STFC/UK
Policies and Programs

David Corney, STFC e-Science centre

First Workshop on Data Preservation and
Long Term Analysis in HEP
DESY 26th – 28th January 2009

Talk Overview

- STFC overview (incl UK Tier1 and e-science)
- Digital Preservation
- Bottom up:
  - STFC core interest projects
  - Other relevant projects
- Top down - UK Research Councils' approach
- Conclusions

About STFC...

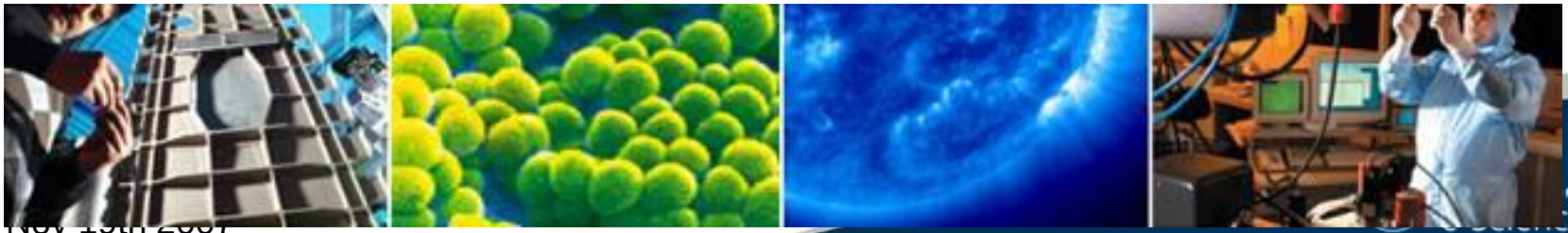
The Science and Technology Facilities Council (UK)

Created on April 1, 2007 (1 of 7 UK research Councils)

Responsible for:

- fundamental research in particle physics, nuclear physics, astronomy, space
- major UK facilities for the physical and life sciences
  - synchrotrons, light sources, lasers, neutrons
- national laboratories at RAL, Daresbury, UKATC
- international science projects
  - CERN, ESO, ESA, ILL, ESRF…

Over 2000 staff and an annual budget of over £700M



Rutherford Appleton Laboratory

STFC e-Science Centre

Exploit e-Science technologies throughout STFC's programmes, the research communities they support and the national science and engineering base.

- Especially ISIS (Neutron Spallation Source), DLS (Synchrotron X-Rays), CLF (Lasers), CERN
- LHC via the UK Tier1 centre
- Grid, HPC, Data storage, Libraries, Data Management, Visualisation
- UK Digital Curation Centre
- ~80 staff

