PaNdata-ODI

Deliverable D4.3

Deployment of cross-facility metadata searching

Grant Agreement Number	RI-283556
Project Title	PaNdata Open Data Infrastructure
Title of Deliverable	Deployment of cross-facility metadata searching
Deliverable Number	D4.3
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Demonstrator
Contractual Delivery Date	01 Jul 2013 (Month 21)
Actual Delivery Date	18 October 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 work reported in this deliverable summarises the data cataloguing activities of the project till the current stage. These include data ingestion, cross-facility searching and partial integration of services among multiple facilities. Regarding data-ingestion we developed a set of tools for the cataloguing of VLab NeXus datasets. This required the generation of a large number of files that served as a proof of concept that the system can scale in real world cases. The cross-facility search has been designed as a cross-ICAT query results aggregator. It is simple, effective and high-performing as it queries ICATs in parallel while the results are actual ICAT objects instances. A substantial effort has been put on the Service Verifications where multiple facilities of the PaNdata consortium participate on deployment and testing of various services in the context of data cataloguing.

Keyword list

Data catalogue, metadata, data management.

Document approval

Approved for submission to EC by all partners on Oct. 18, 2013.

Revision history

Issue	Author(s)	Date	Description
0.1	Milan Prica	26 July 2013	First draft version
0.2	Milan Prica, George Kourousias	30 Aug. 2013	Summary of Service Verifications Reports
0.3	Milan Prica, George Kourousias	11 Oct. 2013	Data-ingestion library and cross-facility searching
0.4	Milan Prica, George Kourousias	16 Oct. 2013	Corrections
1.0		17. Oct. 2013	Final version

Acknowledgements: Alistair Mills (STFC)

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.

PaNdata WP4 is aiming to provide use cases and requirements for the ICAT development to produce a product which meets the functionality of the proposed virtual labs. Important tasks of PaNdata WP4 include the deployment of ICAT and service verification; this work is progressing well. Currently, ICAT is in production at ILL, ISIS and DLS and several labs (e.g. ALBA, DESY, ELETTRA and ESRF) have deployed ICAT prototype instances, which have been tested during service verification.

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, an open source meta-data management system designed for large facilities. ICAT development is a collaboration involving ISIS, the Scientific Computing Department of STFC, ILL and the Diamond Light Source. In the D4.2 the main focus was on the deployment of ICAT in a number of partner facilities. Detailed reports on service verification actions conducted to enable the cross facility data accessibility were included. Additionally, we discussed issues related to the NeXus sample files from VLabs (WP5's Virtual Laboratories) and presented a plan for their ingestion into the ICATs.

This document is linked to the task 4.3: Provide remote API access to the individual catalogues and integrate to provide a single search capability across the collaborating facilities. A highly customizable, generic set of tools has been developed for the ingestion of NeXus/HDF-5 files into ICAT catalogues. The set includes a library for fast generation of VLab compatible files and a module for searching over multiple catalogue instances. Furthermore, we report on the work done in deployment of ICAT catalogues in the partner facilities, ingestion of sample data and authentication. Three service verifications were held since the last deliverable to check the progress of the work and we will present the results in detail.

2. Data Ingestion and Cross Searching Software

A mature ecosystem of a Data Cataloging solution like ICAT may include a large set of software tools. Its backbone is a database system and a suitable API. Often a front-end like Topcat[1] is provided as an example. Such a system may be further customized according to the user requirements or inspire the development of domain specific solutions. Finally additional software tools and specialized development libraries may assist both developers and end users. Smaller utilities may be of importance too; the scripts developed for SV1-6 demonstrate this. In this section we focus in two types of tools: a data ingestor[2] and a cross facility searching system. The data ingestor populates the ICAT DB with metadata while the cross facility search permits the execution of queries across multiple facilities and returns the combined results. Due to the nature of such software multiple approaches and systems exist. We present the systems developed by Elettra with focus on the needs of WP4 (i.e. compatibility with VLab files from WP5) and additional requirements of two specific facilities of the PaNdata consortium, the synchrotron Elettra and its Free Electron Laser Fermi.

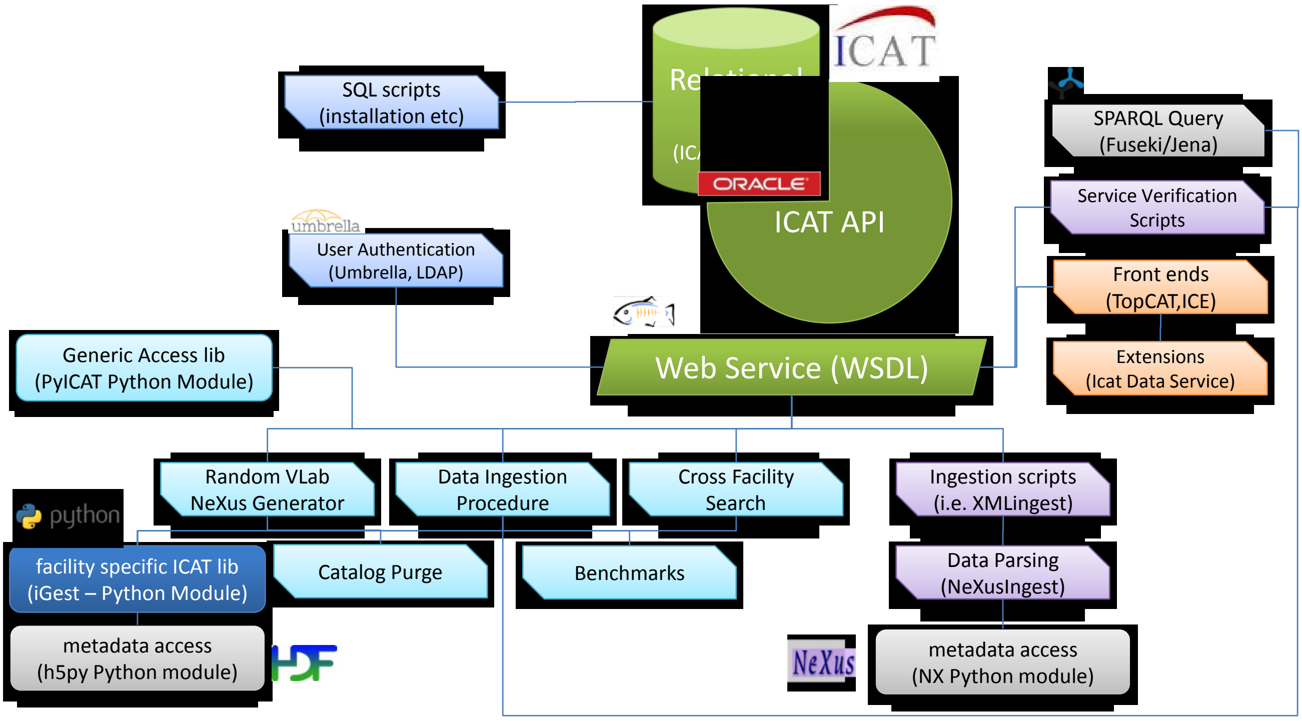


Figure 1: ICAT-based data cataloguing ecosystem

[1] A web based GUI able to search across multiple ICAT instances. http://code.google.com/p/topcat

[2] Software that performs data ingestion.

2.1. Virtual Lab Data Ingestion with iGest

The architecture of a Data Ingestion[3] System has been described in D4.2 and multiple approaches exist for ICAT (i.e. PyICAT[4], nxingest[5], xmlingest[6], NexDaTaS[7]). D4.2 and D4.3 require the ingestion of data produced from the VLabs WP5 as described in D5.2. A suitable system, iGest, is in development[8] that provides various functions to generated VLab files (NeXus[9]) and ingests them into an ICAT. iGest can be used as an external library (as Python module[10]) in larger programs.

The first stage achieved was the generation of a very large number of VLabs files based on a limited number of samples provided by WP5. The files have been used as templates and provided a pool of possible metadata values (e.g. multiple chemical formulas) an iGest method can generate random valid NeXuS HDF5 VLab files. This method returns a dictionary data structure of metadata to be ingested in ICAT. Provided the metadata correspondence between data file and ICAT, the file (actually its metadata) is ingested to ICAT.

The system requires a plain text configuration file with a standard format (RFC 822) `[section] key=value`. It contains 3 sections. `[Template]` pointing to the sample/template data file according to which new random variations are generated. `[h5meta]` contains the keys (paths to HDF5 datasets) with comma delimited potential values. The `[icat]` section suggests the host and login that will be used for the ingestion. The `[icatmeta]` defines the correspondence between the ICAT required tables and the fields in the file that contain them (i.e. `[icatmeta] facility=entry/instrument/source/name`, implies that its value should be found in the H5 dataset name of the H5 group `entry/instrument/source/`). Such tool may be of use to the ICAT project for suitable sample data generation, an observation made during a formal project evaluation[11].

