Meeting a scientific facility provider's duty to maximise the value of data

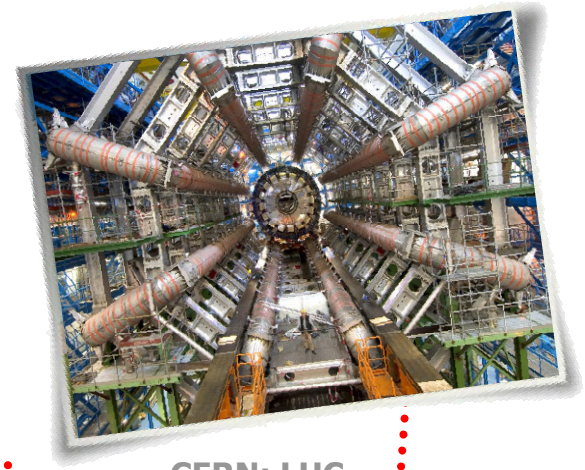
Michael Wilson

Science and Technology Facilities Council, UK

Scientific Computing Department

Science & Technology Facilities Council

ILL and ESRF



CERN: LHC

Big Science

Particle Physics
Earth Observation
Astronomy

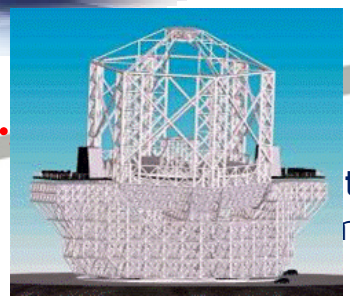
Rutherford Appleton Laboratory

ESA: Top Sat

Small Science

ISIS Neutron Source
Diamond X-ray Source
UK Central Laser Facility

SKA



ESO: E-ELT

ting
nent

Science & Technology Facilities Council

Big Data, Big Computers

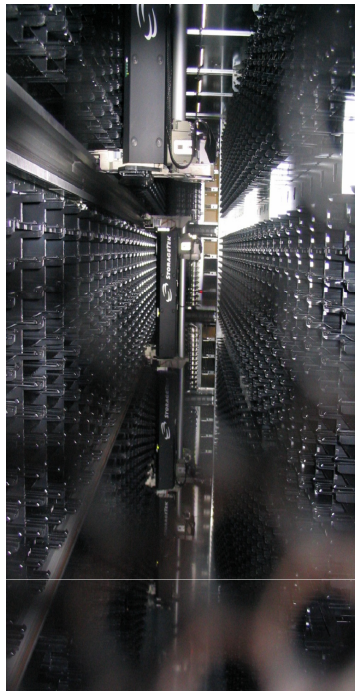
Facility Data Archives

All **ISIS** data (~25 years) > 3,000,000 files

All **Diamond** Data (~5 years) > 100,000,000 files

CERN LHC Tier 1 Data

UK hub for **LHC** data (~3 years: 11PB)



The StorageTek tape robot 100PB Capacity

Computing Architectures:

1 - the UK's most powerful computer (IBM BlueGene/Q :1.4 petaflops)

2 - the UK's most powerful graphics processor computer (190,000 graphics cores : 248 teraflops)

3- large commodity computing server (7000 processor cores)

4 - high throughput super-data-cluster (4.6 petabytes of parallel file storage with 1 terabit per second aggregate bandwidth from the data to the processors)

