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
ISIS Facility Data Collection

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

ISIS Data Archives
All ISIS data (~25 years)
> 8 million files
250,000 datasets
8 TB 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
Data Preservation Infrastructure
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)


Diagram:

- Proposal
- Access Control
- Related Material
- Legal Note
- Study
- Study Description
- Investigator
- Sample
- Data Set
- Data File
- Topic
- Experiment Facility time
Data Preservation Infrastructure
Discovery & Reward: Data DOI

Investigation title: Magnetic moment of EuO in spin filtering magnetic tunnel structures.
DOI: 10.5286/ISIS.E.24066298
Date of Experiment: Thu Feb 19 13:34:31 GMT 2009
Publisher: STFC ISIS Facility
Data format: RAW/Nexus
Data Citation
The recommended format for citing this dataset in a research publication is as:
[author], [date], [title], [publisher], [doi]
For Example:
Abstract

Science and Technology Facilities Council
ISIS User Office: +44 (0) 1235 445592
GLOSSARY : SITE-MAP : ACCESSIBILITY : PRIVACY POLICY : ACCESS TO INFORMATION : TERMS OF USE : WEBMASTER
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

Scientist submits application for beamtime

Facility committee approves application

Facility registers, trains, and schedules scientist’s visit

Facility registers, trains, and schedules scientist’s visit

Scientists visits, facility run’s experiment

Raw data filtered and cleansed

Data analysis

Subsequent publication registered with facility

Scientific Computing Department

Metadata Repository
Discovery: Datacite Search
TopCat: Browse & Search
Data Preservation Infrastructure
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?