# Integrating Research Information: Requirements of Science Research

**Brian Matthews**

Scientific Information Group
E-Science Centre
STFC Rutherford Appleton Laboratory

brian.matthews@stfc.ac.uk

Science & Technology Facilities Council

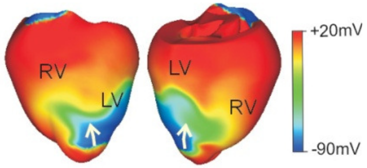
**The science we do**

**Some Integration Drivers**

**Metadata for integration**

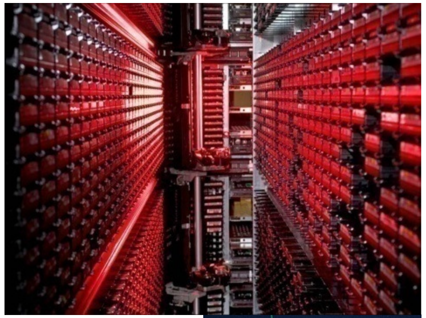
# The science we do

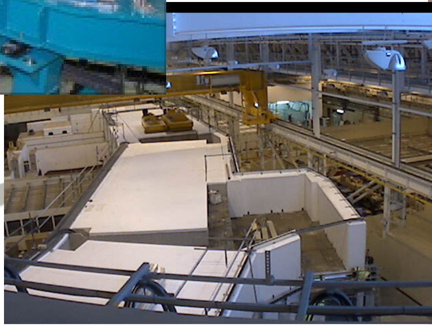
# Science and Technology Facilities Council



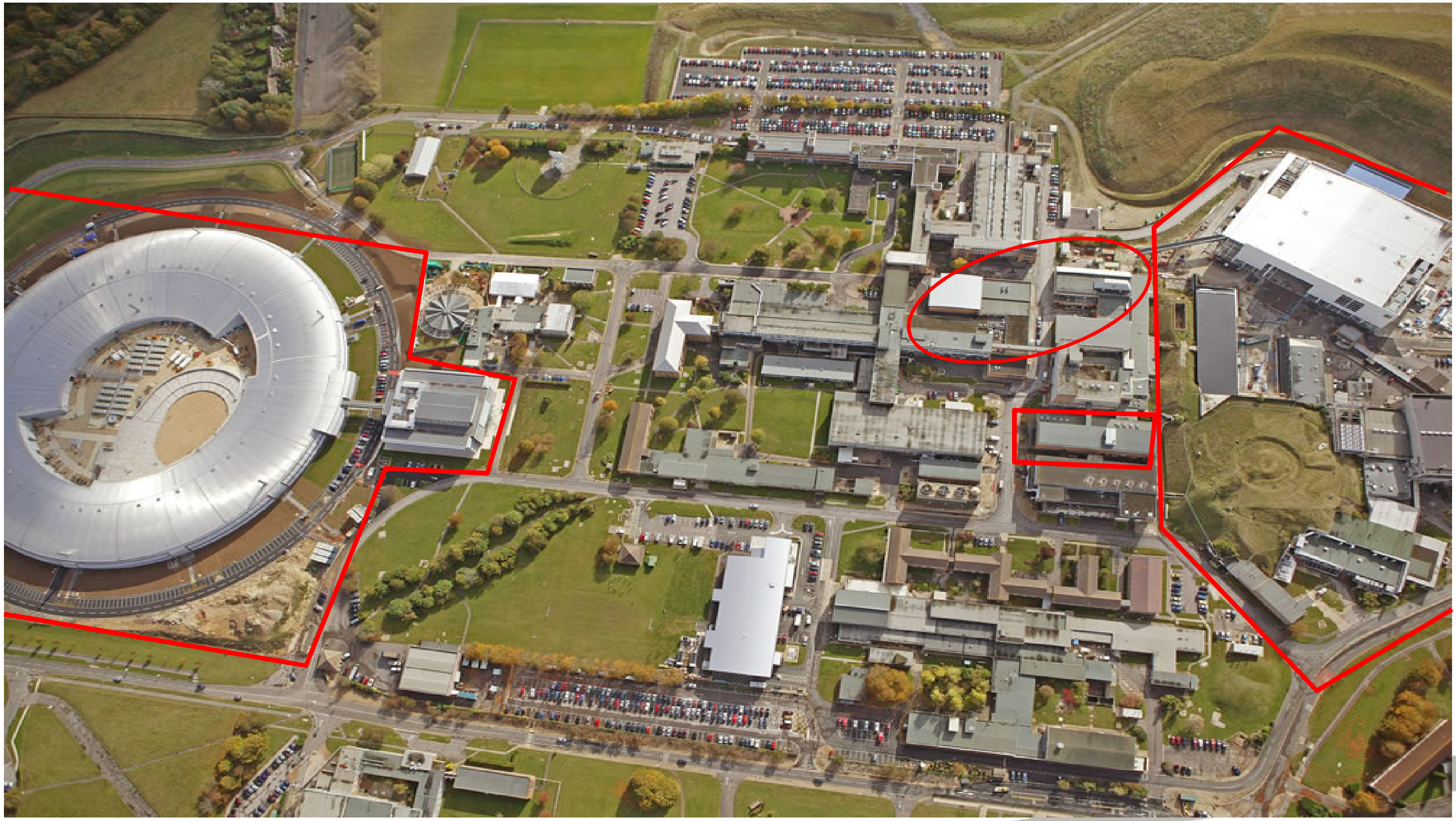
- Provide large-scale scientific facilities for UK Science
  - particularly in physics and astronomy
  - ISIS and Diamond Light Source facilities

- E-Science Centre
  - Provides advanced IT development and services to the STFC Science Programme
  - Strong role in management of our science data





# Large-Scale Facilities

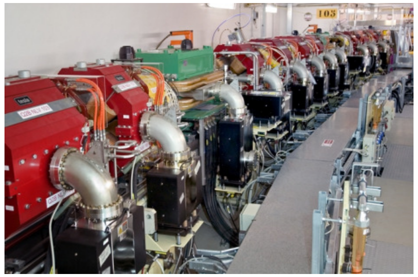


*Big Facilities for Small Science*

Science & Technology
Facilities Council

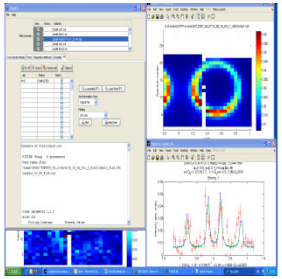
# The Science we do - Structure of materials

