps]
    A --> C[bindings]
    C --> D[libpniio]
    C --> E[libpniGUI]
    C --> F[libpniAlgorithm]
    D --> G[libpniCore]
    E --> G
    F --> G
    G --> H[PNI library stack]
  
```

Fig. 7: The NeXus Libraries (PNI-HDRI)

Data compression is an issue for those experiments that generate high data rates. HDF5 has internal filters for a lossless data reduction. The compression/decompression process is transparent, because the HDF5 I/O routines use metadata which are contained in the data files to decide whether data have to be decompressed, before they are transferred to the application. However, vendors are currently implementing compression algorithms in FPGA or ASIC electronics in order to achieve the highest possible throughput, thereby bypassing the HDF5 procedures. This would lead to the case that the applications themselves have to be aware of the compression state of the data and act accordingly. Such a situation must be avoided. Therefore DESY has placed an order at the HDF5 group to integrate external filters. After this feature has been implemented, HDF5 file I/O will be transparent w.r.t. compression. The external filter feature will also allow for discipline-specific compression algorithms.

## 4.2 NeXus Metadata

Metadata is included in the NeXus files for different reasons:

- Experiment description: data can only be analyzed, if all relevant parameters of the measurement are documented. This includes information about the radiation source, the beam-line components, the sample environment and the raw data.

- File description: files are catalogued after they have been produced. The ingestion procedure creates entries in the catalogue using metadata from the files and information from digital user office databases which is retrieved on the basis of the file metadata.
- Recording the data continuum: the verification of published results requires not only the raw data and the experiment description, but also the description of the data processing steps.

It should be common practice to select metadata from the controlled vocabulary which is currently being developed within WP6. Such a common dictionary supports provenance aspects and cross-facility file searches. Furthermore it helps to locate relevant software in the PaN-data Software<sup>13</sup> catalogue and it allows application programmers to identify variables within the files.

The greatest challenge is the complete description of the measurement. It depends on the experimental technique and it reflects the complexity of all components that have an influence on the experiment: the source parameters, the optical components of the beamline and the sample environment. The configuration of the source and certain parts of the beamline are rather static. But the situation in the experimental hutches may change with each new user group on a daily basis or even faster. Even minor changes like rearrangements of a detector have to be taken into account. The information that is sufficient to describe the entire setup cannot be provided manually by the users groups or members of the beamline staff. Instead an automated procedure has to be prepared that supports the generation of the NeXus metadata.

The way how NeXus application definitions (AD) are implemented is an important aspect of metadata management. ADs allow application programmers to locate variables within the files independent of where the files have been created. ADs are specific to experimental techniques. The NeXus International Advisory Committee established a procedure how an AD is approved for a specific discipline. It seems that the process of standardization is much more advanced on the neutron science community compared to the photon science field. In the photon science community there are still lengthy discussions about the comprehension of the ADs and about other details. However, for scattering experiments there exists an AD which has been developed within the High Data Rate Initiative of the PNI institutes of the Helmholtz Society: "Application Definition of a Scattering Experiment"<sup>14</sup>. This AD can be applied to the virtual laboratories VL1 (powder diffraction) and VL2 (small angle scattering). The deployment process will show, whether the scattering ADs are sufficient to describe the experiments. If necessary the AD will be extended.

### 4.3 The NeXus Writer

Several components of the data acquisition system are involved when a NeXus file is created. Fig. 8 shows the relationship. The experiment control client (ECC) is the interface to the user. At PET-RA III Tango<sup>15</sup> is used as the communication layer between the experimental equipment and online applications. The NeXus writer and the configuration server are also implemented as Tango server. Unfortunately Tango is not the only control system within PaN-data ODI which limits the applicability of the NeXus writer and the configuration server.

