ICAT Workshop: Towards ICAT 4.3

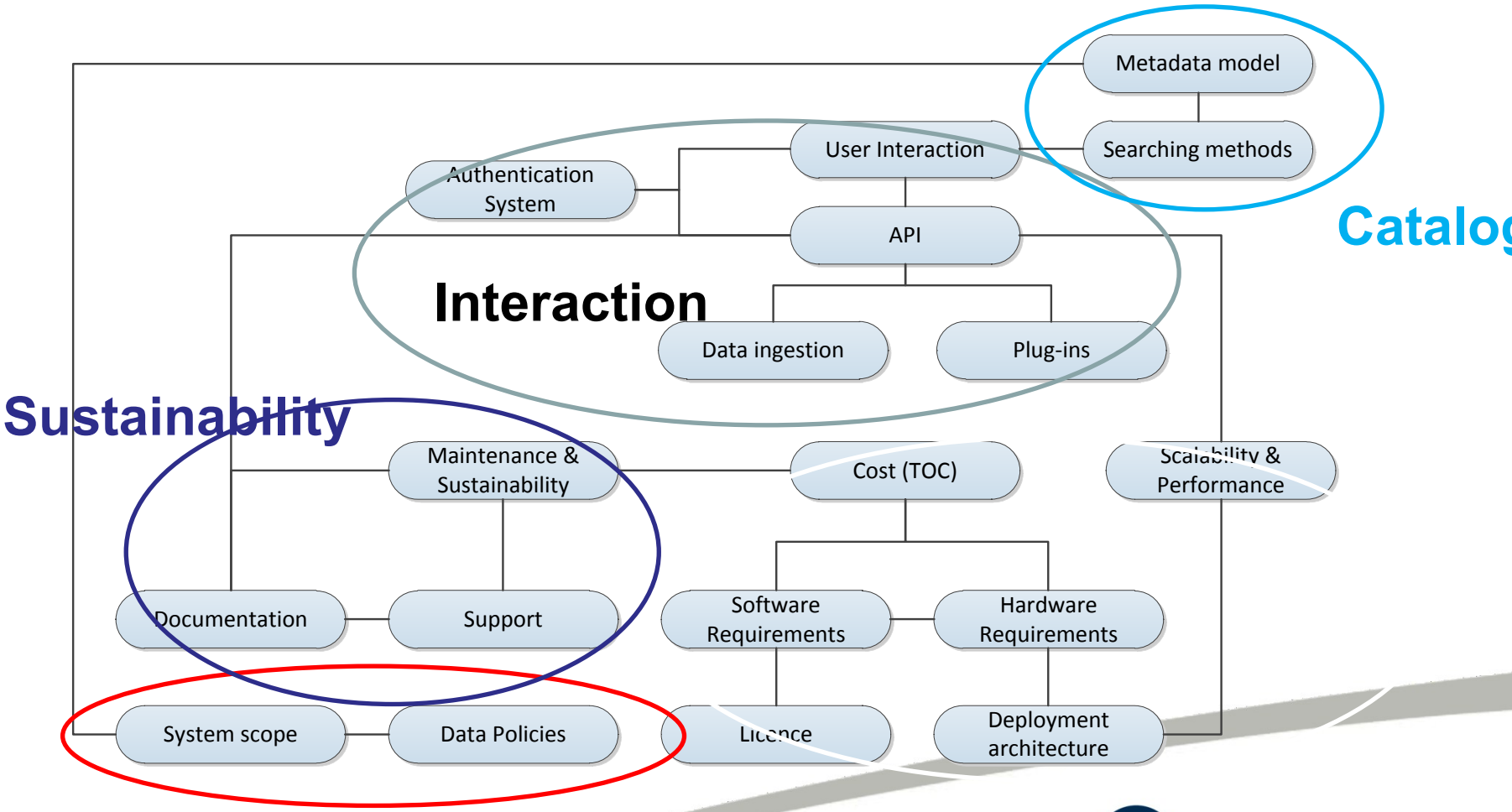
Brian Matthews

Scientific Computing Department

Science & Technology Facilities Council

Evaluation Criteria

- 18 criteria for evaluating a data catalogue



Science & Technology Facilities Council

ICAT Evaluation

Authentication System	The authentication system is a plug-in. A suitable one can be done for Umbrella. Searching across multiple ICAT instances is possible.
Metadata model	The current one was designed with x-ray and neutron experiments in mind (captures the “Beamtime” concept). NXarchive [ref. to later text] has been designed specifically for ICAT.
Querying/Searching methods	Permits keyword based searching. Free-text is supported too.
Software Requirements	Enterprise Java technologies and Oracle RDBMS. The latest version can be deployed on MySQL too as requested by a PaNdata member facility.
User Interaction	The ICAT project has produced an interactive web-frontend to the system (Topcat).
Service API	The ICAT4 API is a layer on top of relational DBMS. The database is wrapped as a SOAP web service so that the tables are not exposed directly. When the web service interface definition (WSDL) is processed by Java then each data structure results in a class definition.
Hardware Requirements	According to the Software Requirements.

