PaN-data ODI

Deliverable D4.2

Populated metadata catalogue with data from the virtual laboratories

Grant Agreement Number	RI-283556
Project Title	PaN-data Open Data Infrastructure
Title of Deliverable	Populated metadata catalogue with data from virtual laboratories
Deliverable Number	D4.2
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	01 January 2013 (Month 15)
Actual Delivery Date	15 February 2013

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

Abstract

This document describes the deployment of the chosen metadata catalogue solution in the legacy context of the collaborating facilities.

Keyword list

Data catalogue, metadata, data management.

Document approval

Approved for submission to EC by all partners on Feb. 15, 2013.

Revision history

Issue	Author(s)	Date	Description
0.1	Milan Prica, George Kourousias	4 Jan 2013	First draft version
0.2	Antony Wilson	10 Jan 2013	Description of TopCAT
0.3	Milan Prica, George Kourousias	5 Feb 2013	Nexus Tomography file and ingestion
0.4	Alistair Mills, Milan Prica, George Kourousias	7 Feb 2013	Service Verification 2 results
0.5	Alistair Mills, Milan Prica, George Kourousias	11 Feb 2013	Plan for populating data catalogues
1.0		15 Feb 2013	Final version

Table 1: Revision of document

Table of Contents

1 Introduction

The PaNdata consortium brings together thirteen major European research infrastructures to create an integrated information infrastructure supporting the scientific process. PaNdata-ODI will develop, deploy and operate an Open Data Infrastructure across the participating facilities with user and data services which support the tracing of provenance of data, preservation, and scalability through parallel access. It will be instantiated through three virtual laboratories supporting powder diffraction, small angle scattering and tomography.

The project aims at the standardisation and integration of the consortium's research infrastructures in order to establish a common and traceable pipeline for the scientific process from scientists, through facilities to publications. At the core there is a series of federated data catalogues which allow scientists to perform cross-facility, cross-discipline interaction with experimental and derived data at high speeds. This will also deliver a common data management experience for scientists using the participating infrastructures particularly fostering the multi-disciplinary exploitation of the complementary experiments provided by neutron and photon sources.

The WP4 will deploy, operate and evaluate a generic catalogue of scientific data across the participating facilities and promote its integration with other catalogues beyond the project.

In the D4.1 we have conducted a survey of numerous existing data catalogue systems and a set of criteria for their evaluation has been defined. A data catalogue of choice for the PaNdata consortium is ICAT, developed and actively maintained by the STFC. ICAT has been examined in detail against the criteria and the results confirmed our choice.

In this document the main focus is on the deployment of ICAT in a number of partner facilities. Detailed reports on service verification actions performed so far are included. These actions are conducted to enable the cross facility data accessibility. Finally, we discuss issues related to the NeXus sample files from VLabs and present a plan for their ingestion into the ICATs.

2 Data Catalogue

2.1 ICAT

ICAT is a metadata cataloguing system chosen by the PaNdata consortium which has been already deployed in several of the partners' laboratories. ICAT runs on top of a database server (JPA+ support e.g. Oracle, MySql) and it presents a Web Service API (developed in Java, running in Glassfish, usable in various languages).

ICAT allows high-quality:

- Registration of data as it is collected in experiments;
- Discovery of data based on what it was collected for;
- Access to the data, according to some policy;
- Association of data with other resources.

Since November 2012, CRISP project [http://www.crisp-fp7.eu] has also chosen ICAT as the metadata catalogue of choice. PaNdata and CRISP partners actively collaborate with the ICAT development team in defining requirements and establishing a roadmap for future releases.

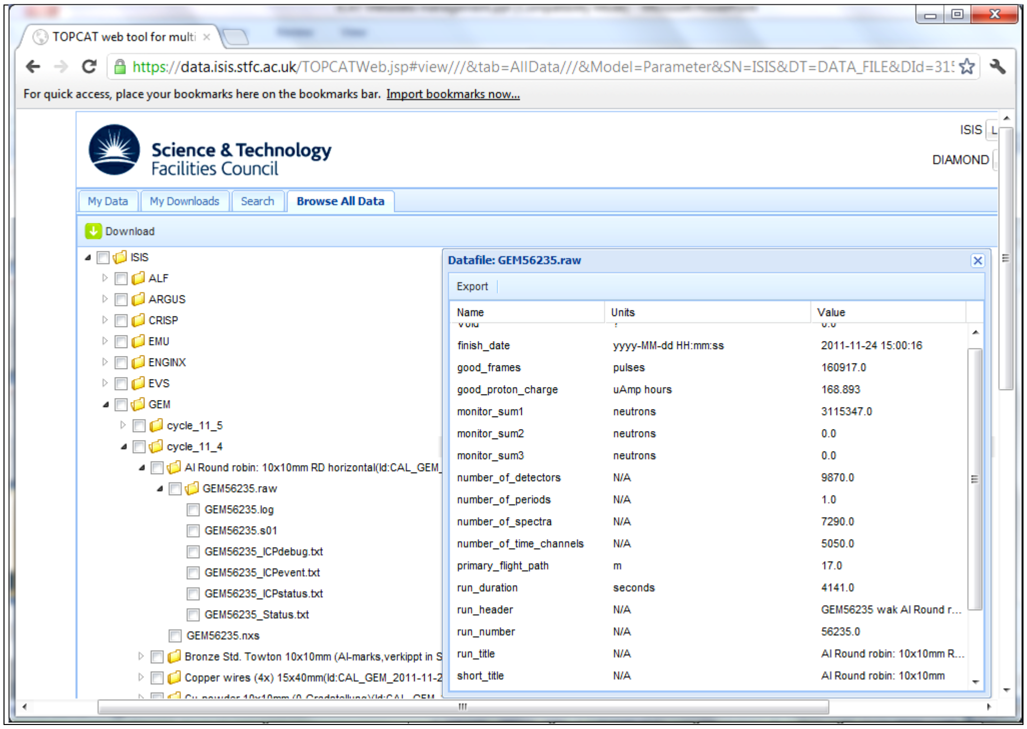
A short history of recent ICAT releases is shown in Table 1. The latest version of ICAT is 4.2.2, a bug-fix release published on January 15th.

Release (date)	Comments	Status
3.3 (2007)	Big API - many variants	End of life in 2013
4.0 (1/2012)	Small API – technology preview	Should no longer be used
4.1 (6/2012)	Production use for new users	Should no longer be used
4.2 (8/2012)	Pluggable authentication	In production use
4.3 (3/2013)	Under discussion	March 2013

Figure 1: History of releases of ICAT

2.2 TopCAT

TopCAT is a web based GUI (Figure 1) that is used to interact with one or more ICATs. It provides a means of browsing data about investigations, data sets and data files. It also enables the user to perform searches across multiple ICATs running at different facilities. It is possible for facilities to provide custom search panels; currently there are custom search panels for ISIS and DLS. TopCAT also gives the user the ability to download the data files. This requires the facility to be running a data service. Due to historical reasons there are a number of different data services already in production that are currently being modified to implement a recently defined common interface. In order to allow the download of data via TopCAT any new data service should implement the ICAT Data Service interface. A reference implementation will be made available after the current data services have been migrated.


