WP6: JRA on Provenance

Brian Matthews

Scientific Computing Department, STFC, Didcot, UK

brian.matthews@stfc.ac.uk

WP6 : Provenance

To research and develop a conceptual framework, defined as a metadata model, which can record the analysis process, and to provide a software infrastructure which implements that model to record analysis steps hence enabling the tracing of the derivation of analysed data outputs.

- To develop a conceptual framework, which can record and recall the data continuum.
- To provide a software infrastructure which implements that model to record analysis steps

- In general
  - A large and complex task
  - Establishing Science benefit

- Start M7 (April 2012), Finish M30
  - STFC (Lead) 18 SM
  - ILL 6 SM
  - ELETTRA 12 SM

Tasks and Deliverables

- Task 1: Requirements for Provenance
- Task 2: Modelling the data continuum
- Task 3: Ontologies for specific instruments/techniques
- Task 4: Tool Support for the Data Continuum
- Task 5: Tracing the Data Continuum
- Task 6: Evaluation

- D6.1: Model of the data continuum in Photon and Neutron Facilities (M12)
- D6.2: Common ontology definition and definition of tools to support the use of provenance for Photon and Neutron Facilities (M18)
- D6.3: Tools for building research objects in Photon and Neutron Facilities (M24)
- D6.5: Evaluation report on provenance management in Photon and Neutron Facilities (M30) .

- Define Data Continuum
  - Define the stages of the facilities process we would need to support for end-to-end coverage
  - Consider who is involved
  - Consider the data and metadata involved
- Costs and Benefits
- Use cases
  - first sketch
  - Derive requirements

Data Continuum

Proposal

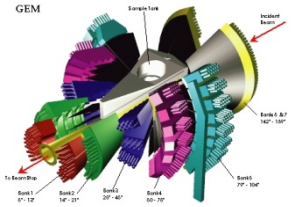
Scientist submits application for beamtime