---

<sup>13</sup> [Software.pan-data.eu](http://Software.pan-data.eu)

<sup>14</sup> [www.pni-hdri.de](http://www.pni-hdri.de)

<sup>15</sup> [www/tango-controls.org](http://www/tango-controls.org)

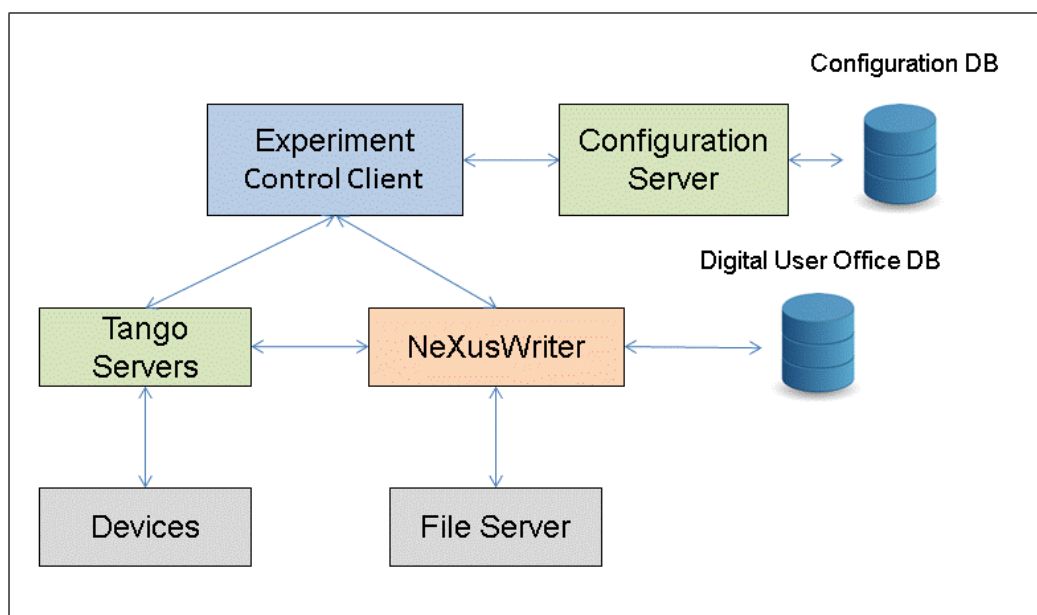


Fig. 8: The NeXus Writer

Let's follow the flow of control to understand the interaction of the components. The entire procedure is initiated by a scientist who starts a measurement through the ECC selecting an experimental technique, including the specific hardware setup. The ECC prompts the configuration server for a suited NeXus configuration. NeXus configurations are NXDL (NeXus definition language) strings with an extended syntax to define data sources and data collection strategies. They are stored in a dedicated database. The ECC sends the configuration to the NeXus writer which uses this information when it opens a NeXus output file. The NeXus writer may retrieve some administrative information from the Digital User Office DB. While the measurement is carried out, the NeXus writer collects data from Tango servers or from the ECC. The output is written to a file server. This can be a local disk or a network file system.

So far, we assumed that the configurations are available from the configuration server. In the following it is described how the configurations are prepared. They consist of components. Components represent simple devices, composite devices or structures. Simple devices are counters, MCAs, 2D detectors and so on. Composite devices have several internal degrees of freedom. Examples are monochromators, diffractometers or insertion devices. Structures represent descriptions of beamlines and similar systems. Components can be inserted or deleted from a configuration without side effects. They may consist of elements which correspond to NeXus fields. Each element is associated with a data source. It has a predicate that determines when the data source is read (INIT, STEP, FINAL, POSTRUN). In addition there is an optional predicate that specifies whether the element has a trigger assigned to it. The possible data sources are Tango servers (attributes, device properties and command outputs), data which is provided by the ECC, database records and data from external sources.

The following list displays the important part of the NeXus writer interface:

- The **Configuration** attribute describes the layout of the NeXus file. It is an NXDL string.
- The **GlobalDictionary** attribute has several optional entries:
  - Client data, e.g. the time stamp

- External data sources which are implemented as Python classes.
- Decoders for encoded data which are implemented as Python classes.

The GlobalDictionary can be updated by the ECC at any time during a measurement.

- The **OpenFile** command uses the attribute filename to create a new NeXus file.
- The **OpenEntry** command uses the Configuration and the GlobalDictionary.
  - Creates the NeXus file structure, data and metadata.
  - Executes an INIT-mode data collection run, i.e. data sources that have an INIT label are evaluated and the data is stored.
- The **StepDictionary** attribute is filled with client data (optional) and with trigger names (optional). The ECC updates this attribute for every step during measurements.
- The **Record** command executes a data collection cycle in STEP mode. It retrieves the client data from the StepDictionary, uses the trigger names to read data sources that have the trigger feature and reads the selected data sources.
- The **CloseEntry** command stores client data from the GlobalDictionary and executes a FINAL mode data collection run.
- The **Close** command closes the file.