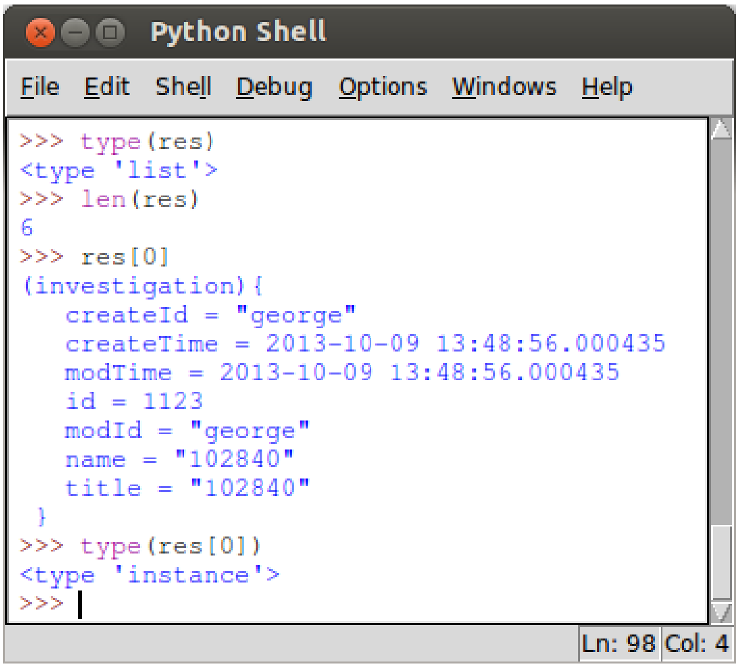
This system allows us to populate multiple ICAT deployments (i.e. different facilities) with thousands of VLab files. Nevertheless iGest has been designed to work with different types of files (not strictly those of VLab) aiming at ingesting existing and heterogeneous HDF5 outputs of data produced by

[3] Data ingestion is the process of obtaining and processing data for storage in a database.
[4] http://apps.jcns.fz-juelich.de/doku/sc/pyicat
[5] https://code.google.com/p/icatproject/source/browse/contrib/scripts/ingestNexus
[6] https://code.google.com/p/icatproject/source/browse/#svn%2Ficat%2Fbranches%2Ficat3-ci%2Ficat3-xmlingest-client
[7] http://code.google.com/p/nexdatas/
[8] Oct.2013 : fully working – essential features - Beta 2
[9] NeXus is a common data format for neutron, x-ray, and muon science. http://www.nexusformat.org
[10] Python file that (generally) has only definitions of variables, functions, and classes.
[11] SO36: "*It is difficult for a new installation to verify an ICAT as there is no sample data in the distributed materials.*" http://www.icatproject.org/index.php/ICAT-R-Software#Future_uptake

ELETTRA and by the FEL FERMI. The architecture has been tested for easy integration with the existing data acquisition pipelines based on the distributed control system TANGO[12].

2.2. Parallel Cross Facility Searching with iGest

An extension of iGest is a module that allows for queries across multiple ICAT instances, thus implementing the cross facility searching. The module is paralellised in multi concurrent processes for high performance. Each process is mapped to an ICAT query and an facility ICAT deployment, and after parallel asynchronous execution, the results are reduced (presented merged). There is sufficient query language[13] for ICAT[14] while the upcoming release v.4.3 will include an advanced platform-independent object-oriented query language[15]. The input is a plain text configuration file like that of the data ingestion system (described above). The `[icat] icatwsdl=` shows the facilities where the query of `[query] icatquery=` is going to be executed.



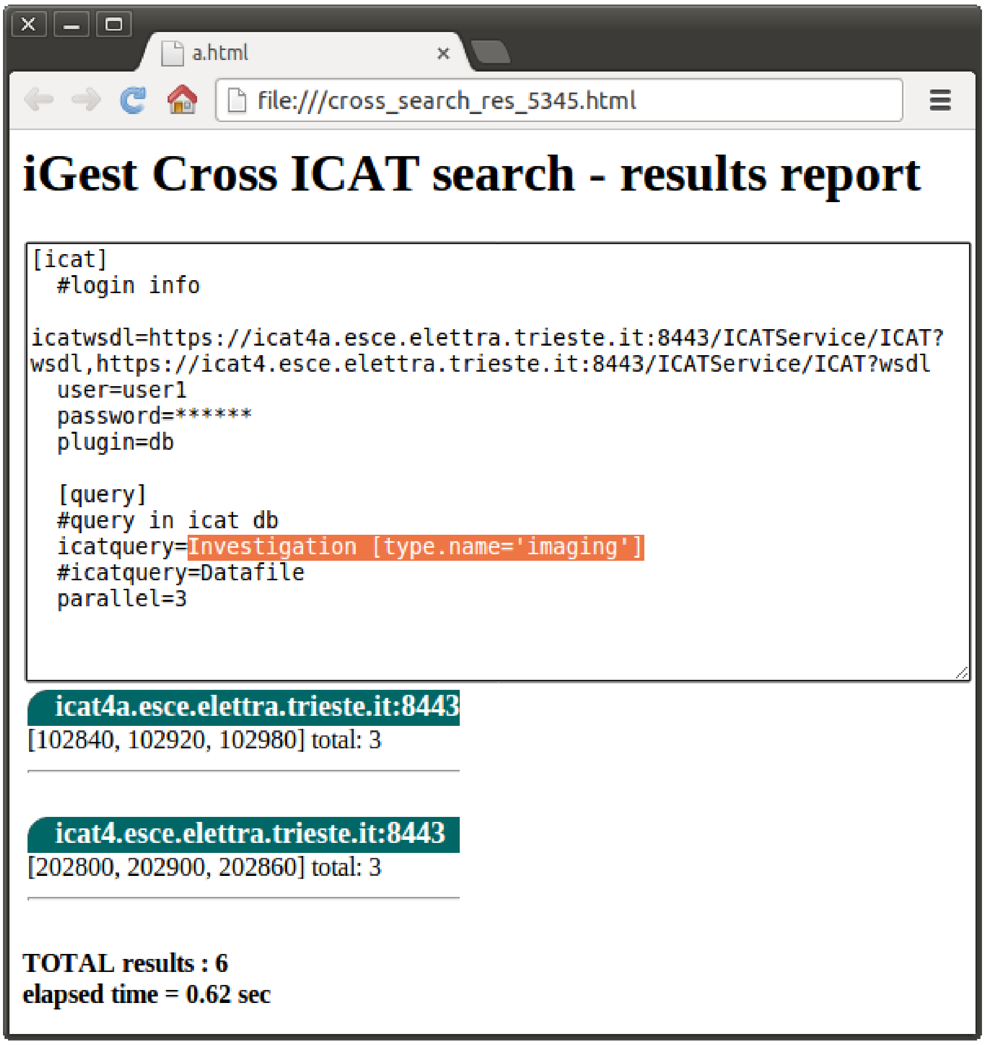
```
>>> type(res)
<type 'list'>
>>> len(res)
6
>>> res[0]
(investigation){
   createId = "george"
   createTime = 2013-10-09 13:48:56.000435
   modTime = 2013-10-09 13:48:56.000435
   id = 1123
   modId = "george"
   name = "102840"
   title = "102840"
 }
>>> type(res[0])
<type 'instance'>
>>> 
```

Figure 2: Cross-facility searching with iGest

The results are complete instances, not just text - thus the module can be integrated in other systems for editing, deleting and other more complex tasks. Preliminary results suggest that

[12] http://www.tango-controls.org
[13] Computer language used to make queries into databases and information systems.
[14] http://www.icatproject.org/mvn/site/icat/4.2.5/icat.client/manual.html
[15] JPQL: http://docs.oracle.com/javaee/6/tutorial/doc/bnbuf.html

parallelism on cross facility searching scales rather well. In depth performance analysis will be the subject of the forthcoming D.4.4.



iGest Cross ICAT search - results report

```
[icat]
  #login info

icatwsdl=https://icat4a.esce.elettra.trieste.it:8443/ICATService/ICAT?
wsdl,https://icat4.esce.elettra.trieste.it:8443/ICATService/ICAT?wsdl
  user=user1
  password=******
  plugin=db

  [query]
  #query in icat db
  icatquery=Investigation [type.name='imaging']
  #icatquery=Datafile
  parallel=3
```

icat4a.esce.elettra.trieste.it:8443

[102840, 102920, 102980] total: 3

icat4.esce.elettra.trieste.it:8443

[202800, 202900, 202860] total: 3

TOTAL results : 6
elapsed time = 0.62 sec

Figure 3: Search results summary

2.3. Follow-up action and preparation for D4.4

Improvements for the iGest modules have been planned. In order to ingest data from current experiments from the facilities Elettra (SRF) and Fermi (FEL) the software has to mature from the current Beta version to production. There is a requirement for importers of archived/past datasets and for integration to the existing control system for providing online ingestion for future data acquisitions. Regarding the cross-facility search, the module should be used by a front end system (like Topcat) to enable end users to perform complex cross facility queries. The aggregated cross facility results are expected to be a challenge for data downloading services like the planned IDS[16]. Eventually the common credentials for such type of data searching should be addressed by the integration of the Umbrella system as described in the tasks of WP3. SV6 will test part of the cross-facility searching and will report its results in D.4.4. Finally for D4.4 it has been planned to extend iGest so that it can provide feedback regarding various performance issues of ICAT (i.e. profiling and benchmarking).