TOPCAT web tool for multi
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Science & Technology Facilities Council
ISIS
DIAMOND
My Data
My Downloads
Search
Browse All Data
Download
Export

https://data.isis.stfc.ac.uk/TOPCATWeb.jsp#view///&tab=AllData///&Model=Parameter&SN=ISIS&DT=DATA_FILE&DId=31

```
ISIS
  ALF
  ARGUS
  CRISP
  EMU
  ENGINX
  EVS
  GEM
    cycle_11_5
    cycle_11_4
      Al Round robin: 10x10mm RD horizontal(Id:CAL_GEM_
        GEM56235.raw
          GEM56235.log
          GEM56235.s01
          GEM56235_ICPdebug.txt
          GEM56235_ICPevent.txt
          GEM56235_ICPstatus.txt
          GEM56235_Status.txt
        GEM56235.nxs
      Bronze Std. Towton 10x10mm (Al-marks,verkippt in S
      Copper wires (4x) 15x40mm(Id:CAL_GEM_2011-11-2
```

Datafile: GEM56235.raw

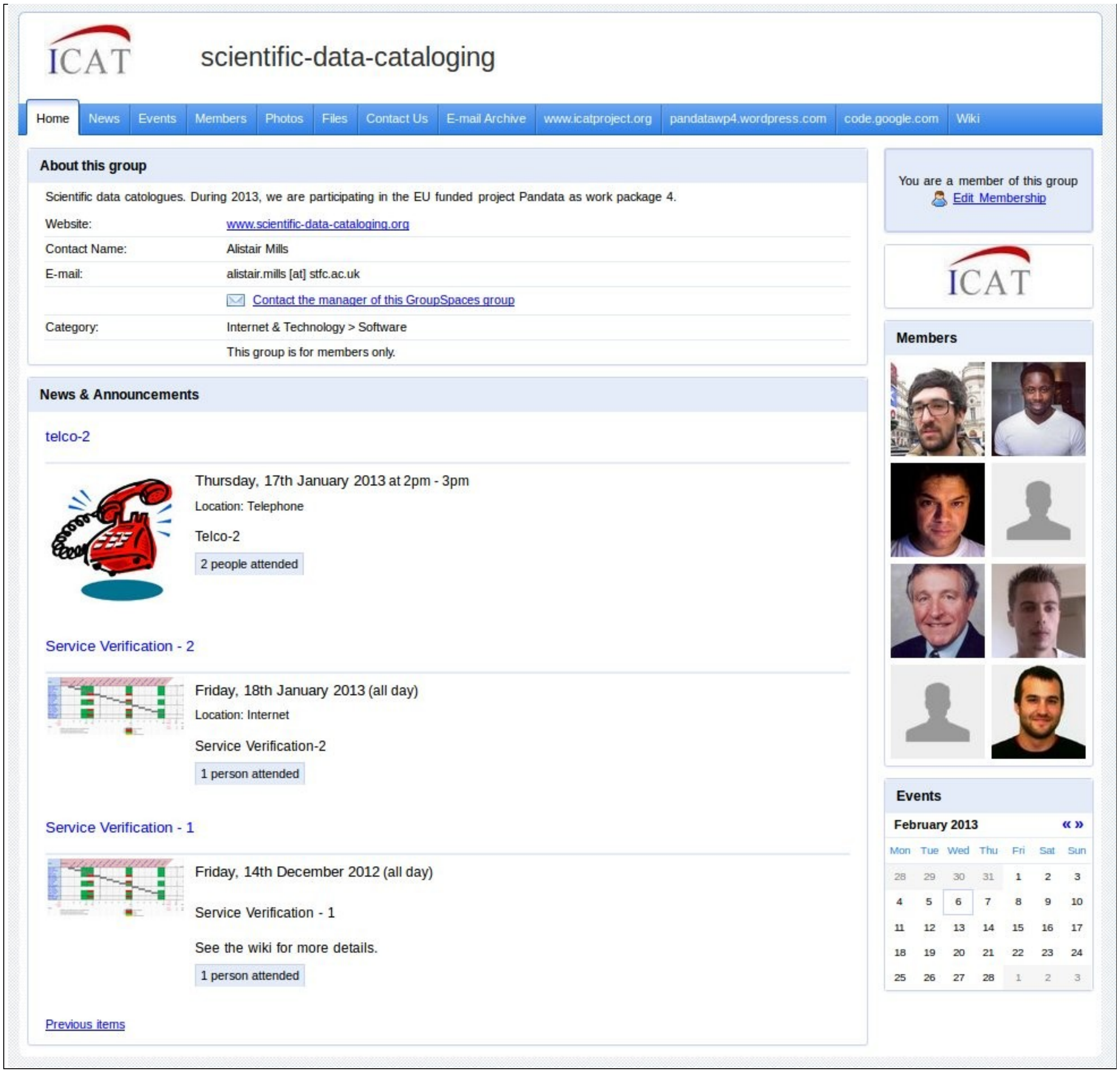
Name	Units	Value
finish_date	yyyy-MM-dd HH:mm:ss	2011-11-24 15:00:16
good_frames	pulses	160917.0
good_proton_charge	uAmp hours	168.893
monitor_sum1	neutrons	3115347.0
monitor_sum2	neutrons	0.0
monitor_sum3	neutrons	0.0
number_of_detectors	N/A	9870.0
number_of_periods	N/A	1.0
number_of_spectra	N/A	7290.0
number_of_time_channels	N/A	5050.0
primary_flight_path	m	17.0
run_duration	seconds	4141.0
run_header	N/A	GEM56235 wak Al Round r...
run_number	N/A	56235.0
run_title	N/A	Al Round robin: 10x10mm R...
short_title	N/A	Al Round robin: 10x10mm

Figure 2: Browsing data with TopCAT

TopCAT is written in Java and makes use of the Sencha GXT Application Framework for the Google Web Toolkit. It makes use of EJB and JPA to store and retrieve configuration and state data. It is able to communicate with different versions of ICAT, these include versions 3.3., 3.4.0, 3.4.1, 4.0, 4.1 and 4.2. However as new functionality is provided it is likely to only be available when communicating with the latest version of ICAT. Currently work is being done to allow users to create their own data sets and upload data. This work has dependencies on ICAT 4.2 and the new ICAT Data Service.

3 Deployment at facilities

ICAT service verification (SV) activities were envisioned as monthly actions, each more complicated than the previous one, in support of the adoption of the ICAT data catalogue in all of the PaNdata consortium facilities. These activities form a step-by-step tutorial for the new adopters of the ICAT and provide feedback to ICAT development team. A Group Spaces web site, scientific-data-cataloging (Figure 2) with login access has been created to share the information regarding the tests.



scientific-data-cataloging

About this group

Scientific data catalogues. During 2013, we are participating in the EU funded project Pandata as work package 4.

