Science & Technology Facilities Council

Managing Research Data for Diverse Scientific Experiments

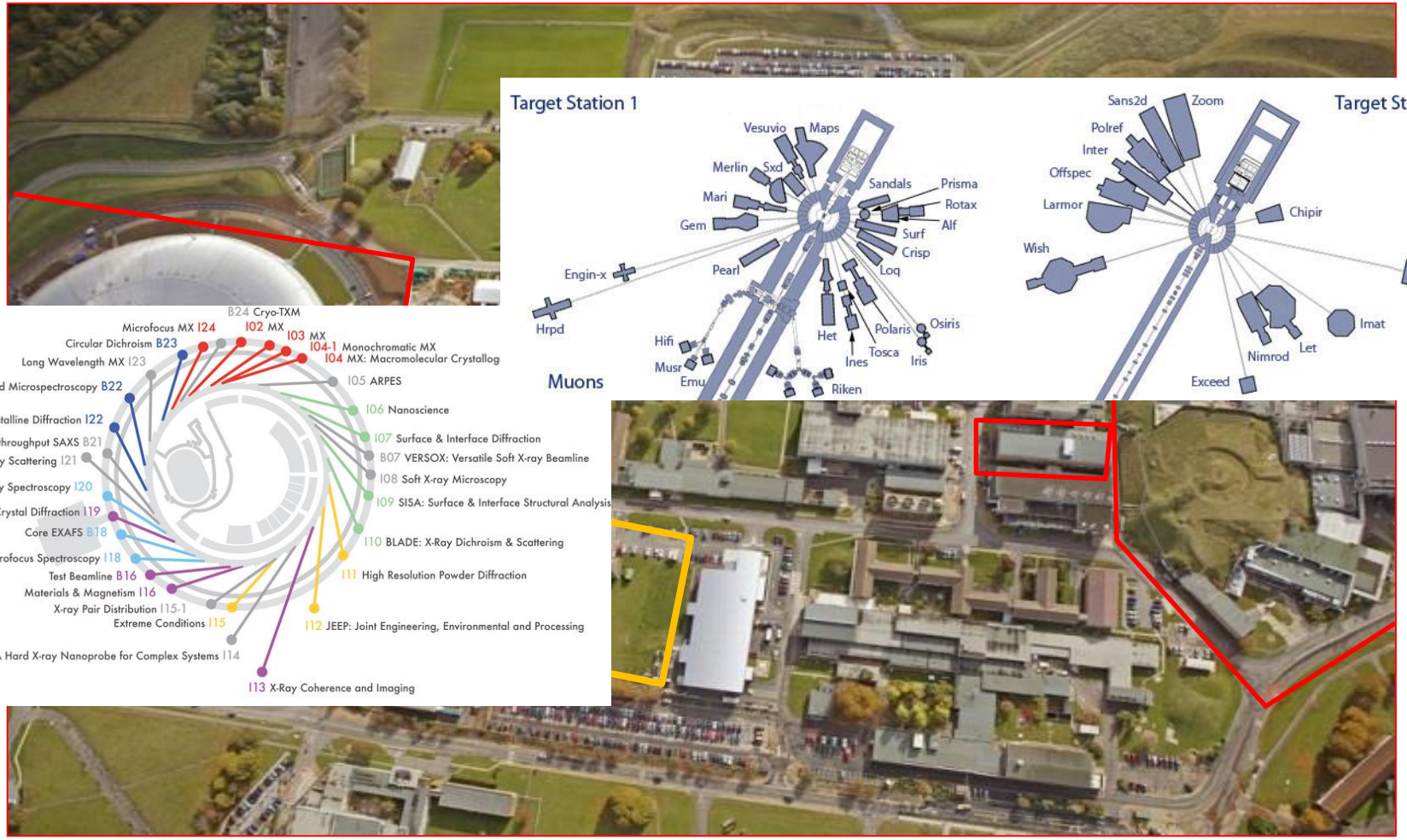
Erica Yang

erica.yang@stfc.ac.uk

Scientific Computing Department
STFC Rutherford Appleton Laboratory

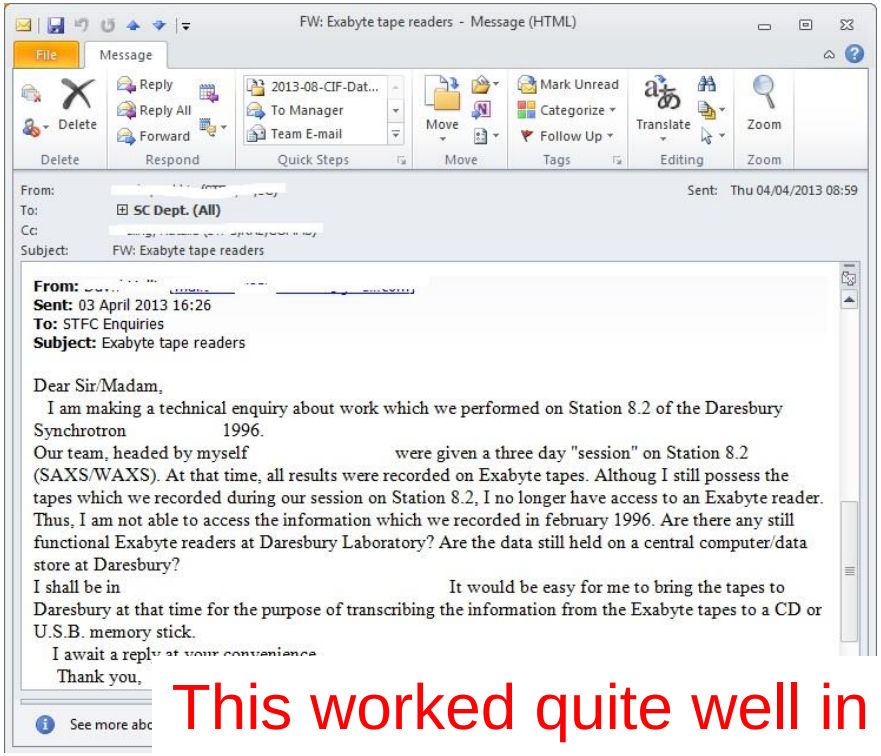
Crystallographic Information and Data Management Symposium
the 28th European Crystallographic Meeting
25 August 2013, Warwick University, U.K.