Visit facility on research campus

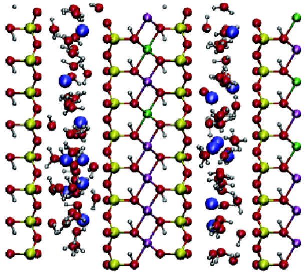


Place sample in beam

Diffraction pattern from sample



Fitting experimental data to model

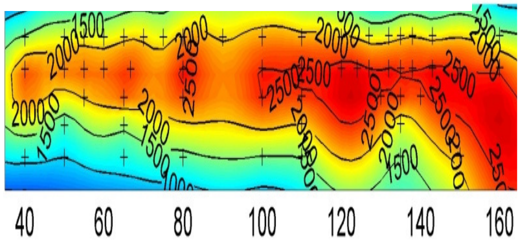


Structure of cholesterol in crude oil

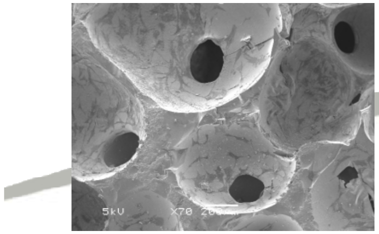
- ~30,000 user visitors each year in Europe:
  - physics, chemistry, biology, medicine,
  - energy, environmental, materials, culture
  - pharmaceuticals, petrochemicals, microelectronics

- 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

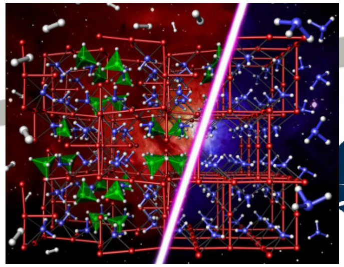
Longitudinal strain in aircraft wing



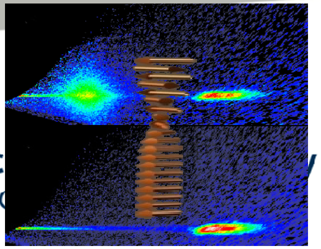
Bioactive glass for bone growth



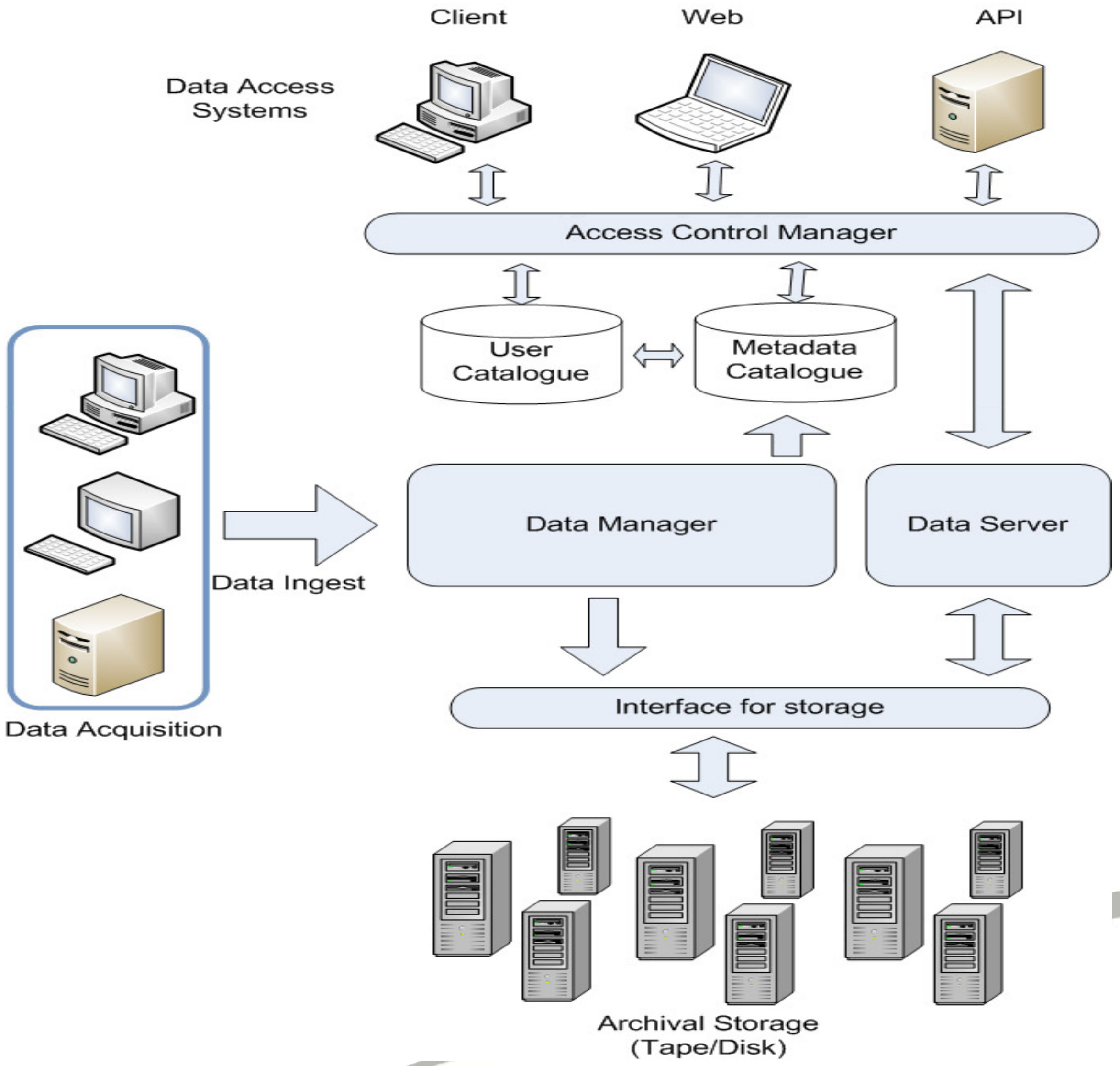
Hydrogen storage for zero emission vehicles