Website:	www.scientific-data-cataloging.org
Contact Name:	Alistair Mills
E-mail:	alistair.mills [at] stfc.ac.uk
	Contact the manager of this GroupSpaces group
Category:	Internet & Technology > Software
	This group is for members only.

News & Announcements

telco-2

Thursday, 17th January 2013 at 2pm - 3pm
Location: Telephone

Telco-2

2 people attended

Service Verification - 2

Friday, 18th January 2013 (all day)
Location: Internet

Service Verification-2

1 person attended

Service Verification - 1

Friday, 14th December 2012 (all day)

Service Verification - 1

See the wiki for more details.

1 person attended

You are a member of this group

Members

Events

February 2013 « »

Mon	Tue	Wed	Thu	Fri	Sat	Sun
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	1	2	3

Figure 3: Group Spaces controlled access web site for scientific data catalogues

With service verification actions we aim to:

- increase participation rates each month with the aim of achieving 100% coverage in April 2013;
- maintain and increase the variety of test connection conditions.
- extend tests to include software from other work packages in PaNdata.

3.1 Service Verification 0

The preliminary service verification was carried out on Friday 9th November 2012. The participation in the service verification was excellent with 11 of the 15 partners providing client tests, and four of the partners providing services for test. Both of our associates from outside Europe participated.

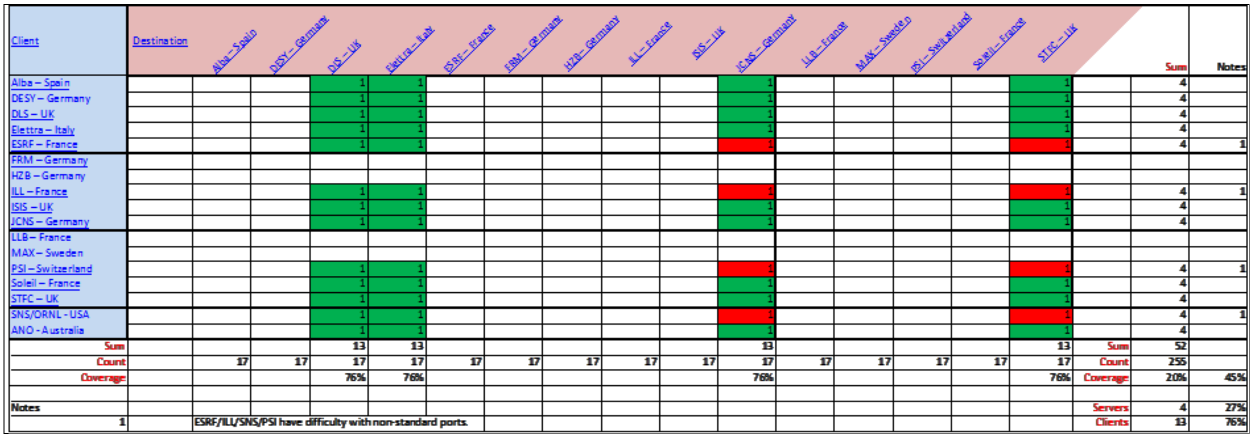
The test involved each client logging on to each of the available servers in turn and logging the session identifier from the server.

The test provided a reasonably large and diverse set of conditions for the tests and included:

- ICAT 4.0, 4.1, 4.2;
- both http and https protocols;
- the db authentication mechanism;
- 11 different firewalls for outbound connections;
- 4 firewalls for inbound connections;
- 2 http connections on ports 8080, 2080, 5080;
- https connections on ports 4081, 8181, 9111;
- connections from 11 countries;
- connections from three continents.

Several of the collaborators ran the tests from home where their connection regime is simpler than within their institution and confirmed that their difficulties in connection were due to their institutions. Summary of the results is shown in Figure 3.

In addition to point-to-point tests, STFC ran 10 continuous instances of the tests for 10 days. This ensured that each of the servers had a base load of 1 connection per second for 10 days. This test passed with 100% success rate.



Client	Destination	Alba – Spain	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		Sum	Notes
Alba – Spain				🟩 1	🟩 1						🟩 1					🟩 1		4	
DESY – Germany				🟩 1	🟩 1						🟩 1					🟩 1		4	
DLS – UK				🟩 1	🟩 1						🟩 1					🟩 1		4	
Elettra – Italy				🟩 1	🟩 1						🟩 1					🟩 1		4	
ESRF – France				🟩 1	🟩 1						🟥 1					🟥 1		4	1
FRM – Germany																			
HZB – Germany																			
ILL – France				🟩 1	🟩 1						🟥 1					🟥 1		4	1
ISIS – UK				🟩 1	🟩 1						🟩 1					🟩 1		4	
JCNS – Germany				🟩 1	🟩 1						🟩 1					🟩 1		4	
LLB – France																			
MAX – Sweden																			
PSI – Switzerland				🟩 1	🟩 1						🟥 1					🟥 1		4	1
Soleil – France				🟩 1	🟩 1						🟩 1					🟩 1		4	
STFC – UK				🟩 1	🟩 1						🟩 1					🟩 1		4	
SNS/ORNL - USA				🟩 1	🟩 1						🟥 1					🟥 1		4	1
ANO - Australia				🟩 1	🟩 1						🟩 1					🟩 1		4	
Sum				13	13						13					13	Sum	52	
Count		17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	Count	255	
Coverage				76%	76%						76%					76%	Coverage	20%	45%
Notes																	Servers	4	27%
1		ESRF/ILL/SNS/PSI have difficulty with non-standard ports.															Clients	13	76%

Figure 4: SV0 results

We had participation rates as follows:

- servers: 27%;
- clients: 76%;
- coverage: 22%.

The lessons learned were the following:

- server services should use standard ports, preferably https:443, alternatively https:8443;
- server services should use certificates from a recognised certificate authority;
- server services should authenticate using their site preferred authentication mechanism.

Many of the clients had difficulties connecting to servers outside their firewall. These difficulties can be eliminated with the use of standard deployment of the services. When the server is appropriately configured, access to the server is handled by the firewall. For example, it is generally not necessary from within a firewall to inform the firewall administrator in order to access secure services such as internet banks.

3.2 Service Verification 1

The first proper service verification tests were run on Dec. 14, 2012. The test material required the servers to run ICAT 4.2. The test included creating an investigation in the ICAT. Remote clients had to log in and search for the facility, investigation, dataset and distinct data files.