STFC Rutherford Appleton Laboratory



Once upon a time …

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



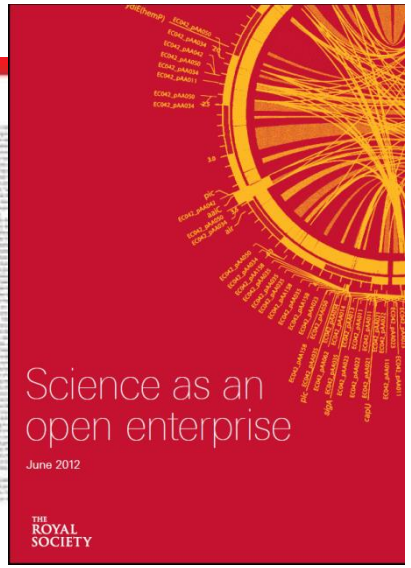
This worked quite well in the first 20 or so years of ISIS.

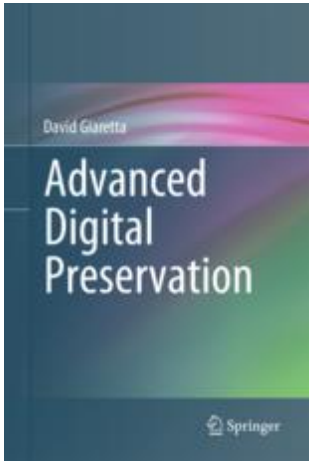
Data Infrastructure

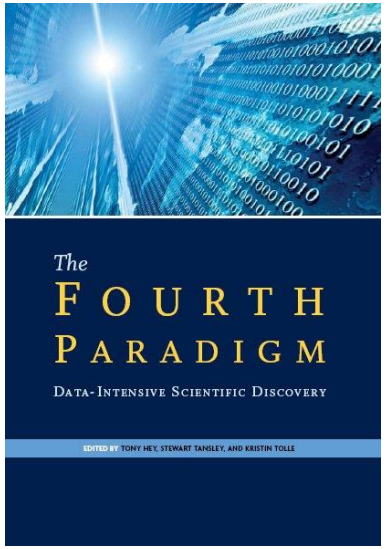
The age of managed data at RAL

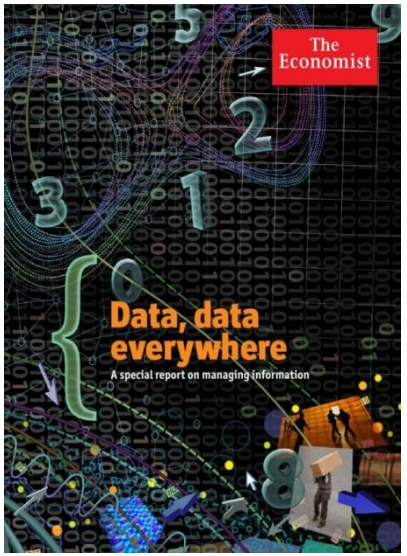
The paradigm, societal, and technological changes over time have made a major impact

- UK eScience programme
- The 4th paradigm: Data-Intensive Scientific Discovery
- Data, data everywhere
- Digital Preservation
- Royal Society Open Data Report
- Continued developments at the facilities

