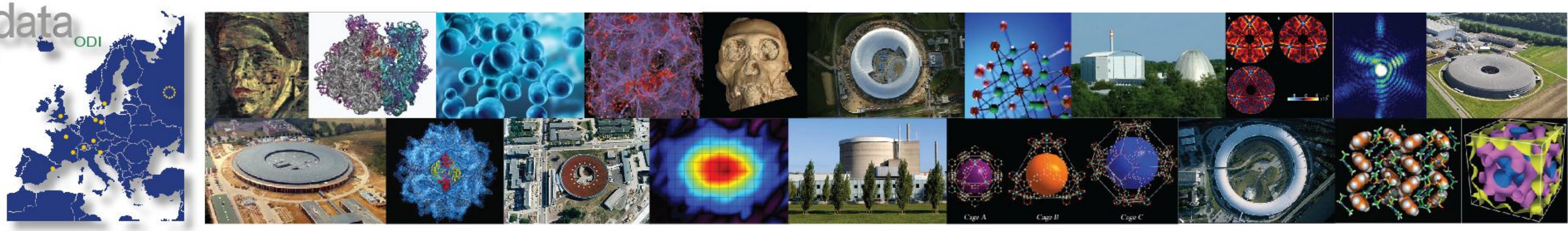


Requirements for data catalogues within facilities

Milan Prica[1], George Kourousias[1], Alistair Mills[2], Brian Matthews[2]

[1] *Sincrotrone Trieste S.C.p.A, Trieste, Italy*
[2]*Scientific Computing Department, STFC, Didcot, UK*

brian.matthews@stfc.ac.uk

ALBA
DESY
diamond
elettra
ESRF
FRM II Forschungs-Neutronenquelle Heinz Maier-Leibnitz
HELMHOLTZ ZENTRUM BERLIN für Materialien und Energie
Neutrons for Science Institut Laue-Langevin
Science & Technology Facilities Council ISIS
JCNS
LLB
MAX-lab
PAUL SCHERRER INSTITUT PSI
SOLEIL SYNCHROTRON

pandata ODI

elettra

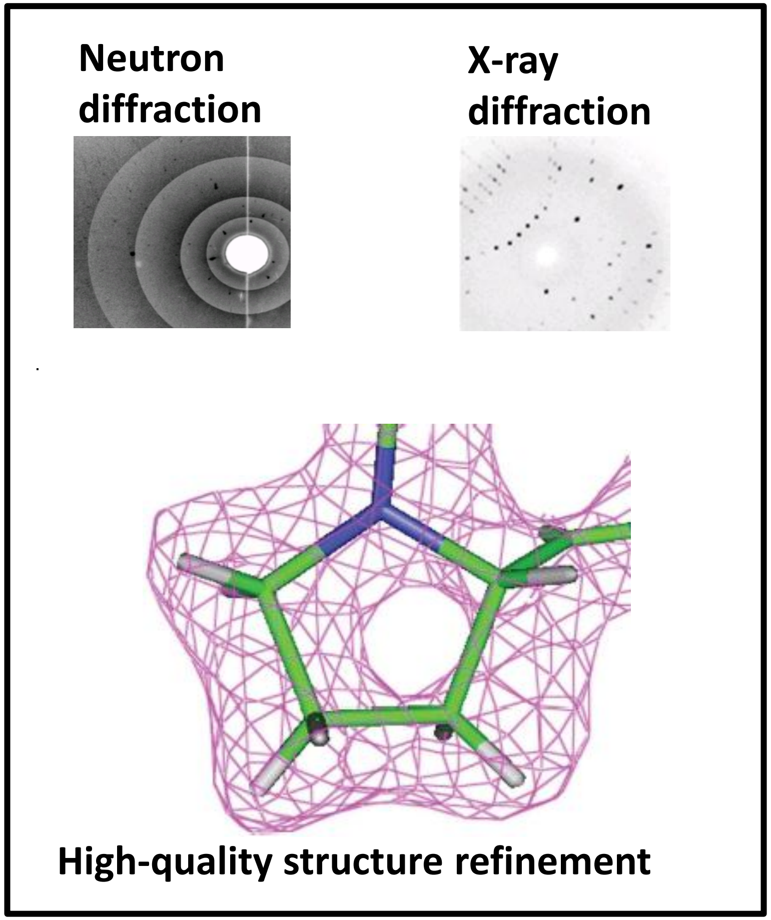
Science & Technology Facilities Council

PaN-data ODI : an Open Data Infrastructure for European Photon and Neutron laboratories

