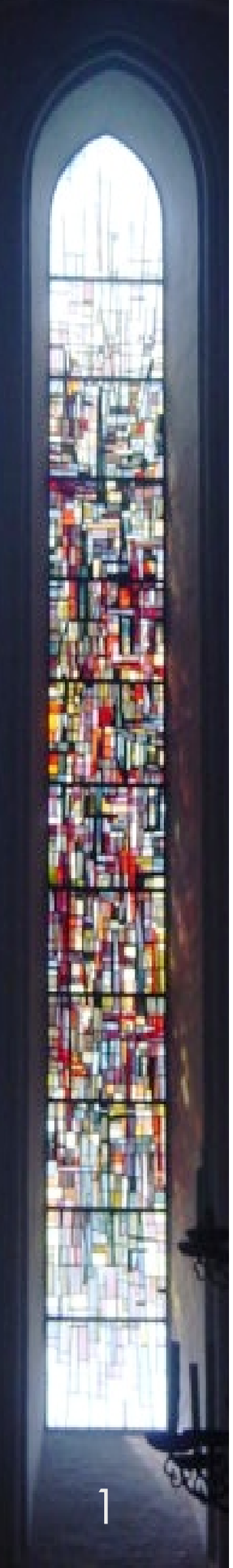
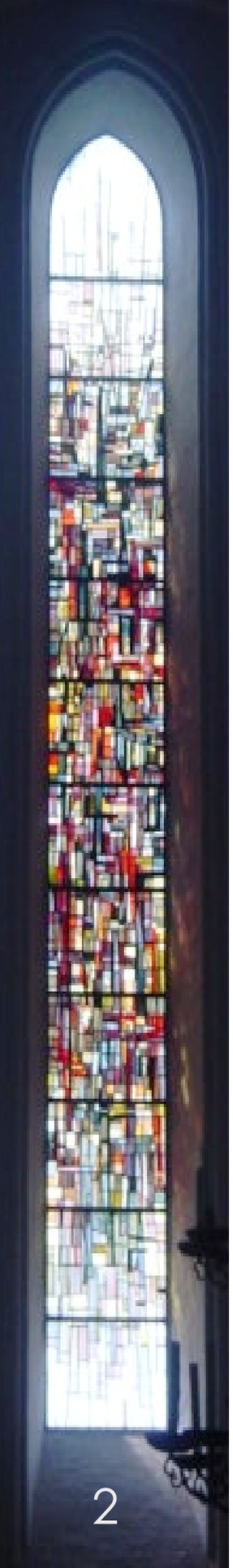
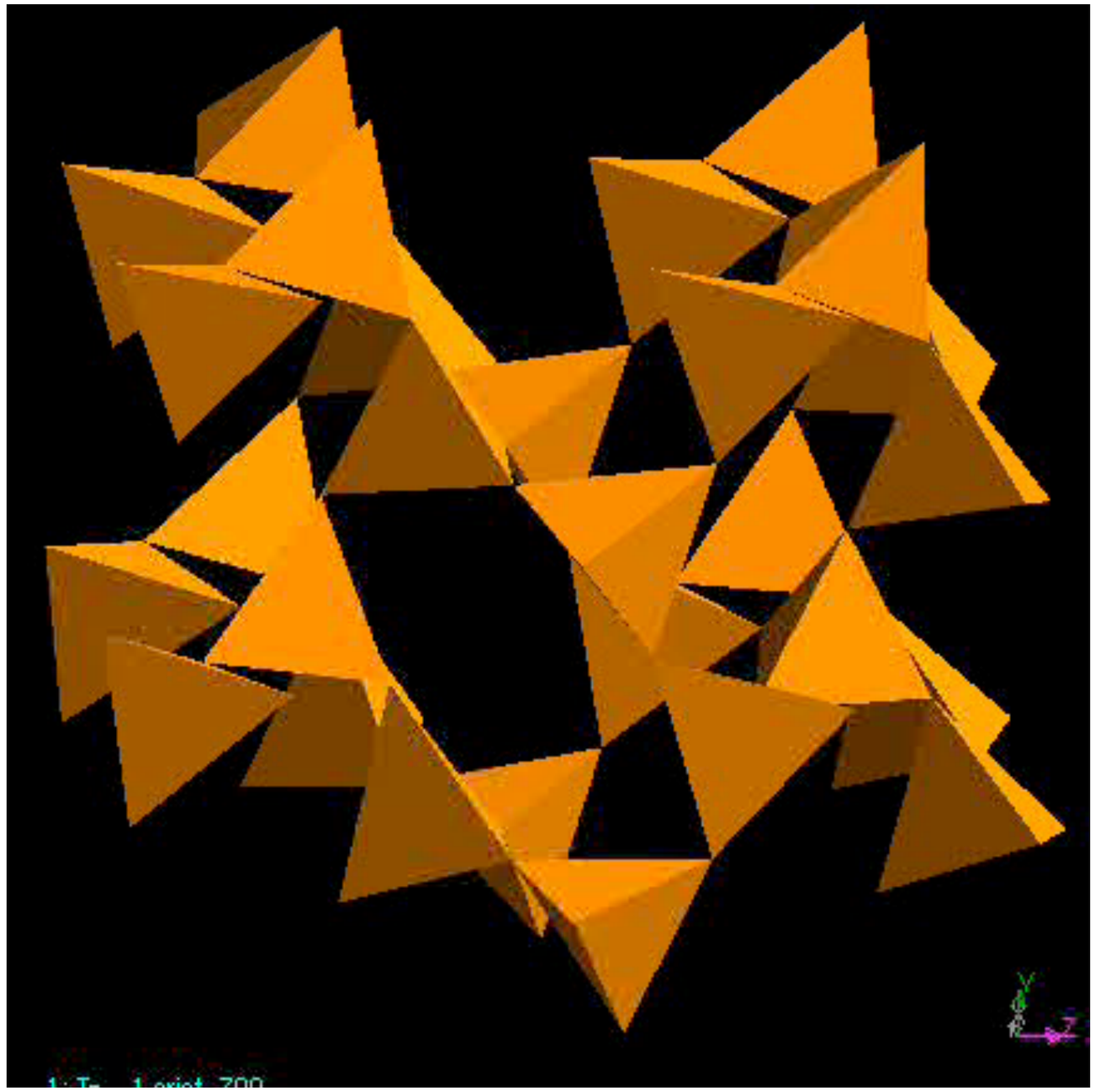


Probing inter-institute issues

Martin Dove (Cambridge) & Brian Matthews (STFC)
(scientist & e-scientist)