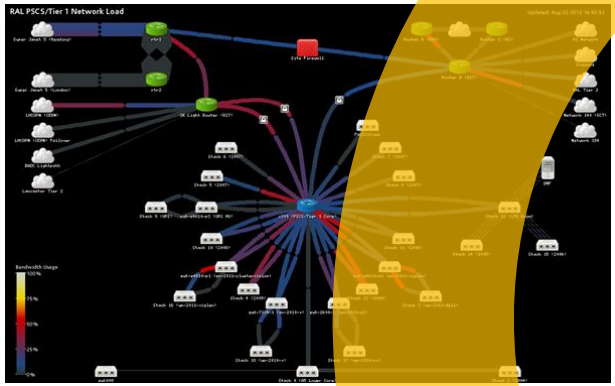




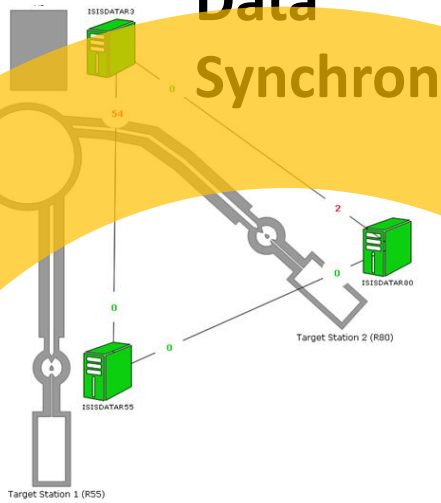


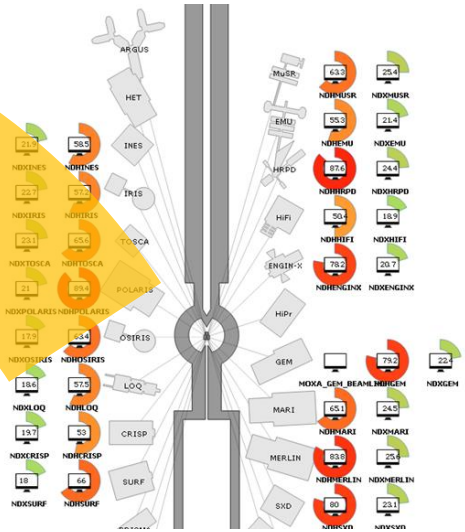


Network monitoring



Data Synchronisation

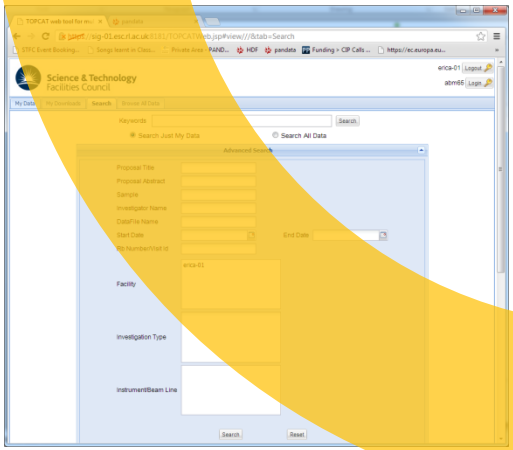




Data monitoring

Now …

Data Cataloguing



Data archive

Science & Technology
Facilities Council

Data Management and Tools

Facility Data Lifecycle

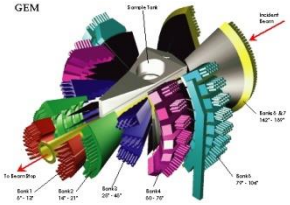
ICAT

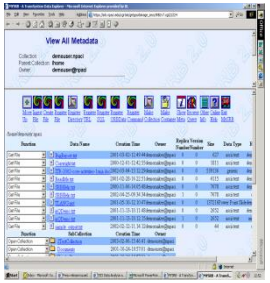
Metadata Catalogue

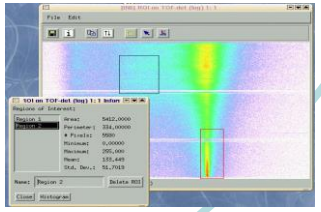
Proposal

Publication

Approval







Scheduling

Experiment

Data reduction

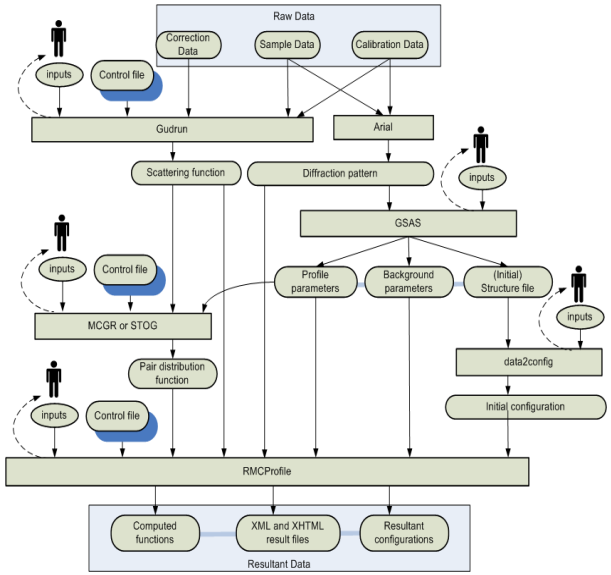
Data analysis

ICAT

http://www.icatproject.org

Traditionally, these steps are decoupled from facilities. However, they are key to derive useful insights.