Federated data catalogues supporting cross-facility, cross-discipline interaction at the scale of atoms and molecules

Provide common tools and user experience across facilities

- Unification of data management policies
- Shared protocols for exchange of user information
- Common scientific data formats
- Interoperation of data analysis software
- Data Provenance: Linking Data and Publications
- Digital Preservation: supporting the long-term preservation of the research outputs



Neutron diffraction

X-ray diffraction

High-quality structure refinement

WP4: Data Catalogue Service

… 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

Data Catalogue: a systematic record of the data generated within facilities including information on the context in which the data was collected, and information on how the data itself is stored.

- Develop a generic software infrastructure to support the interoperation of facility data catalogues,
- Deploy this software to establish a federated catalogue of data across the partners,
- Provide data services based upon this generic framework which will enable users to deposit, search, visualise, and analyse data across the partners‘ data repositories,

Why a Data Catalogue ?

"I know where all the data sets for the experiments on my instruments are, so why do I need a catalogue?"

- Facilities tend to have good infrastructure to support data filing and storage
- Staff and users "know" where their data is
- Back up and archive procedures
- Facilities don't tend to lose data

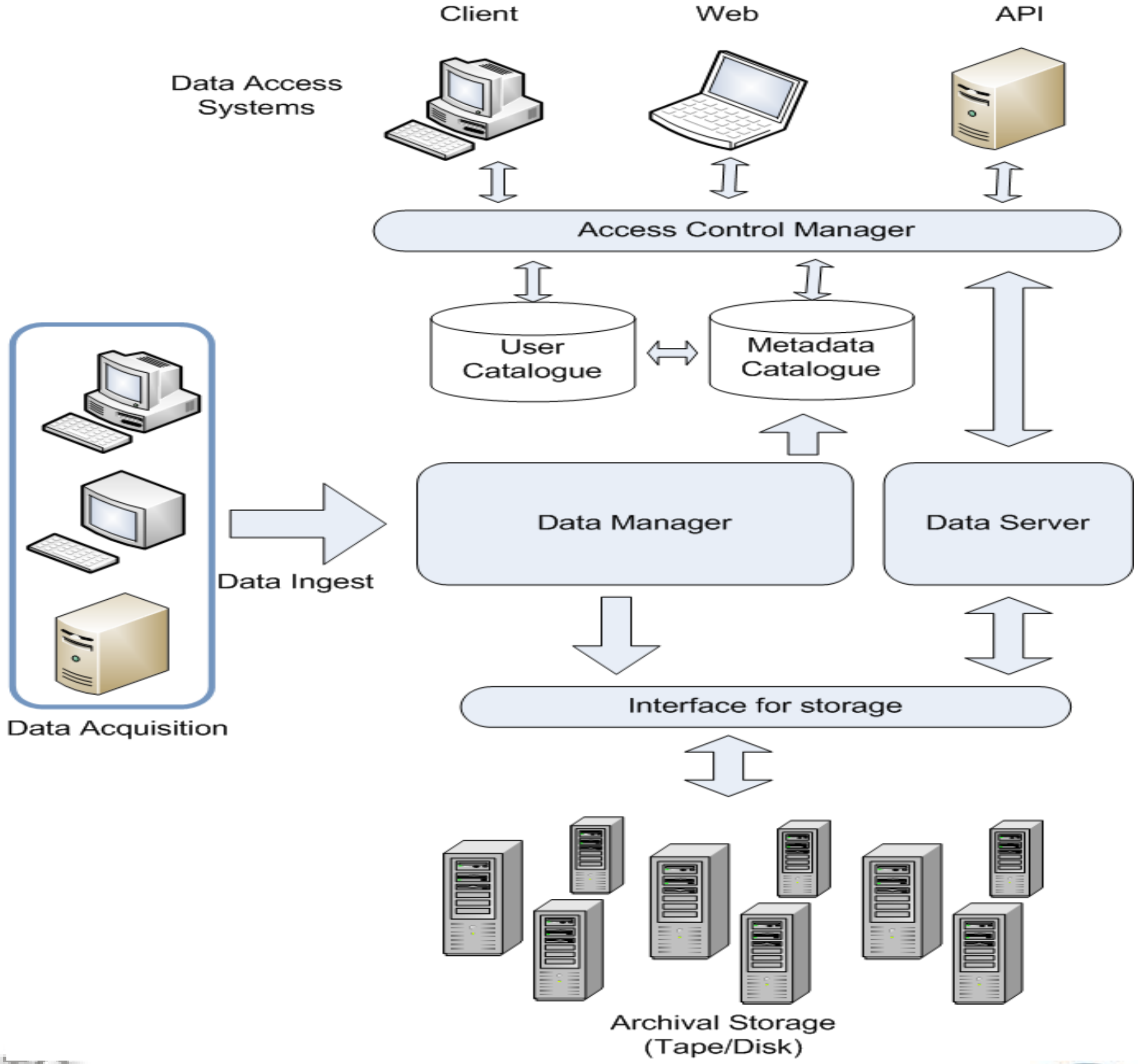
It pays to be systematic

- Track which data set results from which
  - experiment,
  - instrument,
  - proposal
- Indexes data according to the experiment
  - Rather than by a file structure
  - Can move the data around more easily.
  - Automated ingest – cope with the volume of experiments/data
  - Richer contextual information on:
    - the proposal , experimental parameters, calibrations,
  - Relate to other objects
    - data sets, publications
- Manage the quantity of information
- Prime beneficiary is the ***experimental team***

Access and Sharing

- Easier to make data accessible from off site
  - Suitable web-based search/access interfaces
  - Access controlled
  - Data access tools
- Also can be used interface with tools
  - With a suitable search/access API
- Can be federated to find your data in other facilities
- Share data with Your friends : suitable access control
- Data Publication
  - Making data publicly accessible
  - Open data
- Enforce a data policy

