Data Management and Preservation Planning for Big Science

Juan Bicarregui[1], Norman Gray[2], Rob Henderson[3], Simon Lambert[1], Roger Jones[3], **Brian Matthews[1]**

[1]STFC, [2]Glasgow, [3]Lancaster

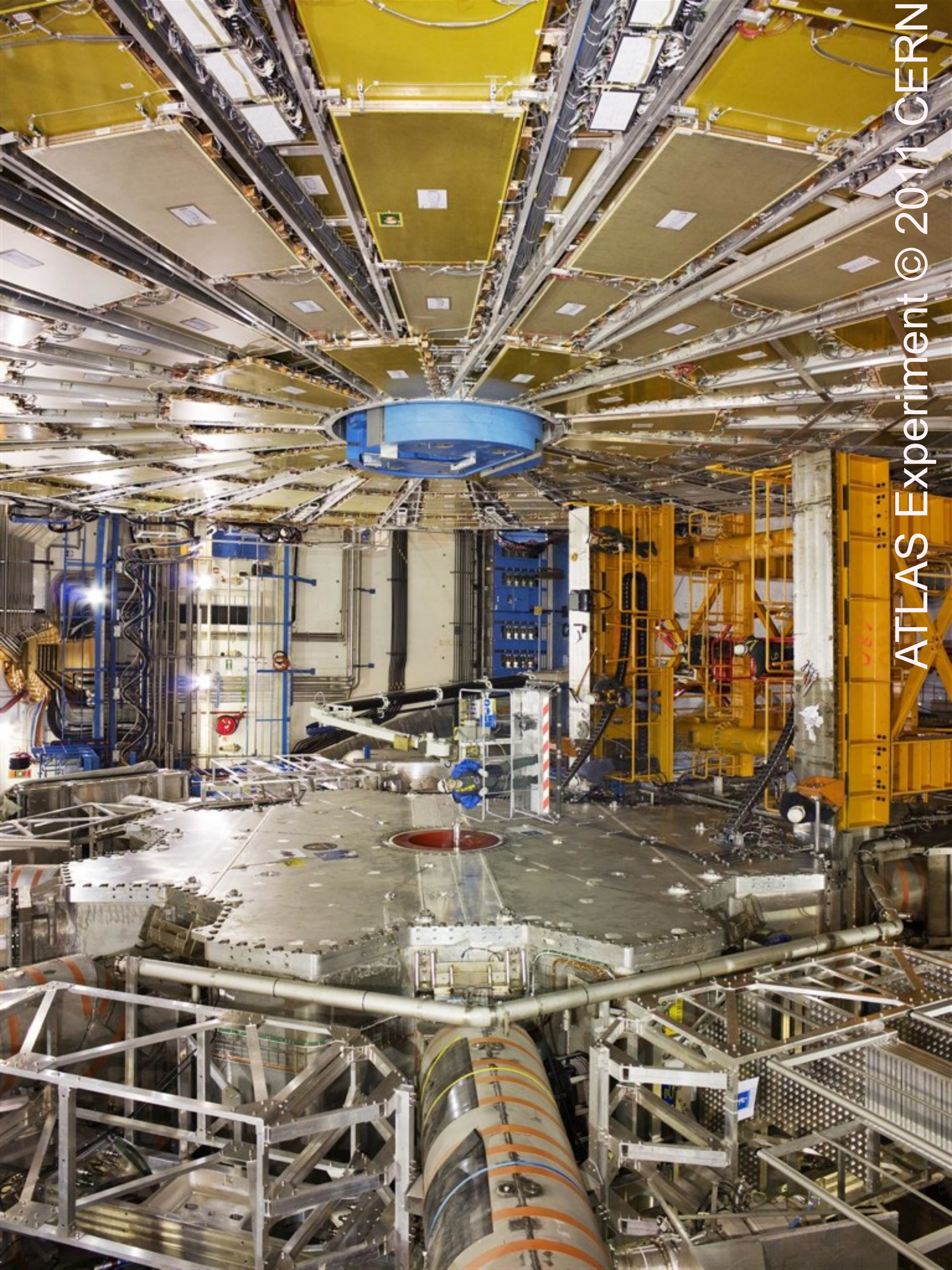
What is Big Science?

- big discoveries
- big money
- big timescales
- big author lists
- big admin
- big infrastructure
- big consensus