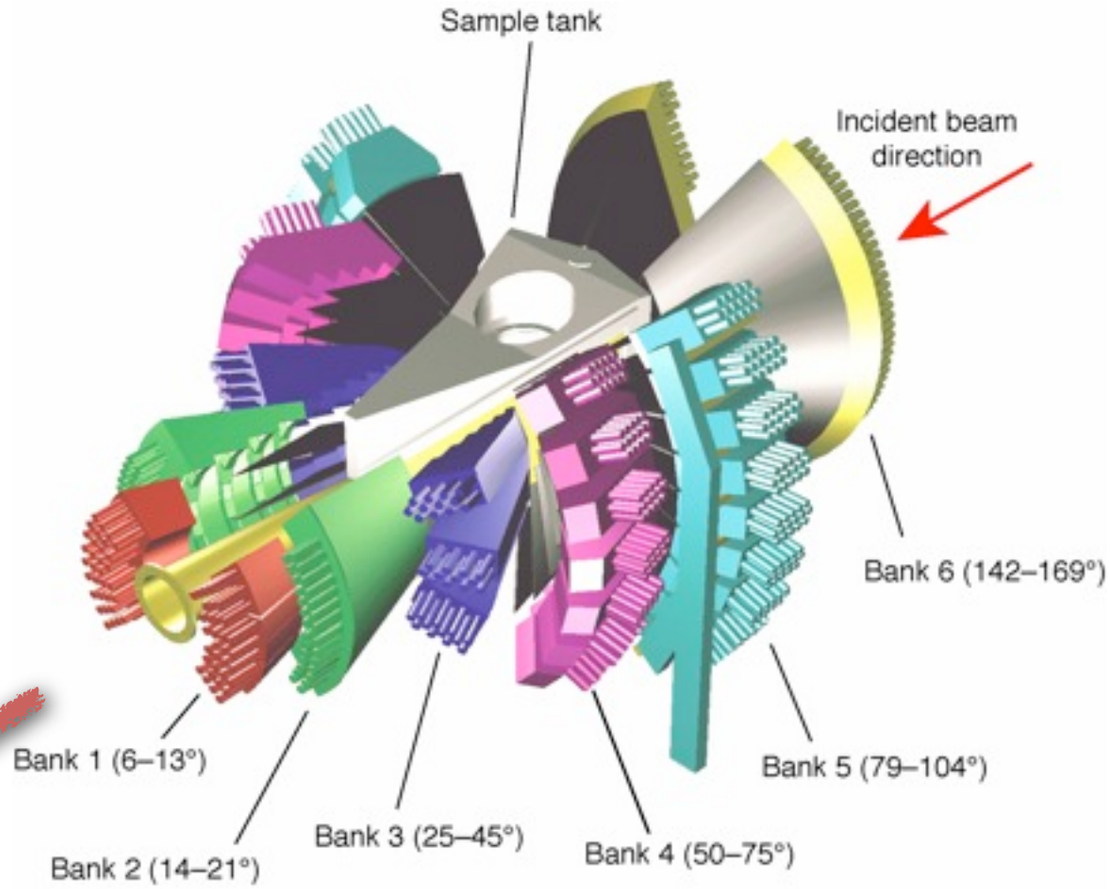
Structural science: example of disordered crystals

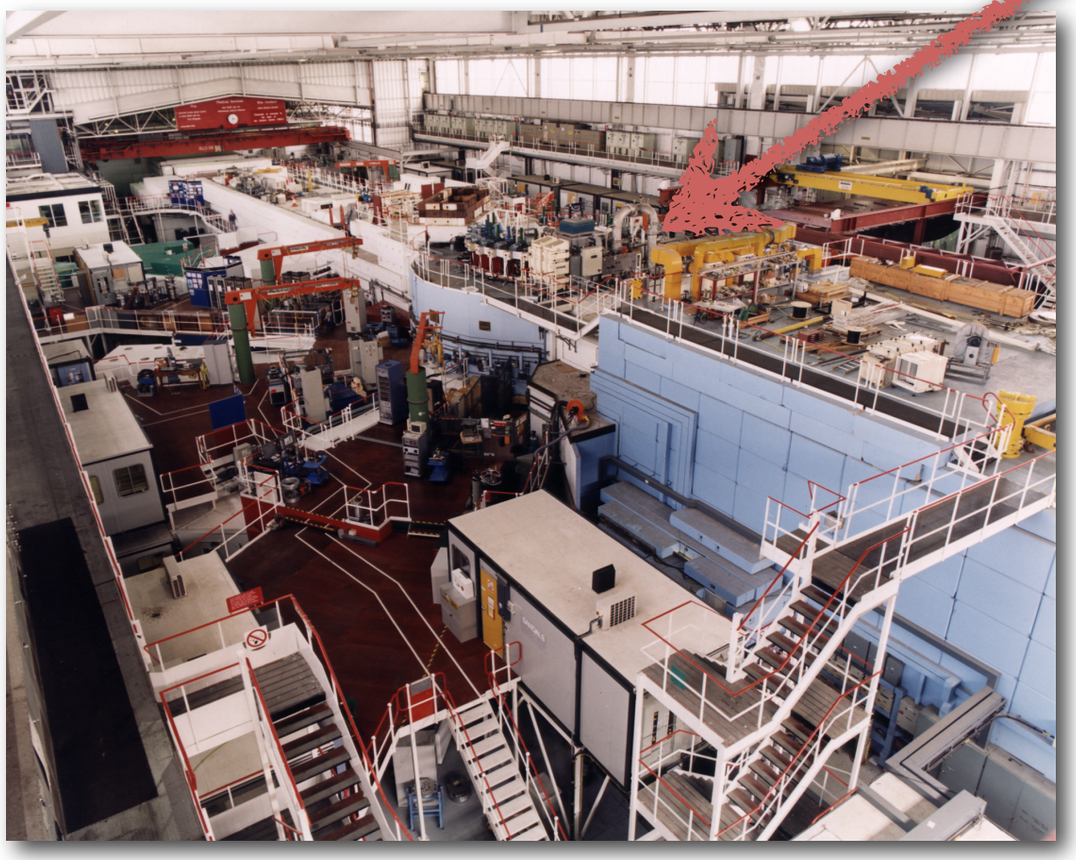


β-cristobalite, SiO_2

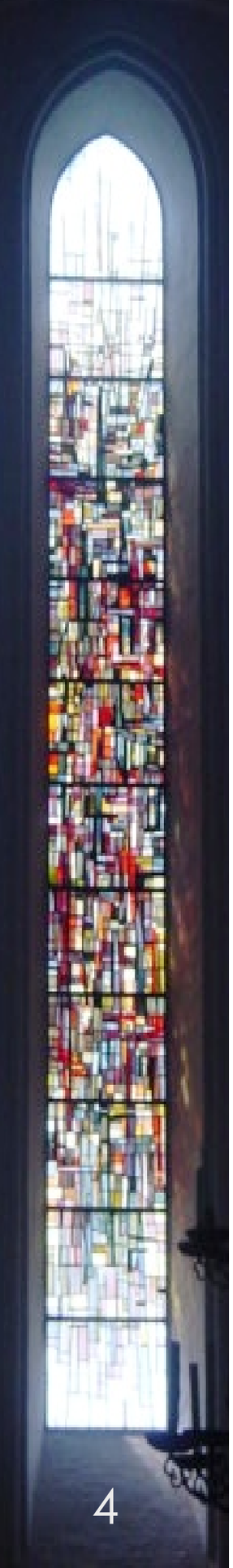
Data collection @ ISIS

- ISIS is the world's second most intense pulsed source of neutrons
- It has over 30 instruments