A Data Management Architecture

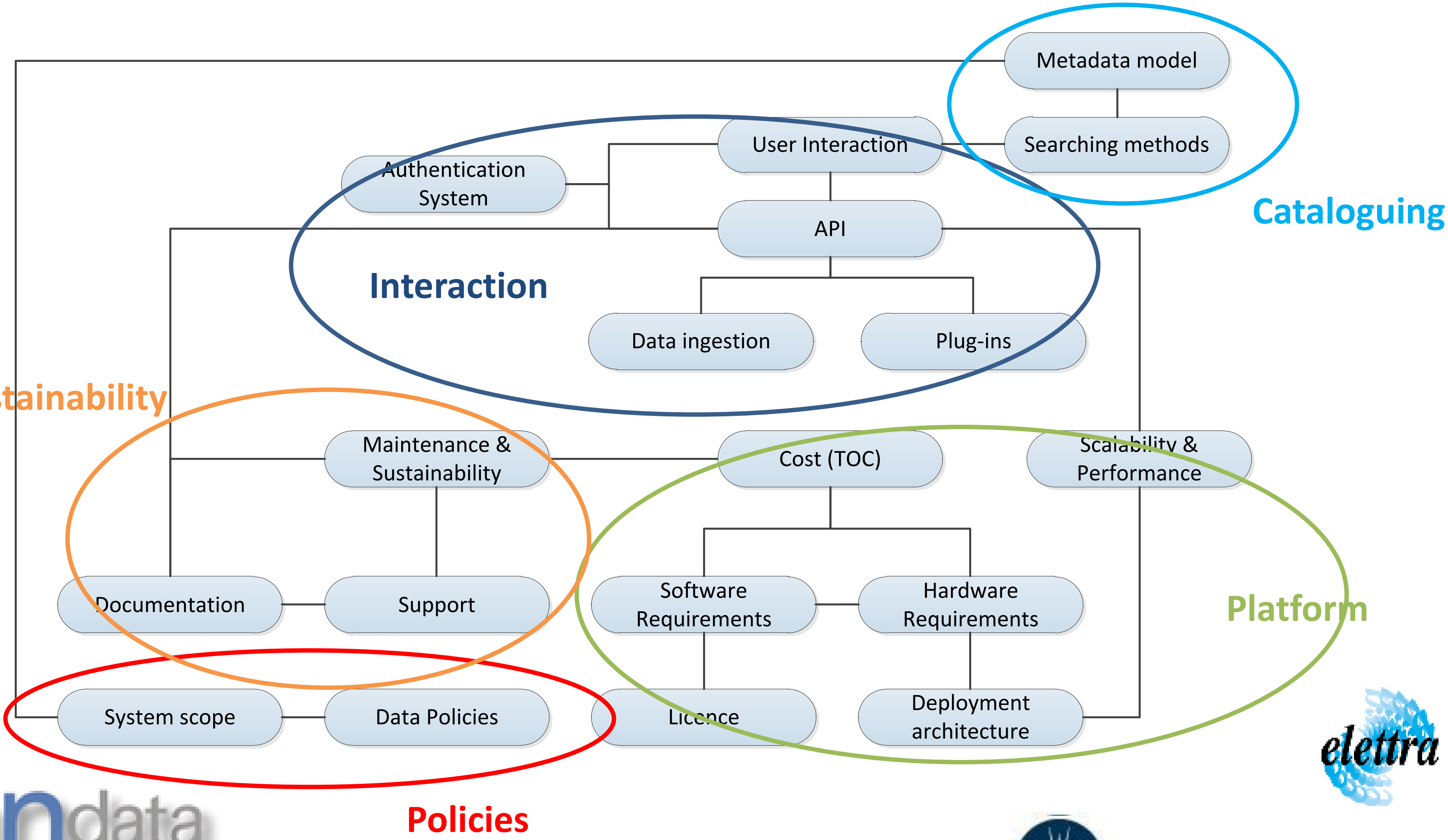


OK, but which ?

- So facilities are investing in Data Catalogues
  - Lots!
    - ICAT
    - DSpace
    - Fedora Commons Repository Software
    - iRODS
    - SRB / MCAT
    - Tardis
    - AMGA
    - Artemis
    - GCube Framework
    - ISPyB
    - Twist
    - Zentity
    - …
- Therefore criteria for making an informed choice

Evaluation Criteria

- 18 criteria for evaluating a data catalogue



Policies

- Specialisation and Systems scope
  - Generic / any type of digital assets
  - Specialised
    - Scientific data
    - Even more specialised
      - Protein Data Bank
- Data policies
  - Compliance
  - Enforcement
    - e.g. IF PaNdata THEN Go Public in 3 yrs

Platform

- Total Cost of Ownership (TCO)
  - Build expertise
  - Maintenance
  - Support
  - Licenses
  - Competitive Advantage
    - choose better catalogue than competition
  - Compliance to standards
    - choose the same for collaboration
- License
  - Proprietary
  - Free
  - Open-source
  - customisations, branching
- Scalability and Performance
  - Scientific data growth
  - PaNdata facilities
  - Enterprise-level expectations
  - Forecast (Free Electron Lasers, new detectors)
- Software requirements
  - Server side
  - OS (Windows/Linux)
  - Techs (Oracle,Java,.Net)
  - Client
  - Browser plugins (Flash, Silverlight, Java)
- Hardware requirements
  - Personal computer
  - Enterprise-level hardware
- Deployment architecture
  - Standalone / single-user
  - Multi-user / client-server
  - Cloud techs
  - Distributed DBs
  - Multiple instances of catalogues
  - Part of Grid comp. platforms

Interaction

- User interaction
  - Console / command line
  - Standalone GUI app
  - Web-portal
  - Integration to existing ones
  - In an ERP
- Service API
  - Bindings
    - Java, Python, C++,...
  - Flexibility
- Authentication system
  - Type of authentication (multi-user env.)
  - Suitability for distributed deployment
  - cross-facility operations