Science & Technology Facilities Council

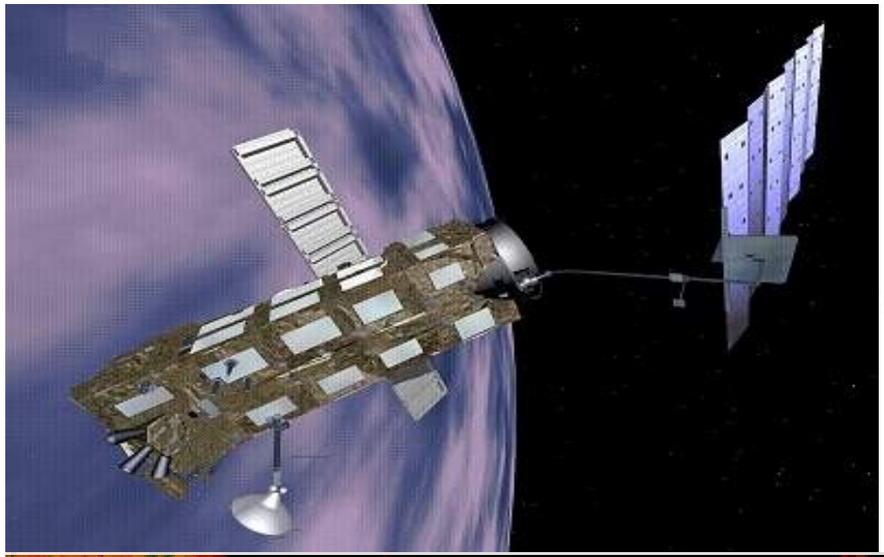
Mauna Kea, in Hawai`i

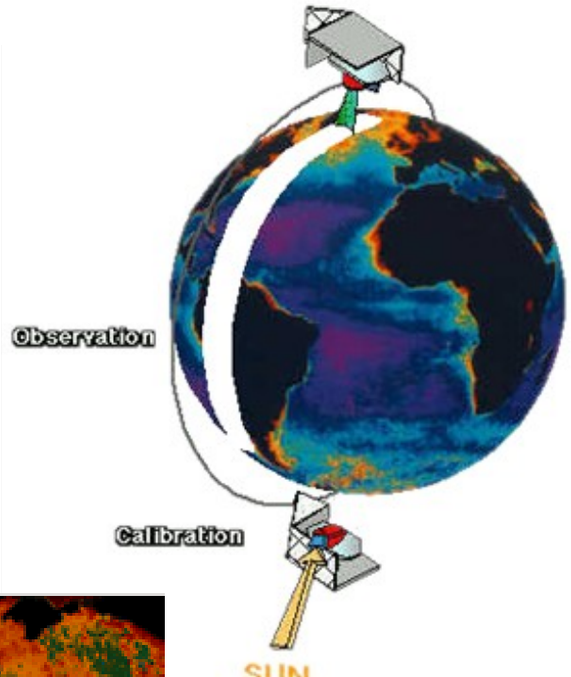
13 telescopes; 11 countries

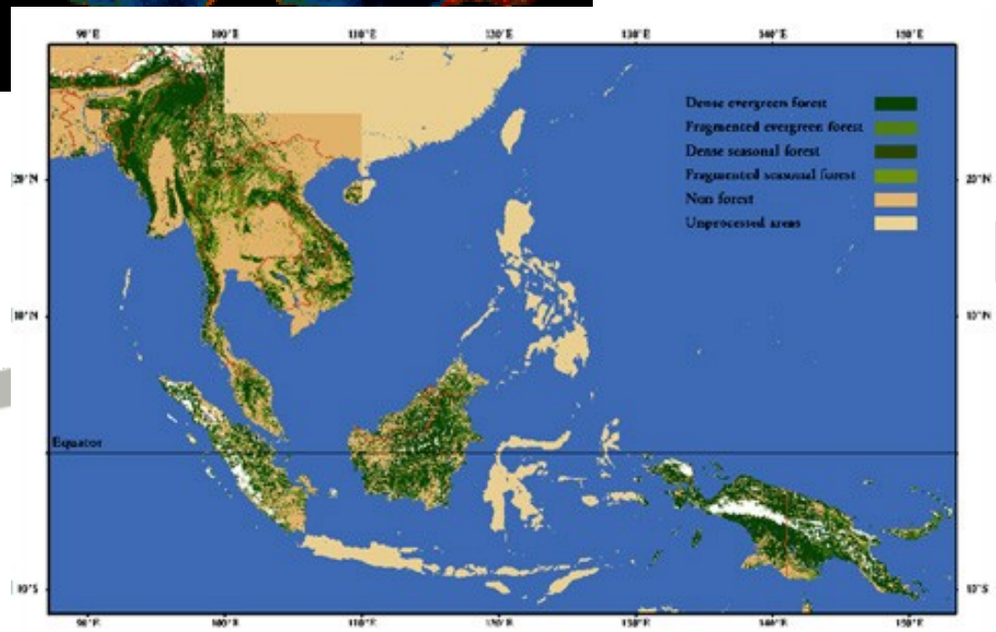


The MERIS Instrument on ENVISAT

- Medium Resolution Imaging Spectrometer (MERIS)
  - an instrument on the ESA ENVISAT EO satellite
- Primarily: sea colour measurement
  - Chlorophyll
  - Suspended sediment
  - Atmospheric aerosol over water
- Also land vegetation
- Understand the carbon cycle
  - How this changes under climate change
  - Also agriculture and fisheries