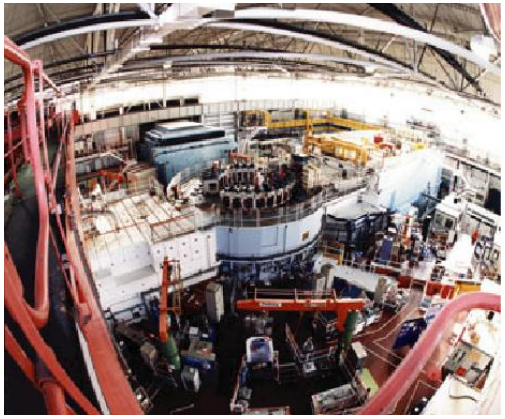
http://www.e-science.clrc.ac.uk/

e-Infrastructure for scientific facilities

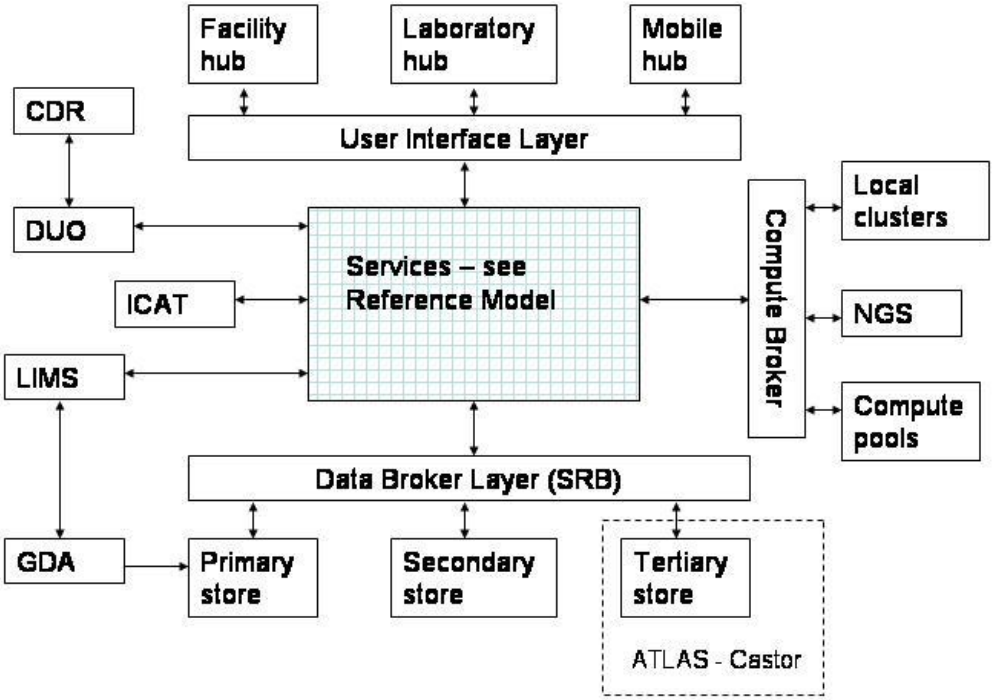
Diamond synchrotron

E-science e-infrastructure provides Data management services to the STFC facilities

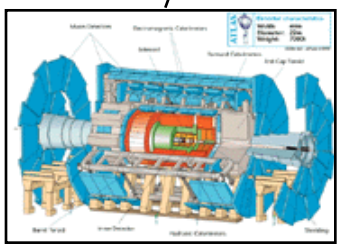
ISIS neutron and muon facility



Vulcan laser facility



Tier Structure



Useful model for Particle Physics but not necessary for others

Tier 0

Tier 1
National centres

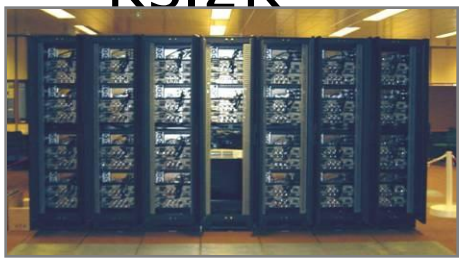
Tier 2
Regional groups

Institutes

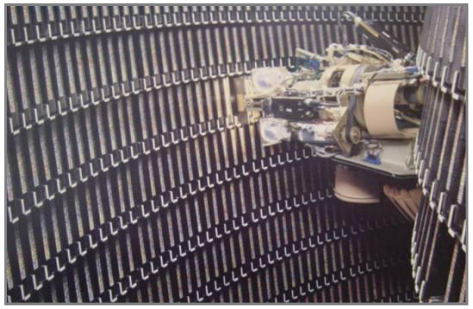
Workstations

Tier-1 Hardware

CPU Power (Reconstruction, Simulation, User Analysis etc).
600 systems, 1250 cores, 1500 KSI2K



Disk Storage (Frequently Accessed)
138 Servers, 3200 drives, 750TB



'Tape' Storage – Long Term retention – write once – read several times a year – 1PB in SL8500 robot + 12 drives

Currently about 45 racks – with a further 25 due to arrive for Xmass

Partial Digital Preservation

1956: IBM350 disk drive - 0.004Gb - $25k/month to lease

Digital Preservation

Preserving knowledge

Implies the need to "migrate": software, hardware, data, processes and *adequate information to allow others to understand and use the data*.

