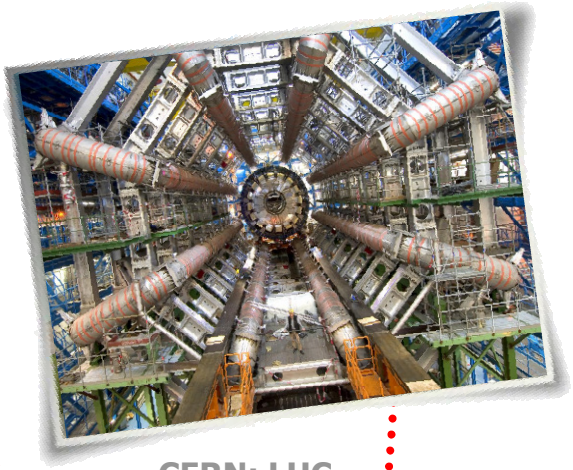
Moving from a scientific data collection system to an open data repository.

Tom Griffin, Brian Matthews, Alistair Mills, Sri Nagella, Arif Shaon, **Michael Wilson**, Erica Yang

Science and Technology Facilities Council, UK

Scientific Computing Department

Science & Technology Facilities Council



Big Science

Particle Physics
Earth Observation
Astronomy

Small Science

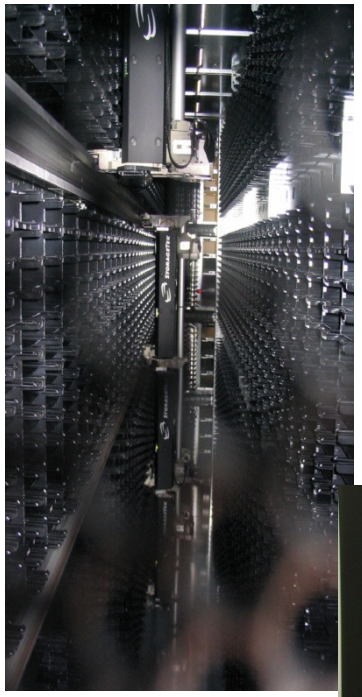
ISIS Neutron Source
Diamond X-ray Source
UK Central Laser Facility

ting
nent

Science & Technology
Facilities Council

ISIS Facility Data Collection

- 2-120 files per experiment
- Format: NeXus, RAW
- 2009:
  - 834 experiments;
  - 0.5million files
  - 0.5TB data



The StorageTek tape robot 100PB Capacity

ISIS Data Archives

All **ISIS** data (~25 years)
>8 million files
250,000 datasets
8TB data

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/theories
5) Used for new science not yet considered
6) Defend patents on innovations derived from science
7) Evidence based policy making

1. Researchers access their own data '98

ISIS PC Controlled Instrument Data File Access

Once you have entered details of the run numbers of the files you require, you will be redirected to a secure connection where:

- You will need to reply **Yes** to the question of "do you want to accept a certificate"
- When prompted, you will need to enter a **username and password** that was issued to you at ISIS - you may use either a Federal ID or VMS account.

If you do not have an account, you should contact the ISIS User Office (isisuo@rl.ac.uk) and request a "Federal ID" - please quote your experiment number (also known as the RB or proposal number) when requesting the account.

If everything is correct then, depending on the option you selected, your files will either be assembled into a ZIP file and you will be prompted to save this to your local disk, or you will be offered a page of links from where you can download them individually. Note that there is a limit on the number of files (100) and, in the case of the ZIP file, total size (300Mb) that can be downloaded in one go. If you are accessing large (e.g. MAPS) RAW files you may need to use the individual file links method for the RAW files and then download all the relevant LOG files in a single ZIP file.

RAW Data File Access

This form gives you access to both the RAW data files (e.g. GEM12345.RAW) and sample environment log files (e.g. GEM12345_Temp1.TXT) collected at ISIS - you just need to select the instrument name, type of files (RAW/LOG) and then enter the numbers of the first and last runs you collected.

PLEASE NOTE: This form only allows you to access data files created on ISIS instruments up to and including cycle 10/2 (August 2010). For data created in cycle 10/3 (March 2011) and beyond you need to use the new STFC data portal.

Instrument ARGUS

First Run Number =

Last Run Number =

Contents ☑ RAW files ☐ NeXus files ☑ LOG files ☐ SAV/S0* (UPDATE-STORE/SaveRun) files

