Supporting scientific processes in National Facilities

Brian Matthews
brian.matthews@stfc.ac.uk

Scientific Computing Department
STFC Rutherford Appleton Laboratory
The scientific data centre at RAL

The data centre at RAL houses >20,000 computer processors and stores >20Pb of data available on-line.

The data centre is used for simulation and data analysis by researchers in all scientific disciplines from throughout the UK and their international collaborators.
Facilities and Resources of The Hartree Centre

Projects and codes developed on state of the art systems:
BlueGene/Q – Fastest UK machine and world’s largest software development platform
Over 5 PB disc and 15 PB tape stores
iDataplex cluster
Data Intensive systems
Visualisation System
Infrastructure for Facilities Science
STFC Rutherford Appleton Laboratory
The science we do - Structure of materials

- Visit facility on research campus
- Place sample in beam
- Diffraction pattern from sample
- Fitting experimental data to model
- Structure of cholesterol in crude oil

~30,000 user visitors each year in Europe:
physics, chemistry, biology, medicine, energy, environmental, materials, culture, pharmaceuticals, petrochemicals, microelectronics

- Billions of € of investment
  - c. £400M for DLS
  - + running costs
- Over 5,000 high impact publications per year in Europe
  - But so far no integrated data repositories
  - Lacking sustainability & traceability

Longitudinal strain in aircraft wing
Bioactive glass for bone growth
Hydrogen storage for zero emission vehicles
Magnetic moments in electronic storage
Data management for the Facilities

Neutrons and photons
Provide complementary views of matter:

Photons “see” electric charge – high atomic number nuclei

Neutrons “see” nucleons – especially hydrogen atoms

Computing Analysis Modelling
Once upon a time …

- Emails, portable disks, a simple web page were all you need.

This worked quite well in the first 20 or so years of ISIS.
Data monitoring

Network monitoring

Data Cataloguing

Data Synchronisation

Data archive

Now …
Scientist submits application for beamtime

Facility committee approves application

Facility registers, trains, and schedules scientist’s visit

Scientists visits, facility run’s experiment

Raw data filtered, and stored

Record publication

Subsequent publication registered with facility

Tools for processing made available

http://code.google.com/p/icatproject/
Experiment Stage

4. Experiment

1. Site Visit
2. Instrument Calibration Run
3. Instrument Setup
4. Sample Setup
5. Instrument Activation
6. Data Acquisition
7. Local Data Storage
8. Experiment close down
• 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

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

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.

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 and CSMD

- The Core Scientific Meta-Data Model (CSMD) is a study-data oriented model which has been developed at STFC since 2004.
- It captures high level information about scientific studies and the data that they produce throughout a facility’s scientific workflow.
- It is a key aspect of the ICAT, a software suite designed to manage the cataloguing and (continuous) access to facilities data.
- OWL Ontology to represent this model

http://www.icatproject.org/mvn/site/icat/4.2.5/icat.core/schema.html
ICAT Tool Suite and Clients

Desktop app

Clusters/HPC

Disk

Tape

ICAT Job Portal

ICAT APIs

TopCAT
(Web Interface to ICATs)

ICAT Data Explorer
(Eclipse Plugin)

ICAT + Mantid
(desktop client)

IDS
(ICAT Data Service)

http://www.mantidproject.org/
http://www.dawnsi.org/
https://code.google.com/p/icat-job-portal/
Supporting Analysis
Traditionally, these steps are decoupled from facilities. However, they are key to derive useful insights.

http://www.icatproject.org
Data Analysis Stage

5. Initial Post-Processing

- Visualise Data
- Analyse Derived Data
- Combine with other Data
- Interpret and analyse results

6. Data Analysis
Managing Data Processing Pipelines

Issues:
1. Valuable data amongst noise
2. Software version
3. Data provenance
4. Distributed analysis
5. Complex and dynamic workflows
6. Usability of tools

Raw data

Derived data

Resultant data

Credits: Martin Dove, Erica Yang (Nov. 2009)
Managing Processed Data

**View**
- Excel
- Standalone web client
- Hosted web client

**Access**
- Restful APIs
- File System

**Model**
- CSV
- JSON
- XML

**Controller**
- Python
- HDF lib
- Nexus lib

**Content**
- HDF files
- Nexus files

**MVC:** Model, View, Controller
Infrastructure for managing data flows

- **Scan**
- **Reconstruct**
- **Segment + Quantify**
- **3D mesh + Image based Modelling**
- **Predict + Compare**

**ICAT**

- **Data Catalogue**
- **Petabyte Data storage**
- **Parallel File system**
- **HPC CPU+GPU**
- **Visualisation**

**Infrastructure + Software + Expertise!**

- **Tomography**: Dealing with high data volumes – 200Gb/scan, ~5 TB/day (one experiment)
- **MX**: high data volumes, smaller files, but a lot more experiments
- Hard to move the data – needs to be handled at the facility?

Some image credit: **Avizo**, Visualization Sciences Group (VSG)
Dawn: Analysis Framework

Diamond/ESRF

Analysis framework

Based on Eclipse

Integrated with ICAT

Uses a workflow tool: Passerelle
Workflow to Provenance
From workflow to provenance

- Change emphasis
  - From workflow
  - To Provenance

- A more Data centred point of view
  - Less planning a process
  - More recording what happened.