Magnetic moments in electronic storage



# A Data Management Architecture



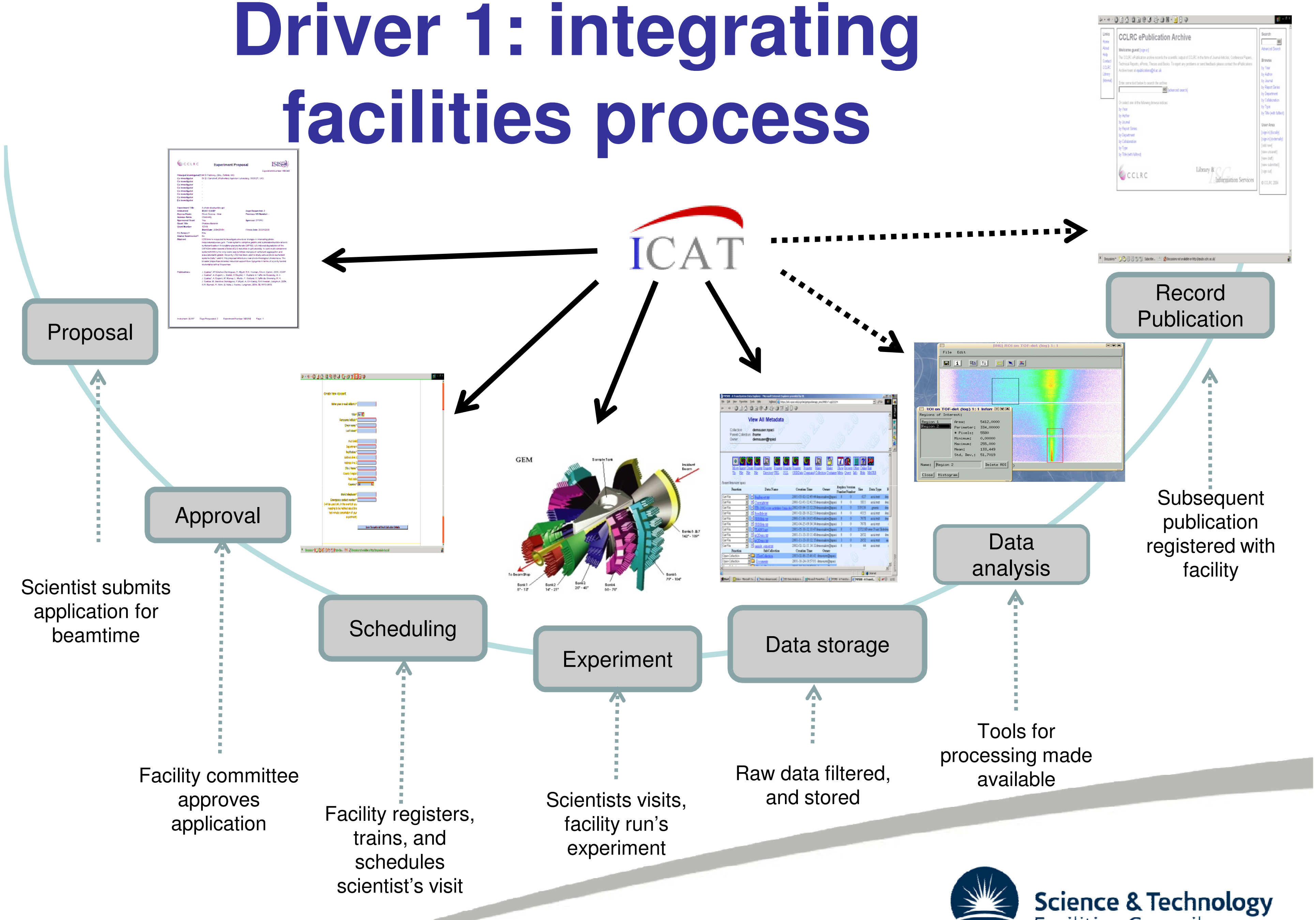
- Generic
  - Can be applied to different customers
- Robust
  - Can be monitored and maintained
- Fast
  - Manages large rates of data ingest
- Scalable
  - Manages the storage of very large amounts of data
- Secure
  - Allows role-based access control to be applied
- Integrity
  - Data Verification at ingest
  - Does not lose or mis-identify data over time
- Monitoring
  - Must generate reports