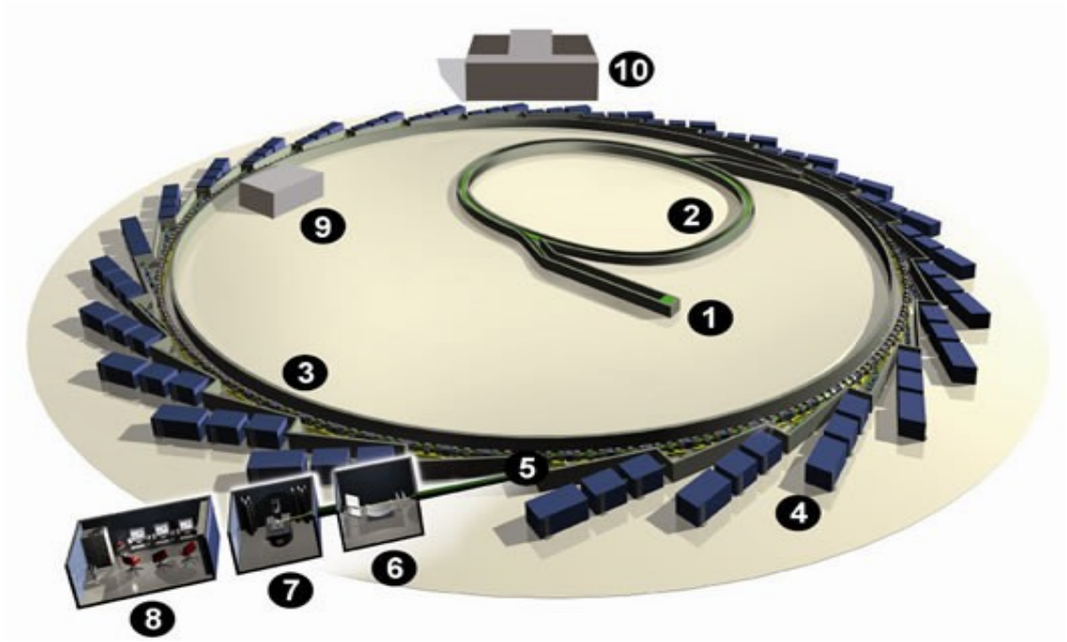




esa

scidip-es

“Source” Facilities: big support for small science



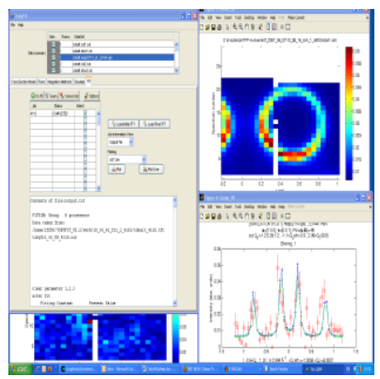
Synchrotron and Neutron Sources

Lots of “small science” experiments
Central “big science” infrastructure

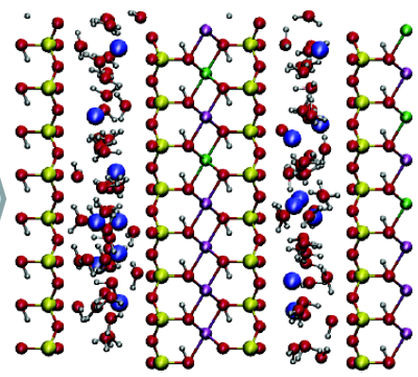
Visit facility on research campus

Place sample in beam

Diffraction pattern from sample



Fitting experimental data to model



Structure of cholesterol in crude oil

Data Management infrastructure

- Lots of data:
  - LHC is 10PB/year;
  - LIGO (gravitational wave) - 1PB/year;
  - SKA will transport 0.5 EB/year intercontinentally (0.05% of total 2015 IP traffic)

- ...but data volume is not the core problem, because...
  - Dedicated Teams
  - Innovative data storage and transport
  - Custom formats and data analysis software
  - Plenty of tacit knowledge (separate curation problem)

- Good data management designed in from conception
  - Data recognised as precious asset of the project

Science & Technology Facilities Council

ROW 8

e-Science

Higgs Boson Candidate at RAL

- On the 18th of May, 2012, the LHC's ATLAS detector picked up the decay of a possible Higgs boson into two electrons and two positrons.
- The data from this collision is stored here in the e-Science building. From this sign, the disk server with the collision data is in the right-hand aisle, second rack on the left, second machine from the top (gdss489).