At DESY, the NeXusWriter will be used in conjunction with Sardana<sup>16</sup> which is currently being rolled-out at the PETRA III experiments to become the future online control system. It has a DevicePool for an abstraction of the device interface, a MacroServer for script execution and SardanaDoors, which are the access points for user interfaces and other clients. The NeXus writer is a separate process which receives data records from a dedicated SardanaDoor. The records are sufficient to control the NeXus writer during data taking. 0D and 1D data are immediately stored in the output files during the measurements. Because of performance reasons 2D data is written to separate file systems. It is planned to develop a DataCollector which inserts the 2D data into the NeXus file after the measurement has ended. This way the NeXus file will contain the complete information of an experiment.

---

<sup>16</sup> [www.tango-controls.org/static/sardana/latest/doc/html](http://www.tango-controls.org/static/sardana/latest/doc/html)

## 4.4 The Component Designer

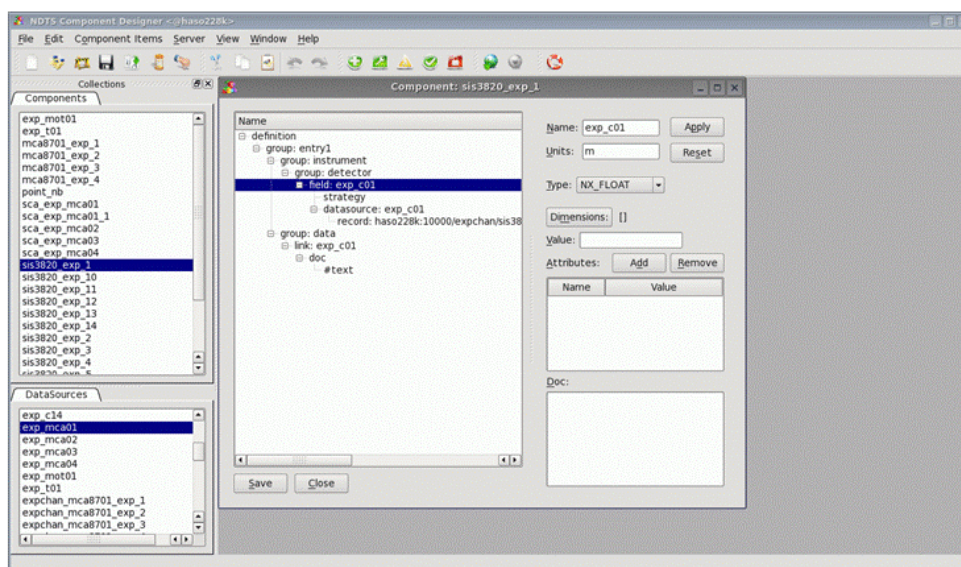


Fig. 9: The Component Designer

The purpose of the ComponentDesigner (Fig. 9) is to create configurations by merging components and by editing their attributes. Components are represented by xml strings which can be produced by the designer or which can be read from an external file. In general, it is most convenient to specify components in an xml file because this procedure can be automatized, e.g. a script reads the device information from a Tango database and generates a list of components with default attributes. The entire NeXus tree can be built with this designer. After a configuration is finished, it is stored in a database.

## 4.5 The CDMA Status

The CDMA<sup>17</sup> project was initiated by SOLEIL and ANSTO to define stable application interfaces for different file formats. It uses a dictionary mechanism to relate application notions to file entities. CDMA is useful also for the PaN-data institutes, because the structure of NeXus files which are created at different facilities is not necessarily identical, even if they belong to the same discipline. CDMA has APIs for C++ and Java. DESY joined the initiative and contributed the Python bindings (Fig. 10). CDMA is an open source project. It is ready to be used.

<sup>17</sup> [code.google.com/p/cdma](http://code.google.com/p/cdma)

# CDMA Python

Python binding to the CDMA C++ library

[Main Page](#)
[Modules](#)
[Classes](#)
[Files](#)

## CDMA Python API developer documentation

This part of the CDMA Python documentation is intended for binding developers only. So if you do not want to contribute C++ code to the Python binding you are entirely wrong here.

This page tries to collate some of the design ideas behind the C++ code creating the Python bindings. For a detailed documentation of the different classes see

- [Utility Classes](#)
- [Wrapper Classes](#)

### Common interface for data objects

While going through the original CDMA C++ API documentation realized that some of the data holding types, namely Array, Attribute, and DataItem, all serve a similar if not equal purpose. However, all these classes showed a different interface which pretty bad because it increases the amount of work I would have to do to create the wrappers. Consequently I packed all these objects in wrapper classes that expose a common interface as shown here.