Science & Technology Facilities Council

# Integration Drivers

Science & Technology
Facilities Council

# Driver 1: integrating facilities process



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 storage

Raw data filtered, and stored

Data analysis

Tools for processing made available

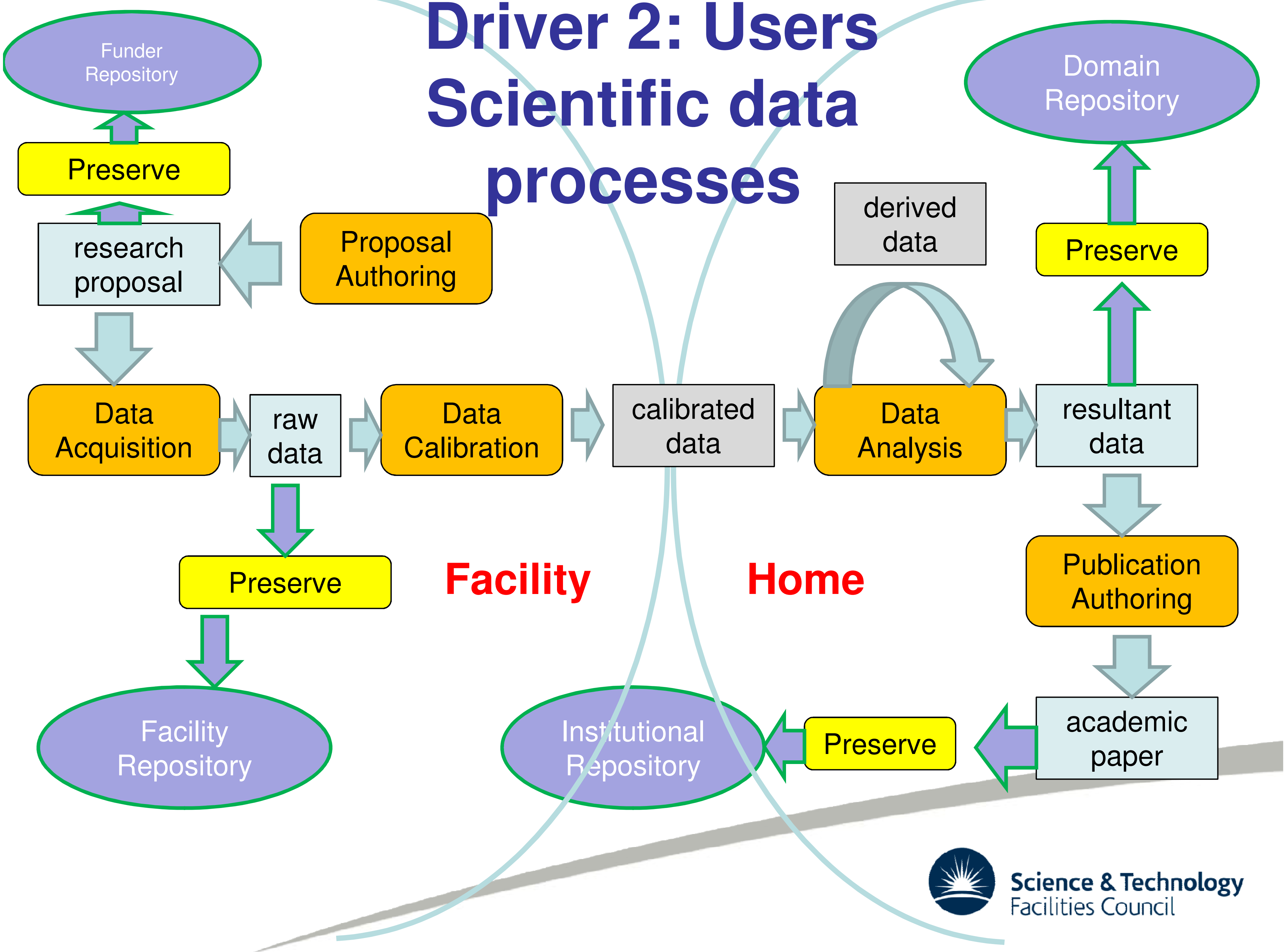
Record Publication

Subsequent publication registered with facility

*http://code.google.com/p/icatproject/*

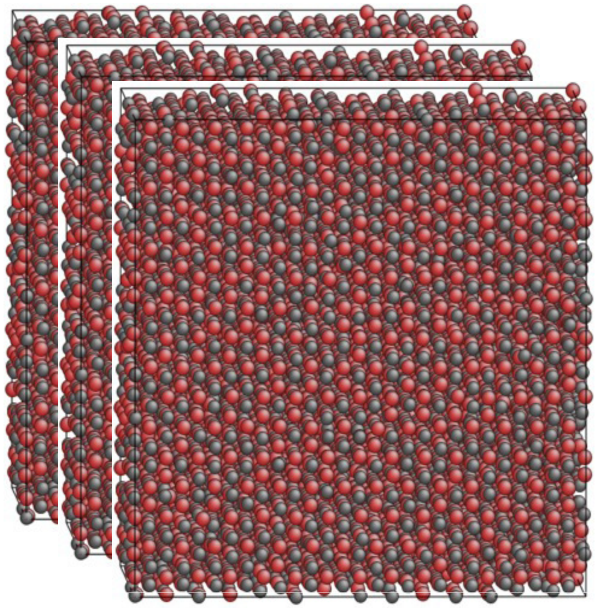
Science & Technology Facilities Council

# Driver 2: Users Scientific data processes



Science & Technology Facilities Council

# Case Study: Earth Sciences, Cambridge



- Seeking construct large scale atomic models of matter that best match experimental data
  - Reverse Monte-Carlo Simulation techniques

- Experiment and data collection conducted at ISIS (SGEM)

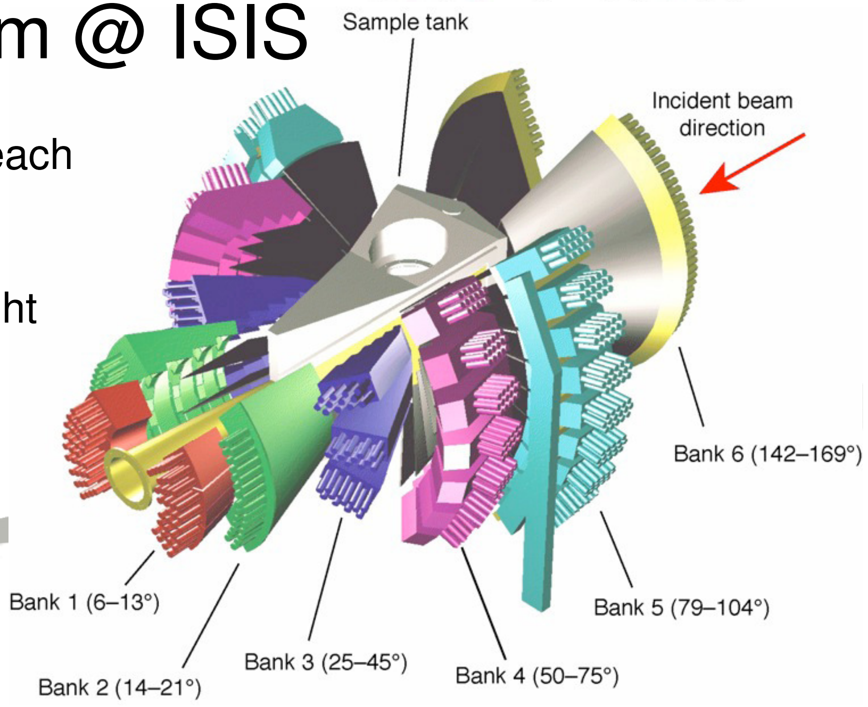
  - ~4000 detectors

  - Each experiment  produces a histogram for each detector

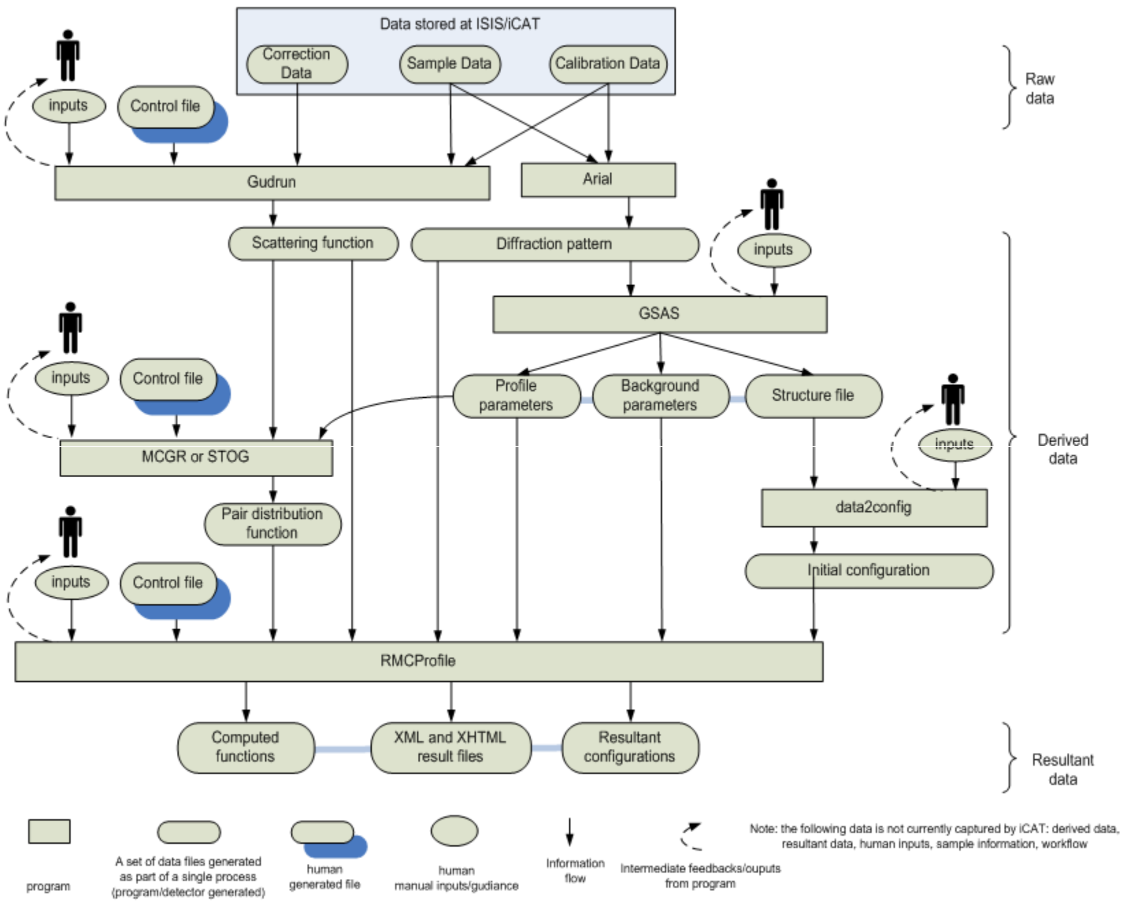
  - Each histogram is a binning of all neutron flight times per pulse, summing all pulses