Managing Data Processing Pipelines



Raw data

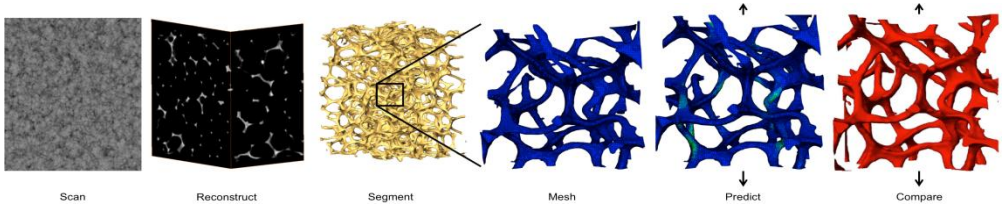
Derived data

Resultant data

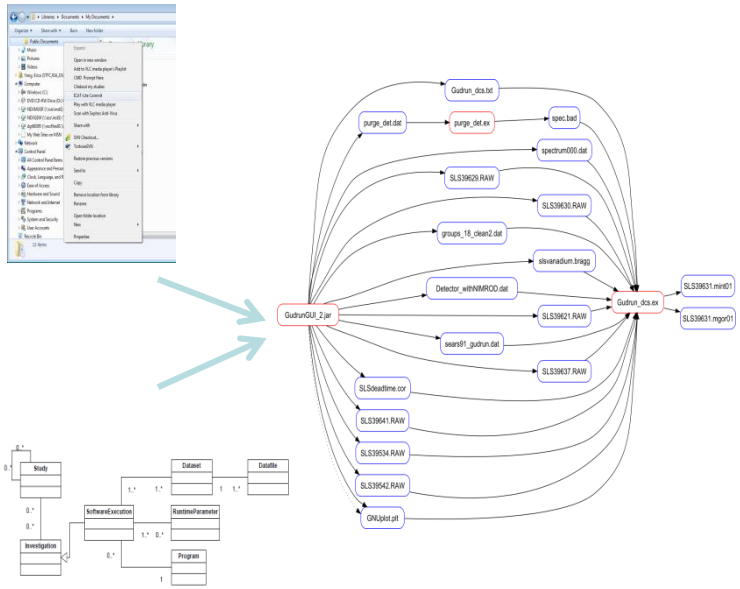
Issues:

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

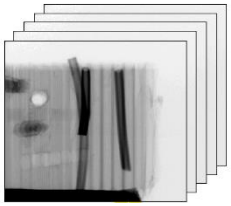
Credits: Martin Dove, Erica Yang (Nov. 2009)



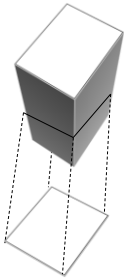
Credit: Phil Withers, Andy Alderson, Sam McDonald



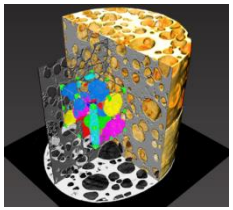
Infrastructure for managing data flows



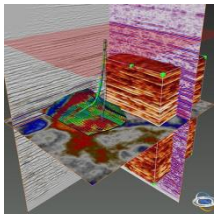
Scan



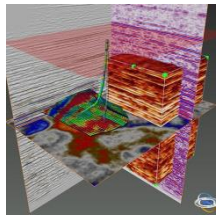
Reconstruct



Segment
+ Quantify



3D mesh +
Image based Modelling



Predict + Compare

ICAT

Data
Catalogue

Petabyte
Data storage

Parallel
File system

HPC
CPU+GPU

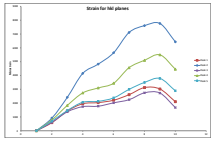
Visualisation

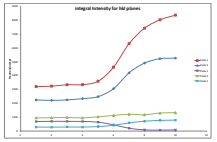
Infrastructure + Software + Expertise!

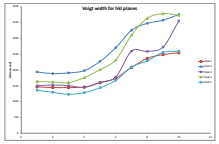
- **Tomography**: Dealing with high data volumes – 200Gb/scan, ~5 TB/day (one experiment)
- **MX**: high data volumes, smaller files, but a lot more experiments
- Hard to move the data – needs to be handled at the facility?

Some mage credit: Avizo, Visualization Sciences Group (VSG)

Managing Processed Data

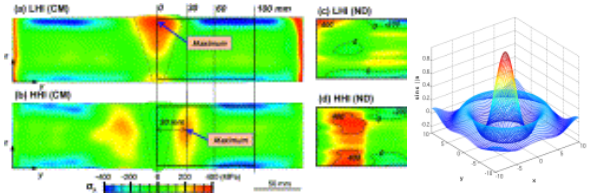






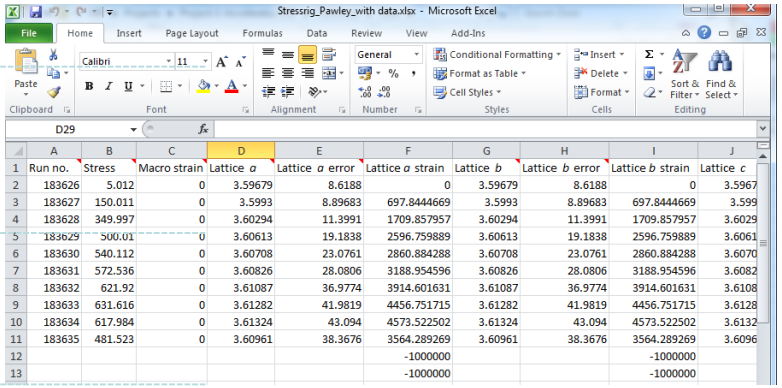
View

Excel | Standalone web client | Hosted web client



(Access)

Restful APIs | File System

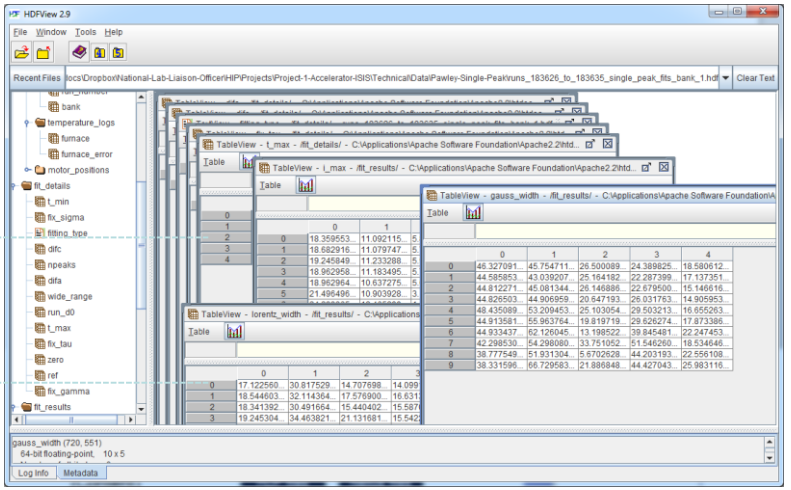


Model

CSV | JSON | XML

Controller

Python | HDF lib | Nexus lib



(Content)