```

class DataInterface
{
public:
    size rank() const;
    size size() const;
   TypeID type() const;
    std::vector<size_t> shape() const;

    //read scalar data
    template<typename T> T get() const;
    //read data from position
    template<typename T> T get(const std::vector<size_t> &pos) const;
    //read region
    ArrayWrapper get(const std::vector<size_t> offset,
                    const std::vector<size_t> shape) const;
};

```

Do not expect to find somewhere a class named DataInterface - this is just a name for the interface I use here.


Generated on Mon Dec 17 2012 09:59:01 for CDMA Python by  1.8.1.2

Fig. 10: Python Bindings for CDMA

## 4.6 Summary

This is the current status of the issues that relate to the standard data format:

- An object-oriented NeXus library with a C++ API and Python bindings exists and is ready to be used.
- A framework for gathering metadata has been designed and implemented. It is currently being tested.
- Python bindings for the CDMA C++ API have been developed and are ready to be used.
- External filters which support discipline-specific compression algorithms will be implemented in HDF5 soon.
- Converters to old formats have not been written yet. They will be provided on demand.

## 5 Scalability

WP8 develops a framework for data storage and data processing that copes with the data rates of modern detectors. It is based on the deployment of pHDF5, the parallel HDF5 library, on parallel file systems. The requirements of a NeXus capable pHDF5 implementation are defined in the report D8.1. The technical performances of systems like FhGFS, GPFS<sup>18</sup> and Lustre<sup>19</sup> are compared in D8.2. A parallel HDF5 writer is currently being developed at Diamond.

The highest throughput of the system is achieved, if the structure of a HDF5 file is adapted to the layout of the file system cache. The relevant parameters depend also on the detector type and the measurement procedure. The optimal configurations will be reported in D8.3. The virtual laboratories VL1 and VL2 will benefit from all these investigations.

However, some adjustments have to be done inside the HDRI-NeXus library to work with the MPI-HDF5 library. At present it is not clear whether the Python bindings can be adapted to MPI. It should also be mentioned that setting up a MPI based environment for writing NeXus files requires detailed knowledge about the MPI procedures themselves and how they have to be configured to make best use of the compute resources.

The implementation of the Virtual Labs is largely independent of the particular choice of underlying file systems and MPI implementations, which will ease the adoption of particular pHDF5 realizations.

---

<sup>18</sup> <http://www-03.ibm.com/systems/software/gpfs>

<sup>19</sup> <http://wiki/whamcloud.com/display/PUB/Wiki+Front+Page>

## 6 Software Catalogue

A central software database, PaN-data Software<sup>20</sup>, has been setup at ILL which makes applications from the field of neutron and photon science publically available. A comprehensive web site is hosting the project (Fig. 11). The software is divided into data analysis, instrument simulation and sample simulation. Each application is accompanied by a set of metadata:

- A web site that hosts the project
- The current version and license type
- The software category
- Beam type
- Institute
- Users, maintainers and contributors
- Instruments
- Software and hardware requirements, supported platforms
- Language
- I/O formats
- Comments

The PaN-data Software is searchable and has a forum to discuss issues. The software database is in a beta state. All requirements which have been listed in D5.1 are fulfilled.

**PaNdata Software** ? Help & Support Social Software Institutes Login Register

**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

PaNsoft is a database of software used mainly for data analysis of neutron and photon experiments. PaNsoft 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.

[Browse software](#)

#### RECENT SOFTWARE

**NAMD**  
NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems.

#### SEARCH FOR SOFTWARE

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

[Search](#)

Fig. 11: The PaN-data Software Site

<sup>20</sup> [software.pan-data.eu](http://software.pan-data.eu)

DawnScience, DPDAK and MANTID are good candidates for analysis programs that will be employed within VL1 and VL2. They will be available via PaN-data Software. Some information about these frameworks can be found in section 7.

## 7 Workflow Management Systems

The importance of supporting data processing with a powerful workflow management system has already been pointed out in D5.1. Here are some key issues:

- Workflows are well suited to provide users with compute resources who don't have the sufficient IT infrastructure at their home institutes or the necessary expertise.
- The data continuum can only be recorded, if the data processing is supported by automated procedures. Workflows are good way to implement this requirement.