Approval

Facility committee approves application

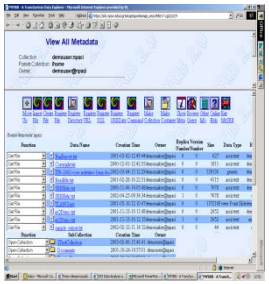
Scheduling

Facility registers, trains, and schedules scientist's visit



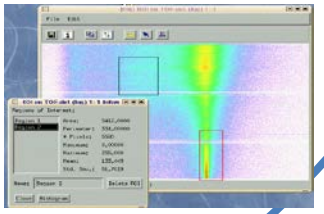
Experiment

Scientists visits, facility run's experiment



Data cleansing

Raw data filtered and cleansed



Data analysis

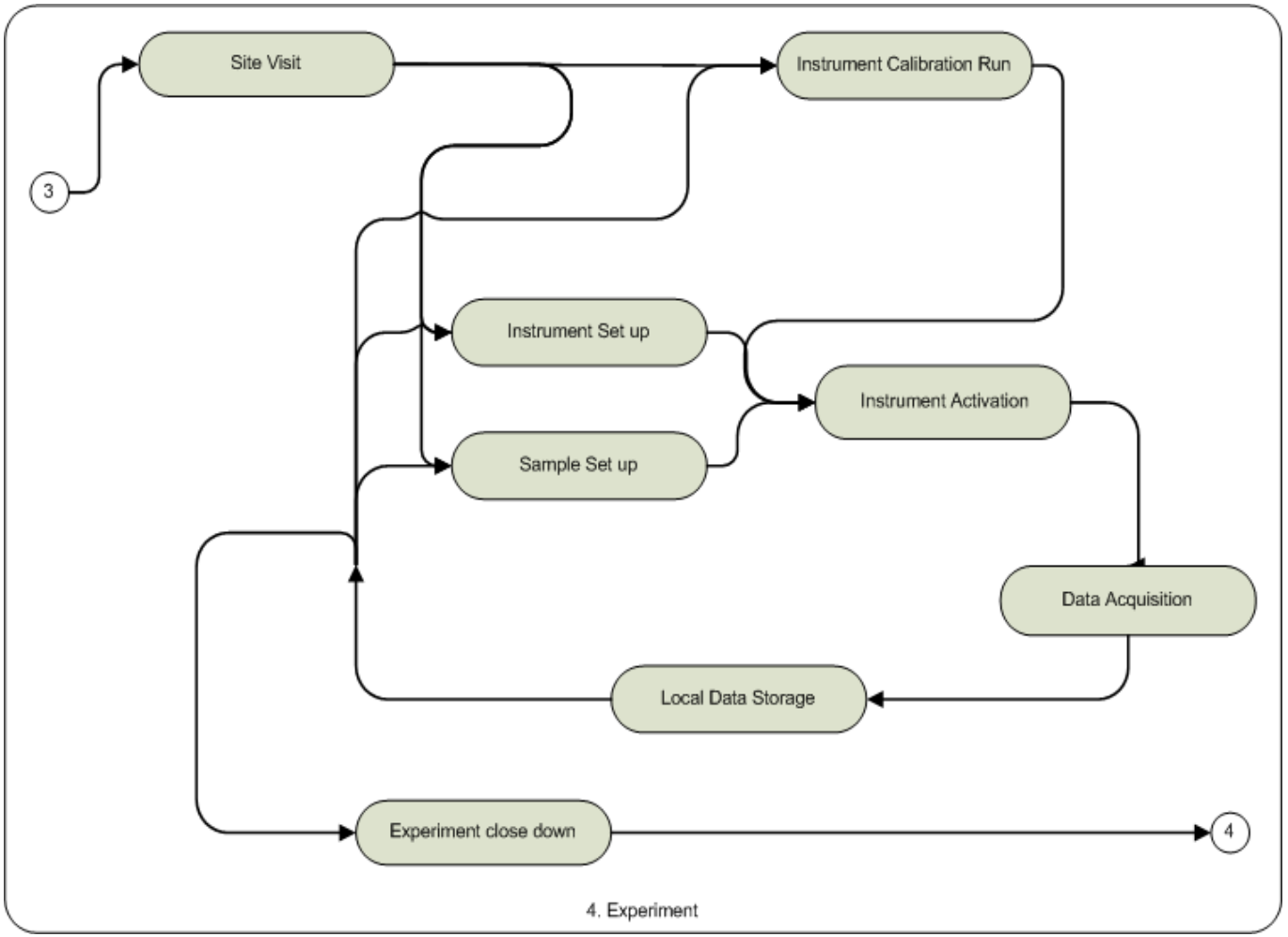
Tools for processing made available

Record Publication

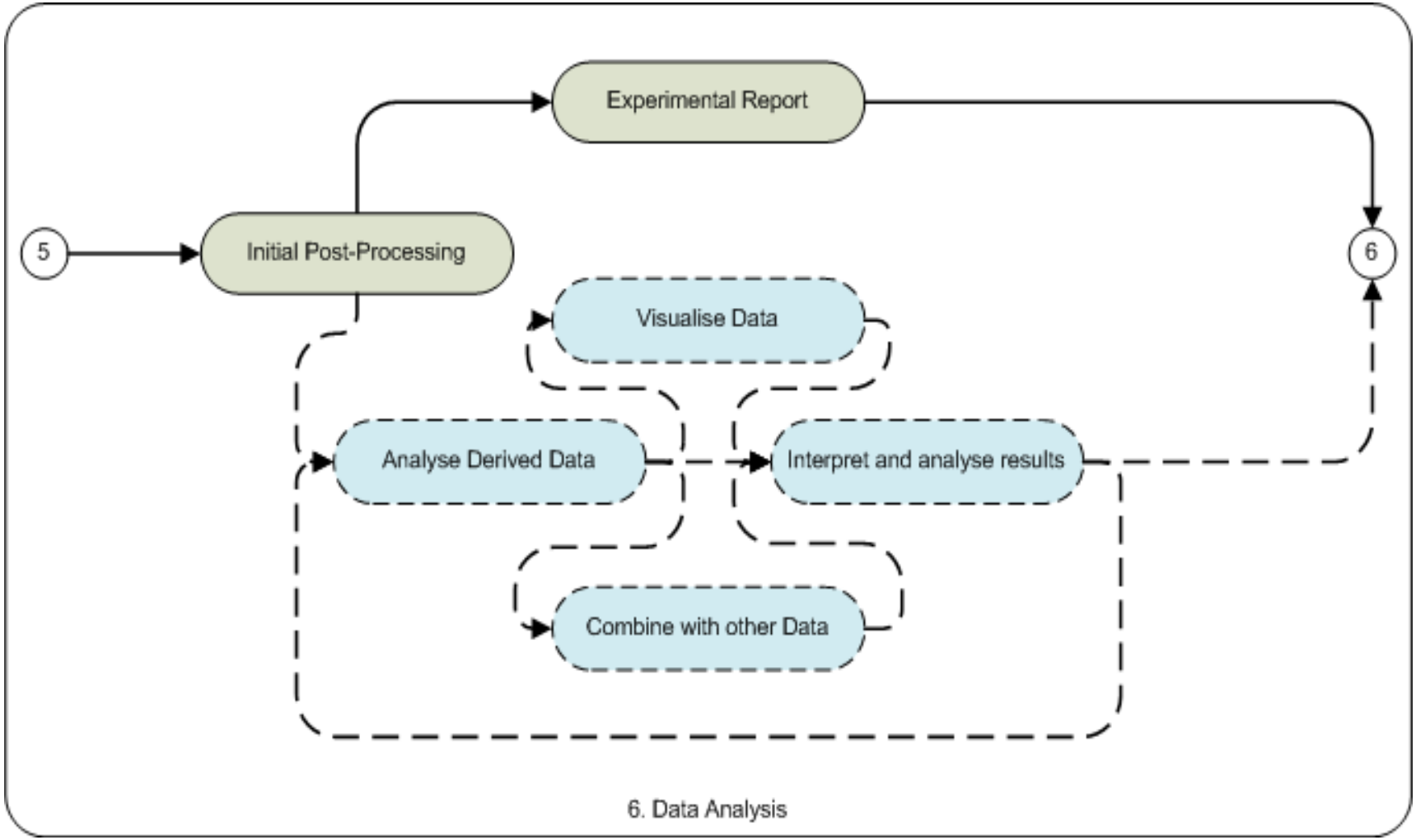
Subsequent publication registered with facility