HDF files | Nexus files

MVC: Model, View, Controller

Science & Technology
Facilities Council

Data Catalogue and Tools

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

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

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.

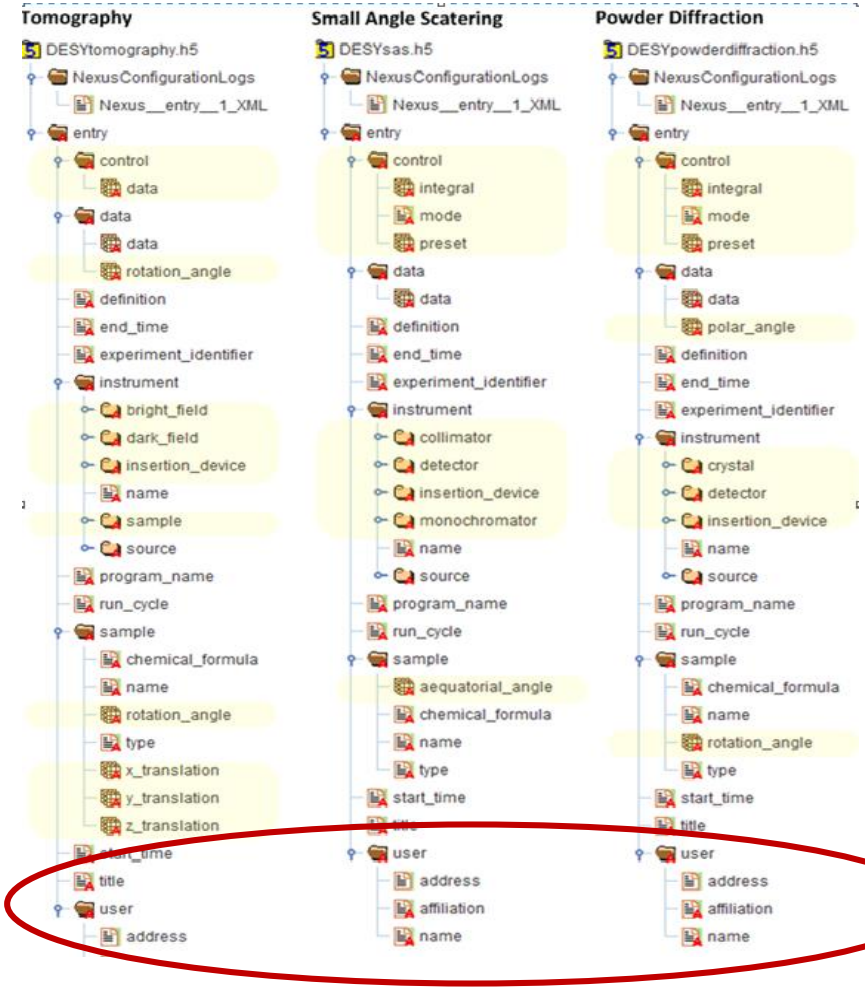
- **Investigation**
- **Investigator**
- **Topic and Keyword**
- **Publication**
- **Sample**
- **SampleParameter**
- **Dataset**
- **DatasetParameter**
- **Datafile**
- **DatafileParameter**
- **Parameter**

ICAT Schema

Application, Datafile, DatafileFormat, DatafileParameter, Dataset, DatasetParameter, DatasetType, Facility, FacilityCycle, Group, InputDatafile, InputDataset, Instrument, InstrumentScientist, Investigation, InvestigationParameter, InvestigationType, InvestigationUser, Job, Keyword, NotificationRequest, OutputDatafile, OutputDataset, ParameterType, PermissibleStringValue, Publication, RelatedDatafile, Rule, Sample, SampleParameter, SampleType, Shift, Study, StudyInvestigation, User, UserGroup