steamlin
computin

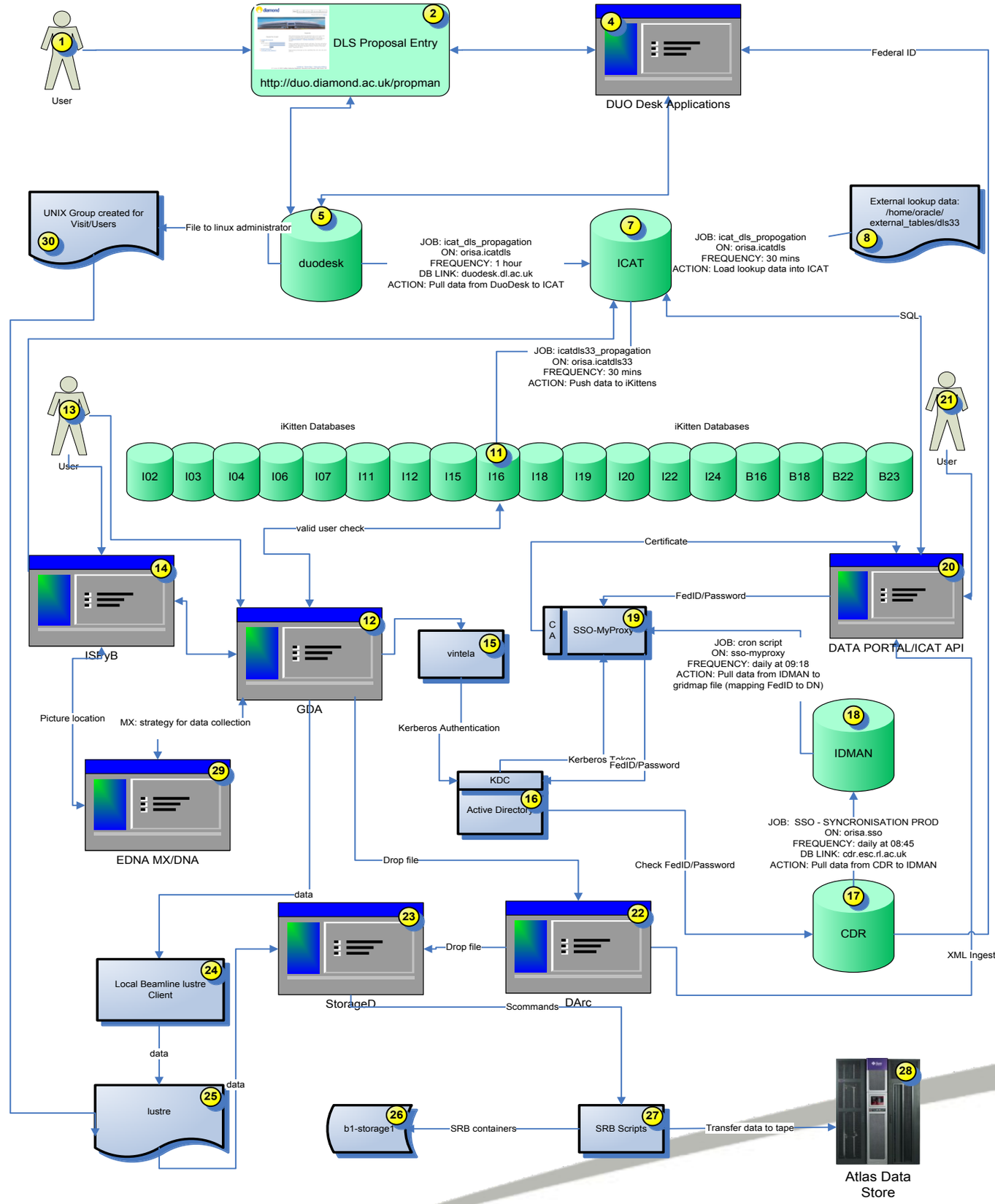
- On the 18th of May 2012, the LHC's ATLAS detector picked up the decay of a possible Higgs boson into two electrons and two positrons.

- The data from this collision is stored here in the e-Science building. From this sign, the disk server with the collision data is in the right-hand aisle, second rack on the left, second machine from the top (gss489)

Science & Technology Facilities Council
e-Science

Diamond e-Infrastructure

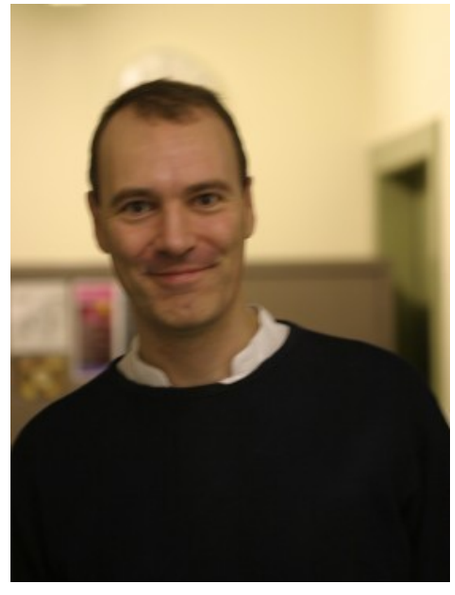
diamond



- Architecture use for DLS
- Scaling is a constant concern
- Data rates keep increasing
  - 70TB per month and rising

Science & Technology Facilities Council

mrd-gw project



JISC funded project – Norman Gray, University of Glasgow

- Data management planning for big science
- The language of ‘data products’ and explicit ‘proprietary periods’ is useful
- Funders should simply require that a project develop a high-level DMP as a suitable profile of OAIS
- Funders should support projects in creating per-project OAIS profiles
- STFC should develop a costings model matched to the data challenges of the big-science community

JISC

Science & Technology Facilities Council

Claim

- The demand for principled data management and data sharing and data preservation is a reasonable and shared one;
- a reasonable framework for at least approaching the problem already exists in OAIS;
- the OAIS recommendation is (just) concrete enough that it is not merely waffle; and
- there is a bounded set of resources which will allow DMP planners to produce a practical project DMP plan, reasonably painlessly.