- Data ingestion
  - Manual (GUI)
  - Automated
  - Inside the pipeline
  - Data format parsers
    - HDF5, NeXus (NXarchive)

- Additional services and plug-ins
  - Workflow
    - Taverna, Kepler
  - Provenance
  - Web portal add-ons
  - Download/upload, permissions

Cataloguing

- Metadata model
  - Vocabulary of elements
  - Common
    - too generic & brief
  - Specialised
    - not standard
  - Expandable

- Querying/searching methods
  - Free-text
  - Hierarchical – based common metadata
  - Tags & tag-clouds
  - Numerical through KR
    - where 100>Temperature>80
  - Wildcards & logic

Sustainability

- Documentation
  - Conceptual design
  - API
  - Front-end user guide (web portal)
  - Commented code
- Support
  - From Dev./lead team
  - From local specialists
  - Custom installations

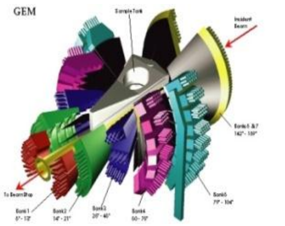
- Maintenance
  - Standards
  - State of the system
  - versioning
  - Troubleshooting processes
  - Future roadmaps
  - Software sustainability practises

ICAT

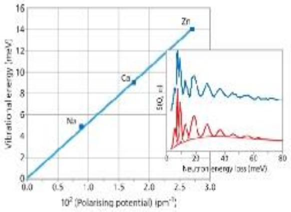
- Secure access to user's data
- Flexible data searching
- Scalable and extensible architecture
- Integration with analysis tools
- Access to high-performance resources
- Linking to other scientific outputs
- Data policy aware