Stage	Actors	Information Systems	Metadata Types
1. Proposal	User Office, Principal Investigator	User office systems, User registration and management, User identity, Proposal systems	User identity, funding sources, institutional information, project description, experiment description, prior art
2. Approval	User Office, Approval Panel	User Office Systems, Proposal Systems	User identity, funding sources, experiment description prior art
3. Scheduling	User Office, Experimental Team, Instrument scientist	User Office Systems, Proposal Systems, H&S systems, Scheduling, Sample tracking	User identity, Sample information, Instrument information, Experiment planning
4. Experiment	Experimental Team, Instrument scientist	Sample tracking, Instrument control, Environmental monitoring, Data Acquisition systems, Data Management systems	User identity, Sample information, Instrument information, Experiment planning, Environmental parameters Calibration information
5. Data Storage	Scientific Team, Instrument scientist, Data infrastructure team	Data Acquisition systems, Data Management systems Data storage systems, Archival Systems	User identity, Data formats, Data set information, File identifiers ….
6. Data Analysis	Scientific Team, Instrument scientist,	Data Storage systems, analysis software packages,	User identity, Data formats, Data set information, File identifiers Instrument parameters …
7. Publication	Scientific Team, Instrument scientist, User Office, Library	User office systems Research Output tracking systems Library systems	User Identity Proposal information Publication information