Science & Technology Facilities Council

Here’s a copy of CCSDS 650.0.

It’s sane.

Get on with it.

Science & Technology Facilities Council

Well, up to a point

- Pointless re-invention
- Awareness of best practise
- Policy framework
- Get to a best solution fast

Guidelines to best-practice for DMP for Big Science
  – For funders and practitioners

MaRDI-Gross : Managing Research Data Infrastructures – Big Science
http://arxiv.org/abs/1208.3754

JISC

Science & Technology Facilities Council

POLICY

Science & Technology Facilities Council

Policy Framework

- Funders Requirements
  - Governmental and inter-governmental policy frameworks, MoUs, treaties
  - Legislation in different countries
  - Funders data policy
- Open Data
  - Funders push towards it
  - Not without costs
    - Maintaining interpretability (Rep Info)
    - Cost may depend on the data product

Science & Technology Facilities Council

Preservation Objectives

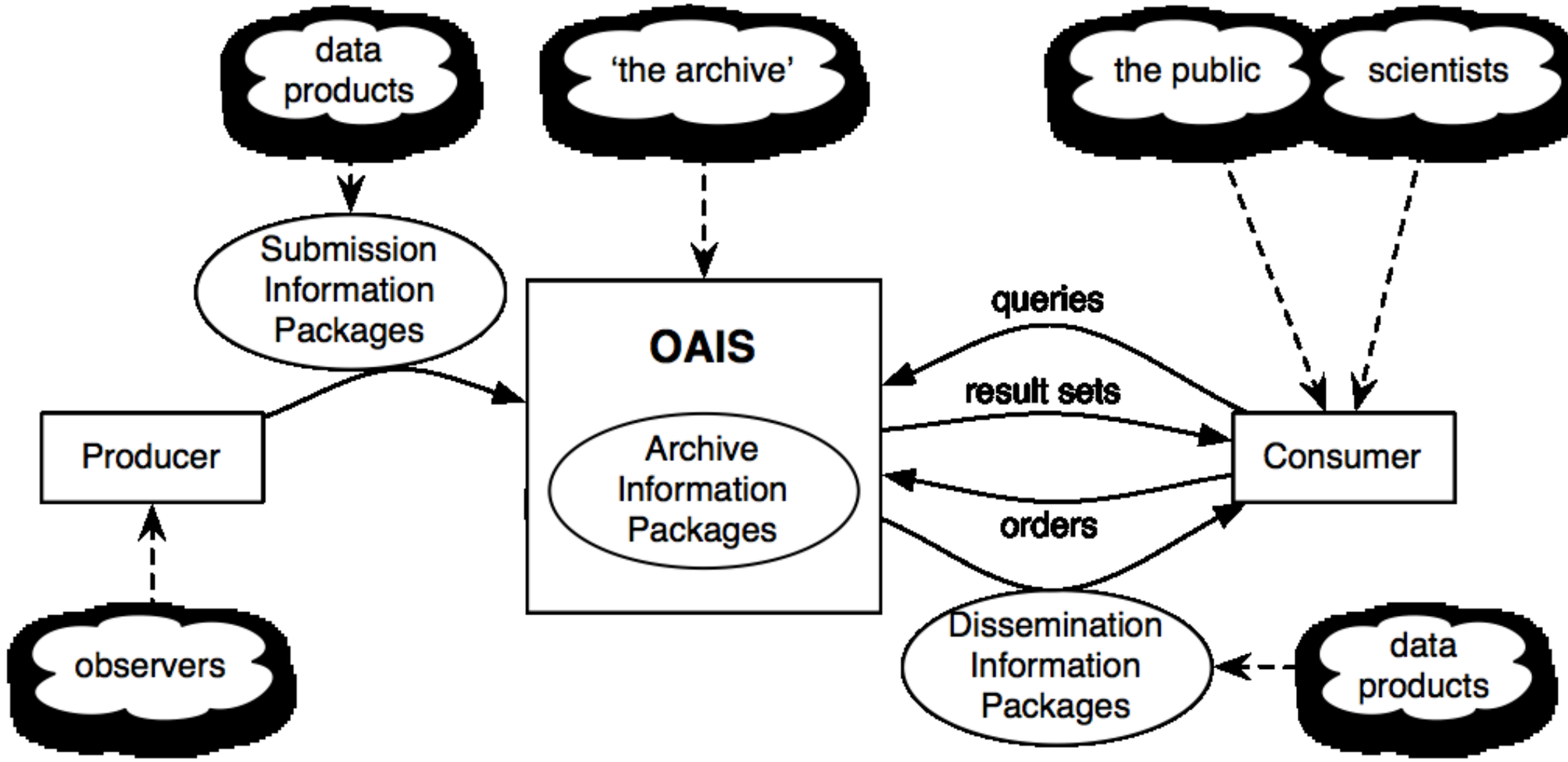
- What are precisely the long-term preservation goals?
- May vary widely between projects
  - Astronomy and Earth Observation
    - potentially reusable indefinitely
  - Particle Physics and Source Facilities
    - Complexity of apparatus
    - Representation information in “tacit knowledge”
    - Obsolescence
    - Cost of preservation – after collaboration complete
- Cost/benefits of long-term preservation
- Policies for data retention and disposal