- The data reduction process has to convert these histograms into meaningful data

Gem @ ISIS



# Earth Sciences: typical workflow



Note: the following data is not currently captured by iCAT: derived data, resultant data, human inputs, sample information, workflow

- Processing dependent on specialised software
  - Sustainability issues
- Context not routinely captured
- Main analysis is reliant on scientist’s knowledge and experience
  - selecting parameters and interpreting data
  - recorded in a lab note book
- Actual workflow not recorded
- Distributed Data - Little shared infrastructure
  - Raw and reduced data stored at ISIS
  - Other data on his/her laptop or WebDAV

Martin Dove & Erica Yang

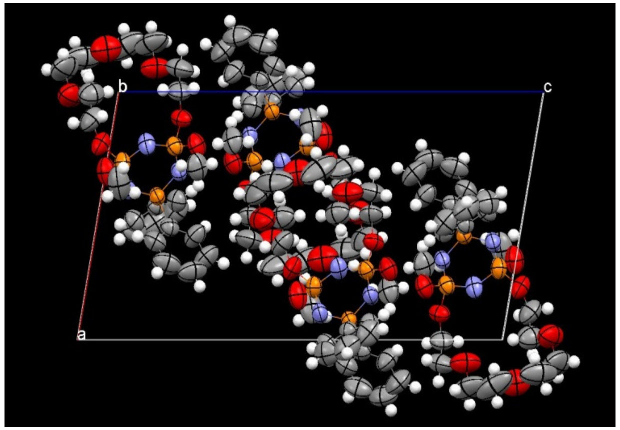
# Driver 3: Publishers

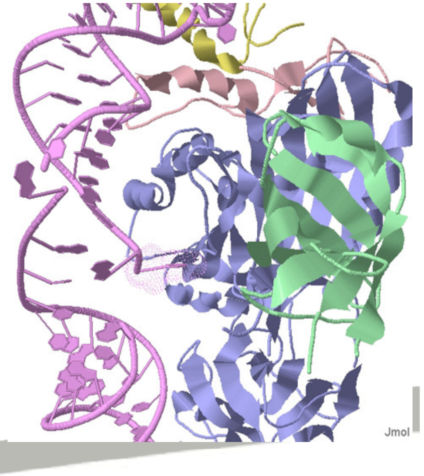
IUCr journal policy - “data” either

- must be supplied in CIF format as an integral part of article submission and are freely available for download or

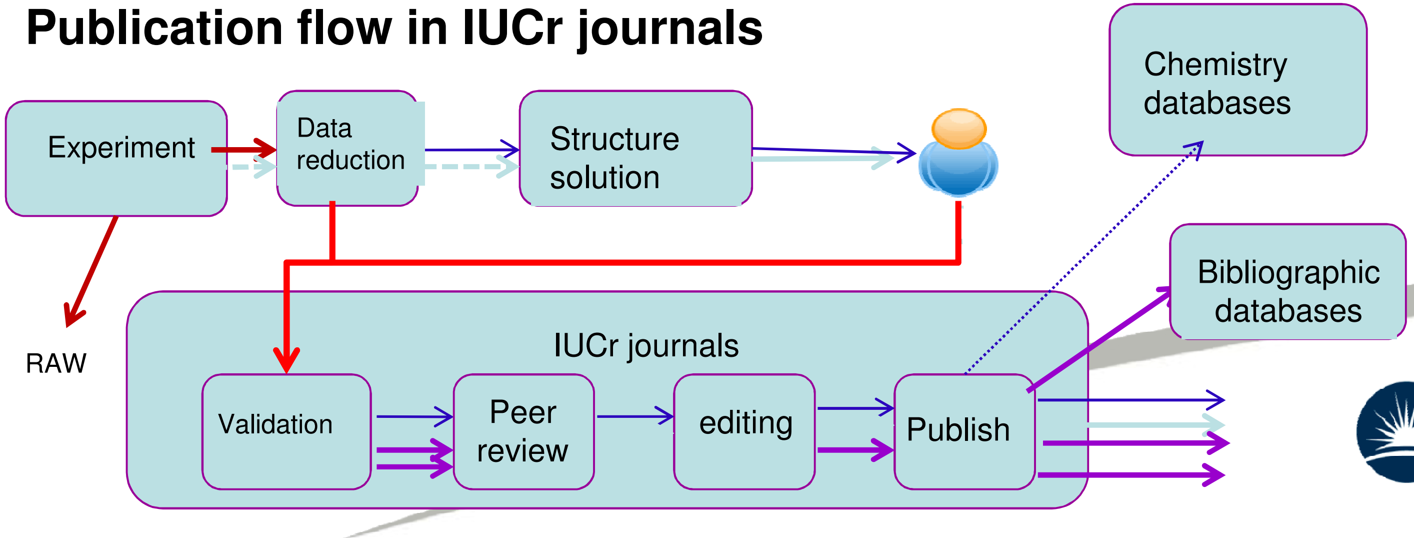
- must be deposited with the Protein Data Bank before or in concert with article publication; the article will link to the PDB deposition using the PDB reference code