- Workflow systems consist of components with standardized interfaces. They can be developed and tested separately. This modular approach facilitates the re-use of program code.

So far, there are at least three projects that are flexible enough to be adapted to computing tasks for Neutron and Photon science

**DawnScience**<sup>21</sup> is an Eclipse/RCP based system for data analysis (Fig. 12). It supports visualization and data processing. DawnScience uses Passerelle, which was created by iSencia<sup>22</sup>, as a workflow engine. The whole system is very modular. Users interact with DawnScience depending on their needs and skills. Beginners take advantage of the standard tools as they are provided through the graphical user interfaces. Experienced users extend the system by writing eclipse plugins (Java) and workflow actors (Python). The graphical user interface can also be customized to meet specific needs. DawnScience is mainly developed by DLS and ESRF.

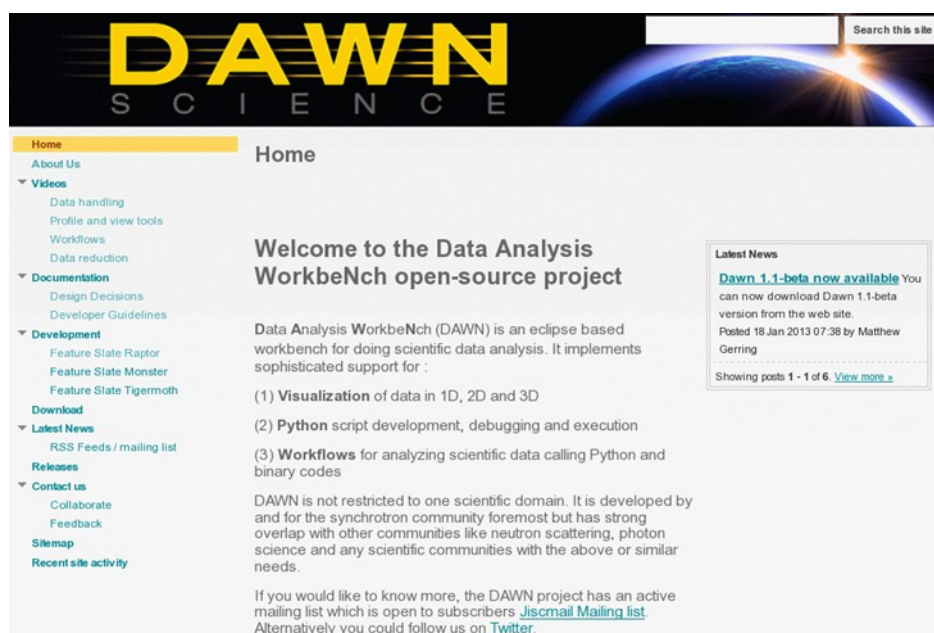


Fig. 12: The DawnScience Homepage


**DPDAK**<sup>23</sup> is a flexible, modular framework for analysing large sequences of small angle scattering data (Fig. 13). An analysis consists of subsequent processing steps which are carried out by

<sup>21</sup> [www.dawnsci.org](http://www.dawnsci.org)

<sup>22</sup> [www.isencia.be](http://www.isencia.be)

<sup>23</sup> [dpdak.desy.de](http://dpdak.desy.de)

plugins. The plugins may be selected from a library of standard modules or they can be customized for specific purposes. DPDAK is written in Python and runs on Windows and Linux. It is developed in a cooperation between DESY and MPIKG.