Central Facility

Example ISIS Proposal

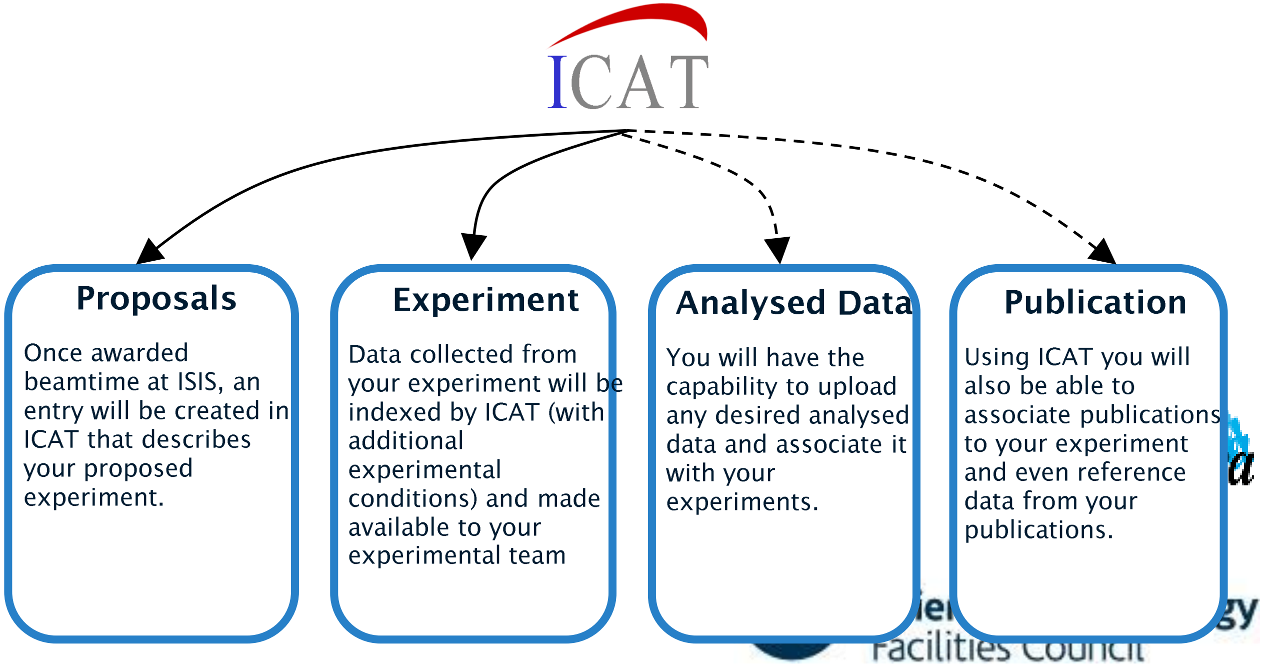


GEM – High intensity, high resolution neutron diffractometer



H2-(zeolite) vibrational frequencies vs polarising potential of cations

B-lactoglobulin protein interfacial structure



ICAT

Proposals