CASPAR and OAIS (Open Archival Information System - ISO 14721)

A framework for long term preservation

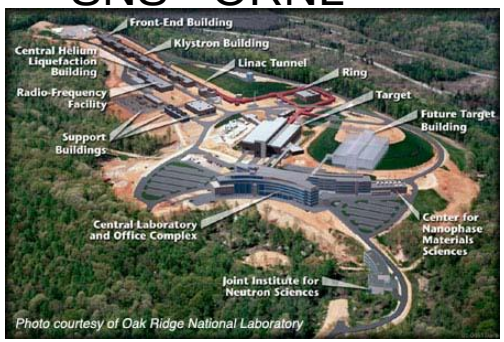
"Representation Information" and "Designated Communities"

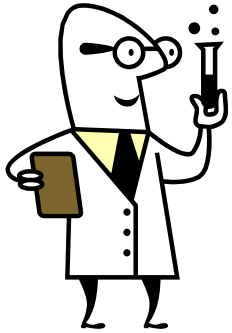
Talk Overview

- STFC overview (incl UK Tier1 and e-science)
- Digital Preservation
- Bottom up:
  - STFC core interest projects
  - Other relevant projects
- Top down - UK Research Councils' approach
- Conclusions

e-Infrastructure – Access to Multiple Facilities

SNS - ORNL





ANSTO - Australia

iCat

ISIS – TS1 + 2

DLS

CLF

EDNP

European Data Infrastructure for Neutron and Photon Sources

Combining European Neutron and Synchrotron Facilities

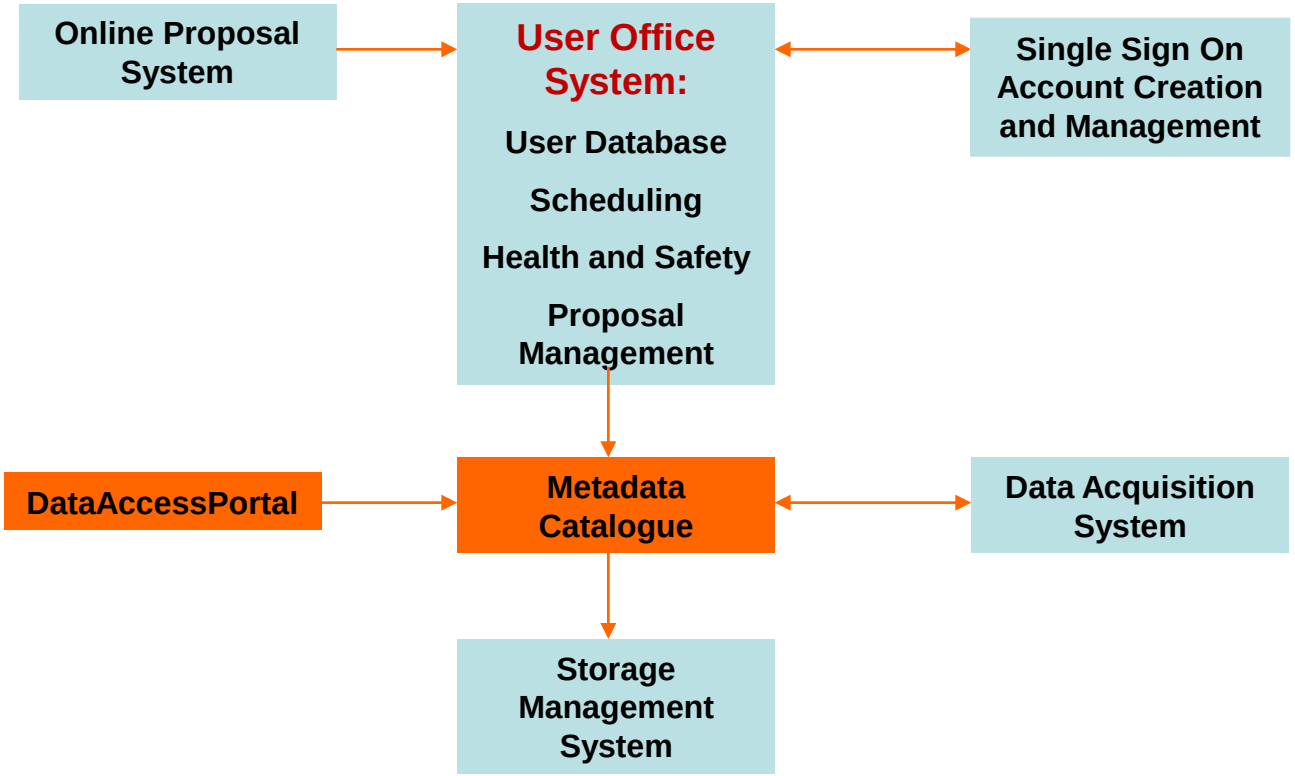
Already a common user community

Across many disciplines

- Materials, chemistry, proteomics, pharmaceuticals, nuclear physics, archaeology …

Science & Technology Facilities Council
e-Science

Underlying Data Infrastructure



ICAT Software Suite, providing the crucial integration of key functions.