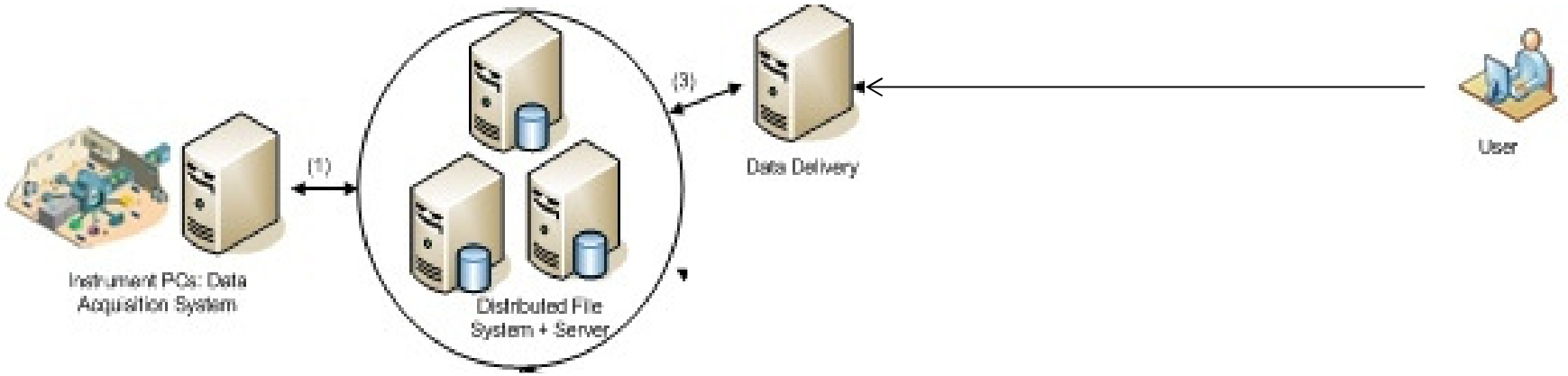
DOWNLOAD via: ◉ ZIP file ○ ZIP file (uncompressed) ○ Web links to individual data files

OR DISPLAY: ○ Windows explorer links to cycle folders (links only work from ISIS network)

OR TRANSFER to: ○ VMS cluster (ISISA/HATHOR/THOTH), files placed in directory SCRATCHSDISK:[ISISDATA] and automatically removed after 1 week