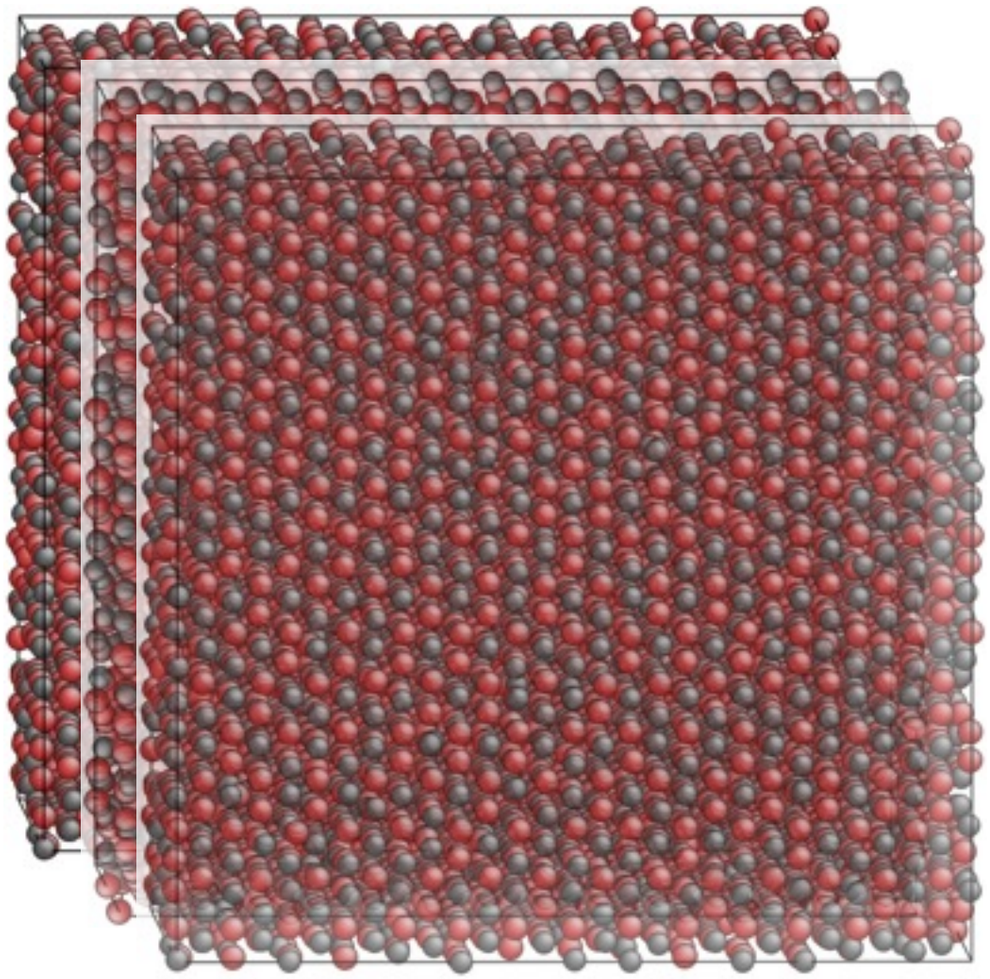




- GEM has ~4000 detectors
- Each experiment produces a histogram for each detector
- These histograms need to be corrected and merged to form data for analysis



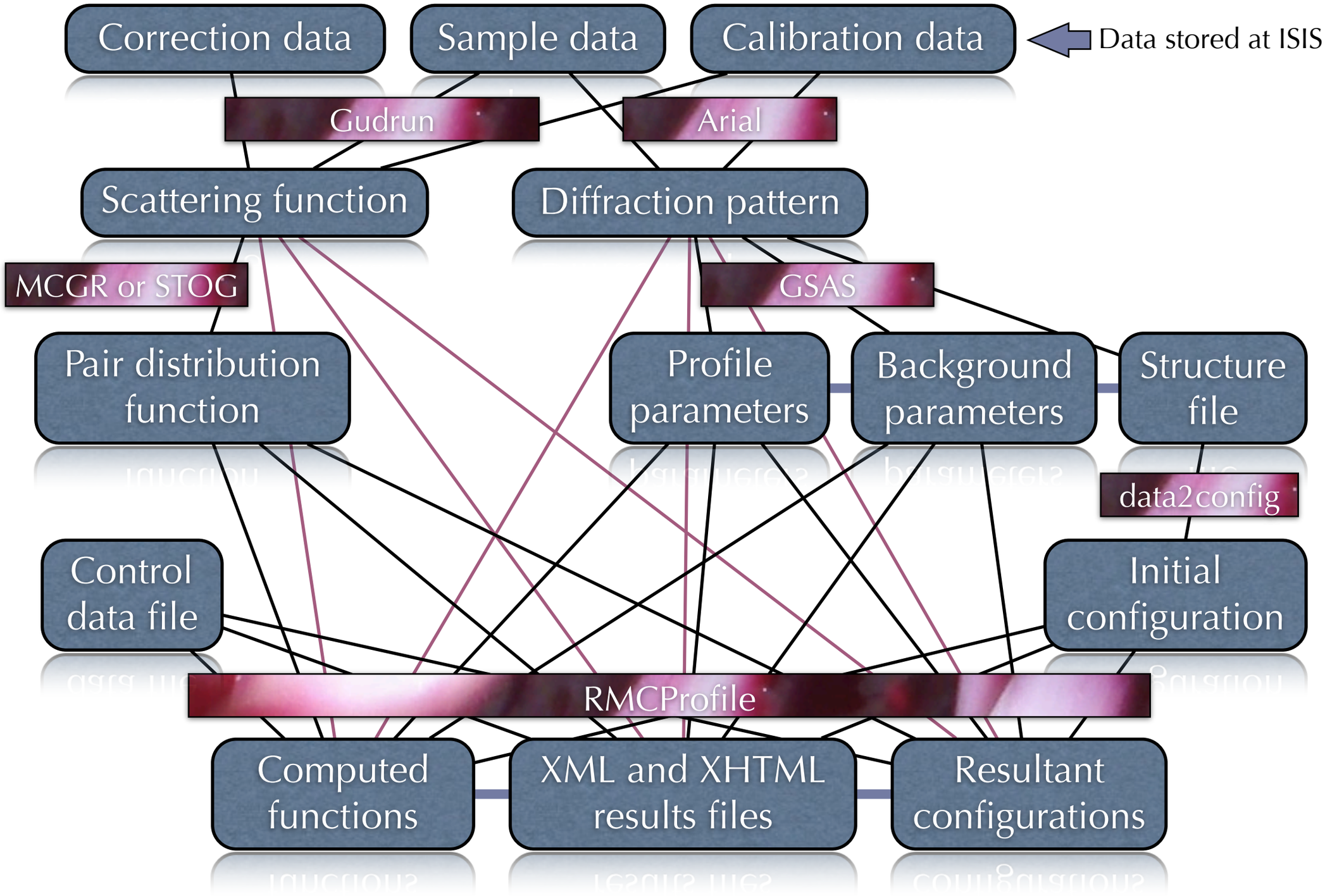
Scientific analysis

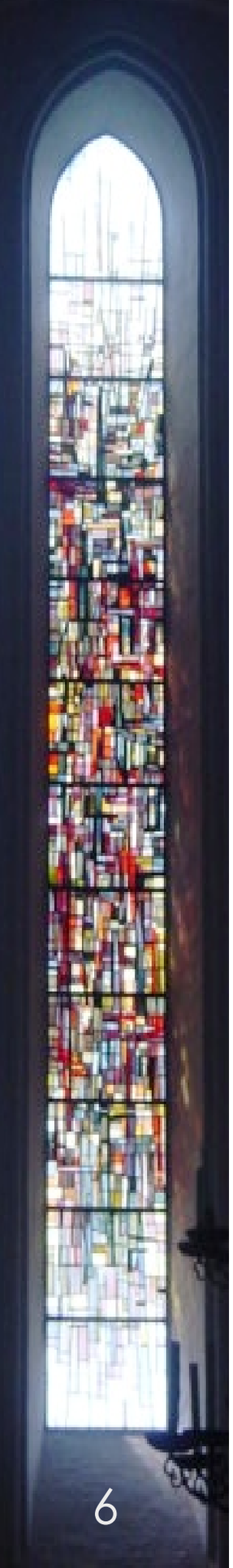


Our aim is to construct large-scale atomic models that best match experimental data for subsequent analysis