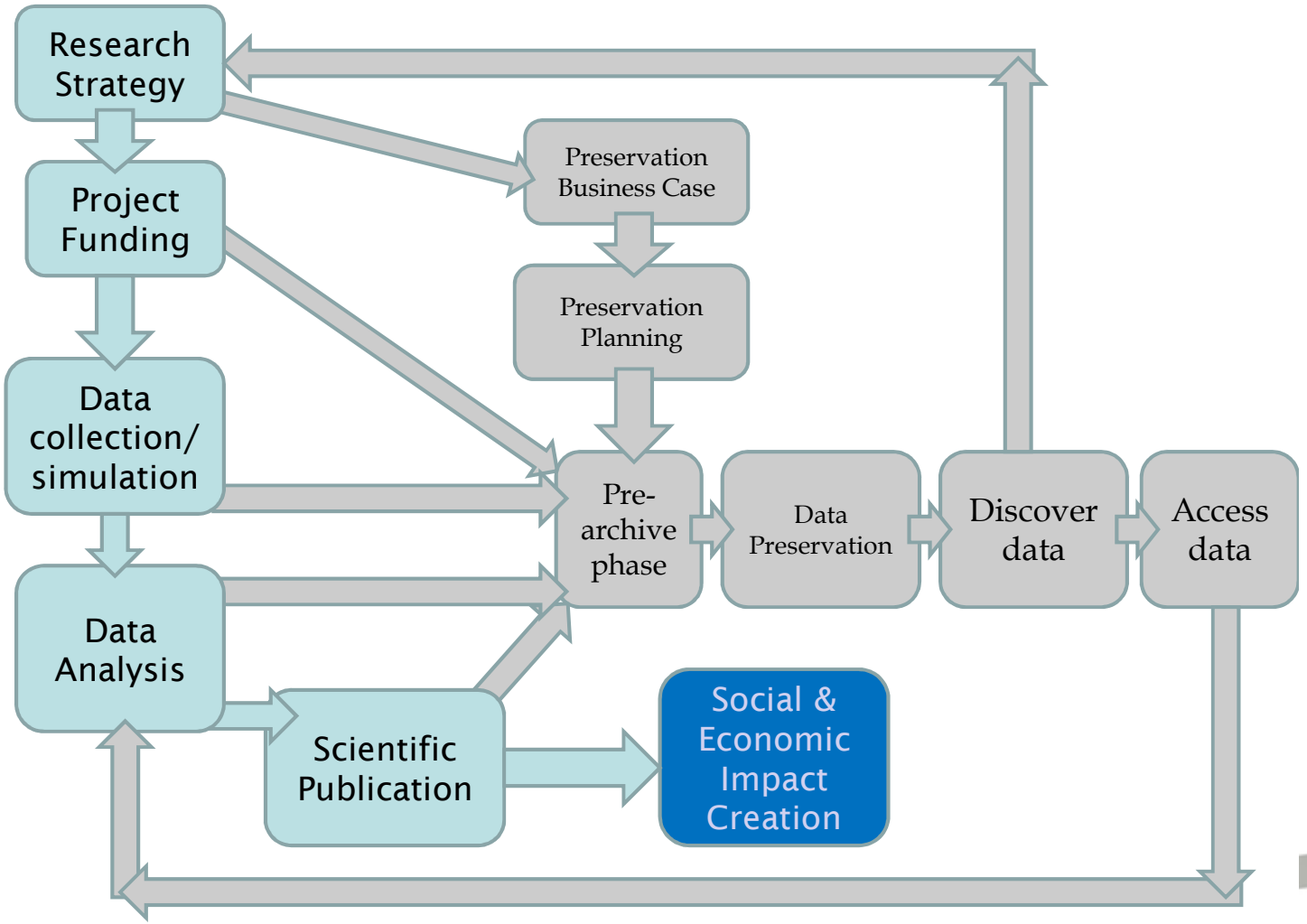
Scientific Computing Department

Science & Technology Facilities Council

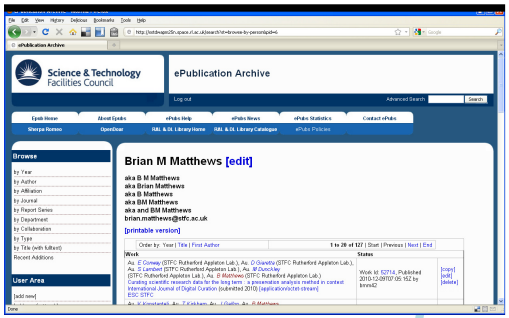
Maximise the value of STFC data

1) Researchers access their own data
2) Other researchers validate published results
3) Meta-studies incorporating data
4) Set experimental parameters and test new computational models
5) Used for new science not yet considered
6) Defend patents on innovations derived from science
7) Evidence based policy making

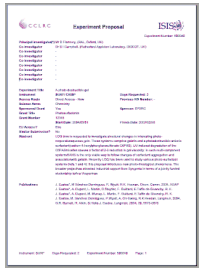
Data Re-use


Research Strategy
Project Funding
Data collection/ simulation
Data Analysis
Scientific Publication
Preservation Business Case
Preservation Planning
Pre-archive phase
Data Preservation
Discover data
Access data
Social & Economic Impact Creation

Facilities Lifecycle



ICAT
Metadata Repository



Proposal

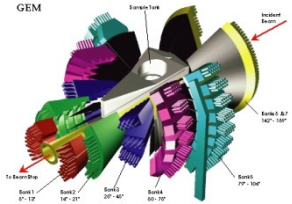
Scientist submits application for beamtime