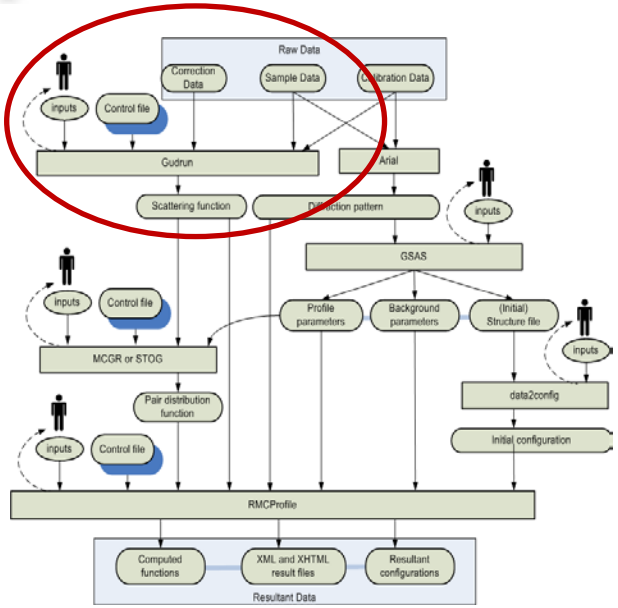
Experiment Stage



Data Analysis Stage



Lessons from GEM



Raw data

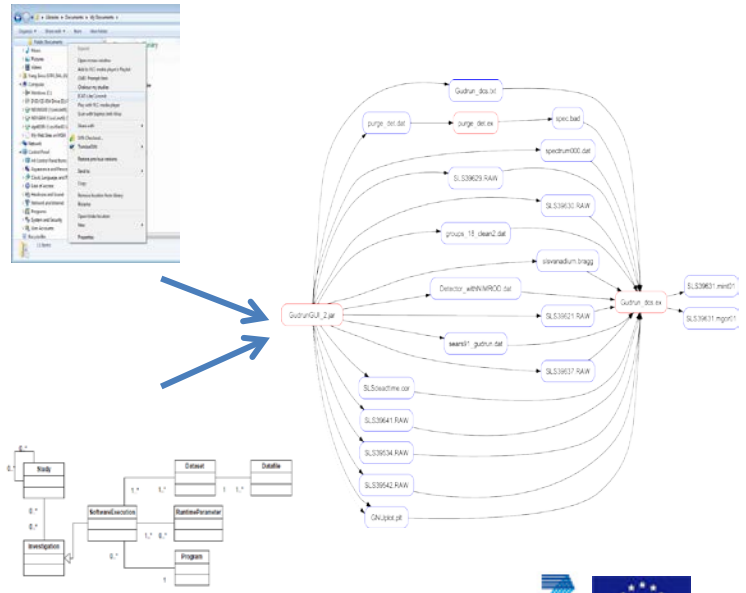
Derived data

Resultant data

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

Issues:

1. Complexity of dependencies
2. Data volumes
3. Valuable data amongst noise
4. Software tracking
5. Distributed analysis
6. Workflow
7. Integration of tools



Researchers are hesitated to change their well established software/practice.

"Why would I change?"

Need to demonstrate the benefits!

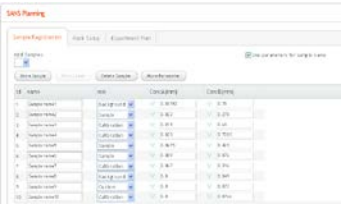
Use Cases for Requirements

- Analyse practical situations where we can gain benefit from recording provenance

- A number of use cases
  - SANS2d (ISIS)
    - Co-ordinating and automating “near to experiment” processes
  - TwinMic x-ray spectromicroscopy beam line (ELETTRA)
    - Co-ordinating multiple experimental runs
  - Tomography beam lines (DLS)
    - Managing high volumes of data
  - Express services (ISIS)
    - Automating “standardised” experiments.
  - Publication linking (ISIS)
    - Tracing research outputs for impact analysis