http://www.icatproject.org/mvn/site/icat/4.2.5/icat.core/schema.html

Nexus and CSMD

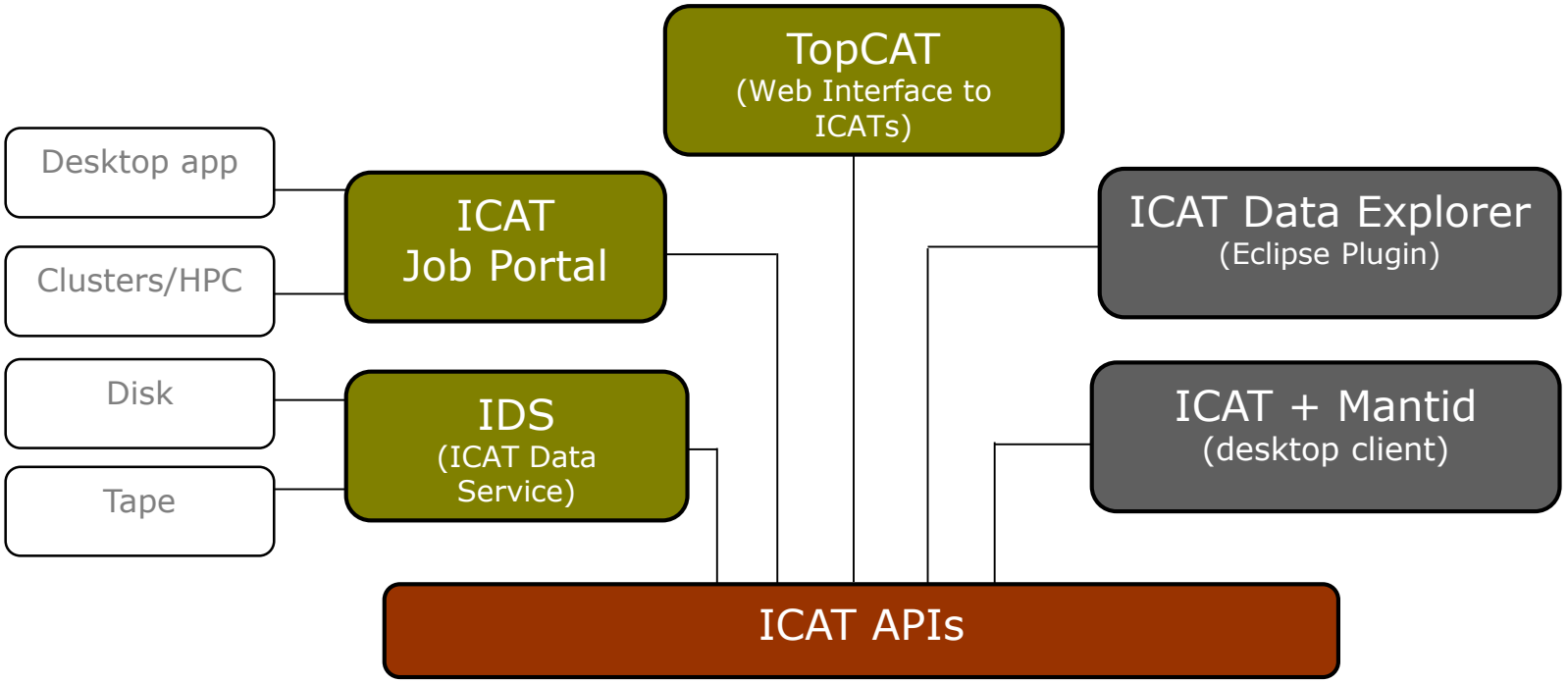


Nexus Application Profile for SAS

http://download.nexusformat.org/

PaNdata-ODI deliverable

ICAT Tool Suite and Clients



ICAT

http://www.mantidproject.org/
http://www.dawnsci.org/
https://code.google.com/p/icat-job-portal/

Science & Technology Facilities Council

diamond

NEUTRONS FOR SCIENCE Institut Laue-Langevin

pandata
Photon and Neutron Data Infrastructure

Ontology for Facility Science

Facilities, instruments, and techniques
(applications: cataloguing, searching, and linking)