<p><b>Welcome</b></p> <p>Dpdak is an open source tool for (online) analyzing large sequences of small angle scattering data. It is written in Python and runs on Windows and Linux. The development is a cooperation between <a href="#">DESY</a> and <a href="#">MPIKG</a>.</p>  <p><b>Max-Planck-Institut für Kolloid- und Grenzflächenforschung</b></p> <p>The SAXS Integration (Fit2D) plugin uses <a href="#">Fit2D</a> software developed by Dr. Andy Hammersley at the <a href="#">ESRF</a>.</p> <p><b>Download</b></p> <p>The latest release is 0.3.0_beta2:</p> <ul style="list-style-type: none"> <li>• <a href="#">Windows</a></li> <li>• <a href="#">Linux</a></li> <li>• <a href="#">Source</a></li> </ul> <p>You can find older releases <a href="#">here</a>.</p> <p><b>News</b></p> <ul style="list-style-type: none"> <li>• 12.09.2012: <a href="#">Requests</a> for next dpdak version 0.4.0</li> <li>• 01.07.2012: dpdak Wiki goes public.</li> </ul>	<p><b>Manual</b></p> <p><b>Getting Started</b></p> <ul style="list-style-type: none"> <li>• <a href="#">Install dpdak</a></li> </ul> <p><b>Plugins</b></p> <ul style="list-style-type: none"> <li>• Base Plugins <ul style="list-style-type: none"> <li>• <a href="#">CHI File Reader</a></li> <li>• <a href="#">Directory Image Logger</a></li> <li>• <a href="#">Image Logger</a></li> <li>• <a href="#">Intensity Correction</a></li> <li>• <a href="#">Intensity Correction (File)</a></li> <li>• <a href="#">Line Integration</a></li> <li>• <a href="#">Line Integration GLAD</a></li> <li>• <a href="#">Peak Fit</a></li> <li>• <a href="#">ROI Values</a></li> <li>• <a href="#">Rho Parameter</a></li> <li>• <a href="#">SAXS Integration (Fit2D)</a></li> <li>• <a href="#">Spec Reader uSport</a></li> <li>• <a href="#">Sum Images</a></li> <li>• <a href="#">T Parameter</a></li> </ul> </li> <li>• Display Plugins <ul style="list-style-type: none"> <li>• <a href="#">Detector Image</a></li> <li>• <a href="#">Multi Plot</a></li> <li>• <a href="#">2D Color Plot</a></li> <li>• <a href="#">Peak Fit Display</a></li> </ul> </li> </ul>
---	--

**Fig. 13: The DPDAK Homepage**

**Mantid**<sup>24</sup> is an open source data analysis framework which was originally created for neutron and muon data (Fig. 14). It can be applied to synchrotron radiation data as well. MantidPlot is the graphical user interface of the framework. It is based on QT. Algorithms are implemented in C++ or Python. The scripting language is Python. Several platforms are supported (Windows, Linux, Mac). Mantid is developed by ISIS and SNS.

<sup>24</sup> [www.mantidproject.org](http://www.mantidproject.org)

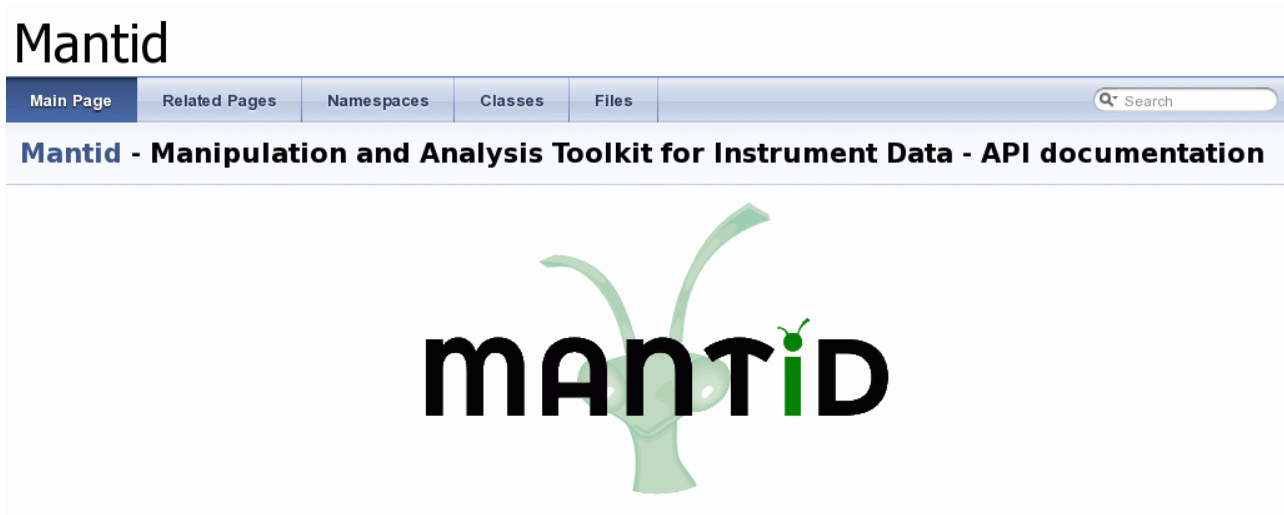


Fig. 14: The Mantid Homepage

The future will show which of these frameworks is best suited for the virtual laboratories. PaN-data actually has submitted a topic proposal aiming at the harmonization of the underlying plugin mechanisms, so that modules can effortlessly be re-used in arbitrary analysis frameworks, which would substantially enhance the analysis framework of choice.

