

# Towards the Preservation of the Scientific Memory

Brian Matthews

Shirley Crompton, Catherine Jones, Simon Lambert

Scientific Computing Department



Science & Technology  
Facilities Council

# STFC Facilities – driving scientific research

Neutron Sources



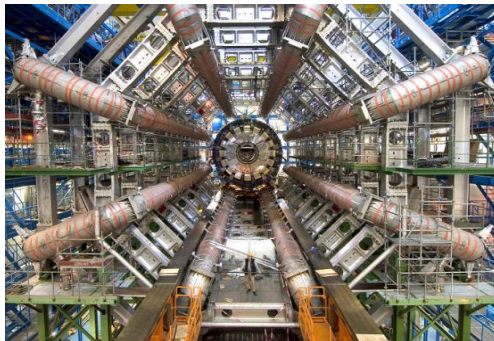
High Power Lasers



Light Sources



Particle Physics



Telescopes



# 10 Years of Curation Research at STFC

- “Curation Coalface Group”
- Claddier : JISC 2005-7
- A programme of projects:
  - CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval. 2006-2009.
  - SoftPres: Tools and Guidelines for Preserving and Accessing Software Research Outputs,, 2008-09
  - ACRID: Advanced Climate Research Infrastructure for Data, 2010-11
  - ODE: Opportunities for Data Exchange, 2010-12,
  - Mardi-Gros: 2011-12
  - SCAPE (Scaleable Preservation Environment) 2011-14
  - SCIDIP-ES (Science Data Infrastructure for Preservation – Earth Science) 2011-14
  - APARSEN, 2011-14
- At least 6 papers at DCC Conferences
- Time to take stock and see how it fits together

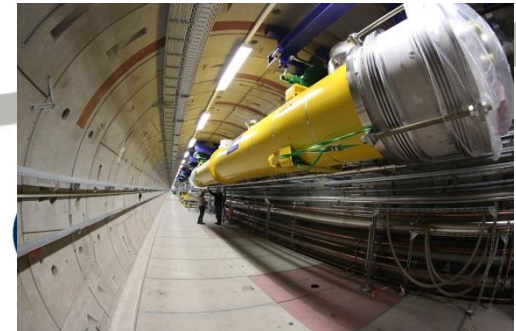


# Diamond Data Rates

- Ever rising data rates
  - Early 2007: Diamond first user.
    - No detector faster than ~10 MB/sec.
  - Early 2009:
    - first Pilatus 6M system @ 60 MB/s.
  - Early 2011:
    - first 25Hz Pilatus 6M system @ 150 MB/s.
  - Early 2013:
    - First 100 Hz Pilatus 6M system @ 600 MB/sec
  - 2015: Latest detectors such as Percival (6000 MB/sec)
- Doubling the data rates every 7.5 months.
- **Tomography**: Dealing with high data volumes
  - 200Gb/scan,
  - ~5 TB/day (one experiment at DLS)
- **MX**: smaller files, but a lot more experiments
- Took first Pb end 2013 (after 6 years of operation)
  - Now up to their 2<sup>nd</sup> Pb and into their 3<sup>rd</sup>
  - Diamond catalogue containing over 600 million files
  - Cataloguing 12000 Files per minute
- EU-XFEL 5000 frame/sec : ~ 50 GB/s