Begin Raw Data Download

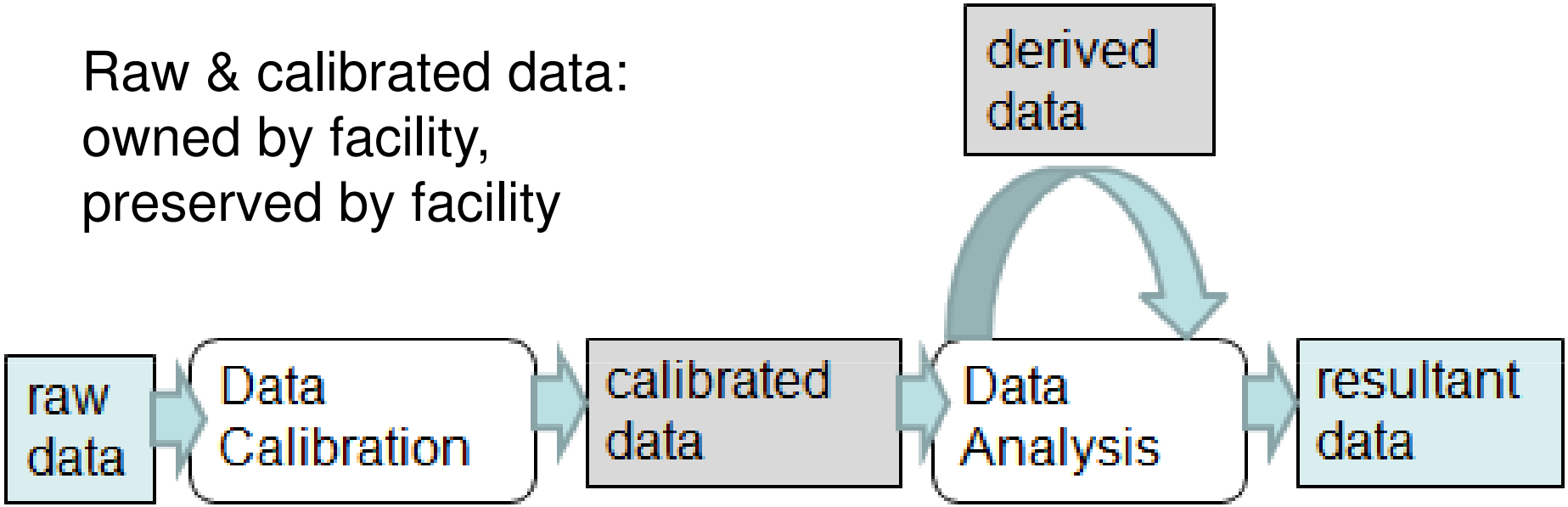
Data Preservation Infrastructure



Existing System

2. Validate published results

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



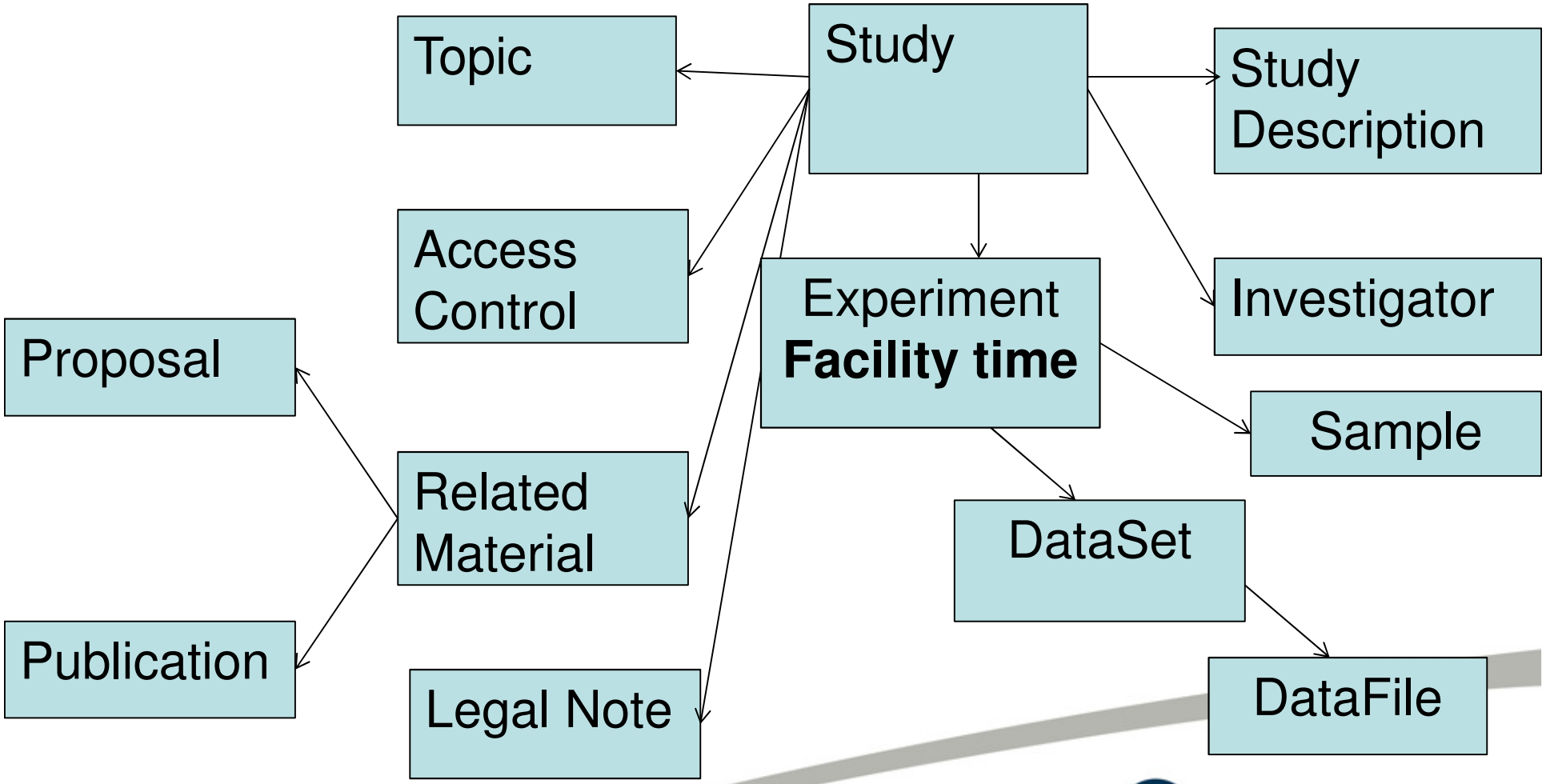
- Derived Data Provenance
- Derived Data Ownership – user, funder
- Analysis software preservation
- DOI for software versions

When to publish data?