Approval

Facility committee approves application

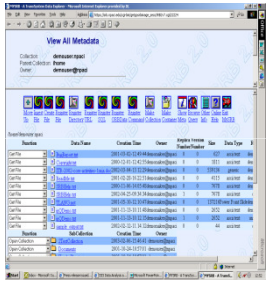
Scheduling

Facility registers, trains, and schedules scientist's visit



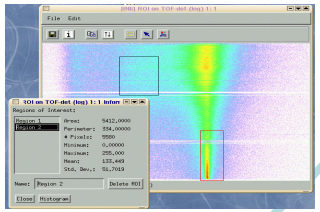
Experiment

Scientists visits, facility run's experiment



Data cleansing

Raw data filtered and cleansed



Data analysis

Tools for processing made available

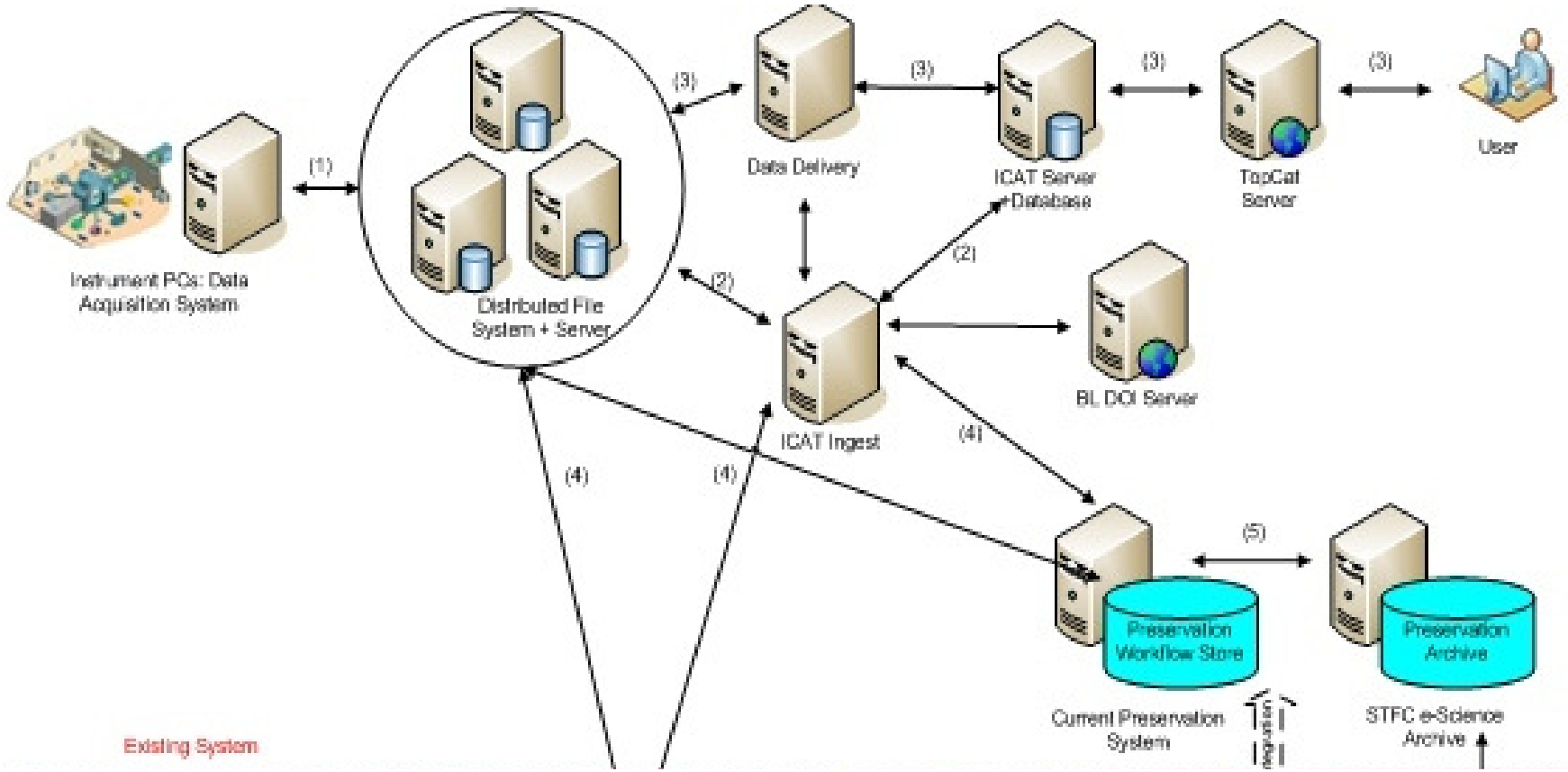
Record Publication