3. VLab Nexus metadata to ICAT mapping

An issue of critical importance in the context of data cataloguing is the establishment of sets for common metadata. Due to the participation of multiple facilities in the PaNdata consortium and the requirement of common cross-facility operations, the agreement on such common metadata names and units is an absolute necessity. The project's WPs (i.e. WP 4, 5, and 6) contribute different aspects to the on-going discussion. Nevertheless it is expected that prior to a final solution, intermediate approaches are necessary in order to provide feedback. Such a solution is the proposal of a minimal yet meaningful (in the terms of cataloguing) metadata set from the Virtual Labs (WP5). These metadata fields should be mapped to the already existing metadata model of the data catalogue ICAT. The current set and the correspondence (VLabs-ICAT) are shown in the table below. Note that the data file format adds to the complexity of the problem as in this case the VLab metadata should be recorded in suitable NeXus paths (HDF5 group followed by datasets or attributes) in a way that maintains compatibility with the standardised NeXus application classes.

Nexus Field	ICAT field	Remarks
/NXentry/experiment_identifier	Investigation	Beamtime unique ID
/NXentry/title	Investigation.Summary	Short description
/NXentry/NXinstrument/name	Instrument.Fullname	Beamline name
/NXentry/NXinstrument/name @short_name	Instrument.Name	Short beamline name
/Nxentry/NXinstrument/NXsource/name	Facility.Fullname	Accelerator name
/Nxentry/NXinstrument/NXsource/name@short_name	Facility.Name	Accelerator short name

[16] https://code.google.com/p/icat-data-service/wiki/IDSMain

/NXentry/NXsample/name	Sample.Name	Sample name
/Nxentry/NXsample/ chemical_forumla	Sampletype.Molecularformula	Chemical formula of the sample
/Nxentry/start_time	Instrument.Start_Date	Start time of the scan
/Nxentry/end_time	Instrument.End_Date	End time of the scan

Table 1: Mandatory metadata mapping

NeXus uses HDF-5[17] files as container files. HDF-5 is a popular scientific data format maintained by the HDF group. While NeXus has been adopted in a number of communities, still the majority of laboratories prefer to save their datasets in customized HDF-5 files. (Such is the case at ELETTRA and FERMI beamlines for example.) However, a flexible ingestor tool for HDF-5 will work just as well with NeXus files as with other HDF-5 files as we prove with the iGest tool.

The WP5 has provided a few sample NeXus files[18] from two Virtual Labs: P03 – Small Angle Scattering (SAS) and P03 – High Resolution Powder Diffraction. We used those as templates to generate an arbitrary number of similar files and populate our ICATs as described in the previous section.

4. Deployment at facilities

ICAT service verification (SV) activities were envisioned as monthly actions of increasing complexity in support of the adoption of the ICAT data catalogue in all of the PaNdata consortium facilities. These activities guide the new adopters of the ICAT, help the testing of new features of the software and provide feedback to ICAT development team.

4.1. Data catalogue

ICAT has been already presented and evaluated in D4.1 and D4.2. The project web site, http://www.icatproject.org, hosts the documentation for the current release of the software while various issues are addressed under the wiki section of the Google code site:

[17] http://www.hdfgroup.org/HDF5/

[18] https://code.google.com/p/pandata/source/browse/#svn%2Ftrunk%2Fdata%2Fp02, https://code.google.com/p/pandata/source/browse/#svn%2Ftrunk%2Fdata%2Fp03

http://code.google.com/p/icatproject. ICAT has many features that fit PaNdata needs, such as the programmable web service interface, the integration of provenance information and the registration of Data Object Identifiers. The developments at the moment include the ICAT Data Service (IDS) and a rich collection of authentication mechanisms. A dedicated web site, http://pandata.org, supports the PaNdata service verifications and the Google code site, http://code.google.com/p/pandata hosts the related code and wiki pages.

4.2. ICAT federations

For the purposes of service verification activities in the WP4, we have divided partners' ICAT installations into three groups (types of ICAT federations) that are described in the following paragraphs.

Bigv federation

A prototype federation using resources on virtual computers takes its name from the cloud which hosts the service. The bigv federation is essentially a test bed for developing and testing the service, and it will be decommissioned at the end of the project. The federation has the following characteristics:

- it has six ICATs called domain5, icat, monash, tng-1, tng-2 and tng-3;
- it is managed and operated by STFC;
- it is hosted on virtual computers provided by bigv.io;
- it has a federated authentication system; all of the nodes accept the same credentials;
- the nodes run a variety of ICATs, but all are ICAT 4.2 or later;
- the nodes have a variety of low grade content using materials used to test ICAT ingestion methods;
- the nodes all have the IDS downloader installed;
- the nodes will evolve over time as new software and services become available.

The bigv federation is active and has been used in service verifications. It serves as a demonstration of the appearance and functionality of a production service. It is however missing the major assets which give the production service value, namely a large volume of good quality data, a reliable authentication and authorization mechanism, and secure access controls.

Development federation

This is a development federation where partners can provide ICATs which are offering a limited service. Most of the partners are participating in the development federation and several of them are expected to move to the production federation when ready. The ICATs in this installation were initially populated with some simple test data and later on with the data from the Virtual Laboratories provided by the WP5.

Production federation

There is a production federation for partners who offer a full ICAT service, including a LDAP or Umbrella authentication and real experimental data.

4.3. Service Verification 3

Service verification 3 was held on April 5th, 2013. It had three parts, the last one a simple questionnaire regarding the deployment intentions of the partners.

Part 1 – Register for ISIS, find and download data

Two persons from each of the partners should have applied to ISIS to be registered as users of the ISIS facility, and then locate some data from the ISIS Topcat. The registration process asks for the institution of the registrant (PaNdata partner institutions are included), and an email address. After an exchange of emails between the User Office and the provided e-mail address, a password is chosen by the registrant and then the registrant is authorized for reading public data in the ISIS catalogue. For the service verification, the participants should have answered to the following question: “What are the names of the NeXus files associated with the LOQ instrument, during cycle_12_5 of the Investigation with title “Aluminium6082 brick 10x10x9mm” on 2013-03-15?”

Part 2 – Register as a user of the demonstrator ICAT and verify access

Users should have registered for access to ICATs on http://www.icatproject.org using a Facebook account (it was recommended to create one for that purpose). Upon registration, the service issued user credentials and inserted them into the authentication database. Users should have tested the new account by login into a Topcat hosted by STFC. This procedure allowed for common db-authentication login into development federation ICATs.

Part 3 – Provide information on your institute's ICAT plan

Partners should have supplied the following information by email.

Name of the institute:
The resource available for this work (0%-100%):
The name of the principal contact:
The grade of ICAT deployment(0-3):
The grade of certificate (0-3):
The grade of authentication (0-3):
The grade of content (0-3):

SV Results

There were three parts to the test, and each of the 12 partners did all three parts: Alba – Spain; DESY – Germany; DLS – UK; Elettra – Italy; ESRF – France; HZK – Germany; ILL – France; ISIS – UK; JCNS – Germany; LLB – France; MAX – Sweden; PSI – Switzerland; Soleil – France. Associate partners were not involved in this service verification.

Part 1 – Register for ISIS, find and download data

All of the partners correctly answered the question about the contents of an investigation in ISIS. This shows that they succeeded in obtaining an ISIS credential, finding the data with Topcat and downloading the data. A number of partners noted minor usability matters; for example the following was noted by Leonardo Sala of PSI:

– if you put a wrong date in the "Advanced Search" mode it searches anyway, and finds everything; this can in principle kill your server, if enough data is stored in ICAT;
– the date format is not iso (e.g. 2013-04-05) but US-style (middle-endian[19]), which I found confusing;
– if I look for "Aluminium", "Aluminium6082", "Aluminium6082 brick 10x10x9mm" in Proposal Title in the "Advanced Search" mode, I cannot find anything. If I remove that condition, and just use the date, "Aluminium6082 brick 10x10x9mm_TRANS" appears in the list. If I simply put

[19] Refers to the American mm/dd/yy style of writing dates.