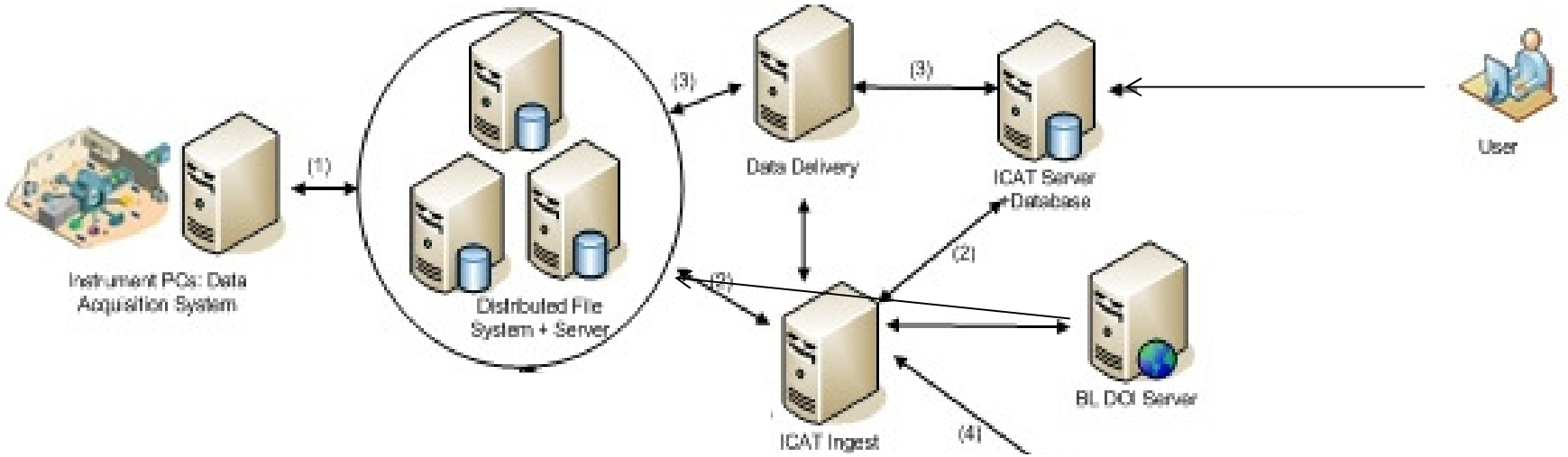
- Commercial data
  - < 1% of data
  - subject to individual contracts
  - Don't publish
- Data Policies – different science, different facilities, different policy
- Data Embargo – PhD period 3 years
- Record who accesses data
- Metadata Embargo – 2004 Haumea example

Core Scientific Metadata Model (CSDM)