- Would explore some of these in more depth in the project – related to VLabs

SANS2d: Experiment coordination



SampleTracks

OpenGenie Script

Data Acquisition

Data Archive

Sample Information

Mantid

Data Processing

raw data

ICAT

(Extended) ICAT Data Catalogue

DOIs

British Library DOI Server

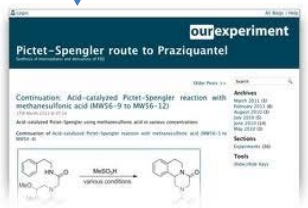
- **Small Angle Scattering VLab**

- **WP4, WP6, WP7**

Outputs

derived data

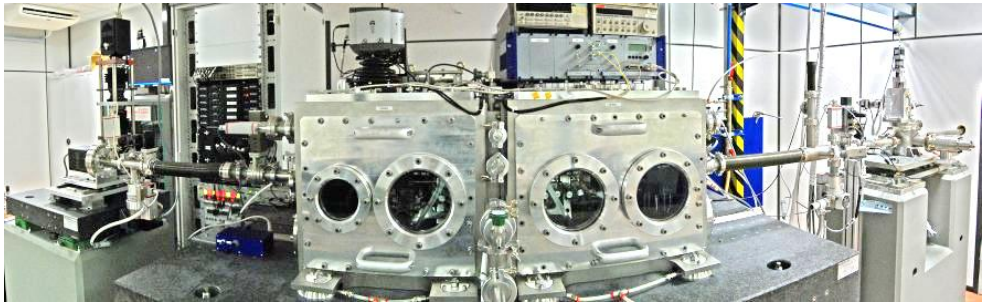
New links



ELN

Publications

TwinMic beamline

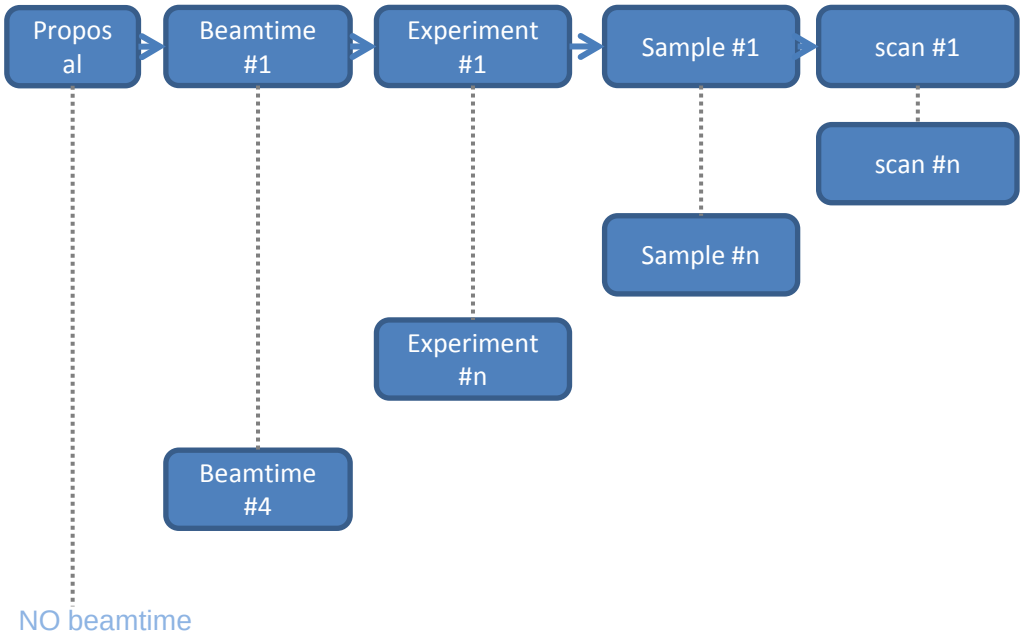


Beam-time Proposal has lots of meta-information

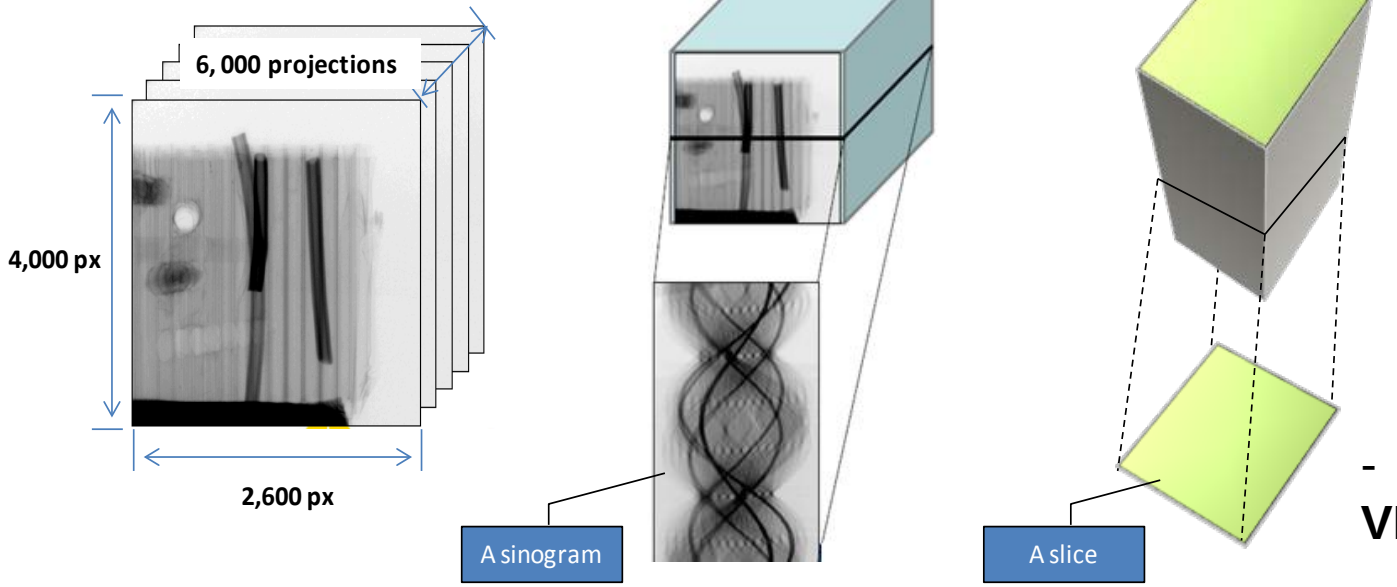
- **1 LT proposal may grant up to 4 beam-times**

- x-ray spectromicroscopy
- 1 instrument →2 modes: Scanning & Full field

- **Modified ICAT to support multiple beam-times**
  - **WP4, WP6**



Tomography Reconstruction



- **Tomography VLab**

- **WP4, WP6, WP8**

Raw images

(2,600 X 4,000 px each image/projection)
(Total: 6,000 images/projections)

Sinograms
from the projections

(Total: 4,000 sinograms
Each sinogram: 2,600 x 4,000 x 1 px)