## 8 Integration

The preceding sections of this report summarize the present status of those PaN-data ODI work packages that are relevant for the virtual laboratories. This section follows a different approach. It focusses on the interchange of the PaN-data activities and on the integration into the virtual laboratories. It is also discussed how scientists benefit from the results that are achieved by PaN-data so far. This document ends by revisiting a table from the deliverable 5.1, the list of tasks that enable the virtual laboratories. The status for each of these tasks is reported.

The requirements for VL1 (powder diffraction) and VL2 (small angle scattering) are almost identical. The only difference between these two virtual laboratories is the data processing algorithm. The other PaN-data services are used in the same way. The situation for VL3 (tomography database) is different. It needs a visualization tool, a portal, a metadata catalogue and access to a file system. As far as the PaN-data activities are concerned VL3 uses the metadata catalogue only.

The remainder of this section discusses VL1 and VL2 related topics.

### 8.1 Standard data format

Data that is analyzed within VL1 and VL2 is stored in NeXus/HDF5 because the PaNdata partners decided to use this format. Since this decision has severe consequences for the whole project and in particular for the integration of the virtual laboratories, the following list has been prepared which addresses the related issues.

- Data that is stored in the output files have physical meaning. To exemplify this, consider a motor that controls a rotational axis which moves a detector. NeXus requires the position of the motor to appear at a certain location in the file structure which is determined by the function of the motor. Furthermore, the stored position has to be in degrees or radian. The raw position of the motor has to be converted by means of a conversion factor and a calibration constant before it is written to the output file. Eventually the angular position of the detector is to a certain level abstracted from the particular experimental environment. All stored quantities are treated in a similar fashion allowing the data in the output files to be interpreted unambiguously.

For scientists it is a clear advantage to carry home NeXus-formatted data because the analysis is much more straightforward and data which have been collected at different beamlines or even at different facilities can be combined easily. Another benefit of NeXus files is that the amount of additional information, like log book entries, is drastically reduced.

The configuration of a NeXus file is a complex task. This is one of the lessons learnt when deploying NeXus at PETRA III (for VL1 and VL2). A typical experiment at this facility uses hundreds of devices. Each of these has to be given meaning which may depend on the specific experimental technique. In addition, the configuration procedure has to cope with rapid re-arrangements of the components. A typical example is that a broken motor has to be temporarily replaced by another motor which serves a different purpose so far. Such a replacement of hardware has to be accompanied by reassignments in the NeXus configuration.

The component designer which has been described in section 4.4 has proven to be a valu-

able tool for preparing and managing NeXus configurations. For VL1 a suitable configuration has been prepared and tested. The configuration for VL2 will follow soon.

- NeXus is able to handle metadata. They are used for the description of the measurement, for file and software searches, for tracing the data continuum and data preservation. Some details about NeXus metadata can be found in section 4.2. Since this section focusses on the integration of the virtual laboratories, we like to point out that the mandatory subset of metadata which is currently been defined by WP6, will be included into the NeXus configurations of VL1 and VL2.
- The NeXus integration of the virtual laboratories requires a truly object-oriented, type-safe and high-performance C++ library, including Python binding. The code has been developed within PNI-HDRI, see section 4.1, as an open source project. VL1 and VL2 create data files by utilizing the NeXusWriter. The code is also publically available but it comes with the restriction that it runs only in a Tango environment (section 4.3).

The NeXusWriter and the NeXus library are ready to be used. It is expected the experiences which will be made while further deploying the virtual laboratories will lead to improvements of the code.

- The scalability work package concentrates on the HDF5 I/O performance. The goal of work package 8 is to develop a framework which is capable to store and process data which are generated by modern 2D detectors. After the decision to choose the HDF5 version of NeXus as the standard data format had been made, WP8 is able to limit the I/O studies to HDF5 related issues, i.e. optimize the throughput of pHDF5, the parallel HDF5 library, on parallel file systems.

The virtual laboratories do not make use of pHDF5 so far. The transition to pHDF5 requires the installation of a MPI based data collection process using optimized configurations for data structures and the file servers. WP8 is currently developing a parallel writer that solves all of these problems. This application will not only be very useful for the virtual laboratories but also of great interest for all facilities that operate modern 2D detectors. The importance of parallelized I/O will increase with the next generation of area detectors.