The following partners provided servers (7/15):

Facility	ICAT version	Protocol and Port
Alba	4.2	https; port 2081
DLS	4.2	http; port 5080
Elettra	4.2	https; port 8443
ILL	4.2	https; port 443
JCNS	4.2	https; port 5080
Soleil	4.2	https; port 8443
STFC	4.2	http; port 5080

Figure 5: Servers in SV1

The following partners provided clients (12/15):

> Alba – Spain; DESY – Germany; DLS – UK; Elettra – Italy; ESRF – France; ILL – France; ISIS – UK; JCNS – Germany; MAX IV – Sweden; PSI – Switzerland; Soleil – France; STFC – UK.

The following associates provided clients (2/2):

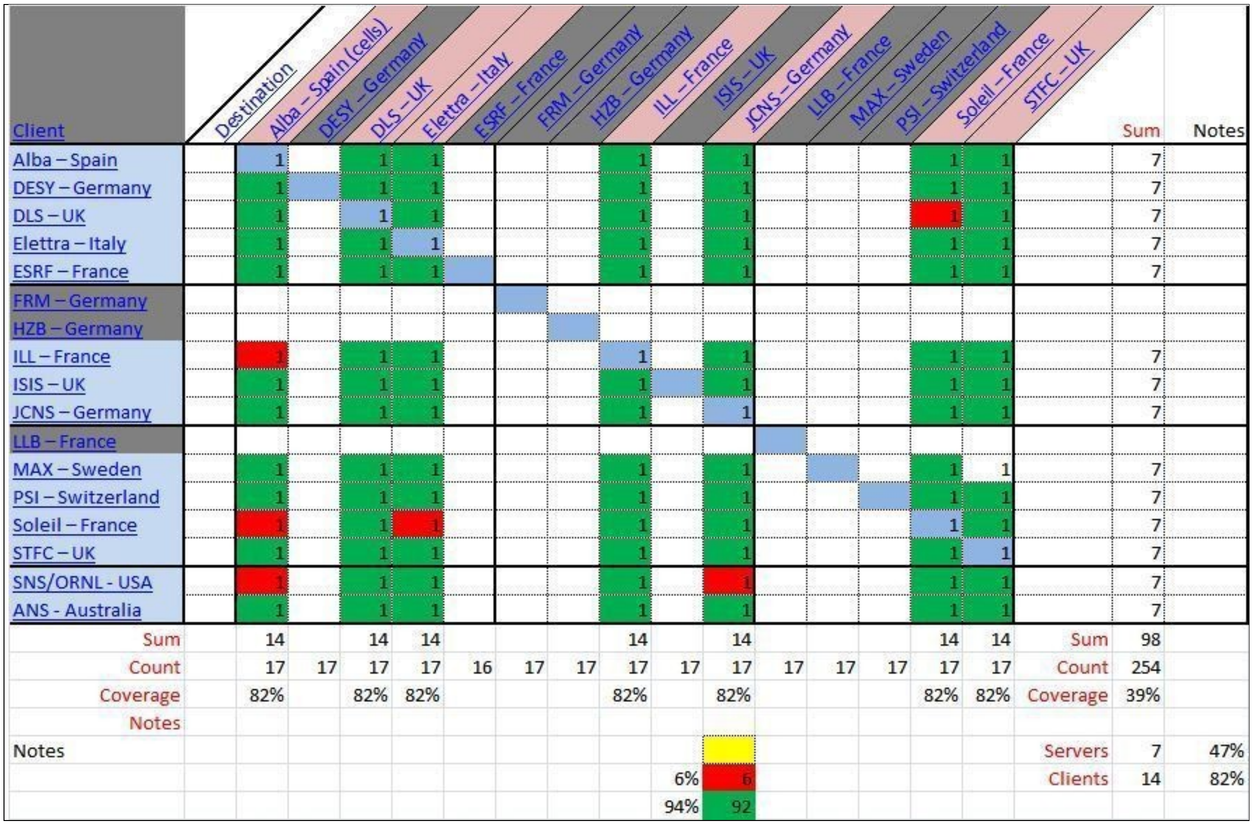
- ANS – Australia;
- SNS/ORNL – USA.

The test involved each client logging on to each of the available servers in turn and logging the session identifier from the server. The test also required that the service providers inject content into the ICAT and that the clients read the content. All of the successful connections received the correct content.

The test provided a reasonably large and diverse set of conditions for the tests and included:

- ICAT 4.2;
- both http and https protocols;
- both db and ldap authentication mechanism;

- 14 different firewalls for outbound connections;
- 7 firewalls for inbound connections;
- http connections on port 5080;
- https connections on ports 443, 2081, 8443;
- connections from 11 countries;
- connections from three continents.



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		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		🟩 1		🟩 1	🟩 1				🟩 1	🟦	🟩 1				🟩 1	🟩 1		7	
JCNS – Germany		🟩 1		🟩 1	🟩 1				🟩 1		🟦 1				🟩 1	🟩 1		7	
LLB – France												🟦							
MAX – Sweden		🟩 1		🟩 1	🟩 1				🟩 1		🟩 1		🟦		🟩 1	1		7	
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		🟥 1		🟩 1	🟩 1				🟩 1		🟥 1				🟩 1	🟩 1		7	
ANS - Australia		🟩 1		🟩 1	🟩 1				🟩 1		🟩 1				🟩 1	🟩 1		7	
Sum		14		14	14				14		14				14	14	Sum	98	
Count		17	17	17	17	16	17	17	17	17	17	17	17	17	17	17	Count	254	
Coverage		82%		82%	82%				82%		82%				82%	82%	Coverage	39%	
Notes																			
Notes											🟨						Servers	7	47%
										6%	🟥 6						Clients	14	82%
										94%	🟩 92								

Figure 6: SV1 results

Several of the collaborators ran the tests from home where their connection regime is simpler than within their institution and confirmed that their difficulties in connection were due to their institutions.

In addition to point-to-point tests, STFC ran 10 continuous instances of the tests for 10 days. This ensured that each of the servers had a base load of 5 connection per second for 10 days. This test passed with 100% success rate.

All of the services correctly injected content in their ICAT.

We had participation rates as follows:

- servers: 47%;
- clients: 82%;
- coverage: 39%.

The following can be observed in the graphic (Figure 6):

- most of the connections were successful;
- some partners had failures connecting to some services.

The lessons learned were the following:

Servers:

- the failures of some connections are still under investigation.

Clients:

- clients should connect from a location representative of a potential user of the ICATs.

Some of the clients had difficulties connecting to servers outside their firewall. These difficulties can be eliminated with the use of standard deployment of the services. When the server is appropriately configured, access to the server is handled correctly by the firewall. For example, it is generally not necessary from within a firewall to inform the firewall administrator in order to access secure services such as internet banks.

3.3 Service Verification 2

The second service verification tests were run on Feb. 1, 2012. It was very similar to the previous one.