Reconstructed slices
in a 3d volume

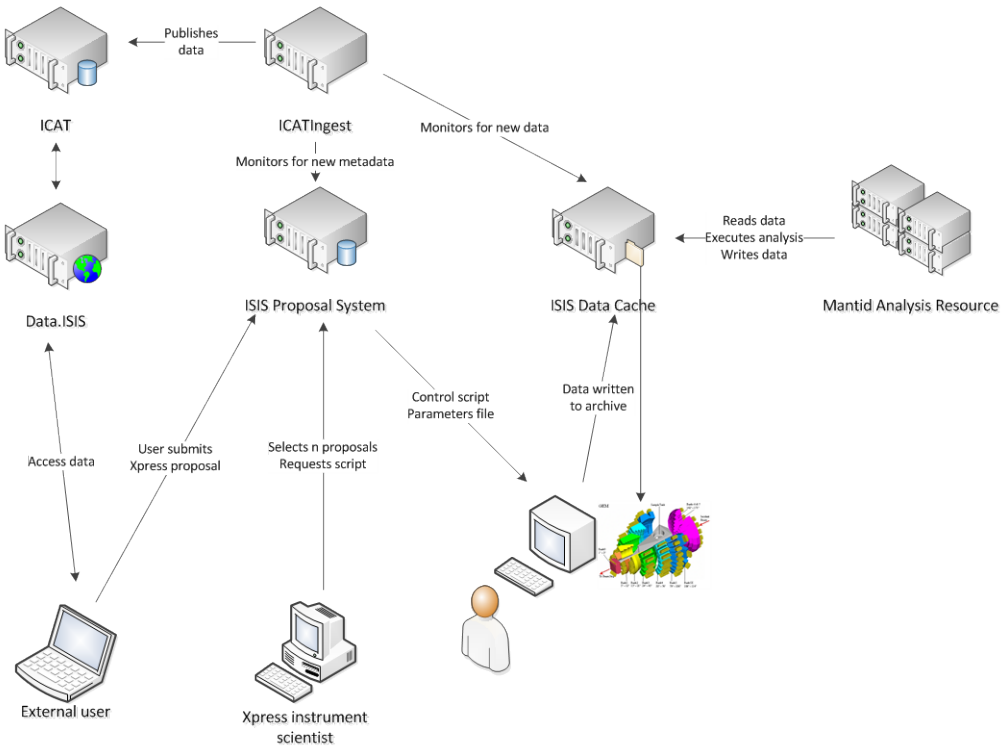
(Total: 4,000 slices)

Dealing with high data volumes – 200Gb per experiment in a reconstruction
- hard to move the data – needs to be handled at the facility

Credit: Dr. Mark Basham, Diamond,

GEM Xpress

- Powder Diffraction
- Experiment by courier
  - Facility staff carry out complete experiment
  - Return Fully reduced and corrected high-quality data to user
- Suitable for automation
- An example of a service which is common in facilities