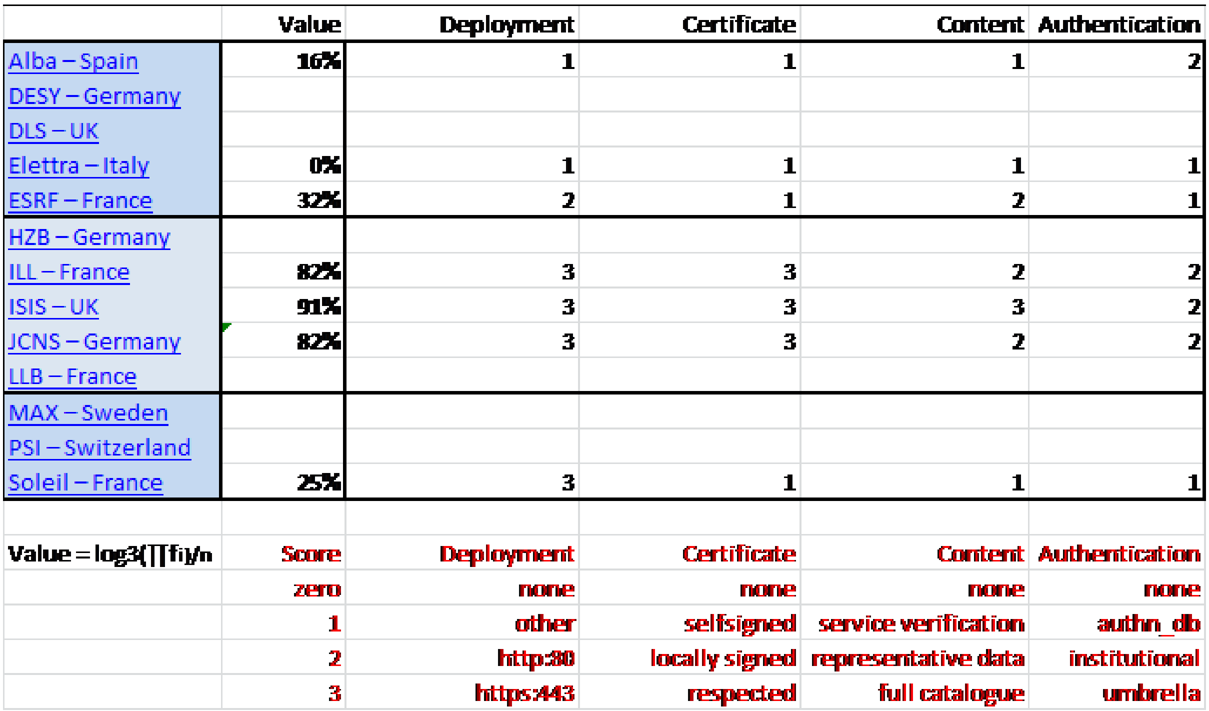
“Aluminium6082″ into the keywords, it works. So, it seems there is some issue with “Proposal Title” search.

Part 2 – Register as a user of the demonstrator ICAT and verify access

All of the partners succeeded with this task. Many of the partners expressed their concerns about the use of Facebook as the authentication mechanism. One partner (HZB) integrated the ICAT authentication service with his experimental ICAT and reported success.

Part 3 – Provide information on your institute’s ICAT plan

All of the sites provided the required information. Many asked if they were to answer the questions for their current ICAT setup, or for the future deployment. The answer is both. The results of the survey are shown in the Fig.4. The value function is a heuristic which provides a value to the ICAT deployment. It provides a simple way to combine four scores into one. The value of 100% can only be achieved with a score of 3/3 in each of the four categories. A score of 50% is required for the service to be considered as a production service. A score of 0% is achieved by an ICAT which has a score of at least one for all four categories. A score of zero for any of the four categories inhibits the ICAT for inclusion in the list of viable ICATs. The value should not be considered like the result of an exam; it does not convey failure!



	Value	Deployment	Certificate	Content	Authentication
Alba – Spain	16%	1	1	1	2
DESY – Germany					
DLS – UK					
Elettra – Italy	0%	1	1	1	1
ESRF – France	32%	2	1	2	1
HZB – Germany					
ILL – France	82%	3	3	2	2
ISIS – UK	91%	3	3	3	2
JCNS – Germany	82%	3	3	2	2
LLB – France					
MAX – Sweden					
PSI – Switzerland					
Soleil – France	25%	3	1	1	1
Value = log3(∏fi)/n	Score	Deployment	Certificate	Content	Authentication
	zero	none	none	none	none
	1	other	selfsigned	service verification	authn_db
	2	http:80	locally signed	representative data	institutional
	3	https:443	respected	full catalogue	umbrella

Figure 4: SV3 results

SV Conclusions

This service verification was successful and was well supported by all of the partners. The highlights include:

1) all of the partners now have access to the public data of ISIS and can use the ISIS Topcat to locate and download data;

2) all of the partners now can access any of the ICATs which use the ICAT Authentication service being offered by STFC;

3) We have a basic plan for the availability of ICATs for service integration for the next 12 months.

Catalogue Service

There are three categories of ICAT services on offer:

1) partners not offering a service;

2) partners offering a limited service;

3) partners who are intending to offer a full service.

At the moment, there is no partner in category 3. There are four partners who are close (ISIS, DLS, ILL, and JCNS). There is a further partner with the intention to offer a full service later in 2013

(ESRF). There are a further three partners who have ICATs which are in category 2 (Alba, Elettra and Soleil). Of the five remaining partners, four have expressed an intention to provide an externally visible ICAT (DESY, HZB, MAX and PSI).

The category of ICAT is not intended to indicate status. The more experienced users of ICAT have substantial operating experience, and can offer a full service. All of the partners have legal and policy constraints and these must be taken into account in a deployment decision. However, as a project, we wish to show that we can create a service embracing about half of the partners. This appears to be possible.

In order to achieve a full service, the following are all required:

1) that the service be deployed on https on port 443 and that it be globally visible;
2) that the certificate associated with the service be globally trusted;
3) that the institution can publish a catalogue of scientific value;
4) that the authentication mechanism be trusted;
5) that the service achieves a value of at least 50% using the value function.

4.4. Service Verification 4

Service verification 4 was held on May 10th, 2013. Its primary focus was on using bigv federation both from Topcat and from client application. Additionally, the partners were asked to update their intentions for service provision and prepare for joining the development federation.

The verification had four parts:

Part 1 – Connect to the Topcat for the bigv federation, locate and download data.

Part 2 – Run a script on a computer at your institution. The script locates and downloads data from an ICAT in the bigv federation.

Part 3 – Provide information on your intentions for provision of ICAT service.

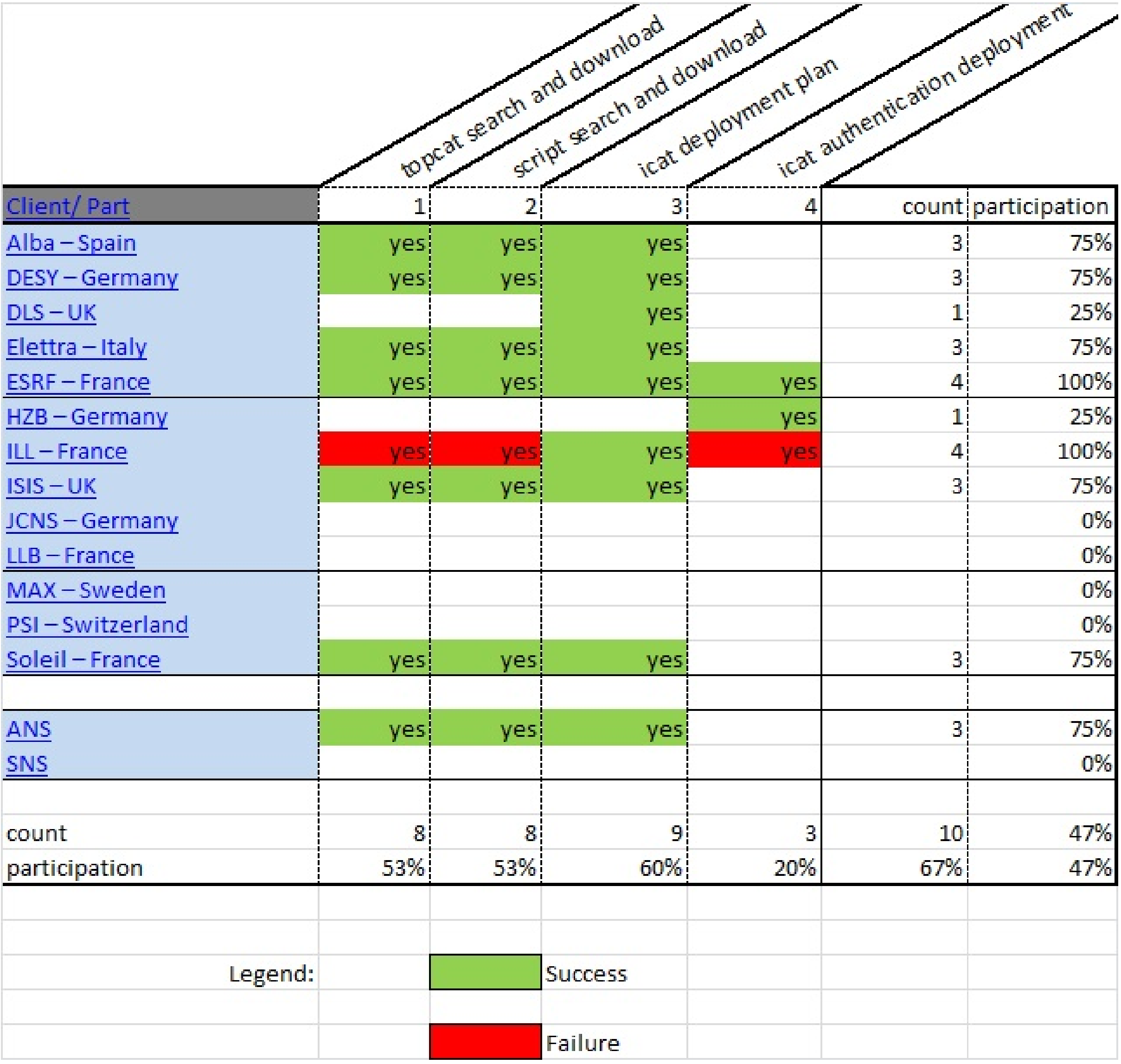
Part 4 – Configure your ICAT to use the ICAT authentication service and verify that it works. This part was optional, aimed at partners intending to participate in the development federation.