Once awarded beamtime at ISIS, an entry will be created in ICAT that describes your proposed experiment.

Experiment

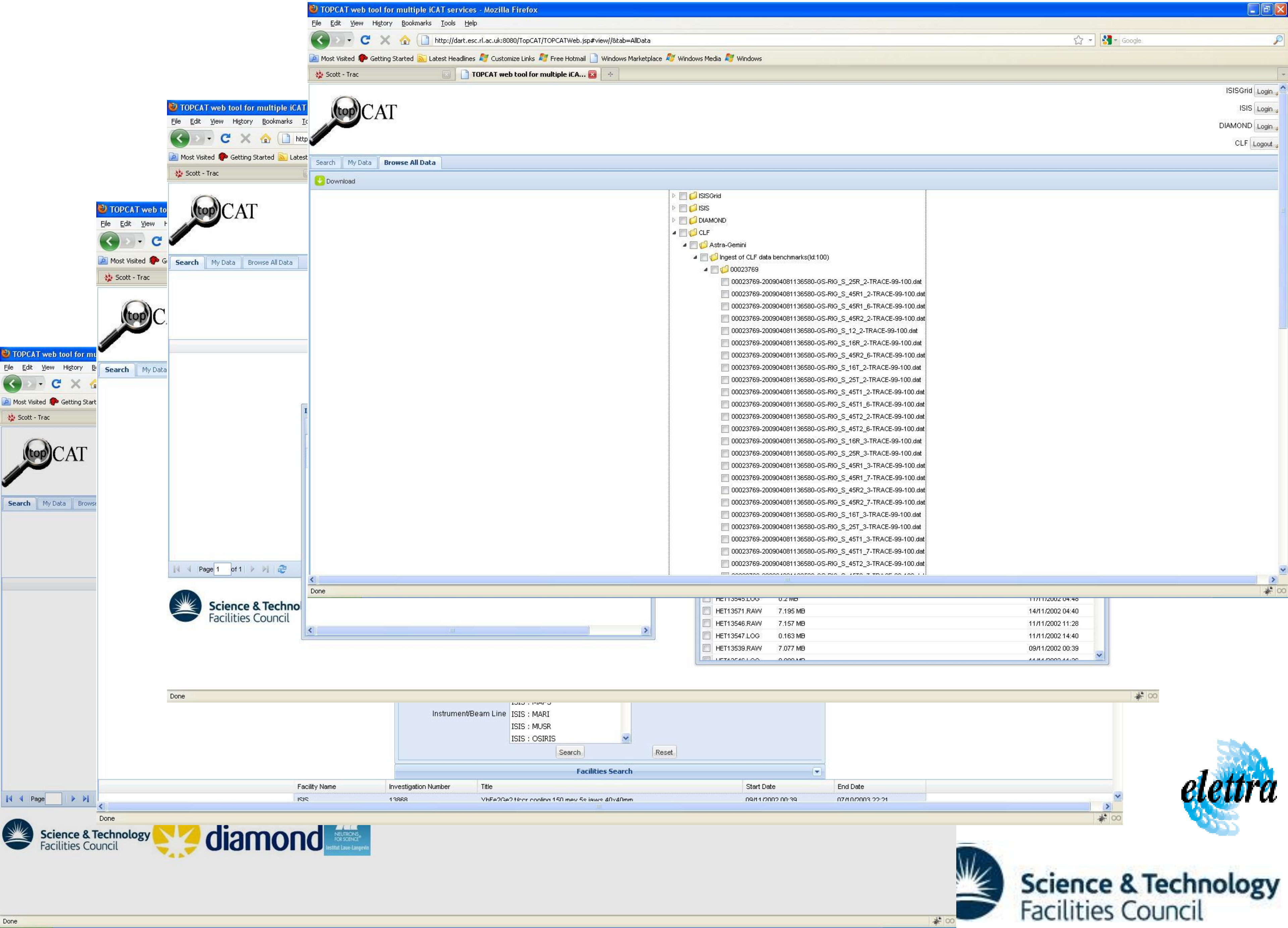
Data collected from your experiment will be indexed by ICAT (with additional experimental conditions) and made available to your experimental team