*PILATUS3 S and X product pages ...*



# A special case?

Number of Users shared between facilities																
	ALBA	BER II	DESY	DLS	ELETT RA	ESRF	FRM-II	ILL	ISIS	LLB	SINQ	SLS	SOLEIL	neutron	photon	all
ALBA	773	7	61	58	51	281	2	51	13	5	10	77	105	69	400	773
BER II	7	1563	115	46	27	179	157	383	198	98	191	62	36	580	329	1563
DESY	61	115	4197	137	222	851	116	255	113	62	95	315	188	469	1294	4197
DLS	58	46	137	4407	102	810	30	267	399	33	52	229	192	546	1130	4407
ELETT RA	51	27	222	102	3167	433	11	77	35	20	18	179	367	141	900	3167
ESRF	281	179	851	810	433	10287	139	900	369	190	174	963	1286	1313	3586	10287
FRM-II	2	157	116	30	11	139	1095	347	137	89	161	33	29	509	259	1095
ILL	51	383	255	267	77	900	347	4649	731	301	395	156	222	1518	1347	4649
ISIS	13	198	113	399	35	369	137	731	2880	89	233	94	56	936	745	2880
LLB	5	98	62	33	20	190	89	301	89	1235	74	39	151	391	323	1235
SINQ	10	191	95	52	18	174	161	395	233	74	1219	224	31	590	415	1219
SLS	77	62	315	229	179	963	33	156	94	39	224	3827	399	371	1470	3827
SOLEIL	105	36	188	192	367	1286	29	222	56	151	31	399	4568	394	1817	4568
neutron	69	1563	469	546	141	1313	1095	4649	2880	1235	1219	371	394	10023	2334	10023
photon	773	329	4197	4407	3167	10287	259	1347	745	323	415	3827	4568	2334	25336	25336
all	773	1563	4197	4407	3167	10287	1095	4649	2880	1235	1219	3827	4568	10023	25336	33025

<http://pan-data.eu/Users2012-Results>

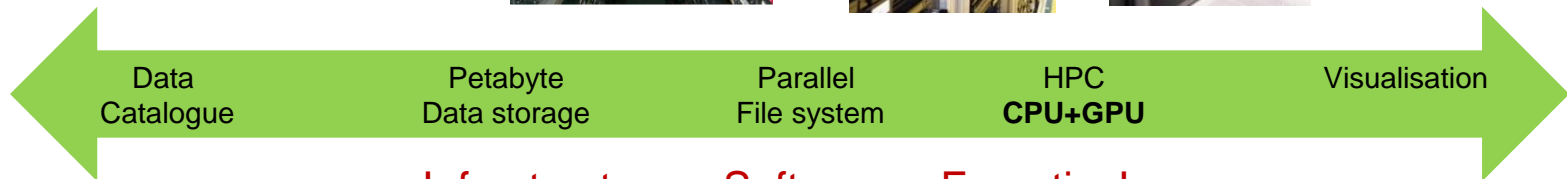
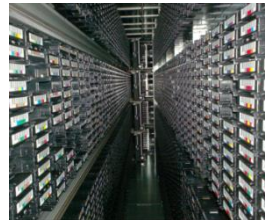
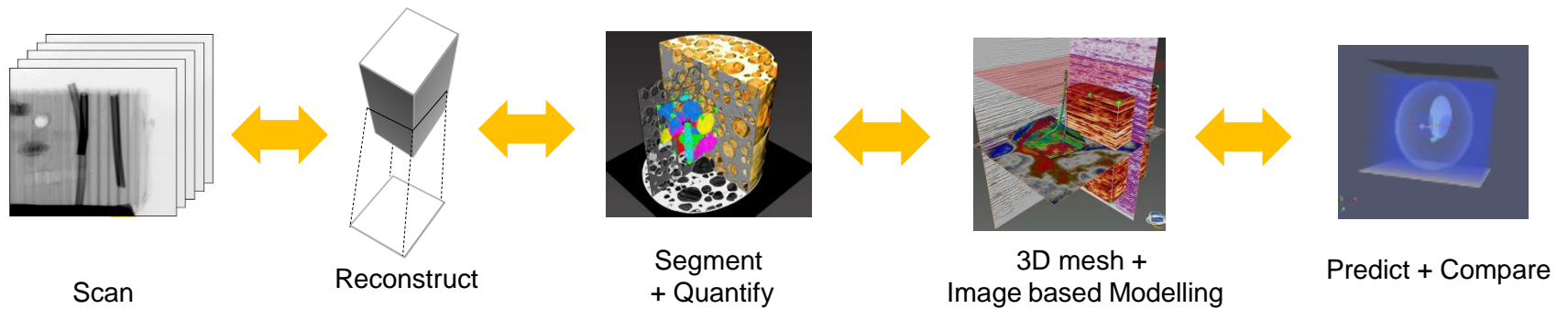
# Implications?