This involves using modelling techniques that are computationally demanding, and which require high-quality data

RMC data flow



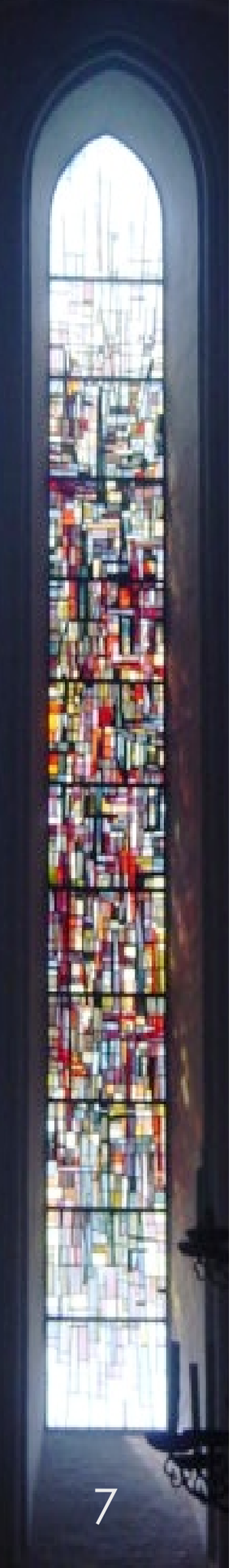


Main challenges

Although ISIS provides good long-term stewardship of the raw data, the important *derived* and *model* data are not managed at all in any collaborative sense …

… new approaches are needed to facilitate sharing data between collaborators, and locating previous data

Described by Neil Beagrie next



Issues in the challenges

Processing pipeline is dependent on a suite of software

- instrument specific (GUDRUN)
- closed (GSAS)
- written in-house (data2config, RMCProfile)

Contextual information is not routinely captured

Analysis is reliant on scientist's knowledge and experience

- in selecting parameters and interpreting data
- not recorded or captured other than in a lab note book

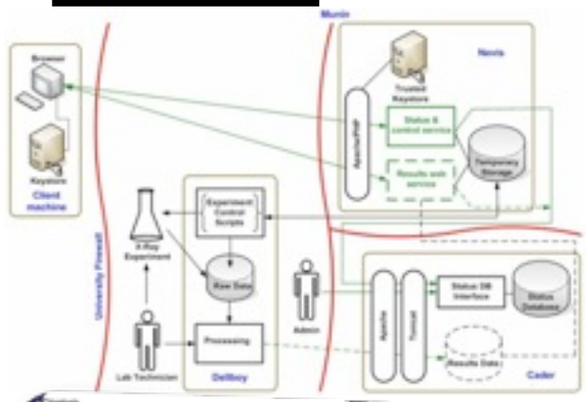
The actual workflow is not recorded

Distributed data - little or no shared infrastructure

- Raw and reduced data are stored at ISIS
- All derived and analysed data are managed and maintained by the individual scientist on his/her computer or WebDAV server

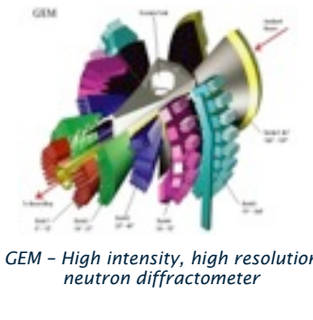
Mapping across organisational infrastructures

Home institution

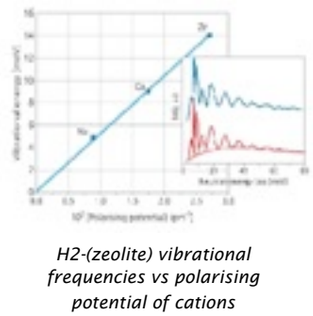


Central facility (eg ISIS)

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

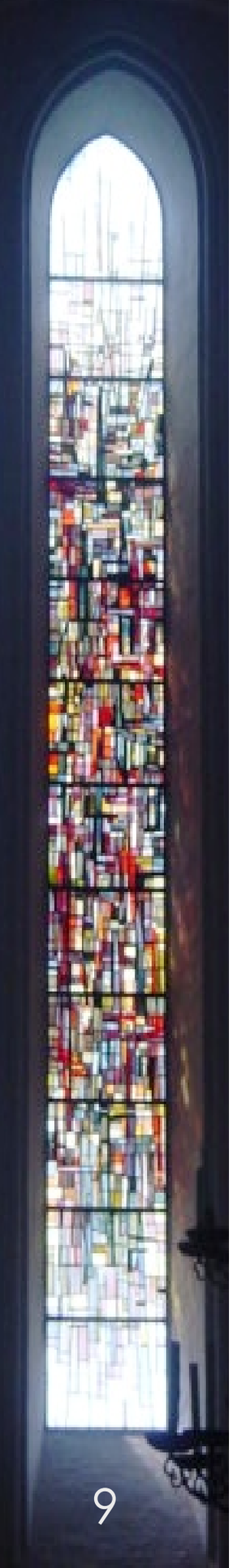
Data collected from your experiment will be indexed by ICAT (with additional experimental conditions) and made available to your experimental team

Analysed Data

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

Publication

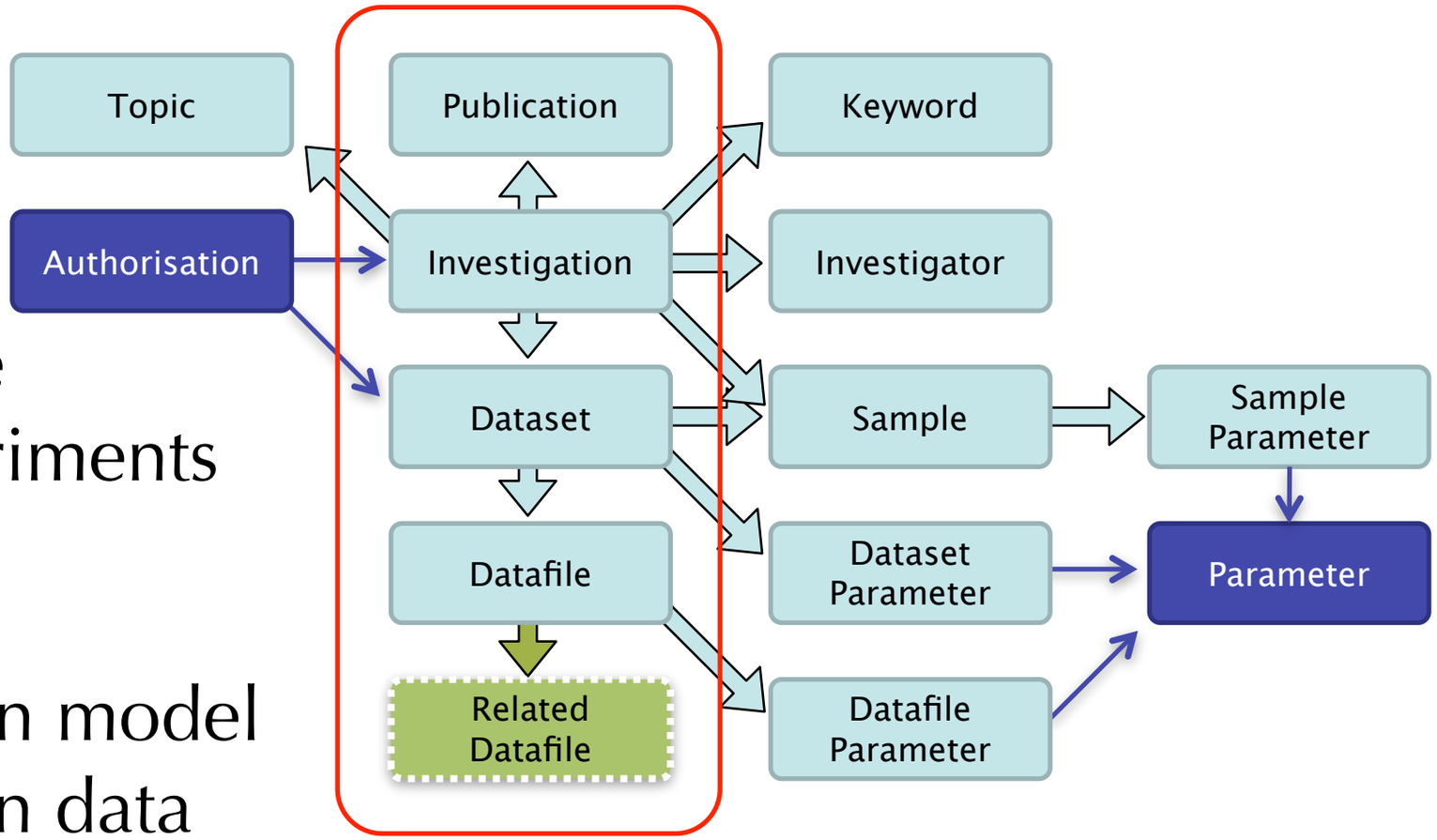
Using ICAT you will also be able to associate publications to your experiment and even reference data from your publications.