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

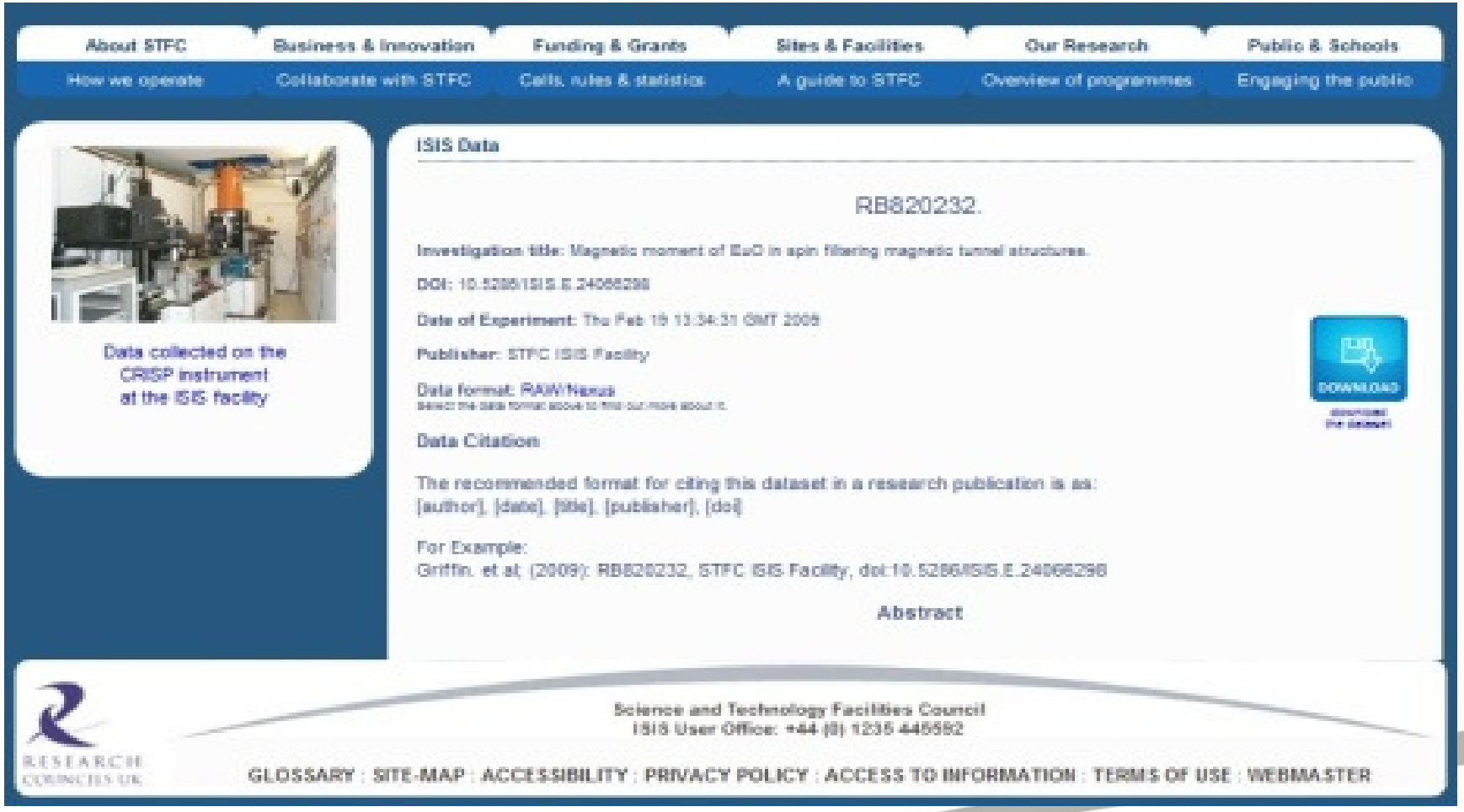


Data Preservation Infrastructure



Existing System

Discovery & Reward: Data DOI

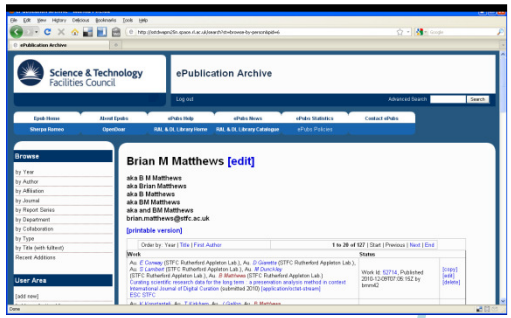


3 + 4 Meta-studies & new models

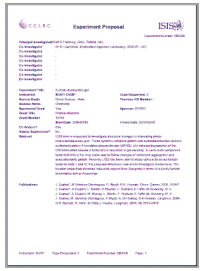
Explanatory information

- 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

Facilities Lifecycle



ICAT
Metadata Repository



Proposal

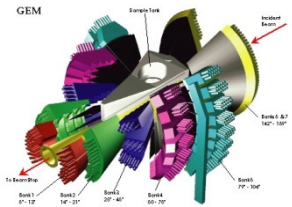
Scientist submits application for beamtime

Approval

Facility committee approves application

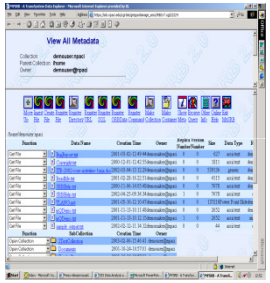
Scheduling

Facility registers, trains, and schedules scientist's visit



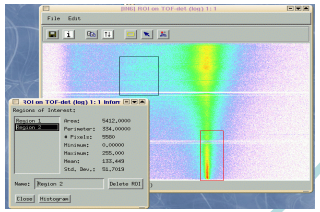
Experiment

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

Discovery: Datacite Search

http://search.datacite.org/ui - Microsoft Internet Explorer provided by STFC

http://search.datacite.org/ui#ui?&q=STFC

Metadata Search beta

STFC | Search

Options | Advanced Search | About Us | Contact | Help

Filter

allocator
datacentre
prefix
resourceType
contributor
creator
publicationYear
publisher
language
refQuality
has_metadata

No active filters. Use the sidebar to filter search results.

9 documents found in 31ms
Page 1 of 1

RB820232: Magnetic moment of EuO in spin filtering magnetic tunnel structures. # 1
doi:10.5286/ISIS.E.24066298
Easton,S • Barnes,C H W • Ionescu,A
BL.STFC - Science and Technology Facilities Council STFC ISIS Facility

RB920486: Electric field effect on the interfacial uncompensated spins in the Co/BiFeO3/STO exchange bias system. # 2
doi:10.5286/ISIS.E.24079627
Steinke,N J
BL.STFC - Science and Technology Facilities Council STFC ISIS Facility

RB1010380: Interaction of the counteracting osmolytes TMAO and urea in aqueous solutions. # 3
doi:10.5286/ISIS.E.24079772
Meersman,F P S
BL.STFC - Science and Technology Facilities Council STFC ISIS Facility

GBS 20.7GHz slant path radio propagation measurements, Sparsholt site [version 1.0] # 4
doi:10.5285/E8F43A51-0198-4323-A926-FE69225D57DD Dataset : Metadata document
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras].
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio

GBS 20.7GHz slant path radio propagation measurements, Chilbolton site [version 1.0] # 5
doi:10.5285/639A3714-BC74-46A6-9026-64931F355E07 Dataset : Metadata document
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras].
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio

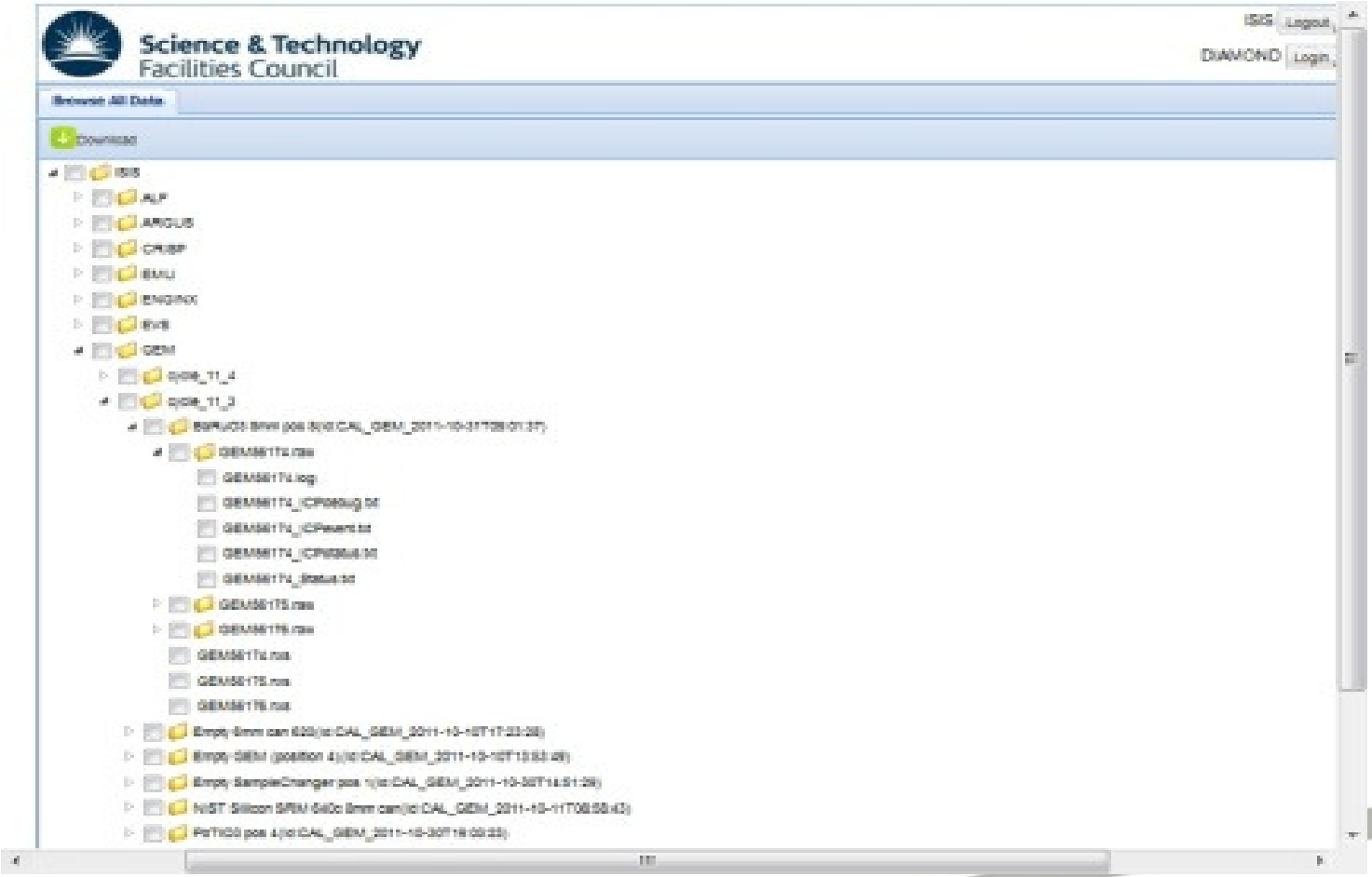
GBS 20.7GHz slant path radio propagation measurements, Dundee site [version 1.0] # 6
doi:10.5285/DB8D8981-1A51-4D6E-81C0-CCED9B921390 Dataset : Metadata document
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras].
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio

ITALSAT radio propagation measurement at 40GHz in the United Kingdom [version 1.0] # 7
doi:10.5285/4A60EE2F-0FD1-4141-9244-7BEBF240BB49 Dataset : Metadata document
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [S.Ventouras, S.A.Callaghan, C.L.Wrench]
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio

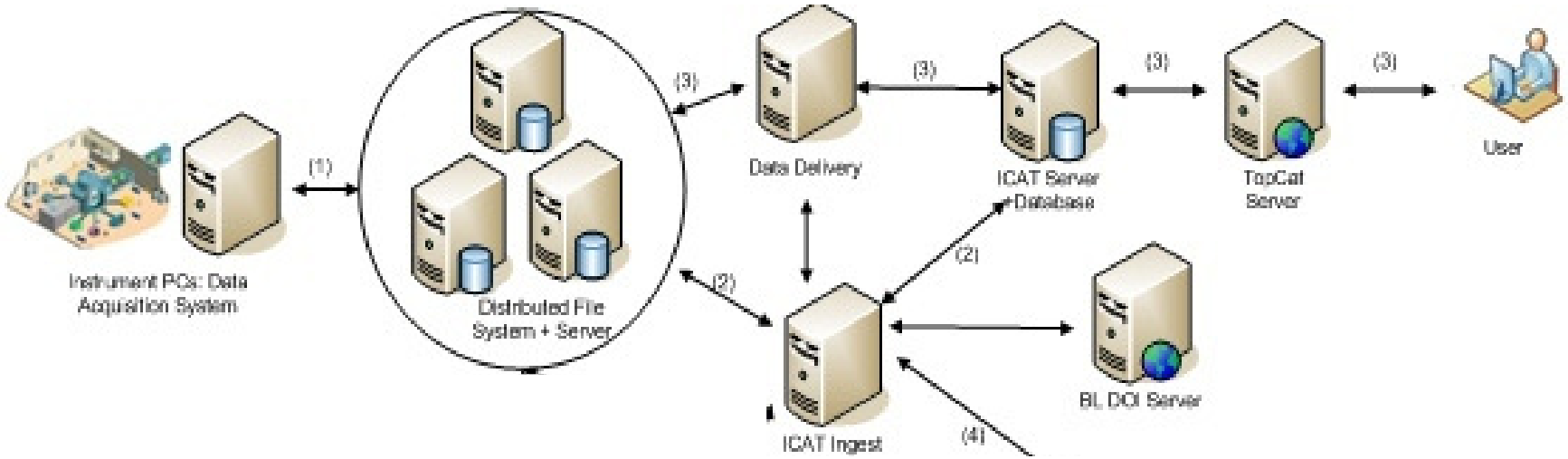
ITALSAT radio propagation measurement at 20GHz in the United Kingdom [version 1.0] # 8
doi:10.5285/3158D138-D592-4045-ADE4-B76CF9F42129 Dataset : Metadata document