The PNI-HDRI library will need some adaptations before it can be linked with the MPI version of HDF5. At this moment it is not clear whether the Python bindings can be ported to MPI. This will be one of the goals for the remainder of the project.

## 8.2 The Metadata Catalogue

It has been stated that the production of NeXus files within the virtual laboratories has begun. The next step is to ingest these files into a metadata catalogue. Section 3 describes the status of the ICAT deployment at the various PaNdata partners. Some of the institutes are already in the position that they run ICAT in a production environment. Others installed ICAT test instances. All partners are actively participating in the monthly service verifications.

However, all PaNdata partners agreed to use ICAT as the metadata catalogue. This is a big achievement which enables scientists to perform cross-facility keyword-based file searches. Furthermore, the PaNdata community also agreed to select keywords from a controlled vocabulary which standardizes the queries. The ongoing work on the development of a mandatory subset of metadata is an initiative that goes in the same direction.

ICAT is never exposed to the general user. Instead it is integrated into front-end applications that serve different purposes like data download, staging, data processing and recording the data continuum. These applications cannot be standardized because they have to respect the particularities of the local IT infrastructures. The Gamma portal is an example for a site-specific ICAT implementation. It will make the DESY data of VL1 and VL2 available for others.

But there are ICAT-compliant applications which are commonly used. The most prominent is TopCAT which is part of the ICAT distribution. It is the most frequently used client for file discovery and file download. In the field of data processing frameworks DawnScience and MANTID have to be mentioned. Both programs are supported by an international community and both programs have an ICAT interface. There is one drawback: DawnScience and MANTID have access to the local IT infrastructure only.

This subsection is concluded by reporting the status of the virtual laboratories with respect to the metadata catalogue. At present, the NeXus files from VL1 and VL2 cannot be ingested into ICAT because the interface to the DESY authorization scheme is still missing. The authorization should be implemented within a few months. Furthermore, a DataServer which is adapted to the DESY file servers also has to be developed and tested. It should also be made available during the remainder of the project.

### 8.3 The Common Authentication System

After Umbrella has been generally deployed, it will be the first time that scientists can access distributed resources like digital user offices, metadata catalogues and the software database with a single set of credentials. The single-sign-on feature will also facilitate cross-facility file searches. Umbrella is currently being deployed at PSI, ILL and ESRF. DESY, HZB, DLS and ALBA will follow this year.

### 8.4 Summary

The deliverable D5.1 “Specific requirements for the virtual laboratories” contains a table that displays the tasks that enable VL1 and VL2:

No.	Task	Work package
1	NeXus configuration	WP5
2	NeXus library	WP5
3	Data collection, VL1 and VL2	WP5
4	Produce reference data	WP5
5	Metadata catalogue installation	All
6	Ingestion tool template	WP4
7	Metadata catalogue API	WP4
8	Mandatory metadata	WP6
9	Common dictionary	WP6
10	Data server	WP4, all
11	Portal (file search, download)	WP4, all
12	Umbrella deployment	WP3, All
13	Umbrella-ICAT integration	WP4, WP3, all
14	PanSoft	ILL

**Table 1: The Tasks that enable the Virtual Laboratories VL1 and VL2, from D5.1**

The following list reports the current status for each of these tasks.

1. The NeXus configuration for VL1 has been implemented. The instrument is fully represented by suitable NeXus component. The mandatory subset of metadata and other metadata are included. VL2 will follow soon.
2. The NeXus library which has been created within the PNI-HDRI project has been successfully tested. The data collection process uses the Python bindings.
3. Data collection has started for VL1. The first data for VL2 will be taken in July 2013.
4. So far reference data has not been made available. This will happen during the next months.
5. The ICAT deployment has begun. Details can be found in the results of the service verifications.
6. A generic ingestion tool has been provided by WP4. It can be customized according to the local needs.
7. The ICAT APIs have been published.
8. WP6 distributed a proposal for the mandatory metadata. The discussion on this topic has started.
9. The common dictionary will be part of D6.2
10. A generic data server exists. It has to be adapted to the particularities at the different institutes. The download part has been implemented, the upload part so far not. This should be done within the remainder of the project.
11. TopCAT is the most frequently used portal.
12. Within 2013 Umbrella will be deployed at PSI, ILL, ESRF, DESY, HZB, DLS and ALBA.
13. The work on the Umbrella-ICAT integration has started.
14. The central software database exists.