- Traditionally the ends up on users disks
- There is a capacity/capability problem
  - Users can't move the data
  - Users can't store the data
  - Users can't process the data
- Experiments combine data from different institutions
- The facility needs to provide more support
- We need to “sweat the assets”
  - Maximise the science extracted from funding
- This needs to be accessible to the user in universities



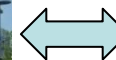
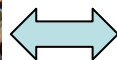


# Post-experimental support



**Infrastructure + Software + Expertise!**

**ISIS:IMAT**



**DLS:I12/I13**



# So what do we need to do?

- Store the data securely
  - Make it available to the right users
- Archive the data
  - Keep it safe for the “long” term
- Keep it usable
  - Maintain the context of the experiment
- Record how it is used
  - Record Provenance
- Record the Science undertaken
  - Data a means to an end





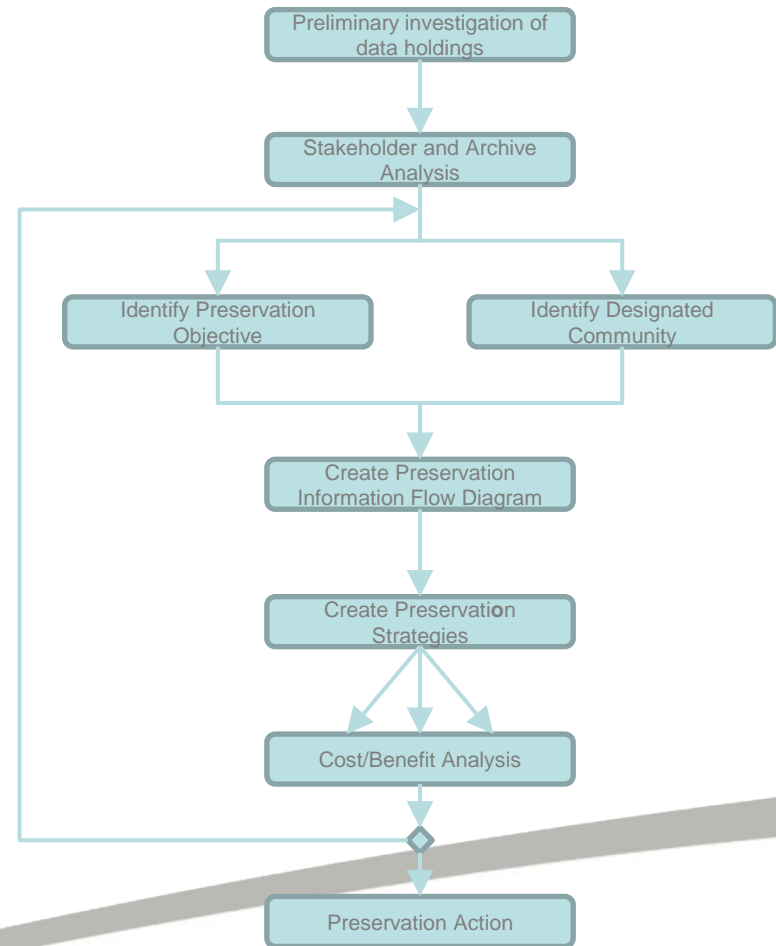
# The Challenges of Preserving the Scientific Memory

- Preservation Analysis
  - What to do with it
- Bit Preservation
  - How to maintain its integrity
- Cataloguing, access and publication
  - How to find and get it
- Preserving the science context
  - Knowing what the science meant
- Preserving provenance
  - Knowing what happened to it
- Preserving the science memory in a distributed environment
  - Knowing where it is



# Preservation Analysis

- Why preserve the data ?
  - Preservation Business Case:
  - Preservation Policy
  - Developing a Preservation Strategy
  - Preservation watch
- Progress on Policy
- Cost models much better understood
  - KRDS
- But we need to make the case to keep the data