Science & Technology Facilities Council

TECHNOLOGY

Science & Technology Facilities Council

OAIS



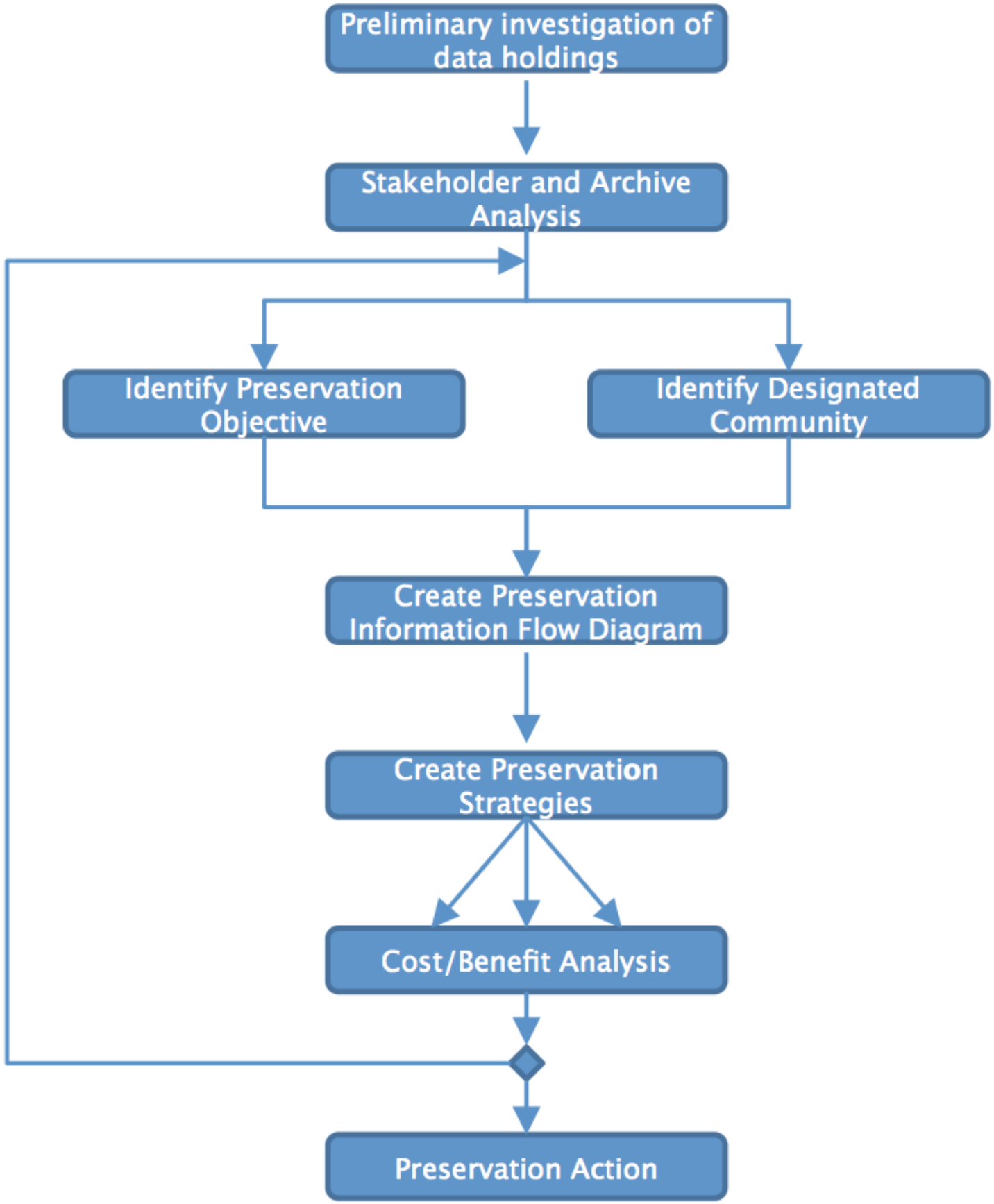
CCSDS 650.0 (2002) = ISO 14721:2003

"almost any system capable of storing and retrieving data can make a plausible case that it satisfies the OAIS conformance requirements"
Rosenthal et al (2005)