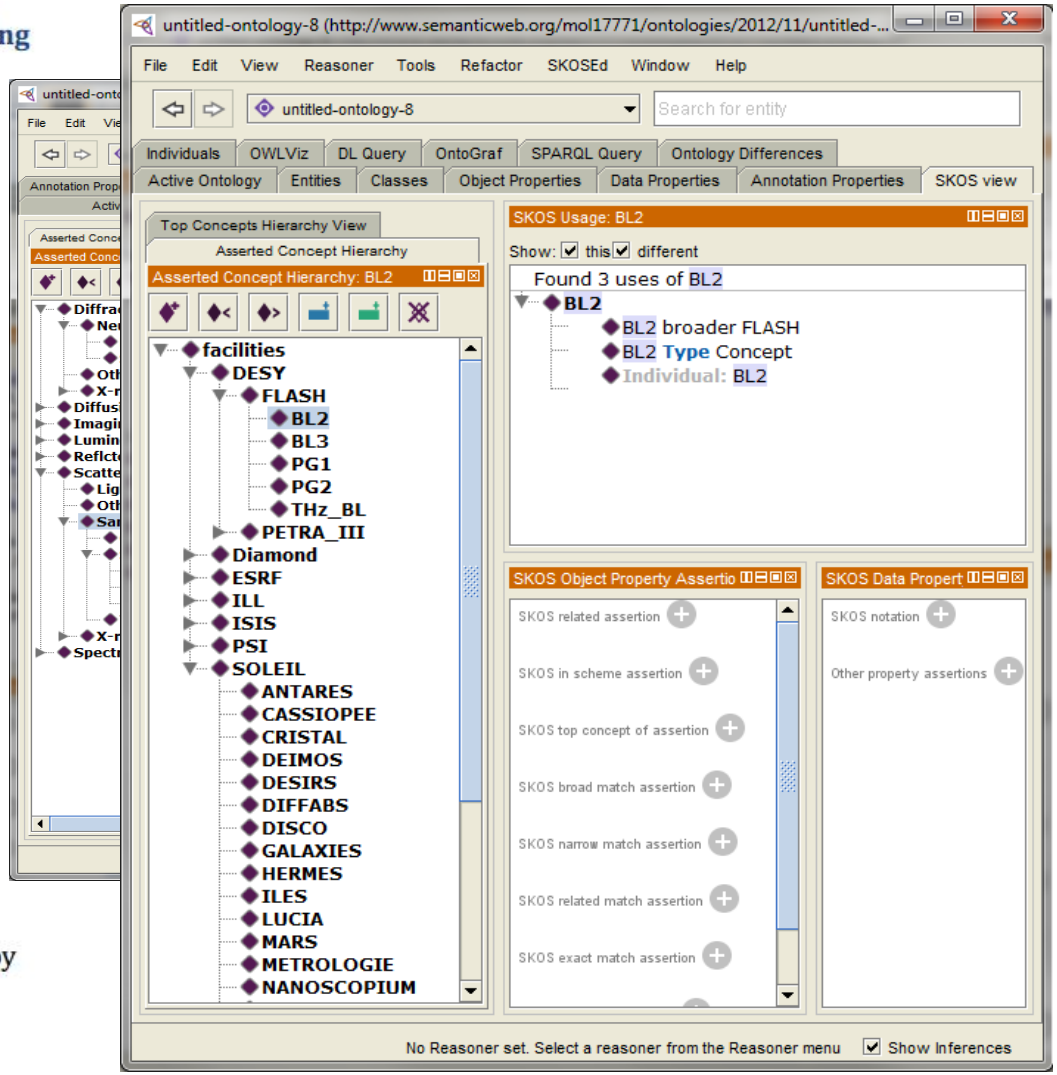
Diffraction

- 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

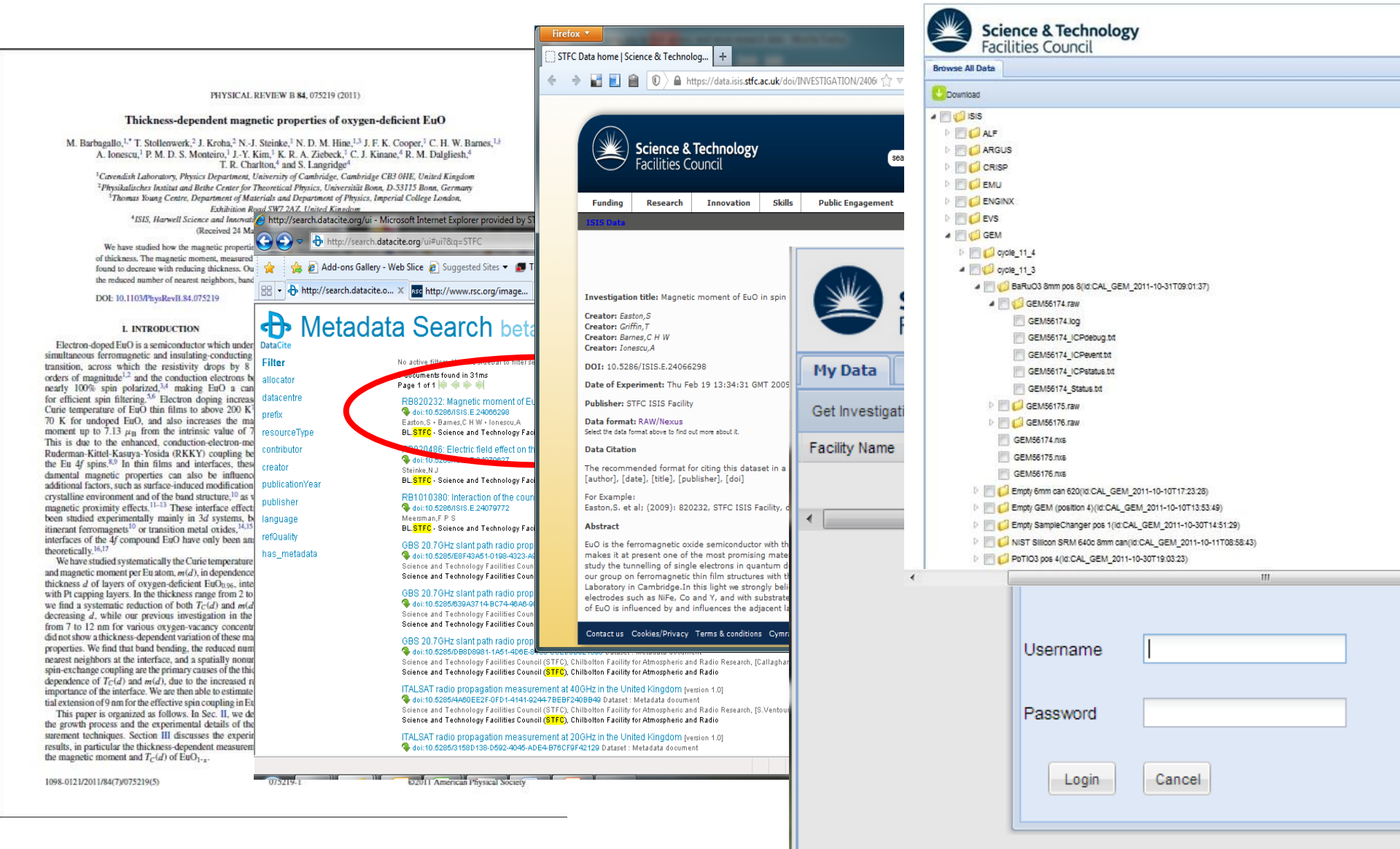


(Open) Data Access

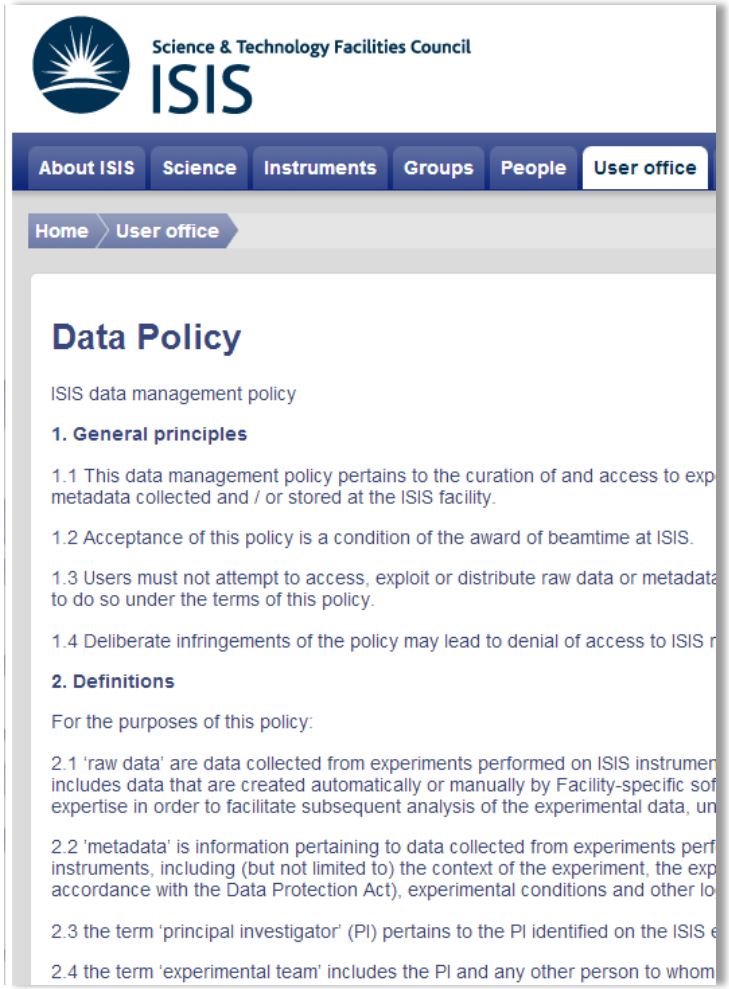
Credit: Brian Matthews

DOI Data Access Process

Paper → DataCite → STFC Page → TopCAT



Data Access and Open Access



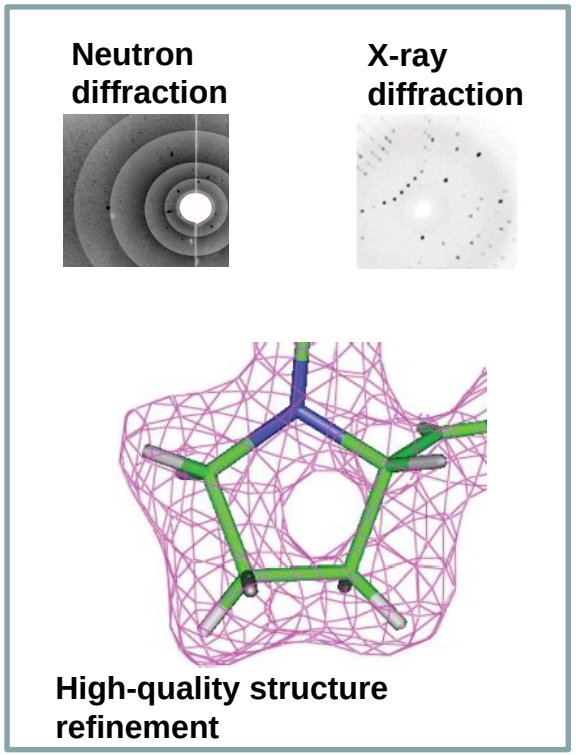
- Access to the on-line catalogue will be restricted to those who **register** with STFC/ISIS as users of the on-line catalogue.
- Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of **three years** after the end of the experiment. Thereafter, it will become publicly accessible.
- The term 'long-term' means a minimum of **ten years**.

http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

Outlooks

Developments that will influence how the data is managed

- Facilities offer complementary experimental techniques for a single beamline (e.g. tomography+diffraction)
- Users increasingly use multiple facilities leading to the need for multi-stream data fusion and processing
- Computational needs of experiments
- The rise of data intensive experiments and computation
  - Real time data processing for live experiments
  - Streaming data processing



Acknowledgement

- STFC (Scientific Computing and ISIS)
  - Brian Matthews, Steve Fisher, Alistair Mills, Kevin Philips, Anthony Wilson, Tom Griffin, Holly Zhen, Juan Bicarregui, Martin Turner, Ronald Fowler
- 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…

Science & Technology Facilities Council

Managing Research Data for Diverse Scientific Experiments

Erica Yang

erica.yang@stfc.ac.uk

Science & Technology Facilities Council
Rutherford Appleton Laboratory

RESEARCH COUNCILS UK

JISC

SEVENTH FRAMEWORK PROGRAMME

Harwell Imaging Partnership
Science & Technology Facilities Council

pandata
Photon and Neutron Data Infrastructure