Core Scientific Metadata model

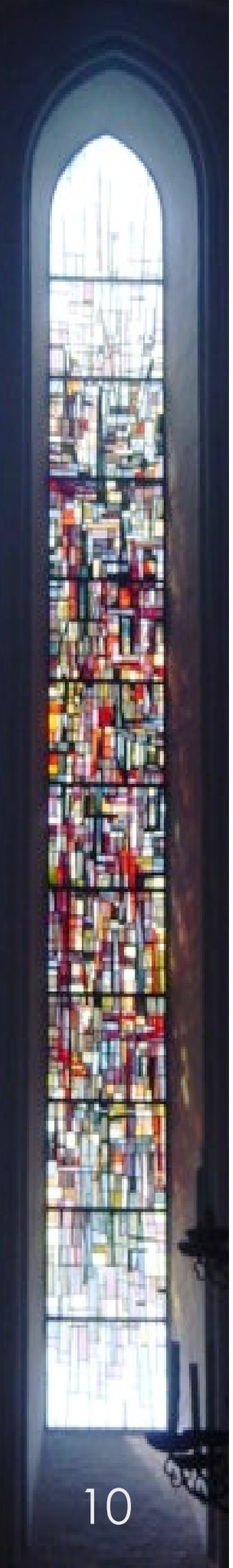
Designed to describe facilities-based experiments in Structural Science

Forms the information model for ICAT, a production data management infrastructure employed by STFC



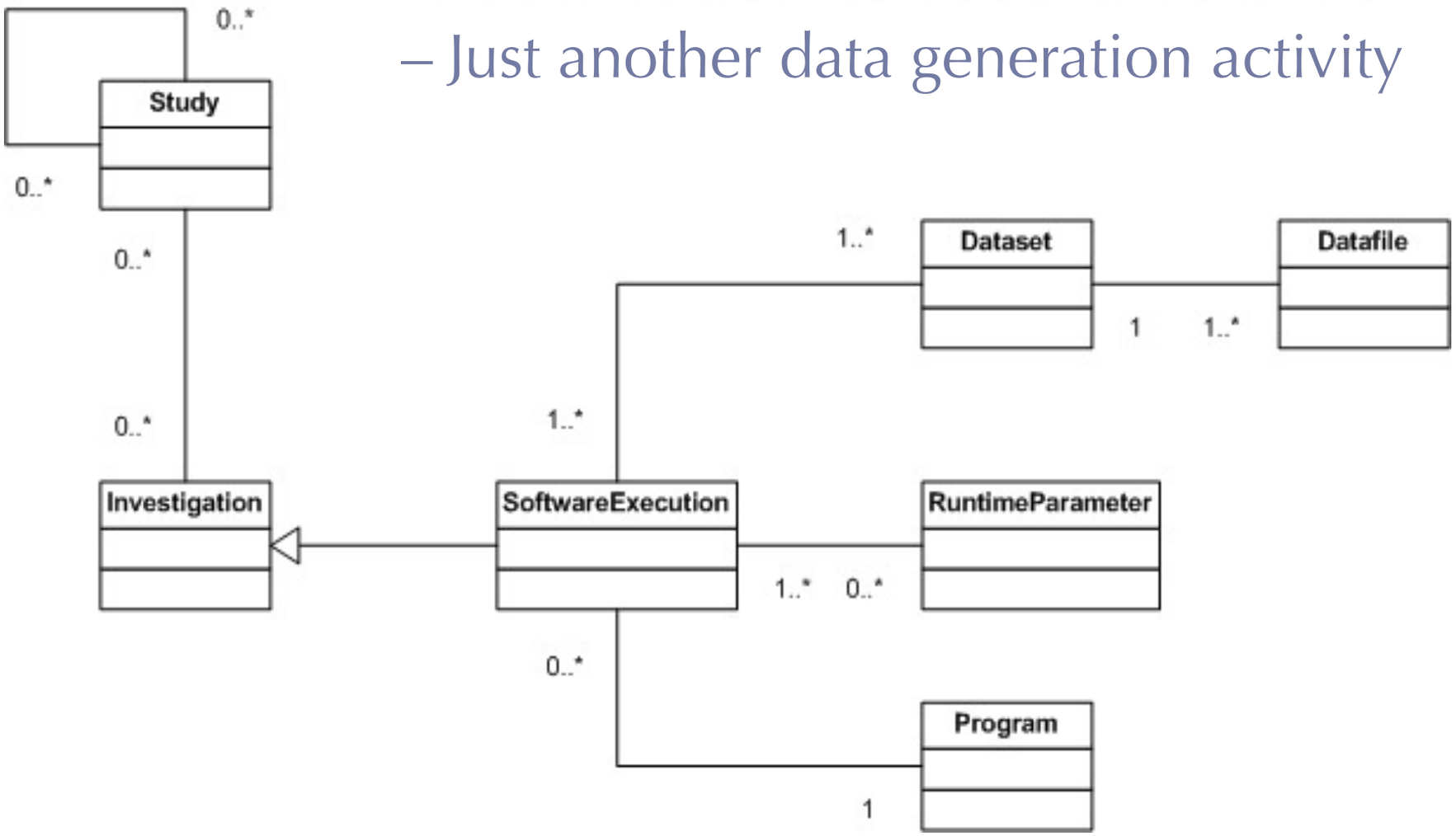
Forms the basis for extensions to:

- derived data
- laboratory based science
- secondary analysis data
- preservation information
- publication data