SV Results

There were four parts to the test, and nine of the partners provided a response to at least one of the parts: Alba – Spain; DESY – Germany; DLS – UK; Elettra – Italy; ESRF – France; HZB – Germany; ILL – France; ISIS – UK; Soleil – France. There was a response from our associate

ANS in Australia.

The participation is shown in the following figure.



	topcat search and download	script search and download	icat deployment plan	icat authentication deployment		
Client/ Part	1	2	3	4	count	participation
Alba – Spain	yes	yes	yes		3	75%
DESY – Germany	yes	yes	yes		3	75%
DLS – UK			yes		1	25%
Elettra – Italy	yes	yes	yes		3	75%
ESRF – France	yes	yes	yes	yes	4	100%
HZB – Germany				yes	1	25%
ILL – France	🟥 yes	🟥 yes	yes	🟥 yes	4	100%
ISIS – UK	yes	yes	yes		3	75%
JCNS – Germany						0%
LLB – France						0%
MAX – Sweden						0%
PSI – Switzerland						0%
Soleil – France	yes	yes	yes		3	75%
ANS	yes	yes	yes		3	75%
SNS						0%
count	8	8	9	3	10	47%
participation	53%	53%	60%	20%	67%	47%

Legend: 🟩 Success

🟥 Failure

Figure 5: Participation of the partners in SV4

Part 1 – Connect to the Topcat for the bigv federation, locate and download data.

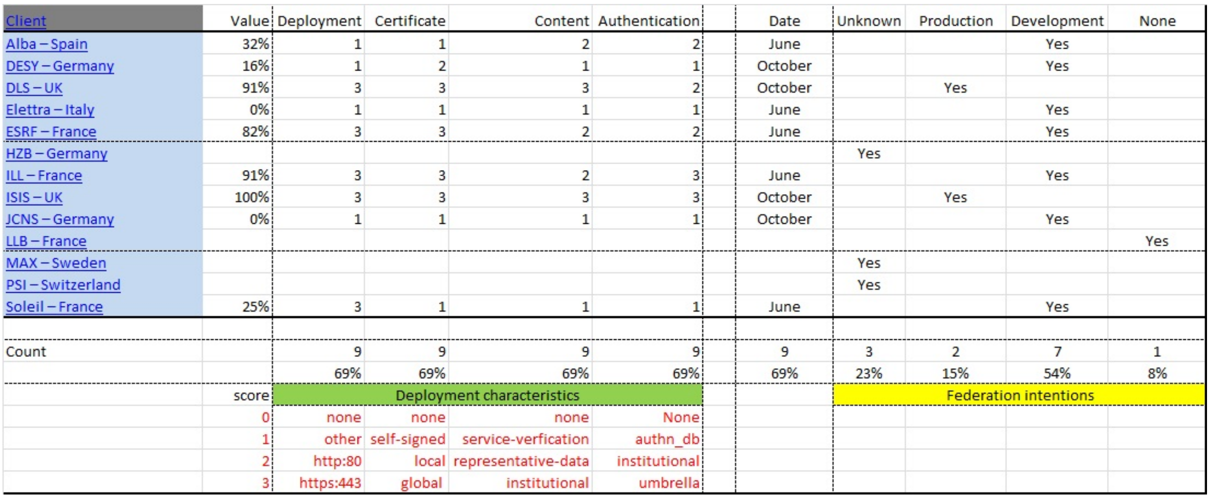
Seven of the 13 partners and one of the associates did this. Only one was unable to make the test work; the firewall at ILL does not allow traffic even on port 8080.

Part 2 – Run a script on a computer at your institution.

The script locates and downloads data from an ICAT in the bigv federation. The results from this part were identical to those for part 1.

Part 3 – Provide information on your intentions for provision of ICAT service.

The results from this part of the test are shown in the following figure.



Client	Value	Deployment	Certificate	Content	Authentication		Date	Unknown	Production	Development	None
Alba – Spain	32%	1	1	2	2		June			Yes	
DESY – Germany	16%	1	2	1	1		October			Yes	
DLS – UK	91%	3	3	3	2		October		Yes		
Elettra – Italy	0%	1	1	1	1		June			Yes	
ESRF – France	82%	3	3	2	2		June			Yes	
HZB – Germany								Yes			
ILL – France	91%	3	3	2	3		June			Yes	
ISIS – UK	100%	3	3	3	3		October		Yes		
JCNS – Germany	0%	1	1	1	1		October			Yes	
LLB – France											Yes
MAX – Sweden								Yes			
PSI – Switzerland								Yes			
Soleil – France	25%	3	1	1	1		June			Yes	
Count		9	9	9	9		9	3	2	7	1
		69%	69%	69%	69%		69%	23%	15%	54%	8%
	score	Deployment characteristics						Federation intentions			
	0	none	none	none	None						
	1	other	self-signed	service-verfication	authn_db						
	2	http:80	local	representative-data	institutional						
	3	https:443	global	institutional	umbrella						

Figure 6: Deployment intentions of the partners

The following partners have yet to make their intentions clear: HZB – Germany; MAX IV – Sweden; PSI – Switzerland. Only one partner has decided not to deploy an ICAT. It is likely that the development federation will have seven partners in October: Alba – Spain; DESY – Germany; Elettra – Italy; ESRF – France; ILL – France; JCNS – Germany; Soleil – France. It is likely that the production federation will have 2 partners in October: DLS – UK; ISIS – UK.

Part 4 – Configure your ICAT to use the ICAT authentication service and verify that it works.

Only three of the partners attempted this test. For two of them, the results were satisfactory. ILL were unable to connect to the database due to a restriction in the ILL firewall.

SV Conclusions

This service verification was successful but it was not well supported by all of the partners. For these tests to be valuable, high participation is required. The lack of participation in the authentication deployment (part 4) could lead to operational difficulties in the future service verifications.

The highlights include:

1) most of the partners can access data on an external ICAT using Topcat to locate and download the data;

2) most of the partners can access data on an external ICAT using a script to locate and download the data;

3) we have a basic plan for the availability of ICATs for service integration for the next 12 months.

4.5. Service Verification 5

Service verification 5 was scheduled on June 21st, 2013. Due to the complexity of the setup and the availability of the partners, the tests were run for entire week, till June 28th.

Goals of the SV5 were:

- To verify that partners can locate ICATs in the development federation using client applications.
- To verify that the partners can use Topcat to find information in an ICAT.
- To verify that partners can access the European Affiliation Database (EAD).
- To verify that the partners providing an ICAT service can deploy the necessary services, ie ICAT, authn, ids and Topcat.

The following tables show the services which were ready for the test, the address of the ICAT, and the address of the Topcat.