Is it worth it?

"Long term data stewardship is an expensive activity. The justification for funding it must lie in the contribution that it will make to science, the creation of wealth, or to improvements in the quality of life…." *National Environmental Research Council. Data Policy. December 2002.*

What framework or processes exist to enable HEP to make their decisions? (Is this up to the experiments?)

The DCC and digital curation
(credits Graham Pryor)

Three fundamental elements of digital curation:

- Preservation
  - Environment, security, migration, standards – a trusted body of data
- Access
  - Metadata, authentication, management of data sharing regimes, legal compliance
- Re-use
  - Tools for consistent curation, annotation/citation, representation information, conversion of original data to new data/information, management of IPR

DCCs Objectives (credits Graham Pryor)

- Sustain an effective distributed organisation to support stakeholders in digital curation
- Strengthen curation networks and collaborative partnerships
- Deliver a programme of events that facilitate the development of curation knowledge and skills
- Assist the research community's acquisition of curation skills by making the necessary tools available
- Identify, create and disseminate information resources in support of curation awareness and knowledge

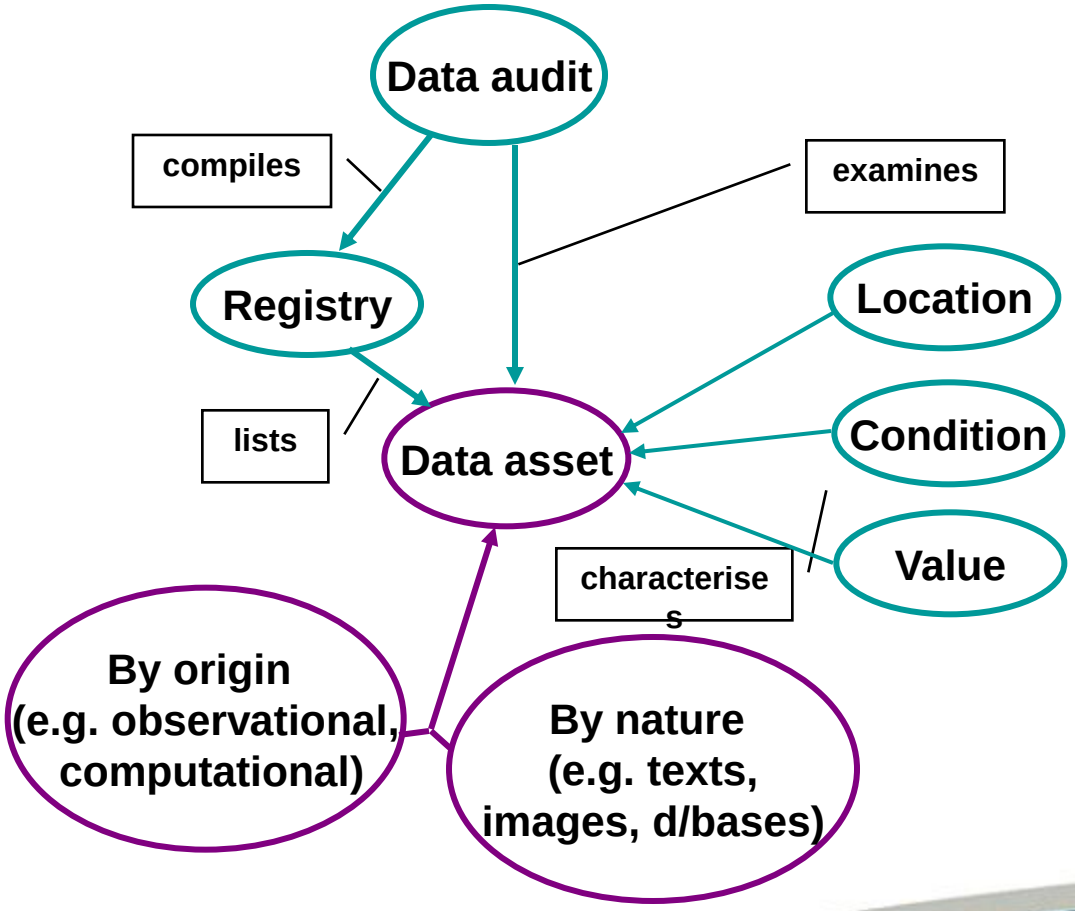
DCC activities & outputs

(credits Graham Pryor)