*Thanks to Brian MacMahon, IUCr*





**Publication flow in IUCr journals**


Experiment

Data reduction

Structure solution

RAW

Chemistry databases

Bibliographic databases

IUCr journals

Validation

Peer review

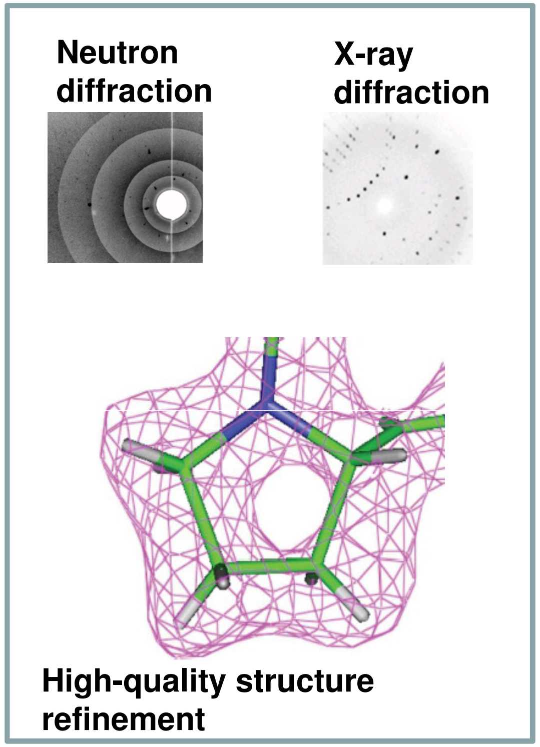
editing

Publish

# Driver 4: Interoperability across Facilities

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

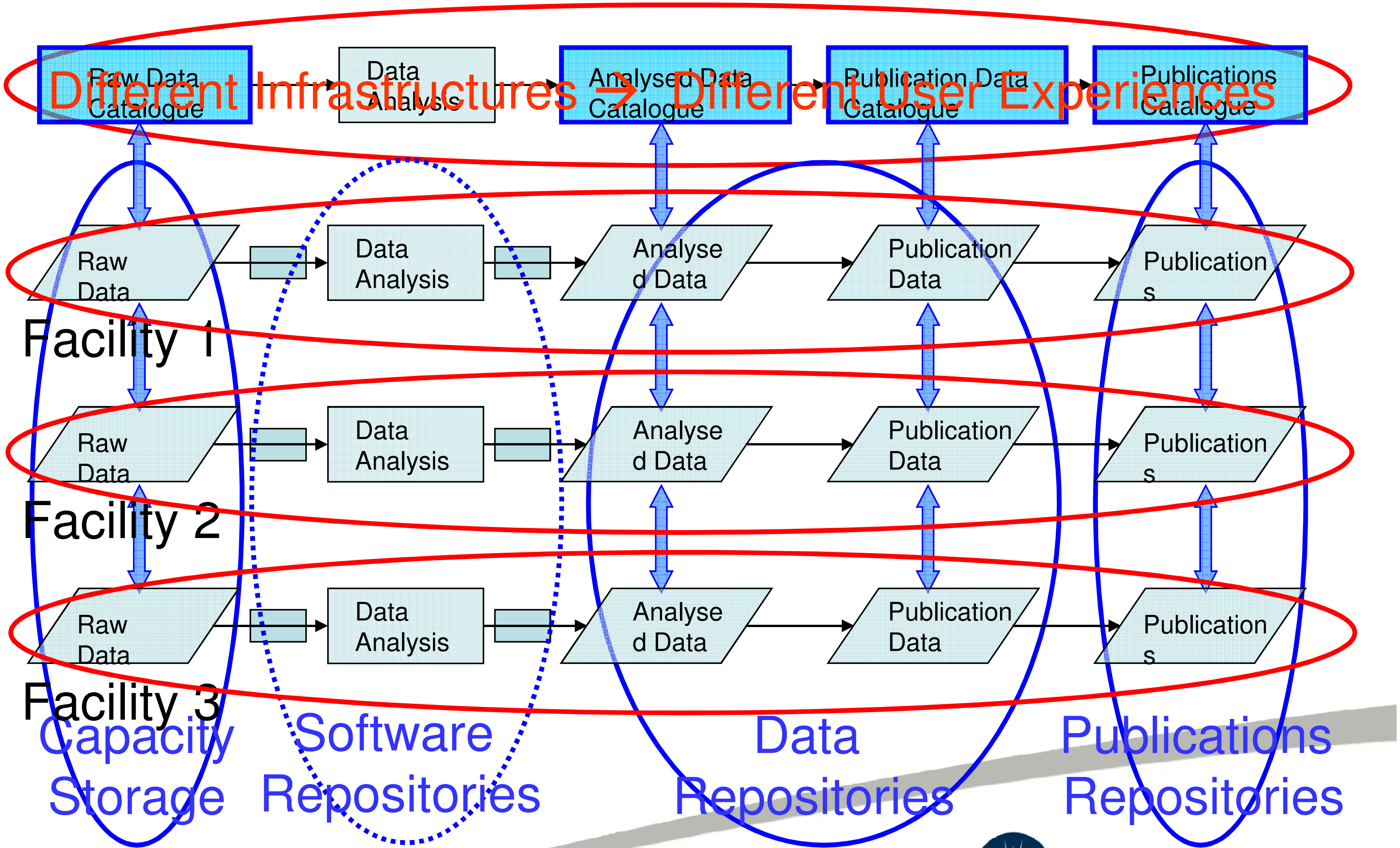
**... to construct and operate a shared data infrastructure for Photon and Neutron laboratories...**



- Common data catalogue
- Integration of users data from different facilities
- Track provenance of data through analysis stages
- Deploy standards for long-term curation
- Support scalability through parallelisation
- Deploy infrastructure in three different techniques

# PaN-Data Vision

Single Infrastructure → Single User Experience



Science & Technology Facilities Council

# Why capture the lifecycle?

From our Drivers (and others) :

- Maintain consistency
  - Don’t need to type stuff in more than once
- Easy for the scientists
  - Infrastructure in university labs is “ad hoc”
  - they lose stuff!
- Provide the evidential basis for research
  - Validate and verify publications
  - Safeguard against error or fraud
- Measure the impact of science
  - E.g. Measure value to service providers, funders and researchers
  - Influence the policy makers
- Reuse of data
  - Get new science from old data
  - Non-repeatable results
  - Value for money
  - Teaching material
  - Comparative studies