Partner	ICAT	Ids	Topcat	Source of credentials	Topcat:icat
DESY	Yes	No	No	http://www.icatproject.org	Not applicable
Elettra	Yes	No	No	http://www.icatproject.org	Not applicable
ESRF	Yes	No	No	http://www.icatproject.org	Not applicable
HZB	Yes	No	Yes	http://www.icatproject.org	HZB
ILL	Yes	No	Yes	http://www.icatproject.org	ILL
JCNS	Yes	No	No	JCNS	Not applicable
Soleil	Yes	No	No	Soleil	Not applicable
STFC	Yes	Yes	Yes	http://www.icatproject.org	Elettra, ESRF, HZB, ILL, STFC

Table 2: Development federation services

Partner	ICAT address
DESY	http://icat.science3d.org:8080/ICATService/ICAT?wsdl
Elettra	https://icat-elettra.grid.elettra.trieste.it:8443/ICATService/ICAT?wsdl
ESRF	https://wwws.esrf.fr/icat/ICATService/ICAT?wsdl
HZB	https://icat.helmholtz-berlin.de/ICATService/ICAT?wsdl
ILL	https://icat.ill.eu/ICATService/ICAT?wsdl
JCNS	https://apps.jcns.fz-juelich.de:5443/ICATService/ICAT?wsdl
Soleil	https://icat.synchrotron-soleil.fr/ICATService/ICAT?wsdl
STFC	http://www.icatproject.org:5080/ICATService/ICAT?wsdl

Table 3: ICAT addresses

Partner	Topcat address
DESY	Not applicable
Elettra	Not applicable
ESRF	Not applicable
HZB	https://icat.helmholtz-berlin.de/TOPCATWeb.jsp
ILL	https://icat.ill.fr/TOPCATWeb.jsp
JCNS	Not applicable
Soleil	Not applicable
STFC	http://www.icatproject.org:5080/TOPCATWeb.jsp

Table 4: Topcat addresses

Client tests

Part 1 – Run scripts on a computer at your institution. The scripts locate data in ICATs in the development federation.

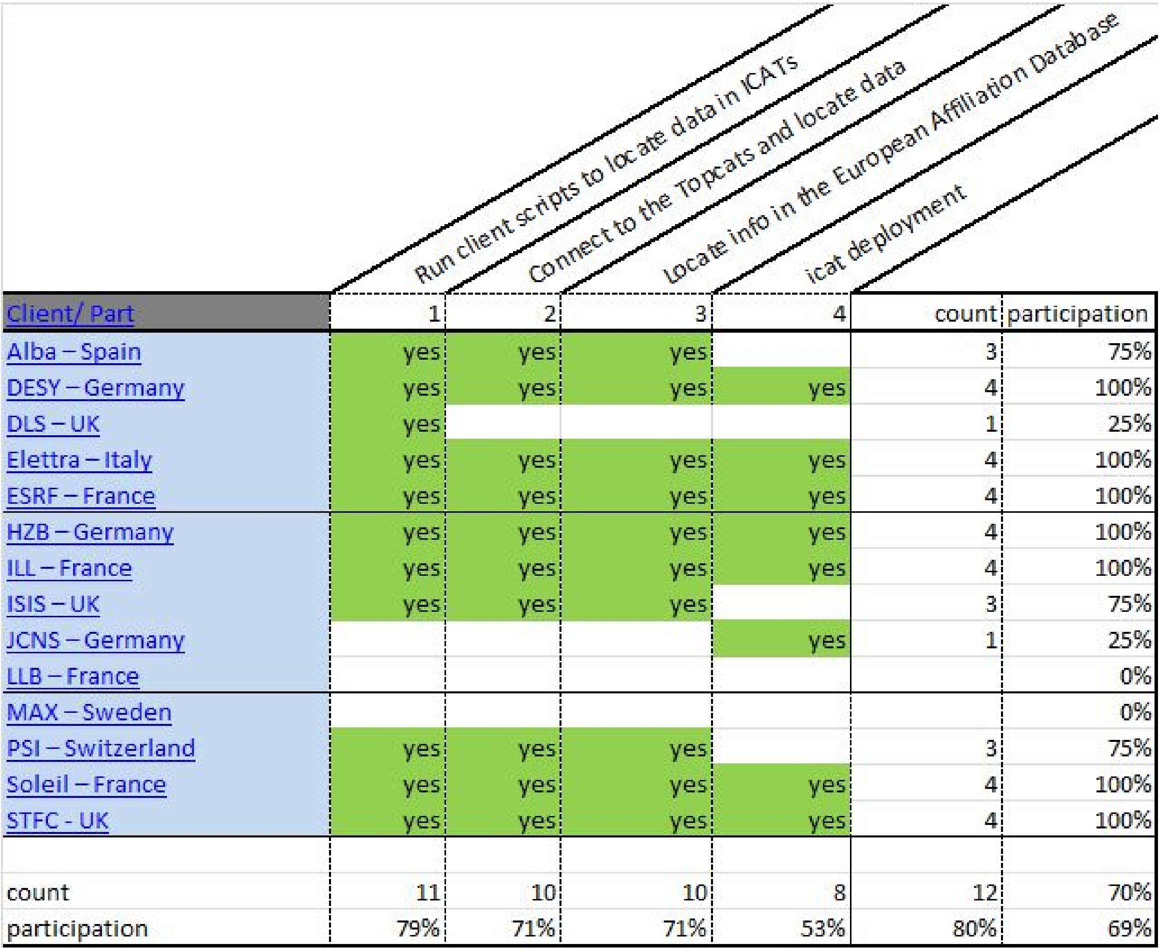
Part 2 – Locate Topcats, logon to their ICAT(s) and locate information.

Part 3 – Run a script which locates information in the European Affiliation Database.

Part 4 – Configure an ICAT for use in the development federation.

Of particular note in this service verification were the ICAT services of DESY, ESRF, and HZB; this is the first time that these services have been part of a service verification. A second notable feature of the service verification was the test of the European Affiliation Database. This is a new service and it performed well. The third innovation in this service verification was the testing of Topcats in the development federation. The success of these Topcats was mixed, and further work will be required to make a production quality service.

There were four parts to the test, and twelve of the fourteen partners provided a response to at least one of the parts: Alba – Spain; DESY – Germany; DLS – UK; Elettra – Italy; ESRF – France; HZB – Germany; ILL – France; ISIS – UK; JCNS – Germany; PSI – Switzerland; Soleil – France. The following partners decided not to take part: LLB – France; MAX IV – Sweden.
The participation is shown in the following figure.



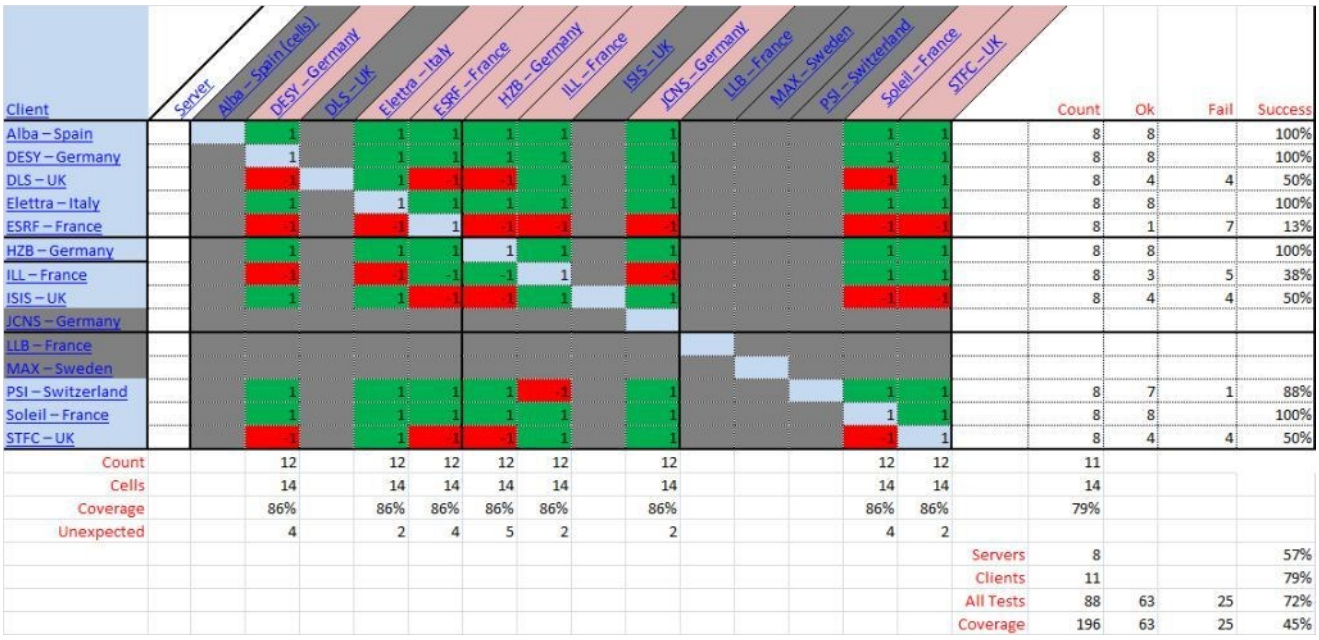
	Run client scripts to locate data in ICATs	Connect to the Topcats and locate data	Locate info in the European Affiliation Database	icat deployment		
Client/ Part	1	2	3	4	count	participation
Alba – Spain	yes	yes	yes		3	75%
DESY – Germany	yes	yes	yes	yes	4	100%
DLS – UK	yes				1	25%
Elettra – Italy	yes	yes	yes	yes	4	100%
ESRF – France	yes	yes	yes	yes	4	100%
HZB – Germany	yes	yes	yes	yes	4	100%
ILL – France	yes	yes	yes	yes	4	100%
ISIS – UK	yes	yes	yes		3	75%
JCNS – Germany				yes	1	25%
LLB – France						0%
MAX – Sweden						0%
PSI – Switzerland	yes	yes	yes		3	75%
Soleil – France	yes	yes	yes	yes	4	100%
STFC - UK	yes	yes	yes	yes	4	100%
count	11	10	10	8	12	70%
participation	79%	71%	71%	53%	80%	69%

Figure 7: SV5 participation

SV Results

Part 1 – Run a client script on a computer at your institution.