Data Audit Framework

Produces an inventory of research data assets with recommended data management plan

- Sample 'boreholes' in selected institutions
- Analysis of data condition
- Data audit collection scenarios
- Development of generic software



Other National solutions?

- Define current and future research data service needs
- Identify priorities for action
- Develop scenarios/options - from “do nothing” to a managed national service
- Develop business plan for preferred option(s), with costs/benefits
- Indicate scale of investment required and estimated ROI

http://www.ukrds.ac.uk/

And more...

- Digital Preservation Coalition:
  - "to secure the preservation of digital resources in the UK and to work with others internationally to secure our global digital memory and knowledge base."
- Alliance for permanent Access
  - PARSE.Insight. "Focus on the infrastructure needed to support persistence and understandability of these key assets over the long term."
- JISC, HEFCE, RCUK,...

Data Policy

- Data Policy (ISIS)
  - 3 year embargo on data (+1 if requested)
  - Commercial data is never made public
  - Instrument Scientists can access all data from their beamline
  - Calibration data is public
  - Any data that involves IPR (e.g. analysed) is private for perpetuity unless explicitly shared by user

- Automatic Enforcement of policy
  - A research area

UK RC's & Data Policies

- No reference to a Data Policy on web site:
  - Arts and Humanities Research Council (AHRC)
  - Engineering and Physical Sciences Research Council (EPSRC)
  - Science and Technology Facilities Council (STFC)
  - (PPARC PPARC DELIVERY PLAN: 2006-08)

- Economic and Social Research Council (ESRC)
  - Applications invited for Demonstrator Scheme for Qualitative Data Sharing and Research Archiving. (2005)
  - No mention of Data policy in EPSRC Delivery Plan 2008-2011

Biotechnology and Biological Sciences Research Council (BBSRC):

- "Data Sharing Policy" (15 Jan 2009):
- BBSRC encourages community development of standards where these do not currently exist...
- BBSRC recognises that different approaches to data sharing will be required in different situations and considers that it is most appropriate for researchers to determine their own strategies for data sharing and outline these within their research grant proposal(s).
- All applicants must include a "statement on data sharing" as part of the case for support within research grant proposals.

Medical Research Council (MRC)

MRC policy on data sharing and preservation:

Our policy builds on the central principles of the Organisation for Economic Co-operation and Development (OECD) in its report "Promoting Access to Public Research Data for Scientific, Economic and Social Development". These are that publicly-funded research data are a public good, produced in the public interest, and that they should be openly available to the maximum extent possible.

Policy applies to all MRC-funded research. It does not prescribe when or how researchers should preserve and share data, but requires them to make clear provision for doing so when planning and executing their research.

Data Sharing can lead to improved science: BIRN

What's in it for me?

Natural Environment Research Council (NERC)

Data Policy V2.2 (December 2002) 18 pages

The NERC data policy details our commitment to support the long-term management of our data. It also outlines the roles and responsibilities of NERC funded scientists, the NERC data centres and NERC management in ensuring that data that are collected using NERC funds are available for the long-term. The main agents for supporting the data policy are the network of NERC data centres.

We have created our data policy to be consistent with legal frameworks, such as the Environmental Information Regulations 2004, and contractual arrangements with other bodies where, for example, NERC holds data on their behalf but without owning the Intellectual Property Rights

Conclusions?

Data Preservation is complex, expensive and unsolved.
"HEP" need to:

- clarify what they are trying to achieve
- understand the costs and potential benefits, and decide if they are worth it

How will this happen?

Significant work into Data Preservation has been going on outside HEP. How can HEP best learn from this?

What is the role of the experiments in this? Who owns the data?

Thanks to:

Andrew Sansum (STFC), Michael Gleaves (STFC), Brian Matthews (STFC), Michael Wilson (STFC), Graham Pryor (DCC)