# Metadata for Integration

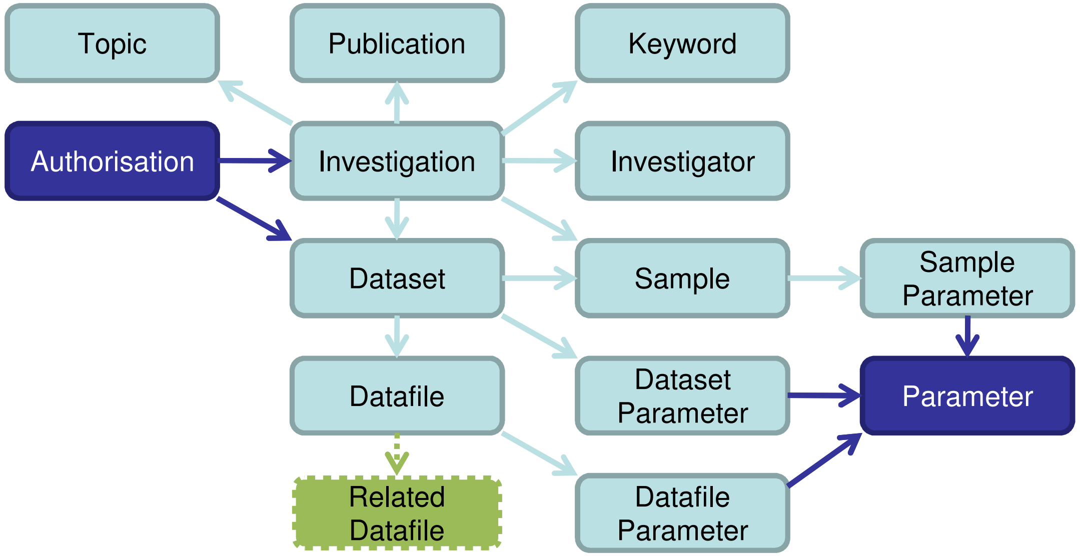
# CSMD

- CSDM: Core Scientific MetaData model

- Designed to describe facilities based experiments in Structural Science

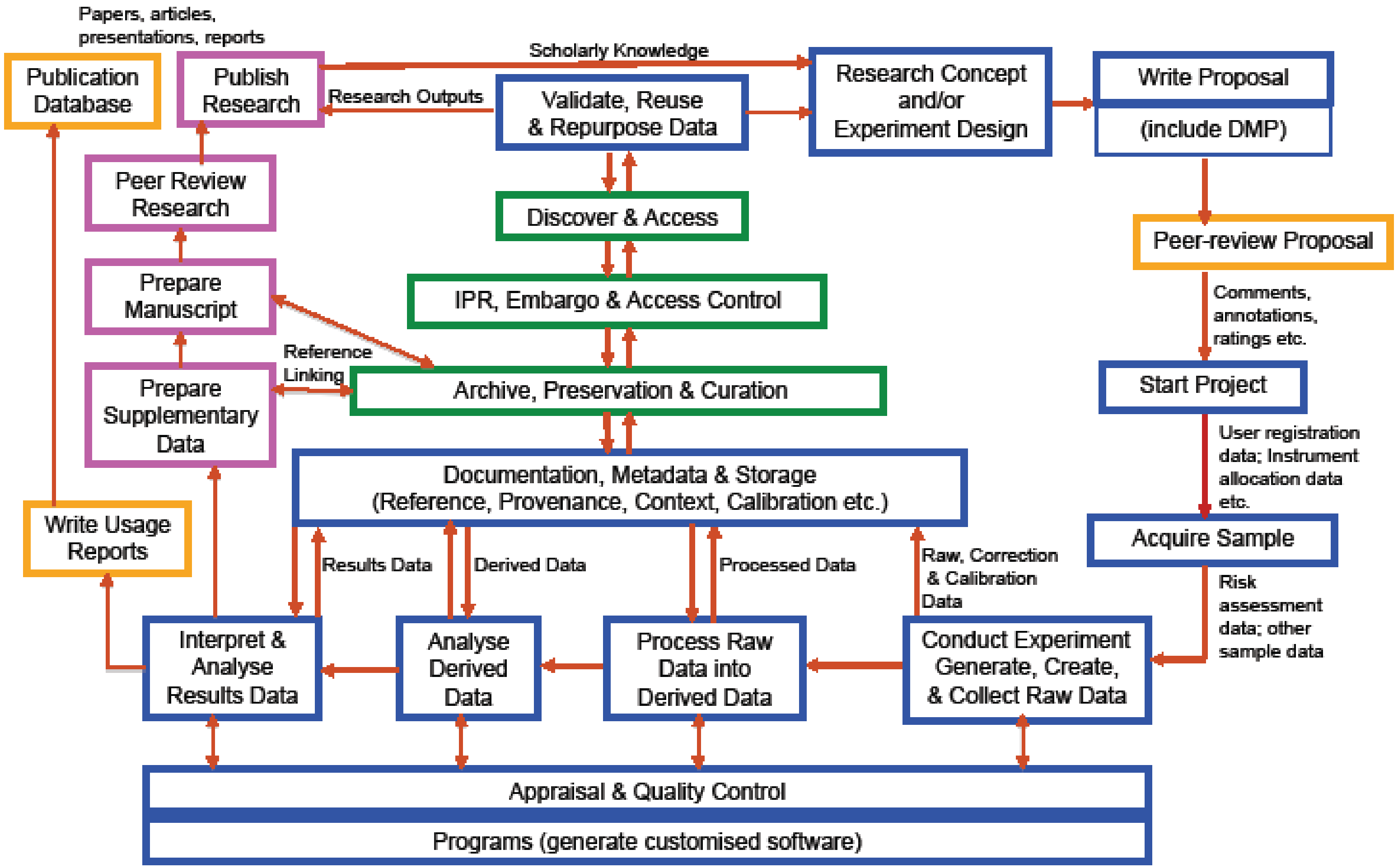
- Forms the information model for ICAT, a production data management infrastructure employed by STFC

- Forms the basis for extensions:
  - To derived data
  - To laboratory based science
  - To secondary analysis data
  - To preservation information
  - To publication data



Science & Technology Facilities Council

# An Idealised Scientific Research Data Lifecycle Model



Papers, articles, presentations, reports

Publication Database

Publish Research

Scholarly Knowledge

Research Outputs

Validate, Reuse & Repurpose Data

Research Concept and/or Experiment Design

Write Proposal

(include DMP)

Peer Review Research

Discover & Access

Peer-review Proposal

Prepare Manuscript

IPR, Embargo & Access Control

Comments, annotations, ratings etc.

Reference Linking

Prepare Supplementary Data

Archive, Preservation & Curation

Start Project

User registration data; Instrument allocation data etc.

Documentation, Metadata & Storage (Reference, Provenance, Context, Calibration etc.)

Write Usage Reports

Results Data

Derived Data

Processed Data

Raw, Correction & Calibration Data

Acquire Sample

Risk assessment data; other sample data

Interpret & Analyse Results Data

Analyse Derived Data

Process Raw Data into Derived Data

Conduct Experiment Generate, Create, & Collect Raw Data

Appraisal & Quality Control

Programs (generate customised software)

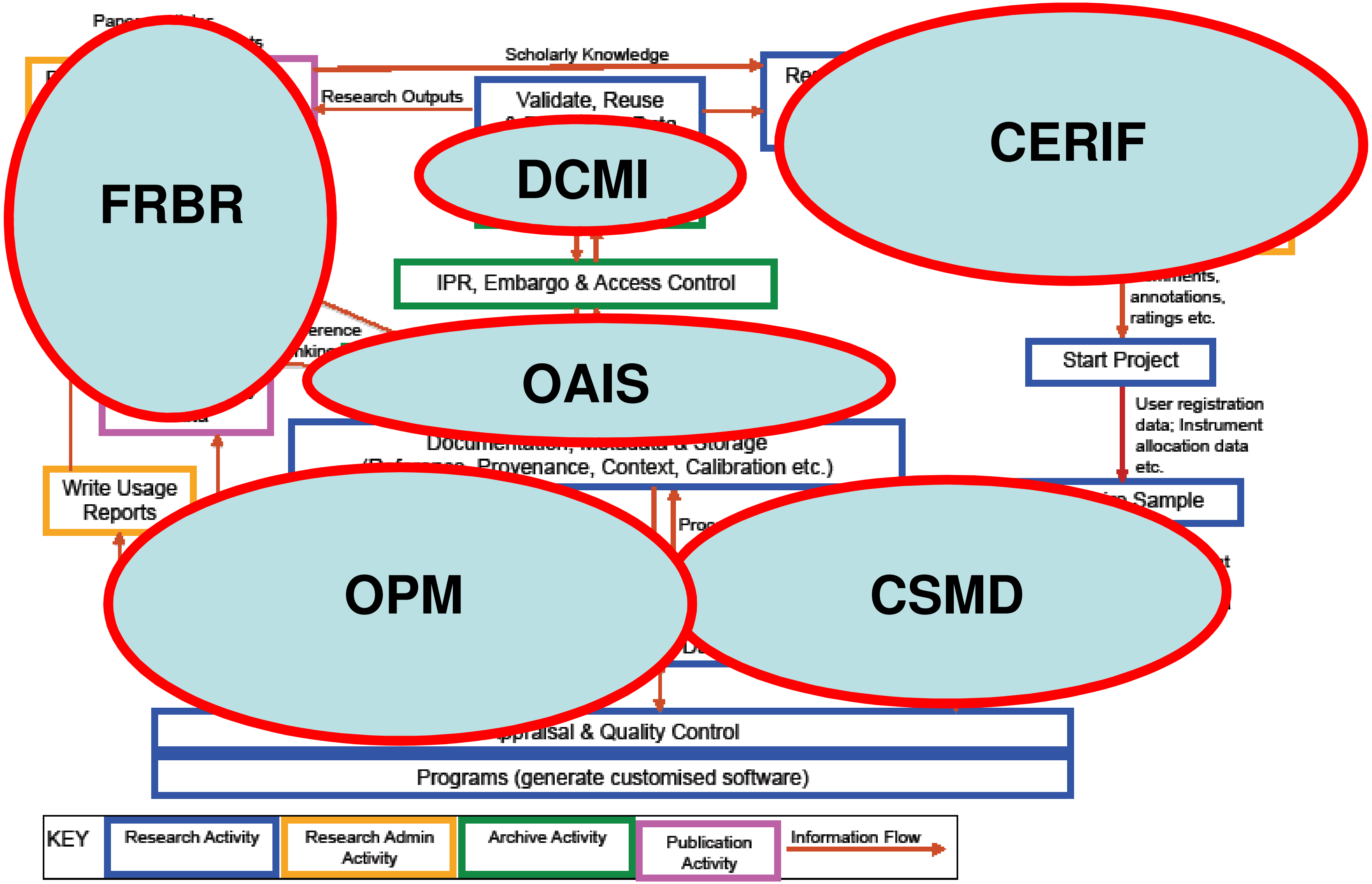
| KEY | Research Activity | Research Admin Activity | Archive Activity | Publication Activity | Information Flow → |
|---|---|---|---|---|---|