ICAT Evaluation

Documentation	**ICAT is well documented. There is no up-to-date user guide for Topcat but there is a website and wiki for the project (http://code.google.com/p/icatproject/).**
Support	The ICAT team is providing the PaNdata consortium with extensive support. Members participate in regular teleconferences and meetings where current and future developments are discussed. There is formal agreement between the ICAT project and PaNdata ODI WP4.
Licence	Open source - FreeBSD.
Data Policies	N/A.
Total Cost of Ownership (TCO)	The system is open source and its current version requires only open-source or free technologies. It offers good and responsive support. It will be used among the PaNdata partners as a common system.
Scalability and Performance	An existing installation of Petabyte scale reports satisfactory performance.
Data Ingestion	The API permits simple data ingestion.
Additional Services and Plug-ins	It is open source, has an API that is SOAP based, and is modular.
Deployment Architecture	Mostly single server instances but a federated distribution of ICATs is possible. This would enable the web portal (Topcat) to be on top of the API.
Maintenance and Sustainability	ICAT follows good software sustainability practices and has been reviewed by the Software Sustainability Institute.
Specialisation and Systems Scope	The system is specialised as it mostly realises a data catalogue service. Its scope is well suited for scientific data.

Further PanData Requirements

Authentication System	Should enable DCS operations across PaNdata facilities. Should be easily integrated with the existing local authorisation and file systems.
Metadata model	Should be extendible and support discipline specific annotations. Unique identifiers to allow for citation (DOIs).
Querying/Searching methods	Should permit keyword based searching.
Software Requirements	The DCS should be based on a widely-used DB.
User Interaction	Should provide the end-user with a web-portal.
Service API	An API that permits at least all the operations that can be done through the portal.
Hardware Requirements	N/A
Documentation	Documentation and user-guide for the API and the web-portal.
Support	N/A
Licence	N/A
Data Policies	N/A
Total Cost of Ownership (TCO)	N/A
Scalability and Performance	Fast enough to permit interactive work (web-portal)
Data ingestion	N/A
Additional Services and Plug-ins	Upload/download/delete of files, change permissions (through the web portal). Support for a Workflow system.
Deployment architecture	N/A
Maintenance & Sustainability	N/A
Specialisation and Systems scope	Virtual labs as described in D5.1 and scientific applications related to the participating institutes.

Further PanData Requirements

Authentication System	Easy integration with the Umbrella system as described in D3.1.
Metadata model	A model that covers the metadata present in an advanced scientific data file format (NeXus).
Querying/Searching methods	Querying value ranges. Use of Tags.
Software Requirements	Server-class Linux OS and reliance on open source solutions.
User Interaction	As in D5.1.
Service API	API bindings for languages as C++, Java, and Python. Ideally, bindings for high level numerical computing environments (Matlab/IDL) so that the users can query the system and index post-processed data too.
Hardware Requirements	Server-class systems.
Documentation	As in D5.1.
Support	Support for both usage and development from the system's development team.
Licence	Open source.
Data Policies	It should not conflict with the PaNdata Europe Data Policy. Ideally should permit automated policy enforcement actions like changing the access of certain data to Public after a predefined time.
Total Cost of Ownership (TCO)	Follow the Compliance to Standards (i.e. using a common system among similar institutes). Pay no fees for support and proprietary software. No high maintenance costs after the end of the project.
Scalability and Performance	Should scale well when additional hardware is added and cope with the data flow of each member facility.
Data ingestion	Sample code for data ingestion through the API.
Additional Services and Plug-ins	Possible interaction with workflow systems (e.g. Taverna, Kepler) and data provenance frameworks.
Deployment architecture	Should permit multiple instances (distributed across the institutions) able to return cross-facility search results.
Maintenance & Sustainability	The system development team should have high maintenance and software sustainability standards.
Specialisation and Systems scope	Software specialised on scientific data without additional technologies like Grid frameworks and complicated certificate based systems.

PaN-Data WP4: Next Steps

- *Task 4.2: Serially, deploy the chosen metadata catalogue solution in the legacy context of the collaborating facilities*
  - ICAT being rolled out across a number of the facilities
  - ISIS, DLS, CLF, ILL, ELETTRA, ESRF, …
  - Regular Service Verifications

- Requirements and implementation influenced by the facilities
  - Open source collaboration: ICAT/PaNData
  - http://www.icatproject.org/
  - http://code.google.com/p/icatproject/

Science & Technology Facilities Council

SSI Review

- Governance
- Communications
- Process
- Policies
- Support
- Infrastructure

Science & Technology Facilities Council

What might go in ICAT 4.3 – and beyond

- Remove the ICATCompat interface
- Improved error messages for less common errors
- Add a Run entity
- Improve notification to include which operation was being used and to make it a better match users needs while imposing minimal overhead
- Allow a rule to specify that an authz plugin should be applied
- Introduce a namespace to relate to types instead of facility. The namespace should have many to many relationship to a facility

Down the line?

- Controlled vocabulary;
  - For searching/reporting
    - keywords, facilities, instruments, techniques
    - parameters and parameter sets
    - samples
- Standard
  - Dataset types
  - Investigation types
  - Datafile formats (mime-types)
  - Groups and their rules (Admin, data ingestors, Instrument scientists …)
- Integration with frameworks
  - Mantid, Dawn
  - Workflow Engine
  - Access to PanSoft software catalogue
- Programming language bindings
  - Java, Python, ??
- Extend the component architecture
  - Light-weight laboratory use
  - Simple web interface
    - ICE
  - Command Line
- XML Ingest
- Example data: script to set up with virtual data, Virtual machine
- Screencast of full installation
- Linked Data
- Notification and reporting examples
- Handling more than one facility within one ICAT.
- Common data server interface (IDS2)
  - Managing large files
  - Handing off to a suitable protocol.

Science & Technology Facilities Council

Feedback

- What would you like to see?
- What can you contribute?

JOIN IN!

http://www.icatproject.org/
http://code.google.com/p/icatproject/
icatgroup@googlegroups.com
icat-developers@googlegroups.com

Science & Technology Facilities Council