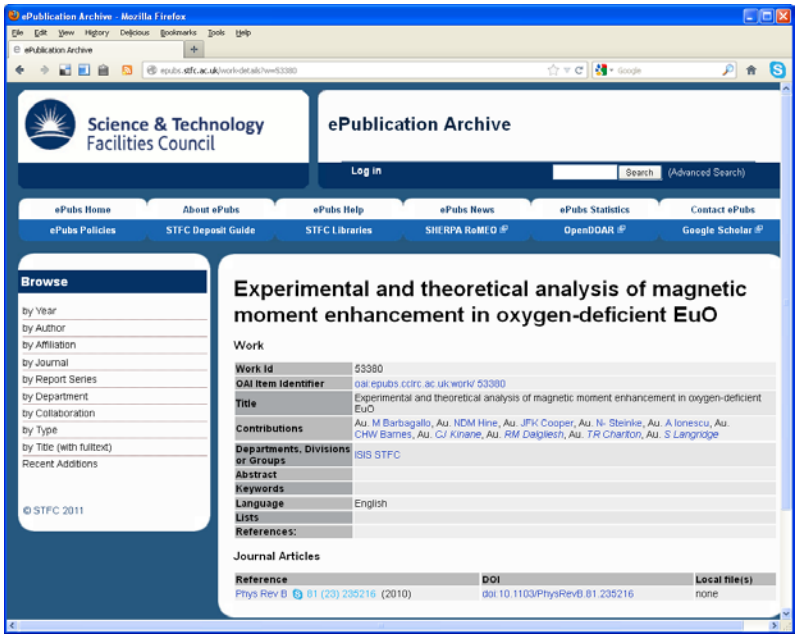


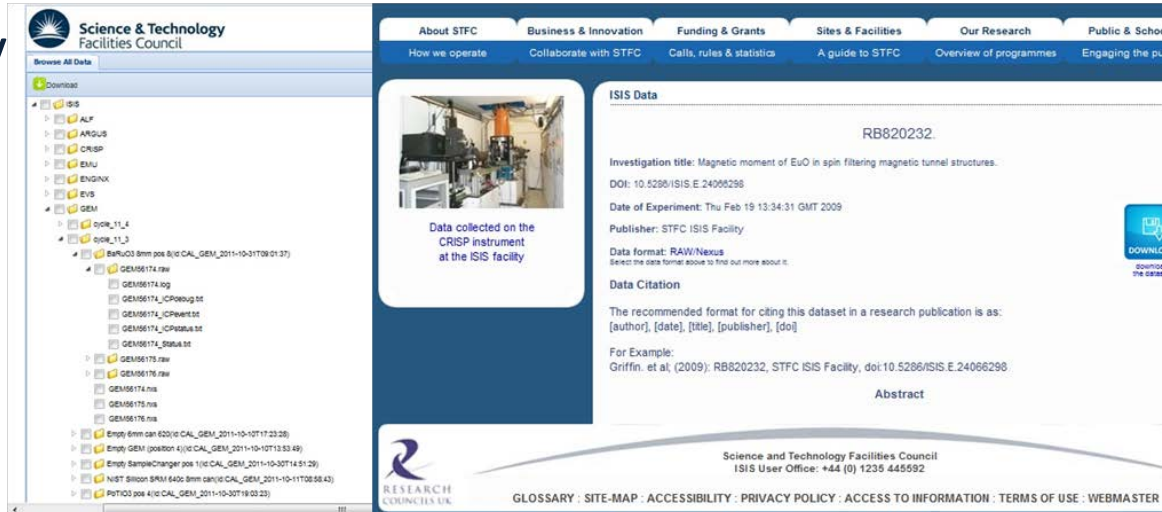
- **Powder Diffraction VLab**

- **WP4, WP6, WP7**

Resultant data and publication tracking

- Providing upload facility to collect resultant data and publication.
- Provide a link back to the experiment proposal and raw data
- Associated the “research object” with a DOI
- Impact tracking and analy
- **WP4, WP6, WP7**





- Controlled Vocabulary/Ontology
- Definition of tools which support Provenance

A Common Vocabulary

Agreed terms to be used to describe a number of different aspects of Facilities science

- Facility and Facility Type
- Analysis technique
- Instrument and Instrument Type
- Environmental parameters
- Sample information
- Measured parameters

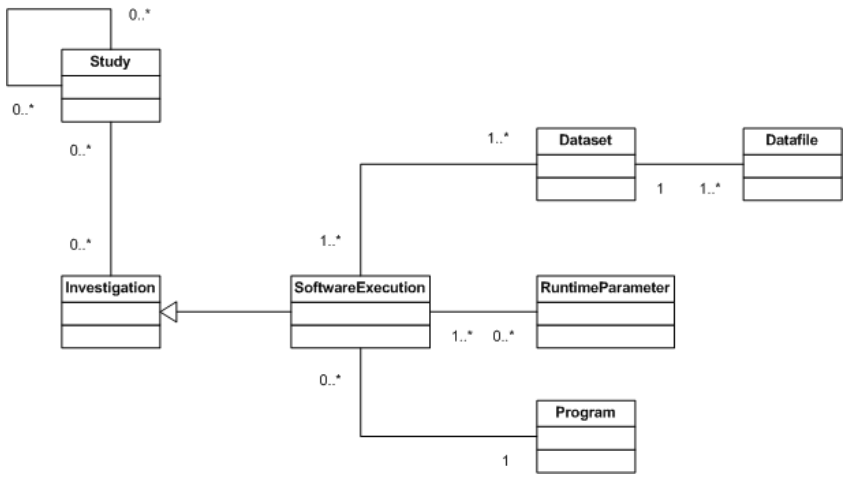
Can use in tools:

- Catalogue and Search tools
- Linking between tools
- Increase precision and reduce ambiguity

- Hard work to get agreed
  - needs community buy-in

Instrument
--- Neutron Instrument
------ ISIS Neutron Instrument
--------- Alf
--------- Crisp
--------- Engin-X
--------- GEM
--------- LOQ
--------- Merlin
--------- OSIRIS
--------- SANS
--- Muon Instrument
------ ISIS Muon Instrument
--------- Argus
--------- Deva
--------- Emu
--------- Hifi
--------- MuSR
--- Pulsed Laser Instrument
------ CLF Laser
--------- Astra-Gemini
--------- Artemis
Scientific Technique
--- Neutron Technique
------ Neutron Diffraction
------ Neutron Spectroscopy
------ Neutron Reflectrometry
------ Small Angle Neutron Scattering
--- Muon Technique
------ Muon Spectroscopy