I<sub></sub>

Infrastructure for Integration in Structural Sciences

Science & Technology Facilities Council

# An Idealised Scientific Research Data Lifecycle Model



FRBR

DCMI

CERIF

OAIS

OPM

CSMD

Scholarly Knowledge

Research Outputs

Validate, Reuse

IPR, Embargo & Access Control

annotations, ratings etc.

Start Project

User registration data; Instrument allocation data etc.

Provenance, Context, Calibration etc.)

Write Usage Reports

Sample

Appraisal & Quality Control

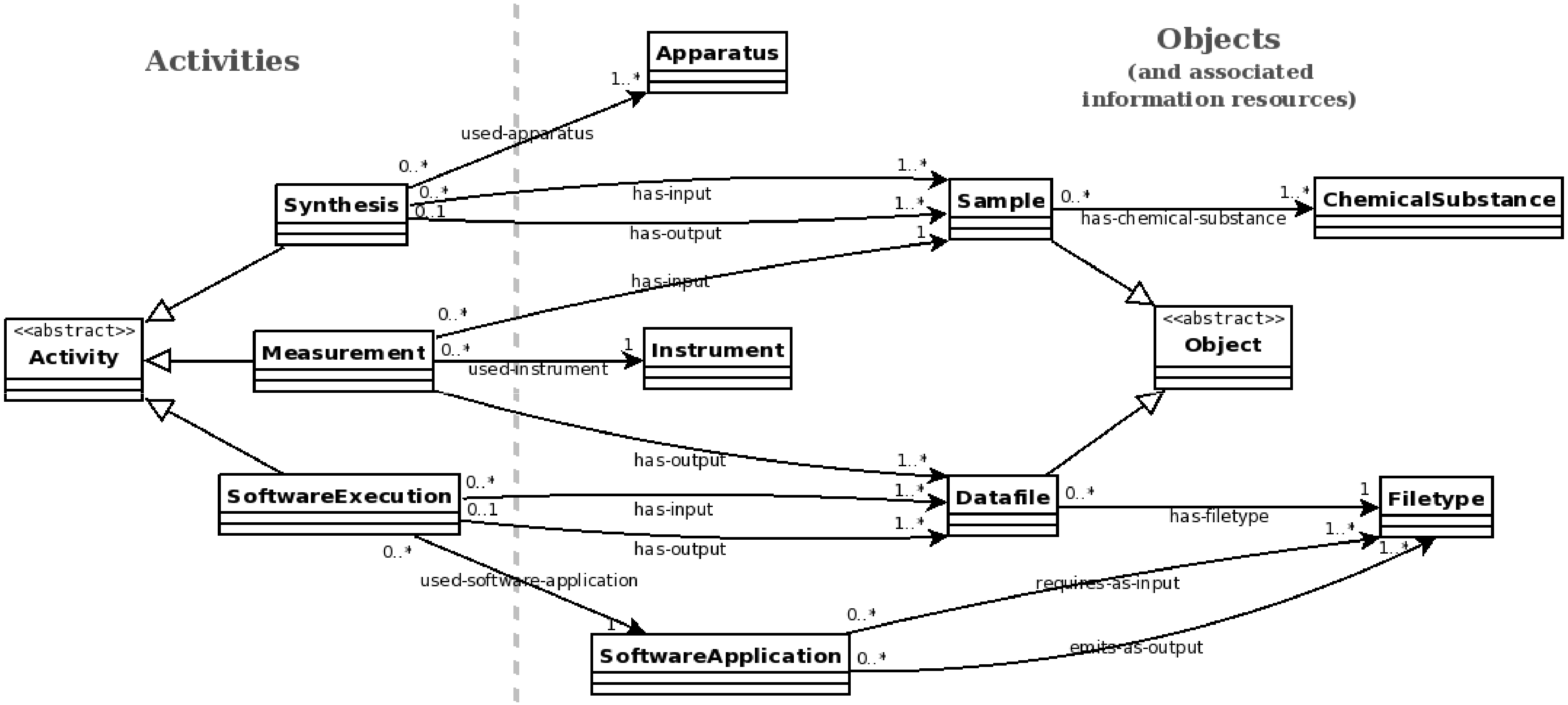
Programs (generate customised software)

| KEY | Research Activity | Research Admin Activity | Archive Activity | Publication Activity | Information Flow |
|---|---|---|---|---|---|

Infrastructure for Integration in Structural Sciences

Science & Technology Facilities Council

# Research Activity Model



A notion of a research activity – a step in the lifecycle model

- Can define different types of activity.

# Interoperability via metadata

The world is Heterogeneous

  - different software, formats, metadata

- Metadata standards for integration
  - In formats (e.g. RDF) and APIs (e.g. Web services)
  - Though not universal – in domains at best

- Need ways of "joining them up"
  - Core standards e.g. DC, CERIF, FRBR
  - Base concepts e.g. "People"
  - Key concept relationships e.g. Owl:EquivalentClass
  - Abstract models to chain metadata together e.g. OPM, RAM
  - Metadata extension e.g. Clarin

# Thank You

# Questions?

*brian.matthews@stfc.ac.uk*

www.e-science.stfc.ac.uk

Science & Technology
Facilities Council