# Benefits analysis

Factors which affect the benefits accrued from keeping data

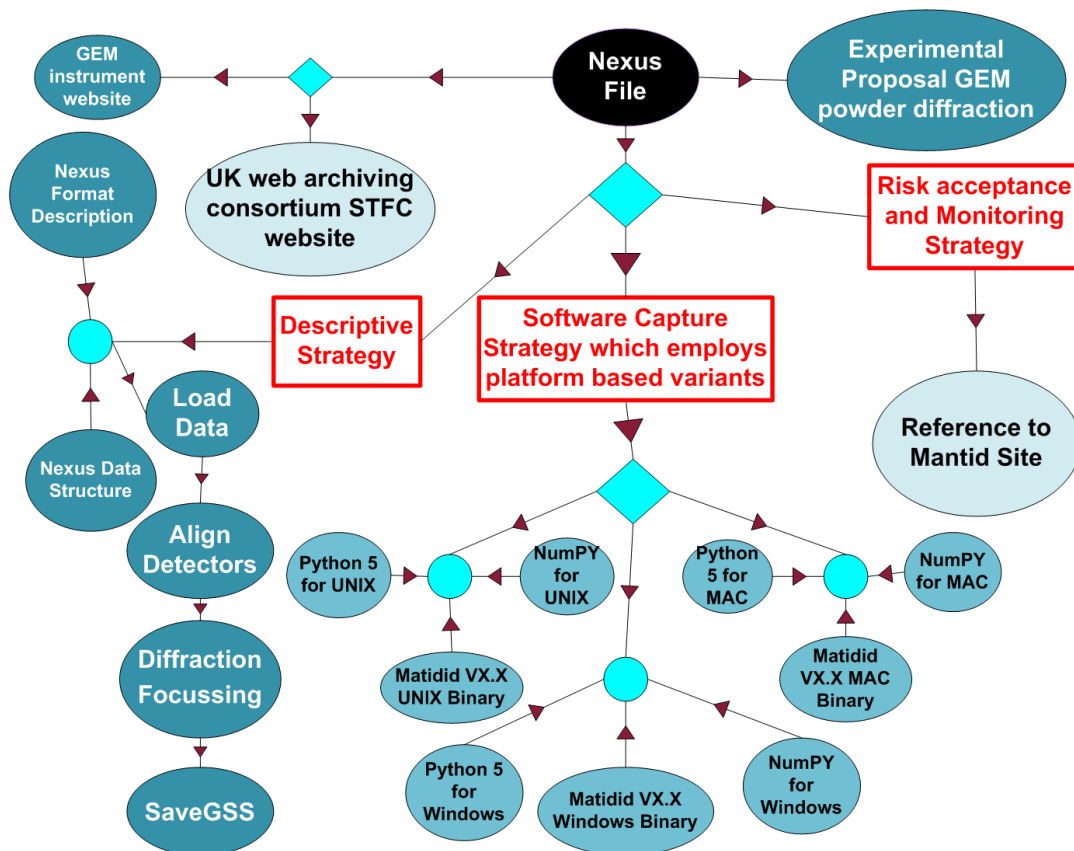
*Utility*  *Substitutability*

- Desirable
  - Is someone using it?
  - Are there measurable impacts ?
- Reusable
  - Is it kept in a state where it can be accessed , understood and reused ?
- Replaceable
  - Can I find an adequate substitute for the data elsewhere?
- Reproducible
  - Can the data be collected again? At what cost?



# Preservation Strategies

- Detailed analysis of the digital assets
  - Inventories
  - Designated community
  - Preservation dependencies
  - Risk analysis
  - Quality assurance
  - Migration and emulation
  - Preservation actions
- Preservation Network Models
  - Esther Conway
- Still needs development into practise



# Bit Preservation

- Storage management
- Replication:.
- Integrity checking:
- Media refresh:,
- “business as usual”
  - We have to do it
  - We know what to do



*Here's a copy of CCSDS 650.0. Its sane. Get on with it.  
Norman Gray*

Challenges of Scale

Challenges of resource control



Science & Technology  
Facilities Council

# Cataloguing, access and publication

- This is core to big science
  - Needed for operationally managing data
- Automated into the process
  - Metadata as middleware
- Levels of metadata
  - Discovery
  - Understanding
  - Usage