The following partners provided servers (7/15):

Facility	ICAT version	Protocol and Port
Alba	4.2	https; port 2081
DLS	4.2	http; port 5080
Elettra	4.2	https; port 8443
ILL	4.2	https; port 443
JCNS	4.2	https; port 5080
Soleil	4.2	https; port 443
STFC	4.2	http; port 5080

Figure 7: Servers in SV2

The following partners provided clients (10/15):

> Alba – Spain; DESY – Germany; DLS – UK; Elettra – Italy; ESRF – France; ILL – France; ISIS – UK; PSI – Switzerland; Soleil – France; STFC – UK.

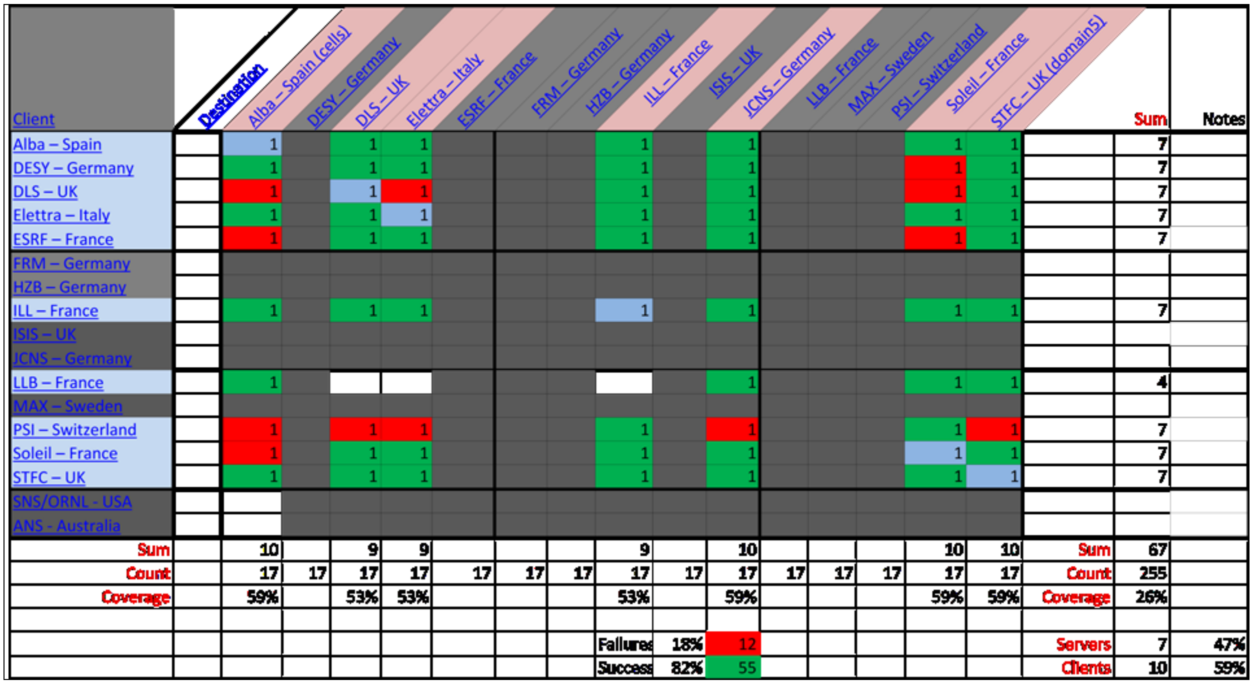
The test involved each client logging on to each of the available servers in turn and logging the session identifier from the server. The test also required that the service providers inject content into the ICAT and that the clients read the content. All of the successful connections received the correct content.

The test provided a reasonably large and diverse set of conditions for the tests and included:

- ICAT 4.2.0, 4.2.1, 4.2.2;
- both http and https protocols;
- both db and ldap authentication mechanisms;
- 10 different firewalls for outbound connections;
- 7 firewalls for inbound connections;
- http connections on port 5080;
- https connections on ports 443, 2081, 8443;
- connections from 9 countries;

Several of the collaborators ran the tests from home where their connection regime is simpler than within their institution and confirmed that their difficulties in connection were due to their institutions.

All of the services correctly injected content in their ICAT.



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 (domain5)		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	17	Count	255	
Coverage		59%		53%	53%				53%		59%				59%	59%	Coverage	26%	
									Failures	18%	🟥 12						Servers	7	47%
									Success	82%	🟩 55						Clients	10	59%

Figure 8: SV2 results

We had participation rates as follows:

- servers: 47%;
- clients: 59%;
- coverage: 26%.

The following can be observed in the graphic (Figure 5):

- most of the connections were successful;
- some partners had failures connecting to some services.

Some of the clients had difficulties connecting to servers outside their firewall. These difficulties can be eliminated with the use of standard deployment of the services. When the server is appropriately configured, access to the server is handled correctly by the firewall. For example, it is generally not necessary from within a firewall to inform the firewall administrator in order to access secure services such as internet banks.

Besides the servers listed in tables 2 and 3, a few other partners (DESY, ESRF) are running ICAT installations that are not yet visible from the outside world due to security constraints on their networks.

3.4 Service Verification 3 and beyond

Considering the outcome of these initial service verification tests we have established a list of goals for the following ones.

Servers:

- ensure that genuine authentication mechanisms are in use;
- ensure that the servers are configured as production services with recognised certificates on standard ports;
- provide representative data in the ICAT;
- remove exceptions which have been added to the firewall for the service verifications.

Clients:

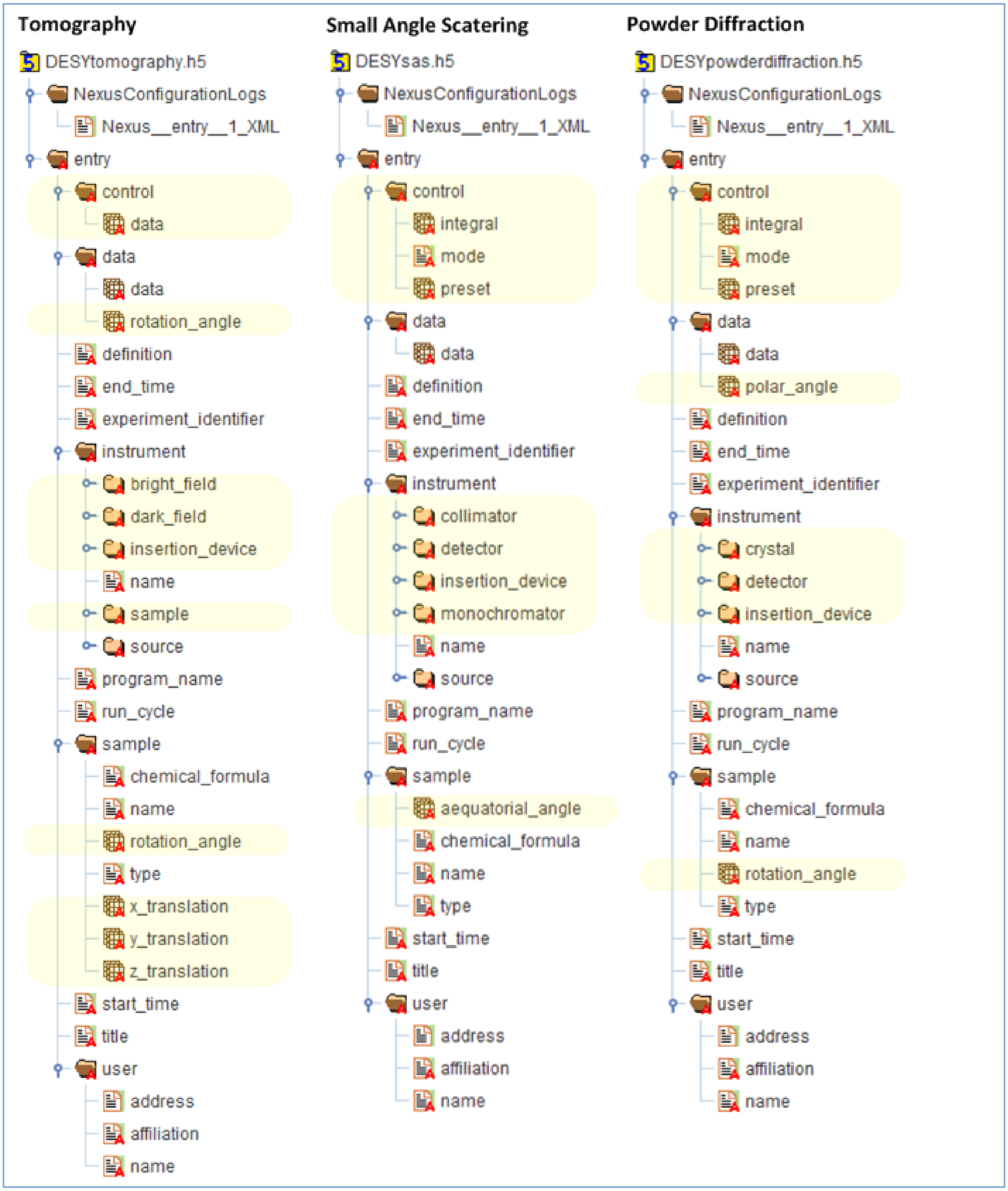
- deploy a Topcat to view the data;
- connect from the work place using standard networking connections;
- remove exceptions which have been added to the firewall for the service verifications.

4 Virtual Labs files and ICAT ingestion

The volume of scientific data is ever increasing. Issues beyond that of archiving are of high importance. Such issues concern the I/O speed, security, cataloguing, provenance, privacy, and common data formats. The latter is of particular interest in the context of PaNdata ODI as it can enable easier data sharing and collaborative research. In practice a common data format acts a standard that enables data exchange among the different facilities and utilisation of relevant services such as data catalogues and data analysis software.

The Virtual Labs (VLabs) of WP5 is the pilot case for PaNdata ODI for scientific data pipelines in three different fields: 1. Tomography, 2. Small Angle Scattering and 3. Powder Diffraction. The rest of the participating facilities, including those of Service Verification 0-1-2, have similar scientific applications. The successful management of VLabs data assures that the chosen approach can be applied to the rest of the partner facilities.

VLabs in DESY have chosen an HDF5 based format which is in line with the suggestions and scope of PaNdata. The file format aims at NeXus compliance. Further information are provided in the deliverables of WP5. The figure bellow (Fig.6) illustrates the main structure of these files including certain differences.



Tomography

```
DESYtomography.h5
  NexusConfigurationLogs
    Nexus__entry__1_XML
  entry
    control
      data
    data
      data
      rotation_angle
    definition
    end_time
    experiment_identifier
    instrument
      bright_field
      dark_field
      insertion_device
      name
      sample
      source
    program_name
    run_cycle
    sample
      chemical_formula
      name
      rotation_angle
      type
      x_translation
      y_translation
      z_translation
    start_time
    title
    user
      address
      affiliation
      name
```

Small Angle Scatering

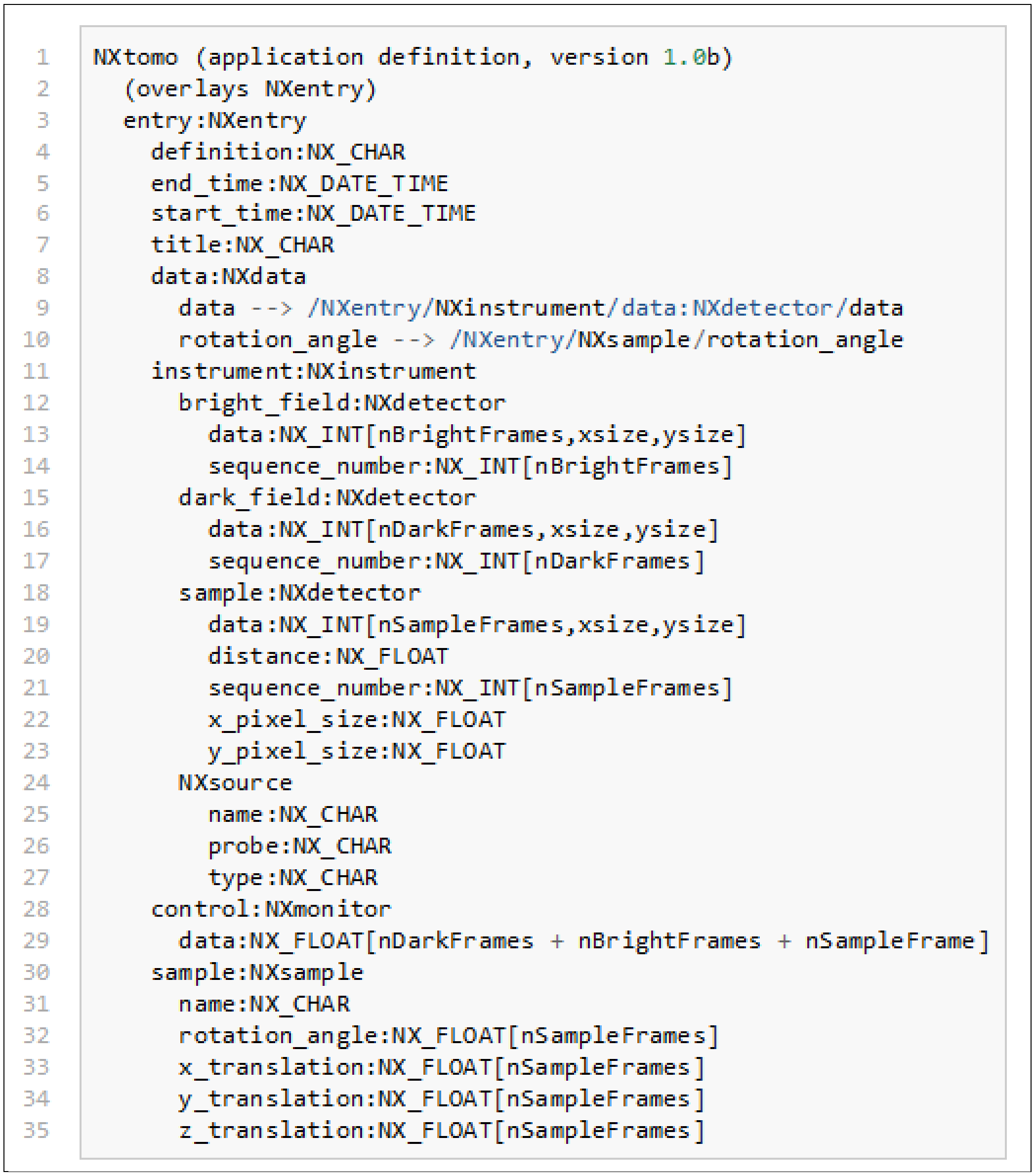
```
DESYsas.h5
  NexusConfigurationLogs
    Nexus__entry__1_XML
  entry
    control
      integral
      mode
      preset
    data
      data
    definition
    end_time
    experiment_identifier
    instrument
      collimator
      detector
      insertion_device
      monochromator
      name
      source
    program_name
    run_cycle
    sample
      aequatorial_angle
      chemical_formula
      name
      type
    start_time
    title
    user
      address
      affiliation
      name
```