Subsequent publication registered with facility

Scientific Computing Department

Science & Technology Facilities Council

Data Preservation Infrastructure

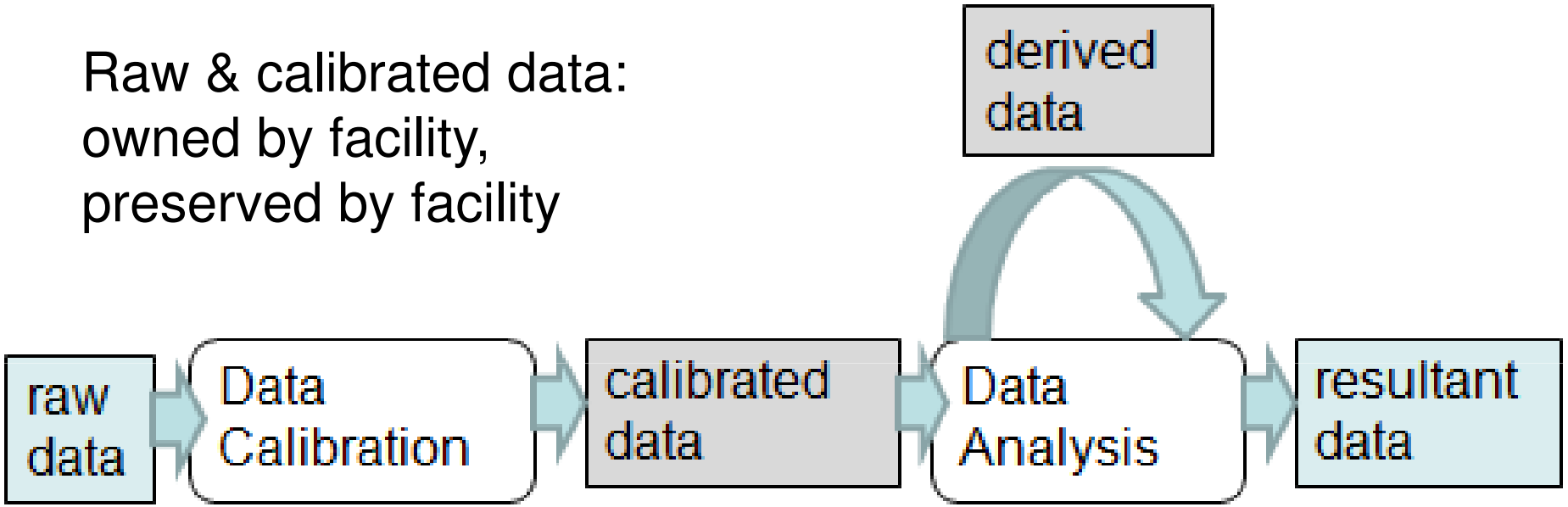


Automated Metadata Collection

- Schedule & Proposal: who, funder, what
  - Except 5% don't do what they proposed
- Instrument: data, instrument settings
- Publication: analysis method, result
- DOI is address for linking

Preserve which data?