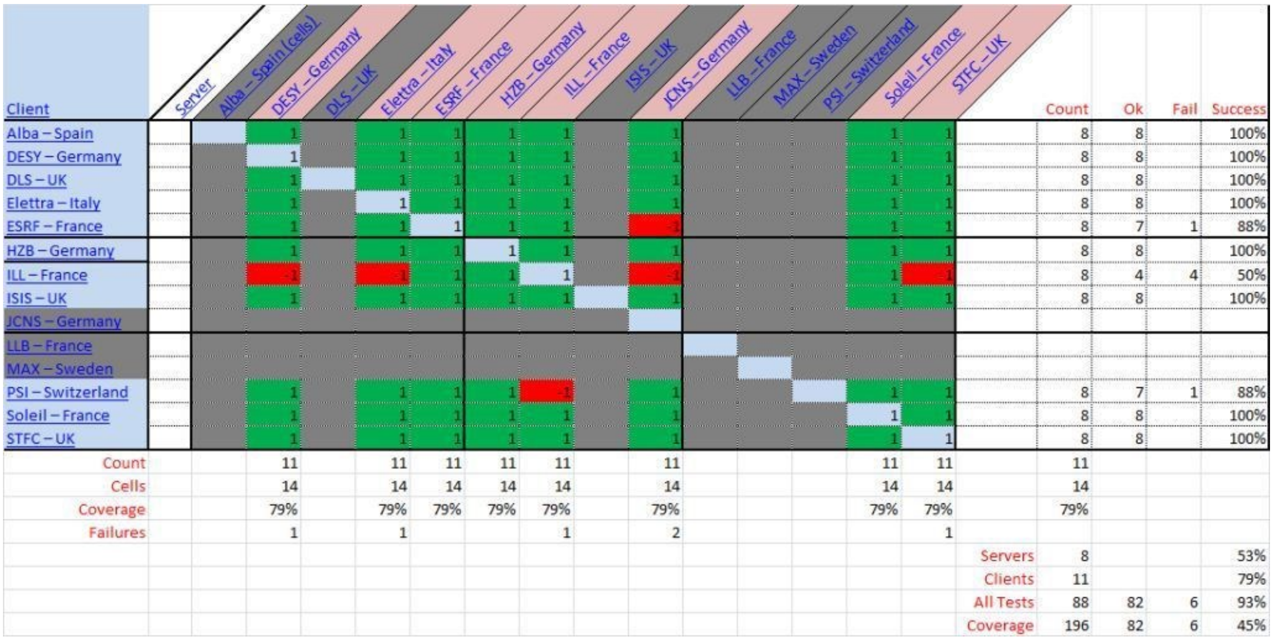
There were two scripts run. Although both scripts were written in bash script, one executed Java client code, and the other executed Python client code. The output from the two scripts was similar. However, the script with the Java programs has been written to make sure that the application is tolerant of security conditions such as a firewall and self signed certificates. The scripts with the Python applications are not tolerant to such conditions.



Client	Server	Alba – Spain (cells)	DESY – Germany	DLS – UK	Elettra – Italy	ESRF – France	HZB – Germany	ILL – France	ISIS – UK	JCNS – Germany	LLB – France	MAX – Sweden	PSI – Switzerland	Soleil – France	STFC – UK		Count	Ok	Fail	Success
Alba – Spain			1		1	1	1	1		1				1	1		8	8		100%
DESY – Germany			1		1	1	1	1		1				1	1		8	8		100%
DLS – UK			-1		1	-1	-1	1		1				-1	1		8	4	4	50%
Elettra – Italy			1		1	1	1	1		1				1	1		8	8		100%
ESRF – France			-1		-1	1	-1	-1		-1				-1	-1		8	1	7	13%
HZB – Germany			1		1	1	1	1		1				1	1		8	8		100%
ILL – France			-1		-1	-1	-1	1		-1				1	1		8	3	5	38%
ISIS – UK			1		1	-1	-1	1		1				-1	-1		8	4	4	50%
JCNS – Germany																				
LLB – France																				
MAX – Sweden																				
PSI – Switzerland			1		1	1	1	-1		1				1	1		8	7	1	88%
Soleil – France			1		1	1	1	1		1				1	1		8	8		100%
STFC – UK			-1		1	-1	-1	1		1				-1	1		8	4	4	50%
Count			12		12	12	12	12		12				12	12		11			
Cells			14		14	14	14	14		14				14	14		14			
Coverage			86%		86%	86%	86%	86%		86%				86%	86%		79%			
Unexpected			4		2	4	5	2		2				4	2					
																Servers	8			57%
																Clients	11			79%
																All Tests	88	63	25	72%
																Coverage	196	63	25	45%

Figure 8: SV5 Analysis-Python

The results from the Java tests show that changes are required to the deployments at DESY, Elettra, JCNS and STFC for these services to be usable by all partners. This will be investigated before the next service verification. The extra failures which are seen in the Python tests are due to the tests being run from behind firewalls. The Python tests do not deal with the firewall. This shows that the Python test programs must be altered to deal with the firewall. This will be investigated before the next service verification.



Client	Server	Alba – Spain (cells)	DESY – Germany	DLS – UK	Elettra – Italy	ESRF – France	HZB – Germany	ILL – France	ISIS – UK	JCNS – Germany	LLB – France	MAX – Sweden	PSI – Switzerland	Soleil – France	STFC – UK		Count	Ok	Fail	Success
Alba – Spain			1		1	1	1	1		1				1	1		8	8		100%
DESY – Germany			1		1	1	1	1		1				1	1		8	8		100%
DLS – UK			1		1	1	1	1		1				1	1		8	8		100%
Elettra – Italy			1		1	1	1	1		1				1	1		8	8		100%
ESRF – France			1		1	1	1	1		-1				1	1		8	7	1	88%
HZB – Germany			1		1	1	1	1		1				1	1		8	8		100%
ILL – France			-1		-1	1	1	1		-1				1	-1		8	4	4	50%
ISIS – UK			1		1	1	1	1		1				1	1		8	8		100%
JCNS – Germany																				
LLB – France																				
MAX – Sweden																				
PSI – Switzerland			1		1	1	1	-1		1				1	1		8	7	1	88%
Soleil – France			1		1	1	1	1		1				1	1		8	8		100%
STFC – UK			1		1	1	1	1		1				1	1		8	8		100%
Count			11		11	11	11	11		11				11	11		11			
Cells			14		14	14	14	14		14				14	14		14			
Coverage			79%		79%	79%	79%	79%		79%				79%	79%		79%			
Failures			1		1			1		2					1					
																Servers	8			53%
																Clients	11			79%
																All Tests	88	82	6	93%
																Coverage	196	82	6	45%

Figure 9: SV5 Analysis-Java

Part 2– Connect to the Topcats in the development federation and locate data.

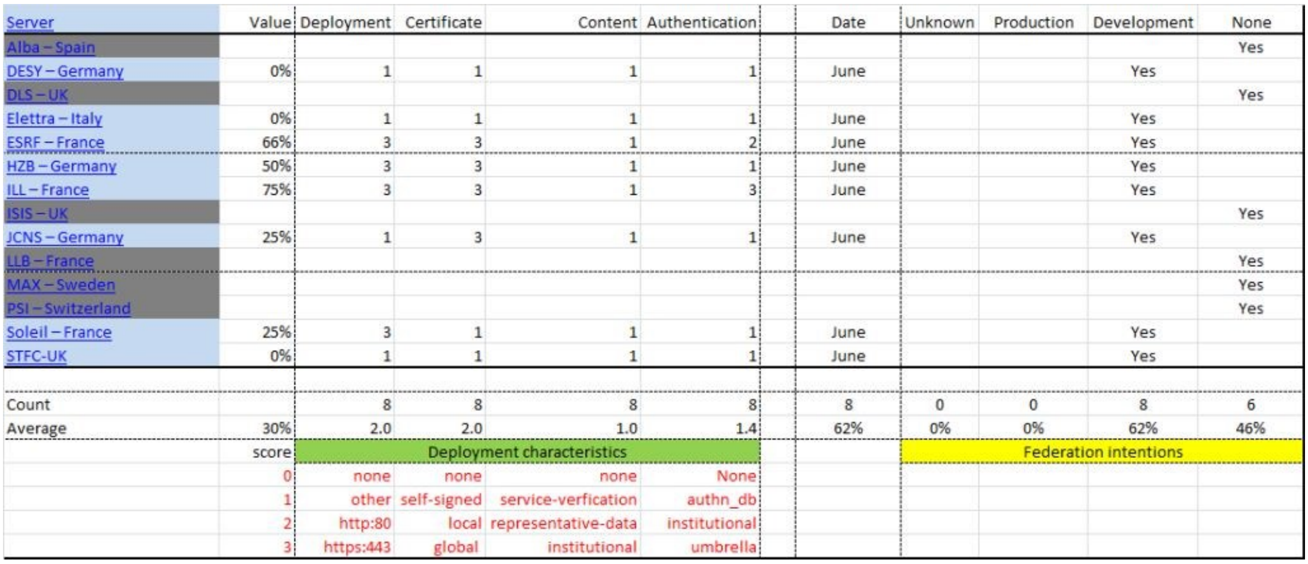
The partners tried this test with mixed results. Some reported success, others reported difficulties. This is the first time that service verification has used Topcats in the development federation; there is clearly a need to improve the quality of the deployments of the Topcats. This will be investigated before the next service verification.