Powder Diffraction

```
DESYpowderdiffraction.h5
  NexusConfigurationLogs
    Nexus__entry__1_XML
  entry
    control
      integral
      mode
      preset
    data
      data
      polar_angle
    definition
    end_time
    experiment_identifier
    instrument
      crystal
      detector
      insertion_device
      name
      source
    program_name
    run_cycle
    sample
      chemical_formula
      name
      rotation_angle
      type
    start_time
    title
    user
      address
      affiliation
      name
```

Figure 9: Virtual Labs (DESY) HDF5 formats (as of Jan,30 2013) for cataloguing of Tomography, Small Angle Scattering, and an Powder Diffraction. The main structural differences are highlighted

Parts of this structure are based on a standard class of application definition provided by NeXus. In the case of tomography is the NXtomo class which is defined in standard XML[1]. In a similar way SAS and powder diffraction can be based on the NeXus classes NXsas and NXmonopd.



```
NXtomo (application definition, version 1.0b)
  (overlays NXentry)
  entry:NXentry
    definition:NX_CHAR
    end_time:NX_DATE_TIME
    start_time:NX_DATE_TIME
    title:NX_CHAR
    data:NXdata
      data --> /NXentry/NXinstrument/data:NXdetector/data
      rotation_angle --> /NXentry/NXsample/rotation_angle
    instrument:NXinstrument
      bright_field:NXdetector
        data:NX_INT[nBrightFrames,xsize,ysize]
        sequence_number:NX_INT[nBrightFrames]
      dark_field:NXdetector
        data:NX_INT[nDarkFrames,xsize,ysize]
        sequence_number:NX_INT[nDarkFrames]
      sample:NXdetector
        data:NX_INT[nSampleFrames,xsize,ysize]
        distance:NX_FLOAT
        sequence_number:NX_INT[nSampleFrames]
        x_pixel_size:NX_FLOAT
        y_pixel_size:NX_FLOAT
      NXsource
        name:NX_CHAR
        probe:NX_CHAR
        type:NX_CHAR
    control:NXmonitor
      data:NX_FLOAT[nDarkFrames + nBrightFrames + nSampleFrame]
    sample:NXsample
      name:NX_CHAR
      rotation_angle:NX_FLOAT[nSampleFrames]
      x_translation:NX_FLOAT[nSampleFrames]
      y_translation:NX_FLOAT[nSampleFrames]
      z_translation:NX_FLOAT[nSampleFrames]
```

Figure 10: Outline of the tomography class as defined in NeXus and used by the DESY VLabs format

The definition of the structure of these files is very important. The ingestion of such files in a data catalogue such as ICAT requires a parsing stage. During parsing, the software involved has to traverse the file and extract from the abovementioned structure the required metadata that need to be ingested in the database of ICAT. The two core APIs that are involved are those for NeXus/HDF5 access and that of ICAT.

[1] http://svn.nexusformat.org/definitions/trunk/applications/NXtomo.nxdl.xml

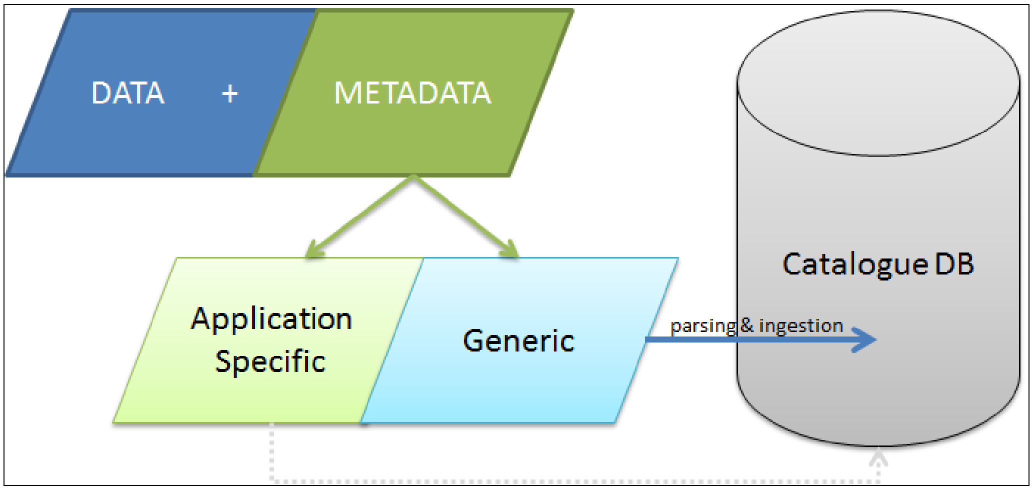


Figure 11: The VLab CT, SAS, and Powder diffraction format contains both data and metadata. Part of the metadata are common among the different applications (i.e. User name) and are ingested in the data catalogue

The current version of the ingestion system for the Virtual Labs data is based on a Python code written by Shelly Ren from Oak Ridge National Laboratory (ORNL). It uses the nexus module for the NeXus binding[2] and Suds[3] as a SOAP python client. Suds is used for accessing the SOAP-based ICAT API since there is no direct Python binding for ICAT 4.2. The roadmap of version 4.3 includes a Python binding. Other than the NeXus binding used, alternative solutions could be that of generic HDF5 modules like H5PY and PyTables[4] for Python based ingestion software. The current version of the ingestion system can be downloaded[5] from the SVN of the ICAT project.