Architecture to support provenance



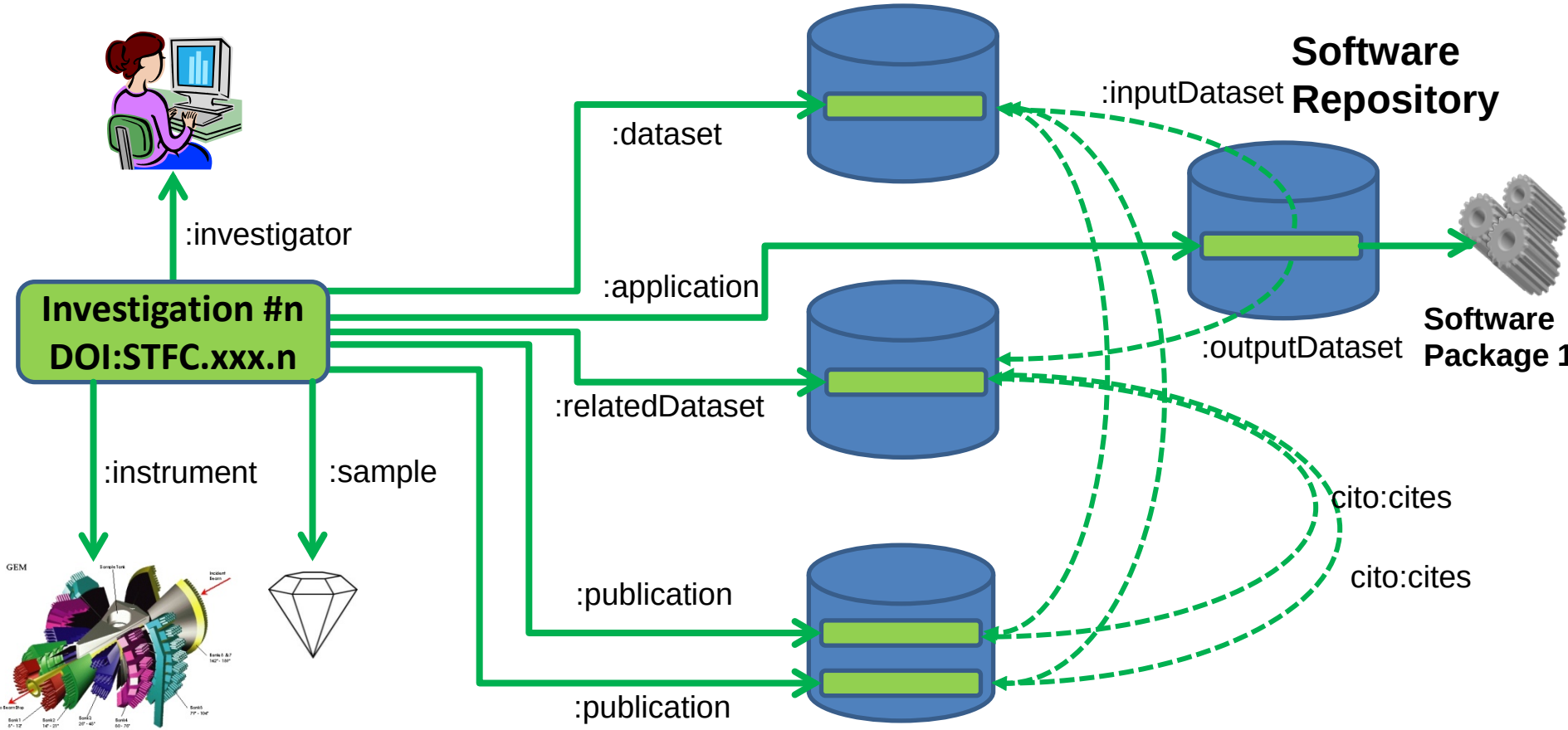
- Modified ICAT to support:
  - Derived data
  - Software, jobs
  - Linking between these
- Modification to the metadata model
- Some early prototypes
- Further develop and use PanSoft for reference software catalogue
  - Would need an API onto PaNSoft
  - Use common information model and vocabularies

Tools from WP6

Would like to see:

- Modified ICAT
  - Manage Applications and jobs
  - Manage dependencies between (raw and analysed) datasets
  - Manage link to Publications
  - Using controlled terms for Keywords
- ICAT Job Portal
- Modifed TopCat for front end
  - Can display “research objects”
  - investigations with all linked components
- PanSoft
- Common Metadata model (extended CSMD
- Controlled vocabulary
  - Instruments, techniques, parameters …

Linking the software application into the research object



- Own metadata format (CSMD)
- OAI-ORE
- W3C Prov ontology
- Assume that the software is in a repository

WP6: Provenance

Brian Matthews

Scientific Computing Department, STFC, Didcot, UK

brian.matthews@stfc.ac.uk