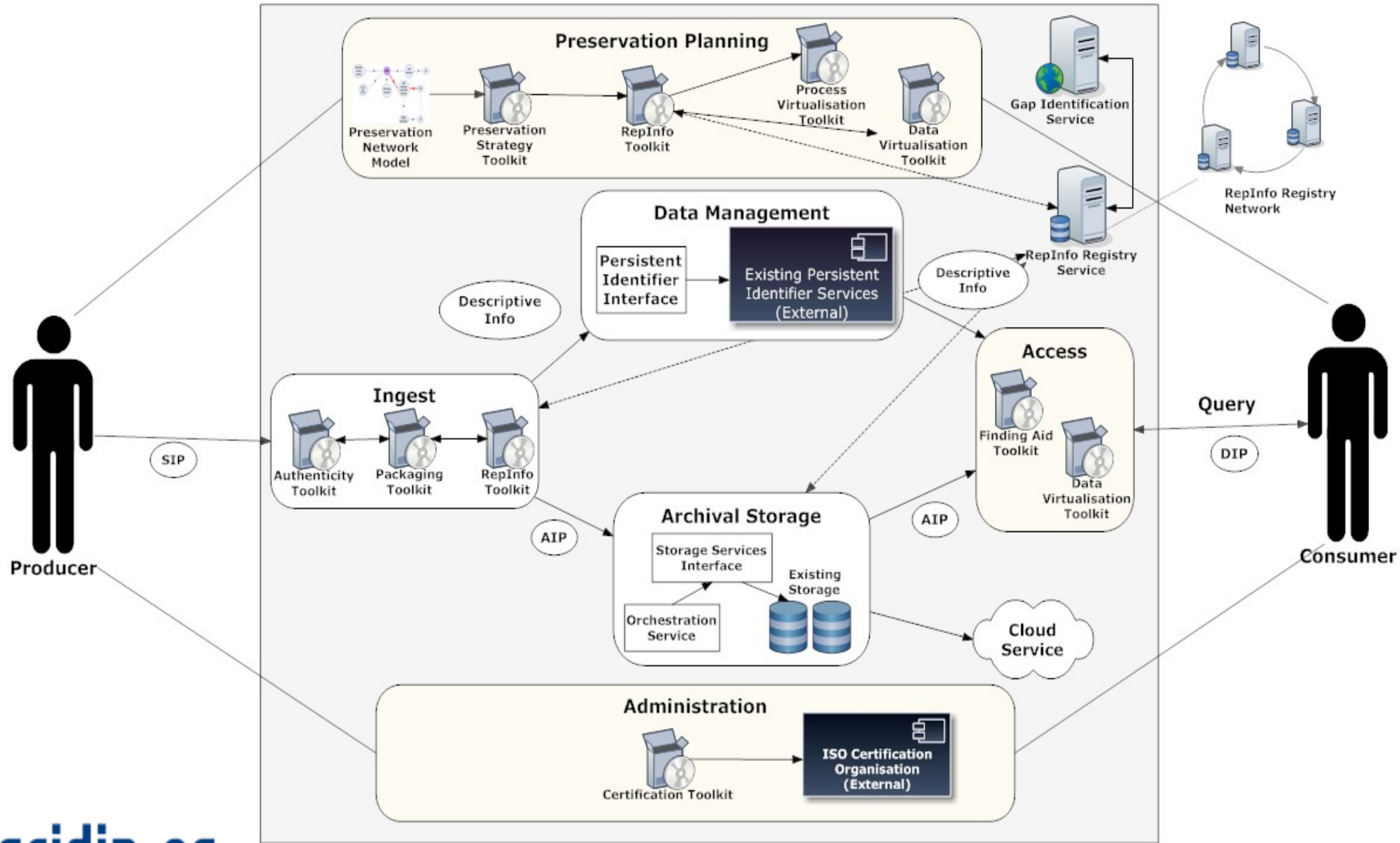
Science & Technology Facilities Council

No general recipe for OAIS

Systematic development of a DMP strategy



SCIDIP-ES Preservation Infrastructure



scidip-es

Science & Technology Facilities Council

Certification

‘Trustworthy repositories audit and certification (TRAC)’

‘Audit and certification of trustworthy digital repositories’
CCSDS 652.0 = ISO-16363:2012

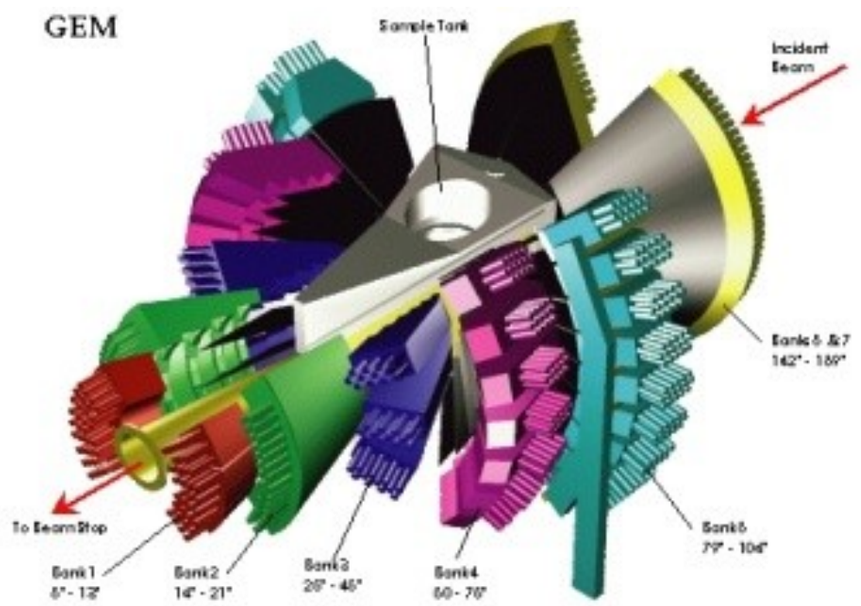
Science & Technology Facilities Council

PRACTICALITIES

Science & Technology Facilities Council

Ingest and Acquisition

- often very expensive, front-loaded and tailored;
- generally absorbed in infrastructure costs for big science
- staffing: expensive but predictable
- Preservation representation information additional
  - but may be relatively small