[2] NeXus binding: http://trac.nexusformat.org/code/browser/branches/4.2/bindings/python/nxs?order=name

[3] Suds: https://fedorahosted.org/suds/

[4] H5PY: http://code.google.com/p/h5py/ and PyTables: http://www.pytables.org/

[5] NeXus Ingestion: https://code.google.com/p/icatproject/source/browse/contrib/scripts/ingestNexus

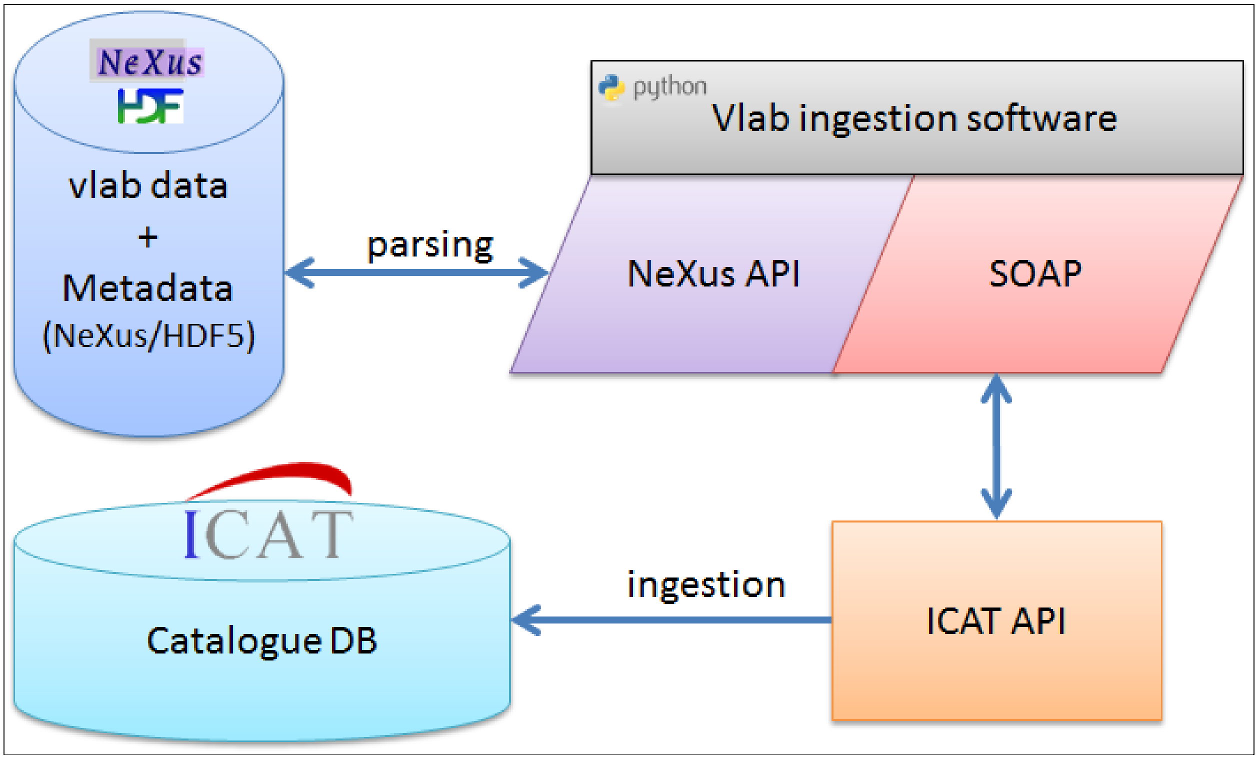


Figure 12: Architecture of the ingestion system for VLab NeXus data into an ICAT data catalogue

Alternative architectures for the ingestion system are possible. The one described here is an off-line approach as it expects all the required metadata to be present in the file. In other scenarios these metadata could be present in the facilities' DBs (i.e. querying user information from the experiment proposal DB) and the data acquisition systems (i.e. collecting metadata by querying a TANGO server). The future work will be towards i) an improved ingestion pipeline which is going to be included in future Service Verifications of ICAT (SV-3 or 4), ii) a study and better definition of the required metadata between ICAT and the VLab requirements (WP5 VLAbs and WP6 Data Provenance).

4.1 Plan for populating the catalogues with data from the Virtual Labs

Through the service verification actions and the collaboration with the ICAT project, the WP4 has achieved a number of goals so far:

- evaluation of data catalogues and confirmation of ICAT choice;
- seven partners have ICAT installations that are ready for ingesting data;
- established a way of collaborating on distributed tests;

- a representative sample of three data types in Nexus format.

The next steps in our planned work activity are:

- analyse how to map the Nexus files into ICAT (STFC);
- develop an implementation of the ingestion process (STFC);
- test the deployment of the ingestion process (ELETTRA);
- provide further examples of the Nexus files (WP5);
- verify operations of ingestion in a future service verifications (ALL).

The estimated amount of work that the abovementioned steps require is:

- analysis - 1 week (minimum) per application definition;
- development - 1 week per application definition;
- testing - 1 week per application definition.

The estimates, when verified by the future work, can be used as indicators of effort and costs for ingesting new data types by each partner facility.

In order to achieve early success, we should select the application definition which appears to be the simplest as the first. It will take 3 weeks before the ingestion is ready for testing by the partners. We plan that the three activities will be carried out by different people, so that it is possible to complete all three data types in 5 weeks. (See Figure 13.)

Week	1	2	3	4	5	6	7
Analysis	▨	▨	▨				
Development		■	■	■			
Test			▨	▨	▨		
Verification						■	■
Preparation	▨	▨	▨	▨	▨		
Provision	■	■	■	■	■		

Figure 13: Timetable for populating the ICATs with data from the Vlabs workpackage

It takes about a week to prepare the materials for normal service verification and we give the sites a week to install and test the materials. However, ingesting Nexus files into a data catalogue requires that the ingestion site has HDF5 and Nexus software installed and operating in a manner which is compatible with the requirements of the ingestion pipeline. If the site is new to HDF5 and Nexus this usually takes about a week. The partners should prepare for this as soon as possible. The next stage of this work requires a large number of datasets produced by the VLabs (WP5).

5 Summary

In the D4.1 we have validated the choice of ICAT as the preferred metadata catalogue for the PaNdata consortium. In this document we describe the activities undertaken to support the deployment of ICAT at our facilities and present a plan for populating the ICATs with actual files from the three VLabs of the WP5. A general architecture for an NeXus/HDF5 ICAT ingestion system is presented and we outline the certain issues regarding the metadata set. This will motivate the forthcoming WP4 activities but also work in WP5 (VLabs) and WP6 (data provenance). The direction of the effort is towards writing data ingestion systems for the scientific instrumentation at a number of PaNdata member facilities. ISIS, DLS and ILL have had ICAT catalogues in production for a past couple of years and are now migrating from ICAT 3 to the latest version 4.2.