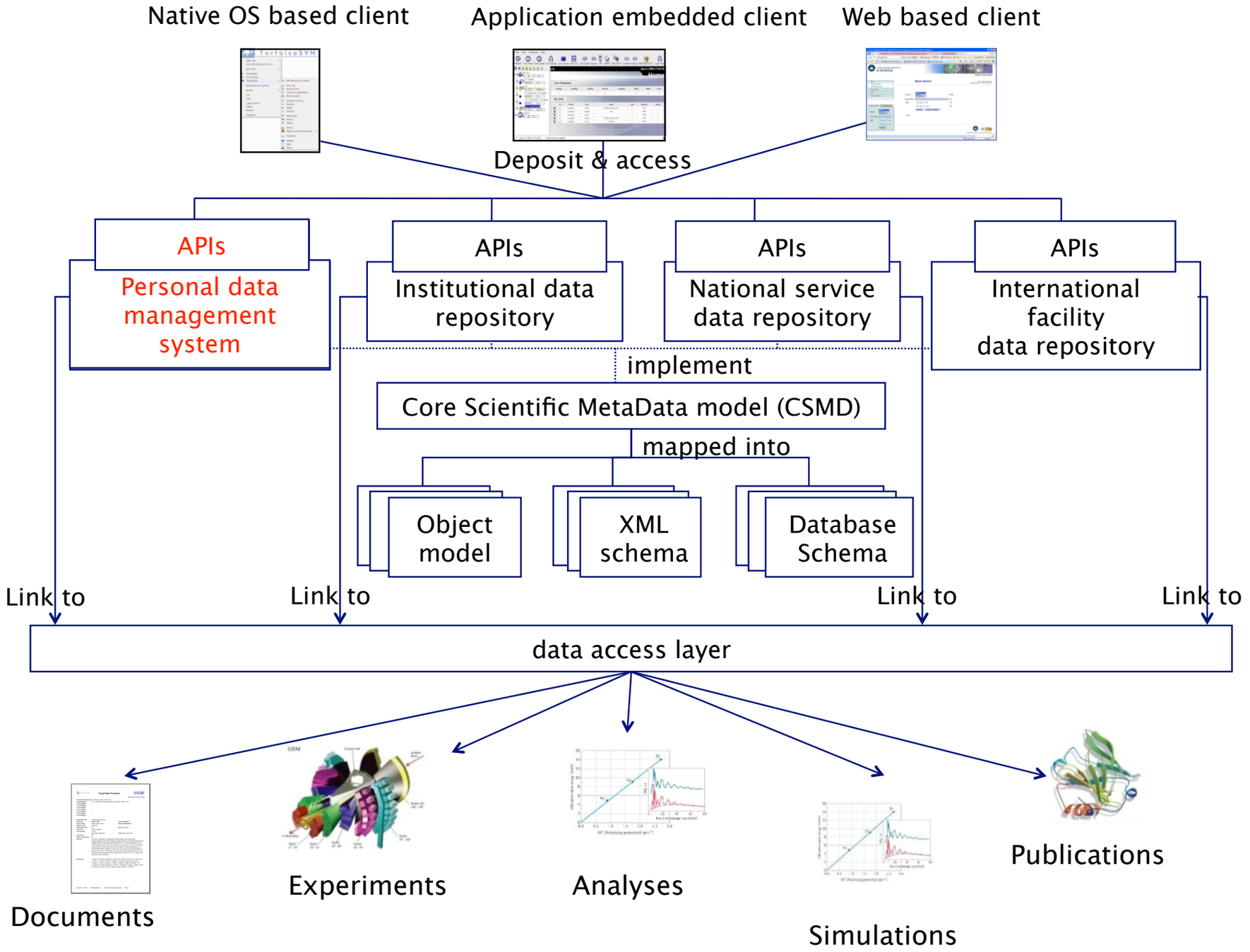


Model with Data Derivation

- Extension to the model to add an alternative Investigation activity type
  - Very straightforward natural extension to the model
- ICAT can be used almost without modification to record data derivation
  - Just another data generation activity



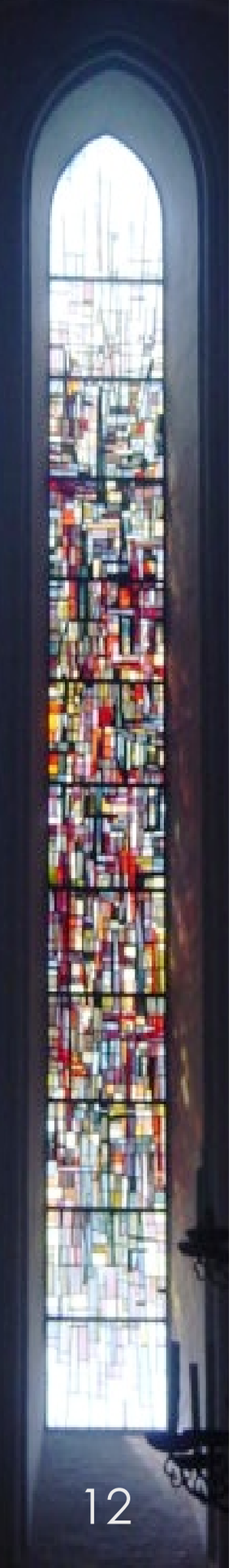
Architecture



Data management client

Metadata repository

Research datasets and files



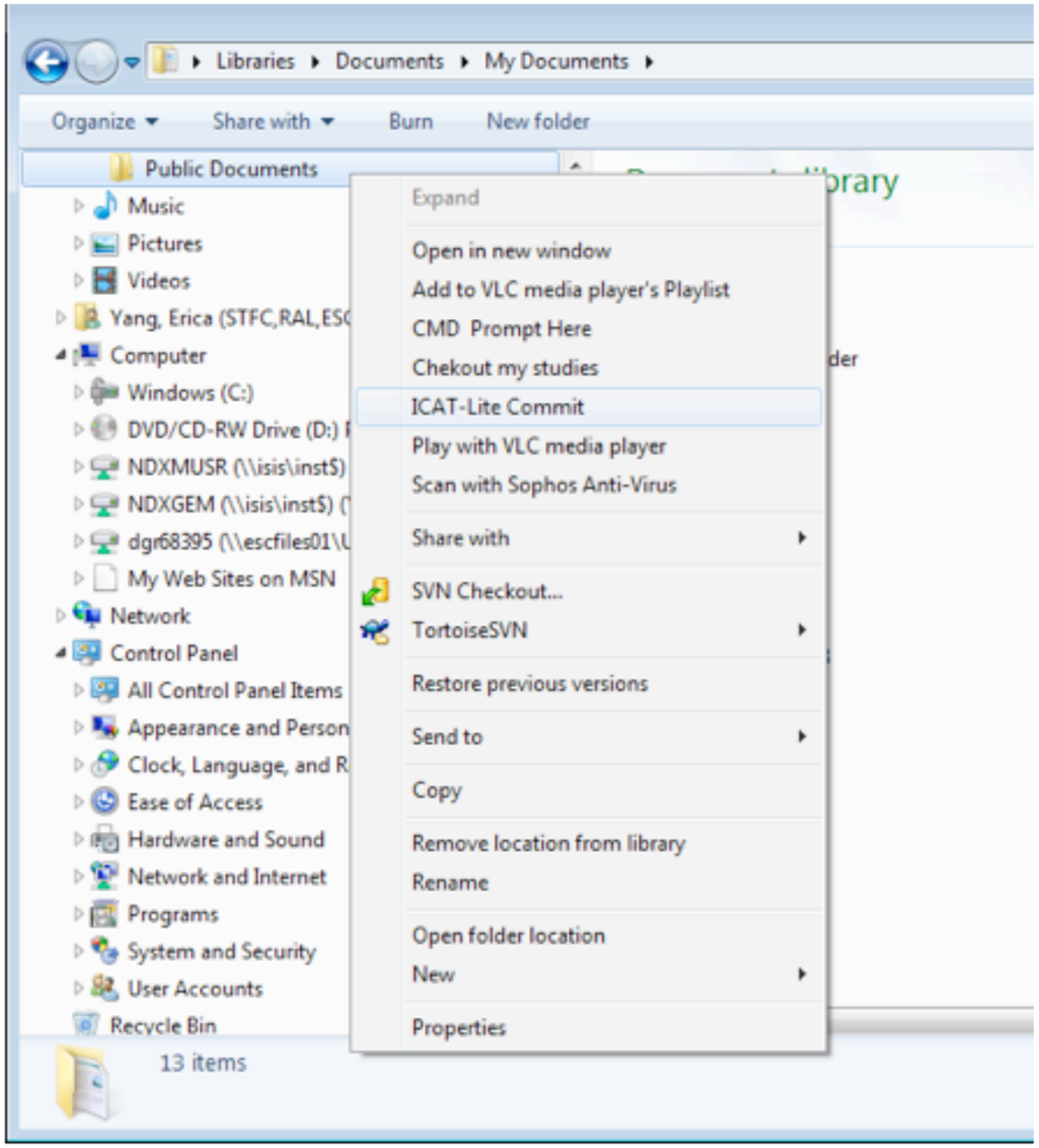
ICAT-lite: a pilot implementation of I2S2 information model