Internet | Protected Mode: Off 75%

TopCat: Browse & Search

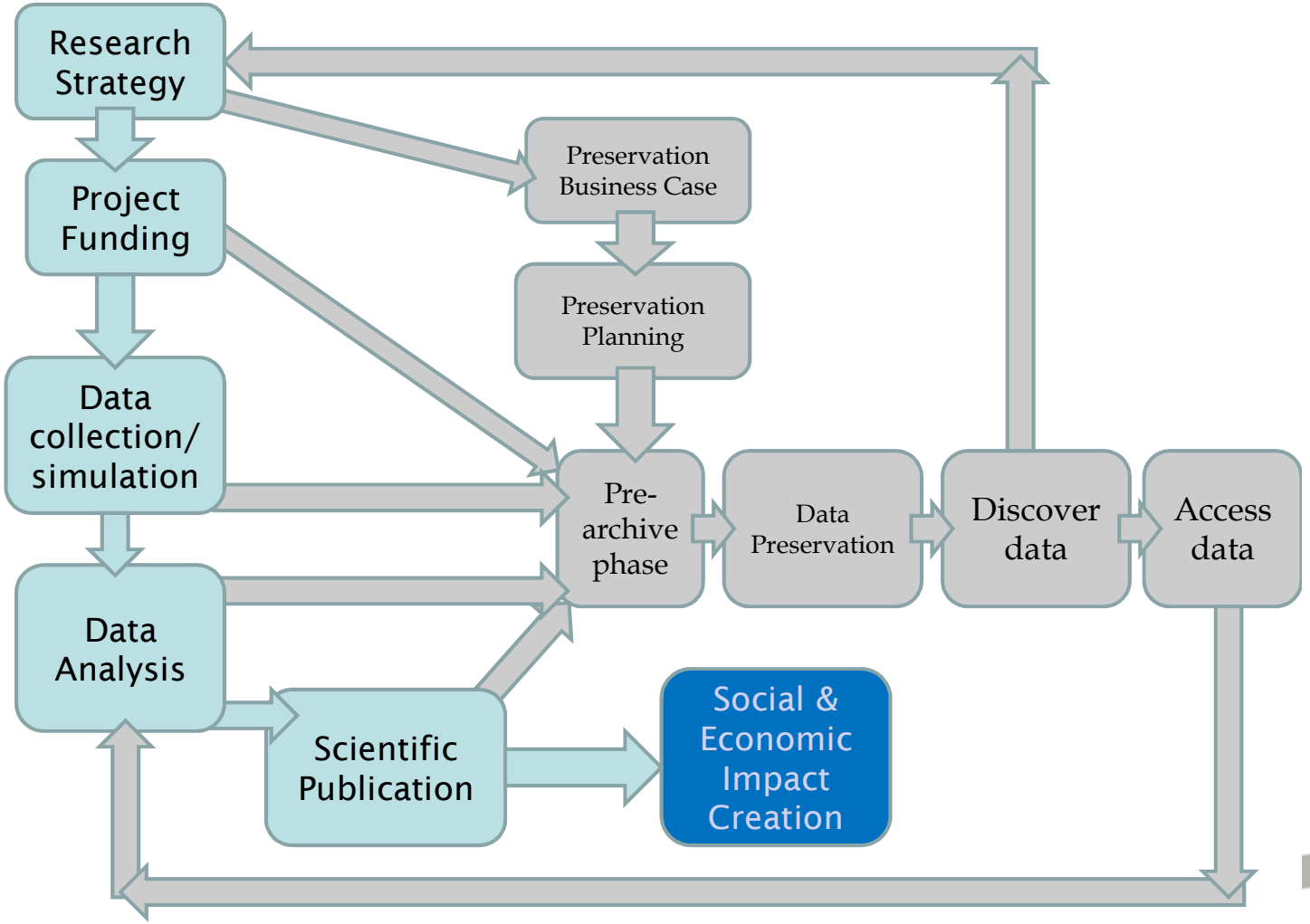


Data Preservation Infrastructure



Existing System

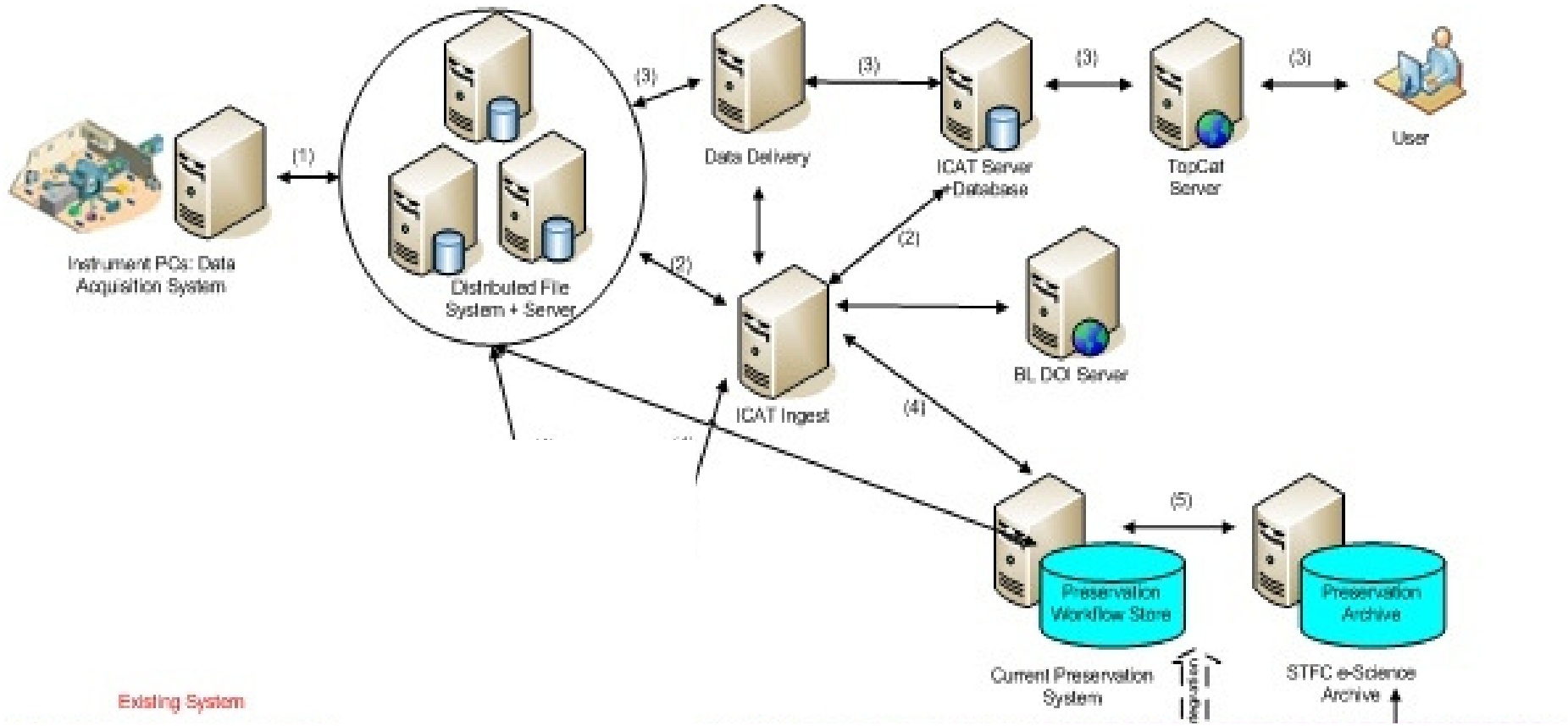
5,6,7 new science, patents, policy



Long Term Preservation

- Tesella Safety Deposit Box
- Fixity Checks
- Data Format Migration
- Long Term archive – Petabyte store

Data Preservation Infrastructure



Conclusion

- Preservation Objectives
- Timescale of objective – short, medium, long
- Designated Communities
- Additional Information
- Security Requirements
- Probability of benefits – low prob., high impact
- Business Case
- Technical Architecture to meet needs

Questions ?


data
LHC
used
research
e-Science
petabytes
particle
experiments
collisions
energy
STFC
Tier-1
first
November
computing
physics
services
users
researchers
beams
construction
experimental
year
new
R89
UK
years
number
Rutherford
March
impact
processing
store
CERN
machine
universe
tier-1
four
events
CMS
Research
LHCb
network
SCARF
centre
world
Atlas
Appleton
Laboratory
expected
storing
started
project
recorded
science
rates
start
move
programme
also
simulating
requirements
time
sites
support
distributed
experiment
activities
plasma
smoothly
initiatives
received
main
Two
TeV
wide
protons
handling
including
high
Scientific Computing
Department
Science & Technology
Facilities Council