Part 3 – Run a script which locates information in the European Affiliation Database.

The partners all reported success with this test. The team at ESRF did a very good job of preparing the test. Future verifications will include additional tests of EAD.

Part 4 – Configure an ICAT for use in the development federation.

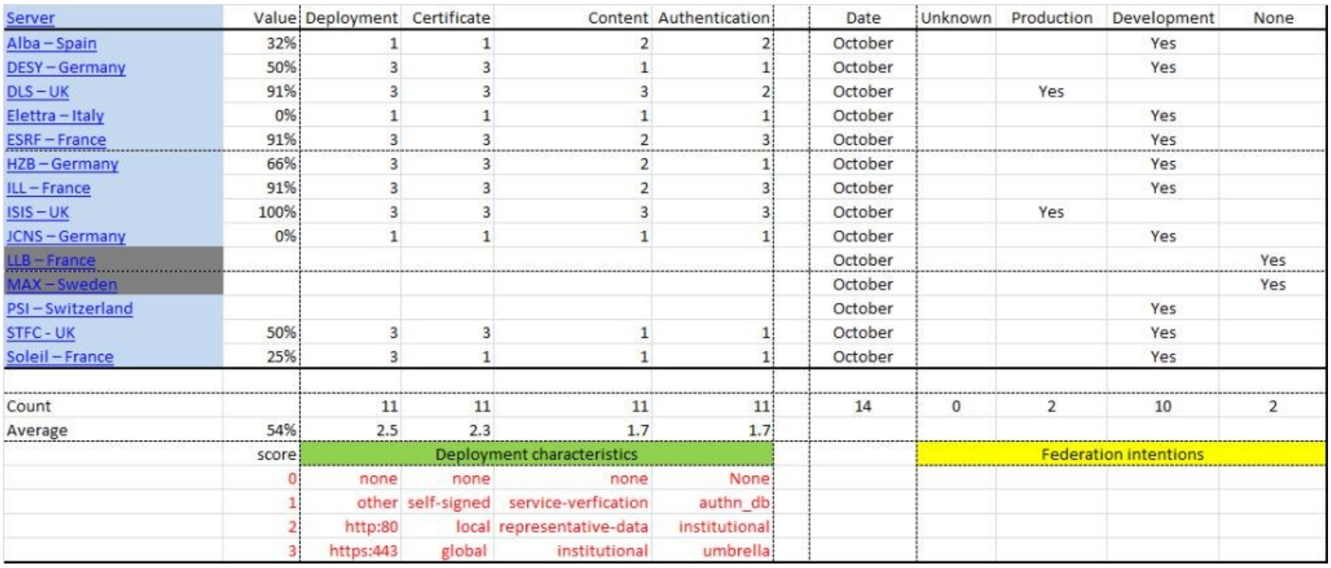
Eight of the 14 partners in the project provided an ICAT service for the development federation. Five partners, Elettra, ILL, JCNS, Soleil and STFC, provided services similar to earlier service verifications. Two of the partners (ISIS and DLS) have production ICATs and will join the Production federation when the production federation is ready. One partner (Alba) did not provide its usual service due to other commitments. PSI has yet to deploy a service, and MAX IV and LLB are unlikely to deploy a service. Six of the partners used an authentication method based on the authentication service provided by http://www.icatproject.org. Two partners (JCNS and Soleil) used a single username and password authentication scheme. The ICAT services are shown in the figure Deployments.



Server	Value	Deployment	Certificate	Content	Authentication		Date	Unknown	Production	Development	None
Alba – Spain											Yes
DESY – Germany	0%	1	1	1	1		June			Yes	
DLS – UK											Yes
Elettra – Italy	0%	1	1	1	1		June			Yes	
ESRF – France	66%	3	3	1	2		June			Yes	
HZB – Germany	50%	3	3	1	1		June			Yes	
ILL – France	75%	3	3	1	3		June			Yes	
ISIS – UK											Yes
JCNS – Germany	25%	1	3	1	1		June			Yes	
LLB – France											Yes
MAX – Sweden											Yes
PSI – Switzerland											Yes
Soleil – France	25%	3	1	1	1		June			Yes	
STFC-UK	0%	1	1	1	1		June			Yes	
Count		8	8	8	8		8	0	0	8	6
Average	30%	2.0	2.0	1.0	1.4		62%	0%	0%	62%	46%
	score	Deployment characteristics						Federation intentions			
	0	none	none	none	None						
	1	other	self-signed	service-verfication	authn_db						
	2	http:80	local	representative-data	institutional						
	3	https:443	global	institutional	umbrella						

Figure 10: Deployments

It is likely that the production federation will have two partners in October. DLS – UK and ISIS – UK. It is likely that the development federation will have eight partners in October: Alba – Spain; DESY – Germany; Elettra – Italy; ESRF – France; HZB – Germany; ILL – France; JCNS – Germany; Soleil – France.



Server	Value	Deployment	Certificate	Content	Authentication		Date	Unknown	Production	Development	None
Alba – Spain	32%	1	1	2	2		October			Yes	
DESY – Germany	50%	3	3	1	1		October			Yes	
DLS – UK	91%	3	3	3	2		October		Yes		
Elettra – Italy	0%	1	1	1	1		October			Yes	
ESRF – France	91%	3	3	2	3		October			Yes	
HZB – Germany	66%	3	3	2	1		October			Yes	
ILL – France	91%	3	3	2	3		October			Yes	
ISIS – UK	100%	3	3	3	3		October		Yes		
JCNS – Germany	0%	1	1	1	1		October			Yes	
LLB – France							October				Yes
MAX – Sweden							October				Yes
PSI – Switzerland							October			Yes	
STFC - UK	50%	3	3	1	1		October			Yes	
Soleil – France	25%	3	1	1	1		October			Yes	
Count		11	11	11	11		14	0	2	10	2
Average	54%	2.5	2.3	1.7	1.7						
	score	Deployment characteristics						Federation intentions			
	0	none	none	none	None						
	1	other	self-signed	service-verfication	authn_db						
	2	http:80	local	representative-data	institutional						
	3	https:443	global	institutional	umbrella						

Figure 11: Intentions

After the service verification in October, some of the partners will migrate from the development to the production federation.

SV Conclusions

This service verification was successful and it was well supported by the partners.

The highlights include:

- Most of the partners can access data on the ICATs of the development federation using a script;
- Some of the partners can access data on the ICATs of the development federation using Topcat;
- All of the partners can access data in the European Affiliation Database;
- We have a plan which shows that a collection of good quality ICAT services will evolve over the next 9 months.

4.6. Future roadmap

SV6 has been scheduled for October 25th while the SV7 will be held in early December. The list of planned activities is ambitious and will encompass services from different work packages.

Among others, the following components shall be deployed and tested:

- ICAT 4.3
- ICE 1.0
- Fuseki server to support Ontology
- Provenance
- EAD
- Topcat 1.9
- iGest modules (cross-facility search and benchmarking)

A workshop will be held at Elettra-Sincrotrone Trieste on November 6 and 7, organized by the WP4, in support of the deployment of these latest developments in the PaNdata software stack.

5. Summary

The work reported in this deliverable summarises the data cataloguing activities of the project till the current stage. These include data ingestion, cross-facility searching and partial integration of services among multiple facilities. Regarding data-ingestion we developed a set of tools (iGest) for the cataloguing of VLab (WP5) NeXus datasets. This required the generation of a large number of files that served as a proof of concept that the system can scale in real world cases such as a deployment in the Elettra-Sincrotrone Trieste accelerators. The cross-facility search has been designed as a cross-ICAT query results aggregator. It is simple, effective and high-performing as it queries ICATs in parallel while the results are actual ICAT objects instances. A substantial effort has been put on the Service Verifications where multiple facilities of the PaNdata consortium (>80%) participate on deployment and testing of various services in the context of data cataloguing. The future plans include the continuation of the SVs, the enhancement of the ICAT software ecosystem, and an analysis regarding performance and efficiency issues of the ICAT in-line with requirements of the forthcoming D4.4.