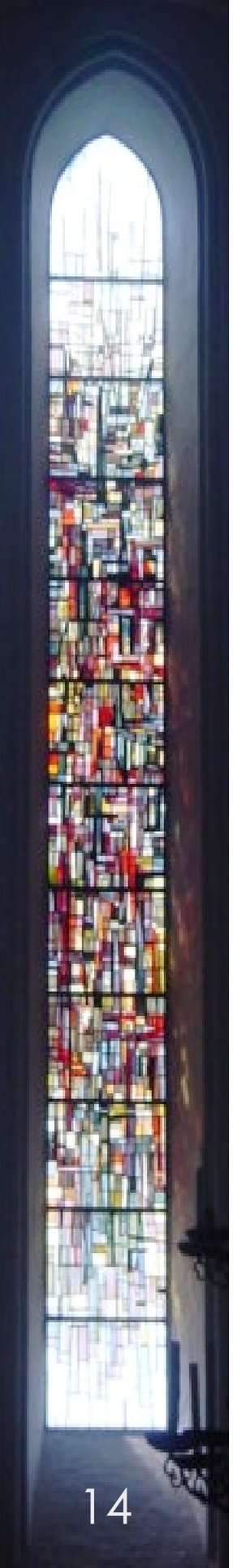
- A proof-of-concept that this model can support bench activities
- ICAT is an enterprise piece of software.
- ICAT-Lite: a cut down version suitable for laboratory usage for
  - Data organisation
    - Archive derived data
    - Annotate derived data
    - Browse archived data
    - Restored derived data
  - Data reuse
    - Secondary analysis
    - Cross analyses study
  - Data sharing
    - Publish archived data
    - Link data
    - View data provenance (graph)
    - Verify data
    - Automated experiment

https://sourceforge.net/apps/mediawiki/icatlite/index.php?title=Main_Page

Implementation of ICAT-lite

- A personal workbench for managing data flows
- Cut down from the facilities ICAT suite
- Allows the user to commit data …
- and capture its provenance.





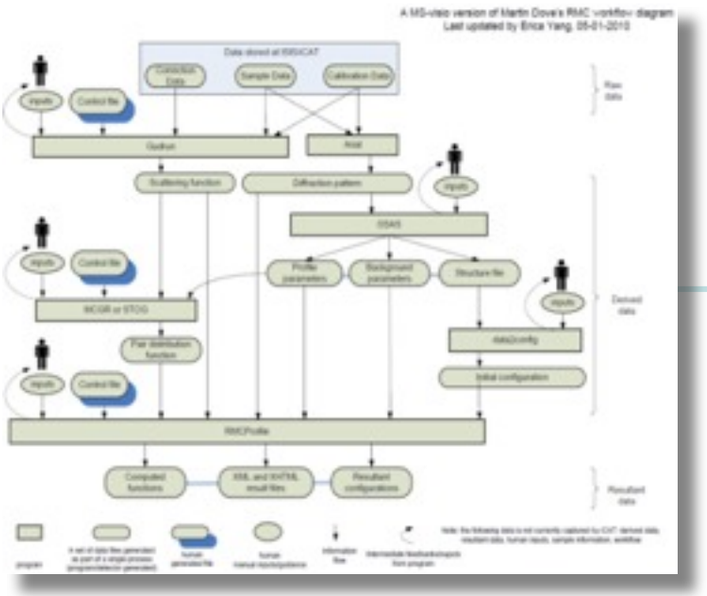
Sample XML pieces

```
<processes>
  <process id="gudrun_java" type="java program">
    <name>GudrunGUI_2.jar</name>
    <directory>GudrunGUI_2</directory>
  </process>
</processes>
<datafiles>
  <datafile id="df1">
    <name>Gudrun_dcs.txt</name>
    <directory>run.SANDALS.Water</directory>
  </datafile>
  <datafile id="df2">
    <name>purge_det.dat</name>
    <directory>run.SANDALS.Water</directory>
  </datafile>
</datafiles>
<datasets>
  <dataset id="d4">
    <datafileref idref="df4"/>
    <datafileref idref="df5"/>
  </dataset>
  <dataset id="d5">
    <datafileref idref="df19"/>
  </dataset>
</datasets>
```

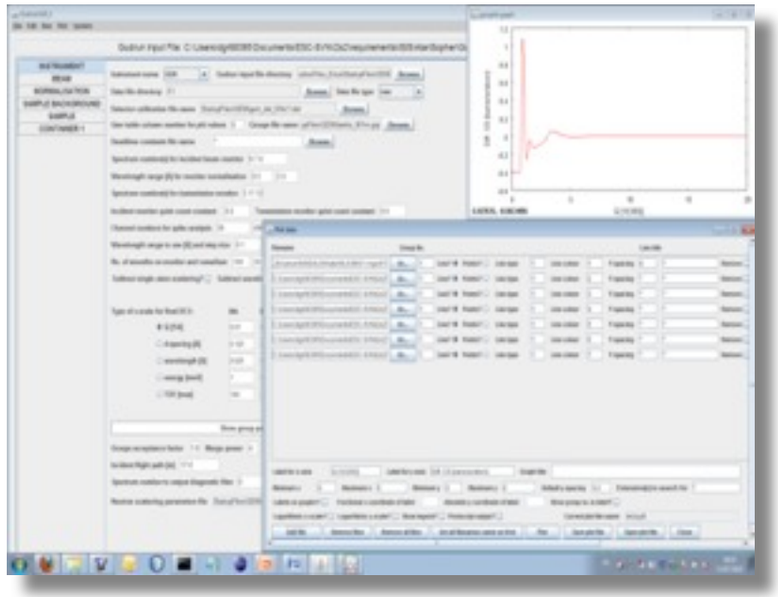
```
<investigations>
  <investigation id="i1" type="analysis">
    <processref idref="gudrun_java"/>
    <datasetref idref="d5" type="others"/>
    <datasetref idref="d2" type="output"/>
  </investigation>
  <investigation id="i2" type="analysis">
    <datasetref idref="d2" type="input"/>
    <processref idref="purge_det"/>
    <datasetref idref="d3" type="output"/>
  </investigation>
  <investigation id="i3" type="analysis">
    <datasetref idref="d3" type="input"/>
    <processref idref="gudrun_dcs"/>
    <datasetref idref="d4" type="output"/>
  </investigation>
</investigations>
<studies>
  <study id="s1">
    <investigationref idref="i1" />
    <investigationref idref="i2" />
    <investigationref idref="i3" />
  </study>
</studies>
```