Analysed Data

You will have the capability to upload any desired analysed data and associate it with your experiments.

Publication

Using ICAT you will also be able to associate publications to your experiment and even reference data from your publications.


TOPCAT web tool for multiple iCAT services - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://dart.esc.rl.ac.uk:8080/TopCAT/TOPCATWeb.jsp#view//&tab=AllData
Most Visited
Getting Started
Latest Headlines
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
Scott - Trac
TOPCAT web tool for multiple iCA...
topCAT
ISISGrid Login
ISIS Login
DIAMOND Login
CLF Logout
Search
My Data
Browse All Data
Download
Page 1 of 1
Done
Science & Technology Facilities Council
diamond
NEUTRONS FOR SCIENCE Institut Laue-Langevin
elettra

- ISISGrid
- ISIS
- DIAMOND
- CLF
  - Astra-Gemini
    - Ingest of CLF data benchmarks(Id:100)
      - 00023769
        - 00023769-200904081136580-GS-RIG_S_25R_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R1_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R1_6-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R2_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_12_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_16R_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R2_6-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_16T_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_25T_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45T1_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45T1_6-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45T2_2-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45T2_6-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_16R_3-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_25R_3-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R1_3-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R1_7-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R2_3-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45R2_7-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_16T_3-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_25T_3-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45T1_3-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45T1_7-TRACE-99-100.dat
        - 00023769-200904081136580-GS-RIG_S_45T2_3-TRACE-99-100.dat

HET13545.LOG	0.2 MB	11/11/2002 04:48
HET13571.RAW	7.195 MB	14/11/2002 04:40
HET13546.RAW	7.157 MB	11/11/2002 11:28
HET13547.LOG	0.163 MB	11/11/2002 14:40
HET13539.RAW	7.077 MB	09/11/2002 00:39

Instrument/Beam Line: ISIS : MAPS, ISIS : MARI, ISIS : MUSR, ISIS : OSIRIS

Search Reset

Facilities Search

Facility Name	Investigation Number	Title	Start Date	End Date
ISIS	13868	YbFe2Ge2 t/ccr cooling 150 mev 5s jaws 40x40mm	09/11/2002 00:39	07/10/2003 22:21

ICAT Evaluation

Criterion	Assessment
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

Criterion	Assessment
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.

pandata ODI

Science & Technology Facilities Council

PaNData Next Steps

- Working with a "reference catalogue"
  - ICAT being rolled out across a number of the facilities
  - ISIS, DLS ILL, ELETTRA, ESRF, …
- Requirements and implementation influenced by the facilities
  - Open source collaboration: ICAT/PaNData
  - http://www.icatproject.org/
  - http://code.google.com/p/icatproject/
- We have a workshop on Thursday pm. - A tutorial

Conclusions

- Data catalogues a key part of the data infrastructure
  - Manage the data explosion
  - Keeping data safe and accessible
  - Sharing
- Lots of options
  - Criteria for judging what is useful
  - But better off sharing on a small number of solutions
  - Make them sharable between each other

http://www.icatproject.org/
http://code.google.com/p/icatproject/
http://pan-data.eu/

brian.matthews@stfc.ac.uk