- Less of an issue for Big Science
  - *It's just what you have to do*

Science & Technology Facilities Council

Data Release

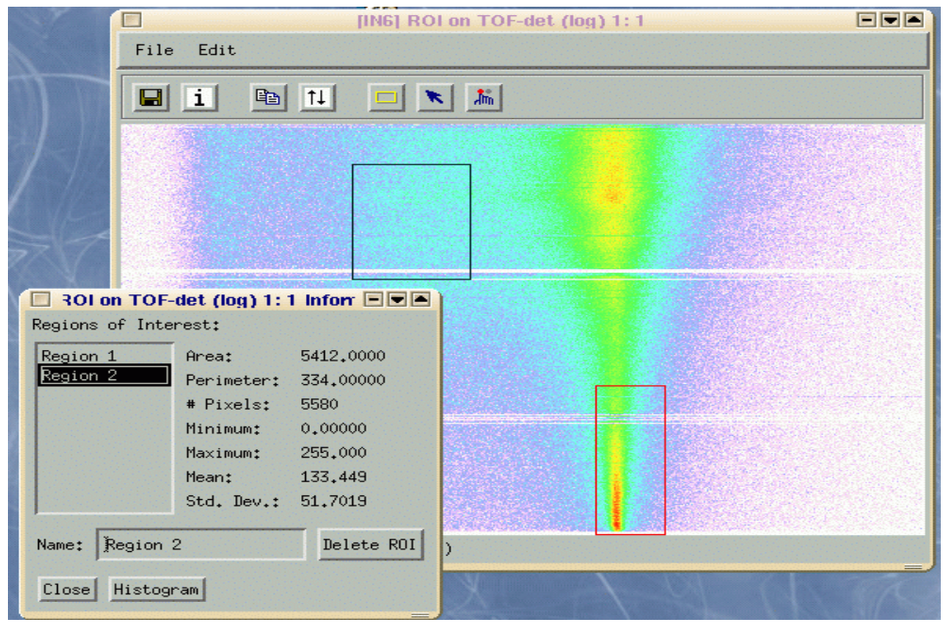
- Data release for large and long term data acquisitions needs to be considered
  - Also needs to take into account the “interpretability” of data

- Examples of practice
  - LIGO: explicit algorithm for timing data release, which is a function of time, amount of space explored, and discoveries
  - ATLAS has ‘RECAST’ service: they don’t release data, but will re-analyse their data with your model
  - Astronomy: either proprietary periods, or surveys have periodic DRs after QA checks
  - Source Facilities: Embargo periods, interaction with “small science”

Science & Technology Facilities Council

Software Preservation

- Specialist software is needed
  - Maintain the interpretability of data
  - Explicit and tacit knowledge
  - Maintaining open source software systems
  - Maintaining data products an alternative



Science & Technology Facilities Council

So what are the lessons for "small science" ?

Big Science

- Ingest part of the infrastructure
- Culture of data management and sharing
- Automated collection of metadata
- "complex" technology solutions to complex and/or large scale problems
- Organisational structures in place
- Lose track of the context after data release

Small Science

- Archives ingesting data from research teams
- Ad-hoc data management, more defensive.
- Metadata collection hard and patchy
- Relatively "established" technological solutions
- Complex organisational and cultural barriers

Science & Technology Facilities Council

Boundary between big and small science

Many cases where boundary is blurred

- Longitudinal studies
  - Long time scales, shared access to scarce research data
- Clinical trials
  - Legal and certification requirements
- Research lab scale equipment
  - E.g. tomography, crystallography, sensor nets
  - Large-scale of data
- Analysis, modelling and simulation
  - Software preservation

- Benefit from experience in big data
  - Tools, methodology, terminology, standards, best practice.

Science & Technology Facilities Council

Conclusion

Big science not simply a matter of scale

- The data acquisition presents significant technical challenges
  - Scale, complexity, dynamism, timescale,
  - Need a dedicated infrastructure to make data available *at all.*

- ‘next generation data management problems’,
  - significant unsolved technical preservation problems.

- *Thus “big-science” repositories become “small-science” repositories as the technical challenges are resolved*
  - *Can then concentrate on the* organisational, cultural, educational and usability issues

Science & Technology Facilities Council

purl.org/nxg/projects/mrd-gw

purl.org/nxg/projects/mardi-gross

http://arxiv.org/abs/1208.3754

brian.matthews@stfc.ac.uk

Science & Technology Facilities Council