Raw & calibrated data:
owned by facility,
preserved by facility



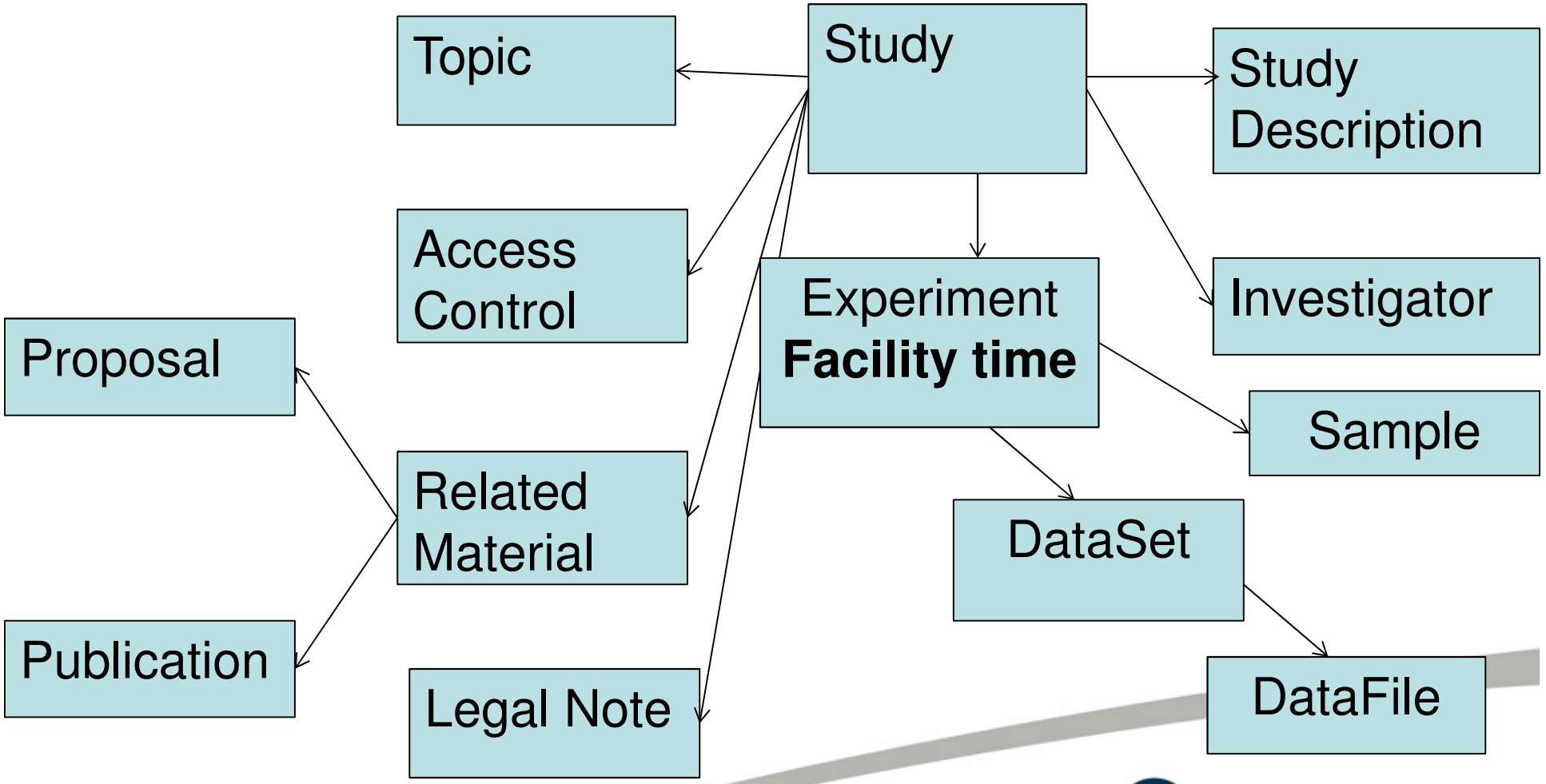
- Derived Data Provenance
- Derived Data Ownership – user, funder
- Analysis software preservation
- DOI for software versions

Cite which data ?

- Facility centric view
  - Investigation is allocated beam time
  - Investigation lasts 5 minutes to 2 weeks
  - Investigation is allocated DOI
  - Investigation is cited
- Metadata levels/model – what level DOI
  - Investigation/experiment
  - Data set(s)
  - Individual file(s)

Core Scientific Metadata Model (CSDM)

- http://code.google.com/p/icatproject/wiki/CSMD



When to publish data?

- Commercial data
  - < 1% of data
  - subject to individual contracts
  - Don't publish
- Data Policies – different science, different facilities, different policy
- Data Embargo – PhD period 3 years
- Record who accesses data
- Metadata Embargo – 2004 Haumea example

Avoid data mis-use

- Sensitive data
  - Satellite images of New York 9/11/2001
- Making large data sets usable
  - UN IPCC Report 5 data
  - CERN LHC experiment data
- Publish samples for teaching
- Avoid time wasting by conspiracy theorists
- Science should be verifiable

ROI in data centres ?

- Cost of data curation
  - ESA mission based approach
- Recent reviews of 2 UK data centres by Neil Beagrie and John Houghton
  - Interview based approach
- ENSURE EU project
  - http://ensure-fp7-plone.fe.up.pt/site/
  - Cloud based costs
  - Value of data for preservation objective

Questions ?

Scientific Computing Department

Science & Technology Facilities Council