Coping with derived data

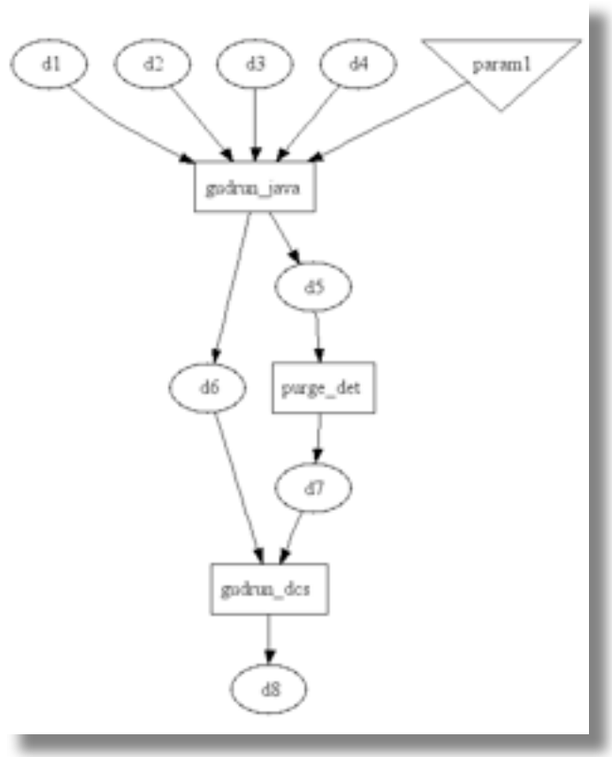
Data analysis workflow



Scientific software: Gudrun



Data analysis folders



Archive
Browse
Restore

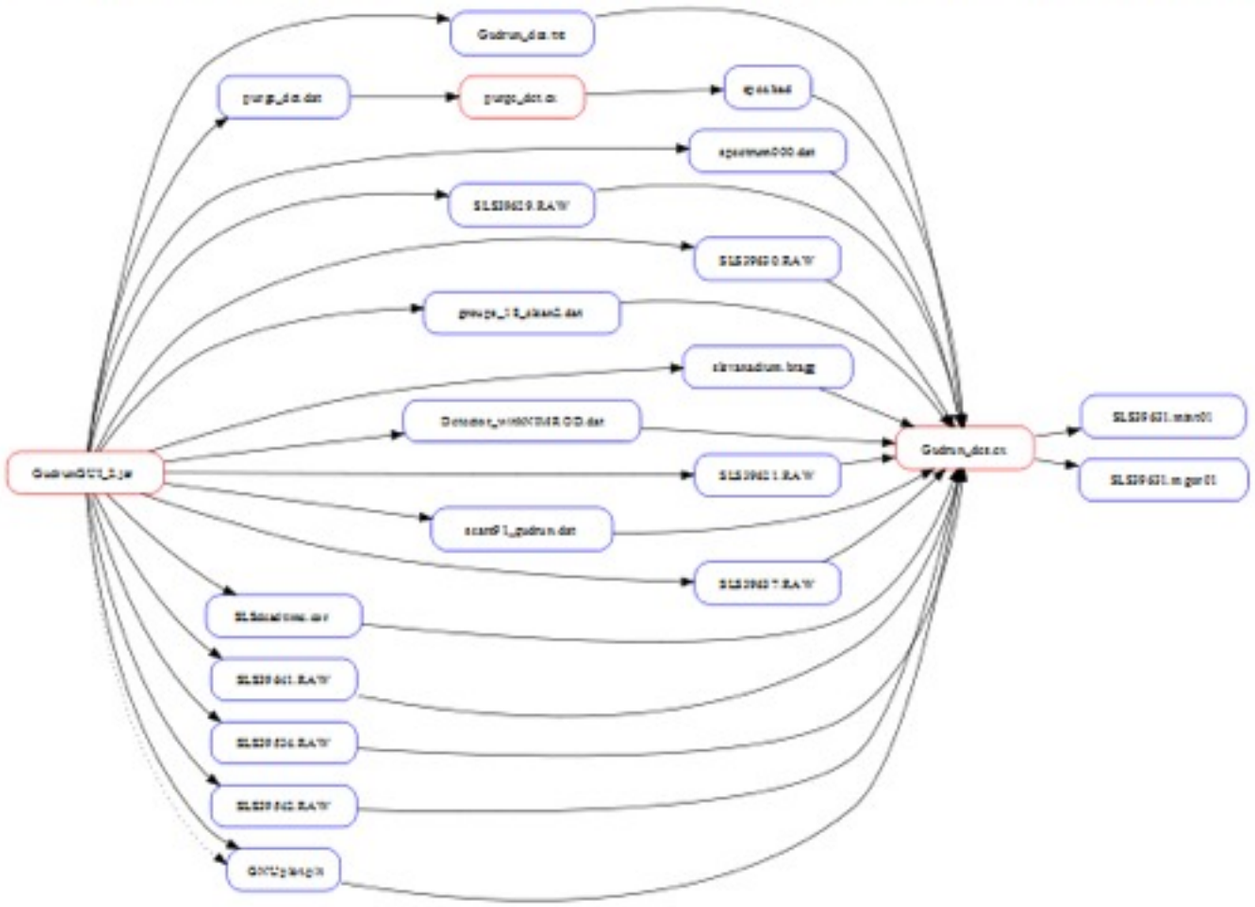
Derived Data

Screen shot

ICAT personal

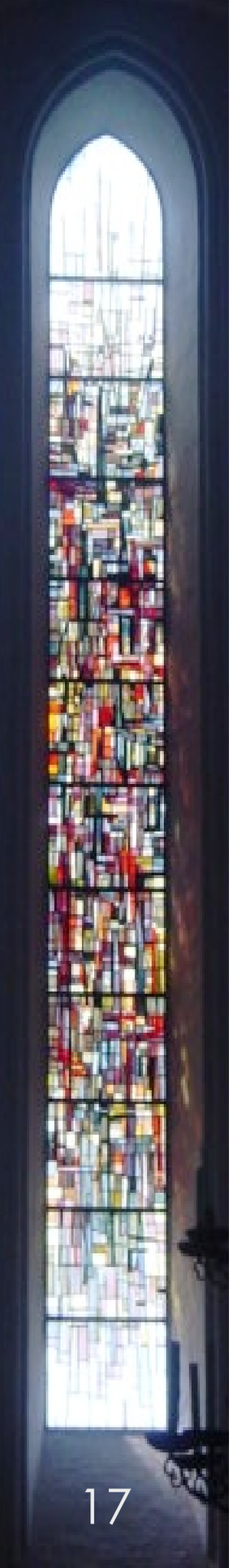
Accelerating Research and Discovery through Better Data Management

| Home | Background | Demo | Data Model | Software | Software Wiki | Project Wiki | I2S2 UKOLN | Project Files | RAL Repository |



Last updated: 2010-10-21 By Erica Yang

Sitemap | Contact Us | Copyright STFC eScience | Picture Credits (Gnuplot_ellipsoid.svg) |



Conclusions

- Straightforward to extend the Core Scientific Metadata model to cover all parts of the science process
- Proof of concept: still a work in progress

Acknowledgements
- Project colleagues in STFC (Erica Yang), Bath (Manjula Patel) & Southampton (Simon Coles)
- Neil Beagrie (consultant); see talk after break
- Funding from JISC