- Provide data in context
  - For maintaining context
  - For providing history
  - For validation
  - For preservation
  - For reuse
### Icat Job Portal Main Panel (Datasets)

#### Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Experiment Type</th>
<th>Instrument</th>
<th>Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>20120524_0002_0001_632c1e9-9f32-4a39-a649-855ed5592c27</td>
<td>Colocalisation</td>
<td>OctopusSM3</td>
<td>Dummy Investigation 1</td>
<td>T47D 3 Affibody 639 nm laser</td>
</tr>
<tr>
<td>20120524_0002_0001_0bb36de-dd79-4c13-84ca-72a6385d0334</td>
<td>T47D 3 Affibody</td>
<td>OctopusSM3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20120524_0002_0001_e421cc0c-3d7eb-4e3f-ba6f-2f6665d316bb</td>
<td></td>
<td>OctopusSM3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20120524_0002_0001_e421cc0c-3d7eb-4e3f-ba6f-2f6665d316bb</td>
<td></td>
<td>Unknown Instrument</td>
<td>T47D 3 Affibody</td>
<td></td>
</tr>
<tr>
<td>20120524_0002_0001_e421cc0c-3d7eb-4e3f-ba6f-2f6665d316bb</td>
<td></td>
<td>OctopusSM3</td>
<td>T47D 3</td>
<td></td>
</tr>
<tr>
<td>20120524_0002_0001_e421cc0c-3d7eb-4e3f-ba6f-2f6665d316bb</td>
<td></td>
<td>Unknown Experiment Type</td>
<td>T47D 3</td>
<td></td>
</tr>
<tr>
<td>20120524_0002_0001_e421cc0c-3d7eb-4e3f-ba6f-2f6665d316bb</td>
<td></td>
<td>OctopusSM3</td>
<td>T47D 3</td>
<td></td>
</tr>
</tbody>
</table>

#### Options

- Show Download URL
- MSMM Viewer Project

---

**Additional Information**

- **endDate**: 2012-11-27T14:18:17Z
- **experiment_type**: Undefined
- **id**: 7201
- **instrument**: OctopusSM3
- **location**: Dummy Investigation 1/20120524_0002_0001_aee07c8e-dc7d-4b6c-a599-6e62eb4829e
- **name**: 20120524_0002_0001_aee07c8e-dc7d-4b6c-a599-6e62eb4829e
- **nchannels**: 1
- **nframes**: 671
- **sampleDescription**: T47D 3 Affibody 639 nm laser
- **startDate**: 2012-01-01T12:00:00Z
- **status**: Completed (No further action required)

---

**Image**

A screenshot of the Icat Job Portal showing a dataset for T47D 3 Affibody 639 nm laser. The image displays the results of the search with 7 datasets found, and a detailed view of one dataset is highlighted.
Ontology for Facility Science
Facilities, instruments, and techniques
(applications: cataloguing, searching, and linking)

Diffraction
- Neutron Diffraction/Elastic Neutron Scattering
  - Powder Diffraction
  - Single Crystal Diffraction
- X-ray Diffraction
  - Grazing Incidence Diffraction
  - Powder Diffraction
  - Resonant Diffraction
  - Small Angle Diffraction
  - Single Crystal Diffraction
  - Soft Diffraction
  - Surface Diffraction
- Other
  - Coherent Diffraction Imaging
  - Diffraction Imaging (Topography)
  - Enhanced Diffraction Imaging

Diffusive - Diffusive MRI

Imaging
- Holography
- Microscopy
  - X-Ray Photoemission Microscopy
  - X-Ray Scanning Microscopy
  - Scanning Transmission X-Ray Microscopy
  - Tomographic Microscopy With CRLs
- Tomography
  - Fluorescence Tomography
Research Objects (ROs) are **semantically rich aggregations of resources** that bring together **data, methods and people in scientific investigations**. Their goal is to create a class of artefacts that can encapsulate our digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge.

www.researchobject.org (WorkFlow4Ever)

Represent Investigation as a Research Object (IRO)

- Build a graph structure for the links in the research object.
- Using an RDF representation, URIs
- Publish as a linked data object
Investigation as a research object

- Own metadata format (CSMD)
- OAI-ORE
- W3C Prov ontology: PROV-O

LCDP 2013
Setting the Boundary

Investigations

Enhanced Publication

E-Portfolio

Making sense of Context in a Linked web of data

LCDP 2013
Challenges

Developments that will influence how the data is managed

• Managing very large amounts of data
• Exploiting large HPC resources available
• Integrating data analysis processes into data management processes
• Integrating data from different sources
• Integrating workflow into data large scale data management processes
• Using provenance information effectively
  • Data tracing
  • Data publication in context
  • Reproducability
Acknowledgement

- STFC (Scientific Computing and ISIS)
  - Erica Yang, Steve Fisher, Alistair Mills, Kevin Philips, Anthony Wilson, Tom Griffin, Holly Zhen, Juan Bicarregui, Martin Turner, Ronald Fowler, Cathy Jones, Vasily Bunakov
- Diamond Light Source
  - Mark Basham, Alun Aston, Kaz Wanelik, Robert Atwood
- Manchester University
  - Philip Withers, Peter Lee
- And many others who have contributed to the development of ICAT, CSMD, and the data infrastructure…