# DLS Archive Architecture

Data Acquisition

Data Storage

Data Access



Lustre file  
store

**StorageD  
Client**

**ICAT  
API**

**ICAT DB**  
Metadata  
Catalogue

*Metadata*

*Metadata*

DB  
Cache

**StorageD**  
Data (De-)  
aggregator,  
Metadata Ingestor

*Data*

*Data*

DB  
Cache

**CASTOR**  
Storage  
System



DB  
Cache

**TopCAT** Web  
frontend

Cache

**ICAT Data Service**

*Retrieved data*

DB

**FUSE** Data  
browser

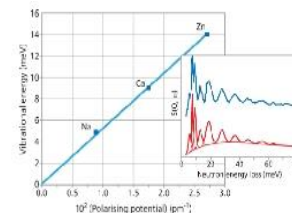
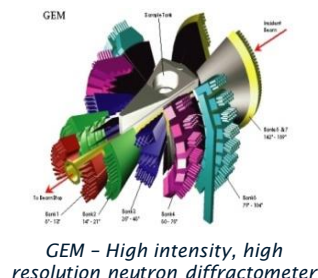
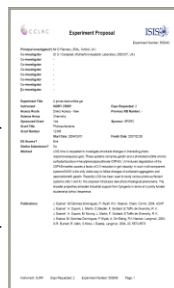


Science & Technology  
Facilities Council

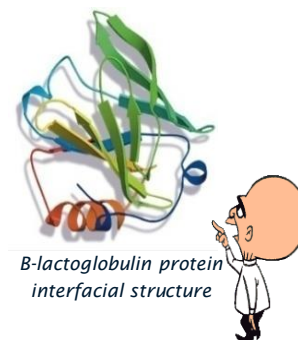
## Central Facility

- Secure access to user's data
- Flexible data searching
- Scalable and extensible architecture
- Integration with analysis tools
- Access to high-performance resources
- Linking to other scientific outputs
- Data policy aware

Example ISIS Proposal



H2-(zeolite) vibrational frequencies vs polarising potential of cations



### Proposals

Once awarded beamtime at ISIS, an entry will be created in ICAT that describes your proposed experiment.

### Experiment

Data collected from your experiment will be indexed by ICAT (with additional experimental conditions) and made available to your experimental team

### Analysed Data

You will have the capability to upload any desired analysed data and associate it with your experiments.

### Publication

Using ICAT you will also be able to associate publications to your experiment and even reference data from your publications.

# DOI Data Access Process

PHYSICAL REVIEW B 84, 075219 (2011)

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

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

Add-ons Gallery - Web Slice Suggested Sites Toshiba Places Web Slice Gallery

http://search.datacite.o...

**Metadata**

DataCite

Filter

- allocator
- datacentre
- prefix
- resourceType
- contributor
- creator
- publicationYear
- publisher
- language
- refQuality
- has\_metadata

**About STFC**

How we operate

**Business & Innovation**

Collaborate with STFC

**ISIS Data**

Investigation

DOI: 10.5288/

Date of Expe

Publisher: S

Data format:

Select the data for

Data Citatic

The recomm

[author], [da

For Example

Griffin. et al;

**Science & Technology Facilities Council**

Browse All Data

Download

- ISIS
  - ALF
  - ARGUS
  - CRISP
  - EMU
  - ENGIX
  - EVS
  - GEM
    - cycle\_11\_4
    - cycle\_11\_3
      - BaRuO3 8mm pos 8(id:CAL\_GEM\_2011-10-31T09:01:37)
        - GEM56174.raw
          - GEM56174.log
          - GEM56174\_ICPdebug.txt
          - GEM56174\_ICPevent.txt
          - GEM56174\_ICPstatus.txt
          - GEM56174\_Status.txt
        - GEM56175.raw
        - GEM56176.raw
        - GEM56174.nxs
        - GEM56175.nxs
        - GEM56176.nxs
      - Empty 6mm can 620(id:CAL\_GEM\_2011-10-10T17:23:28)

**RESEARCH COUNCILS UK**

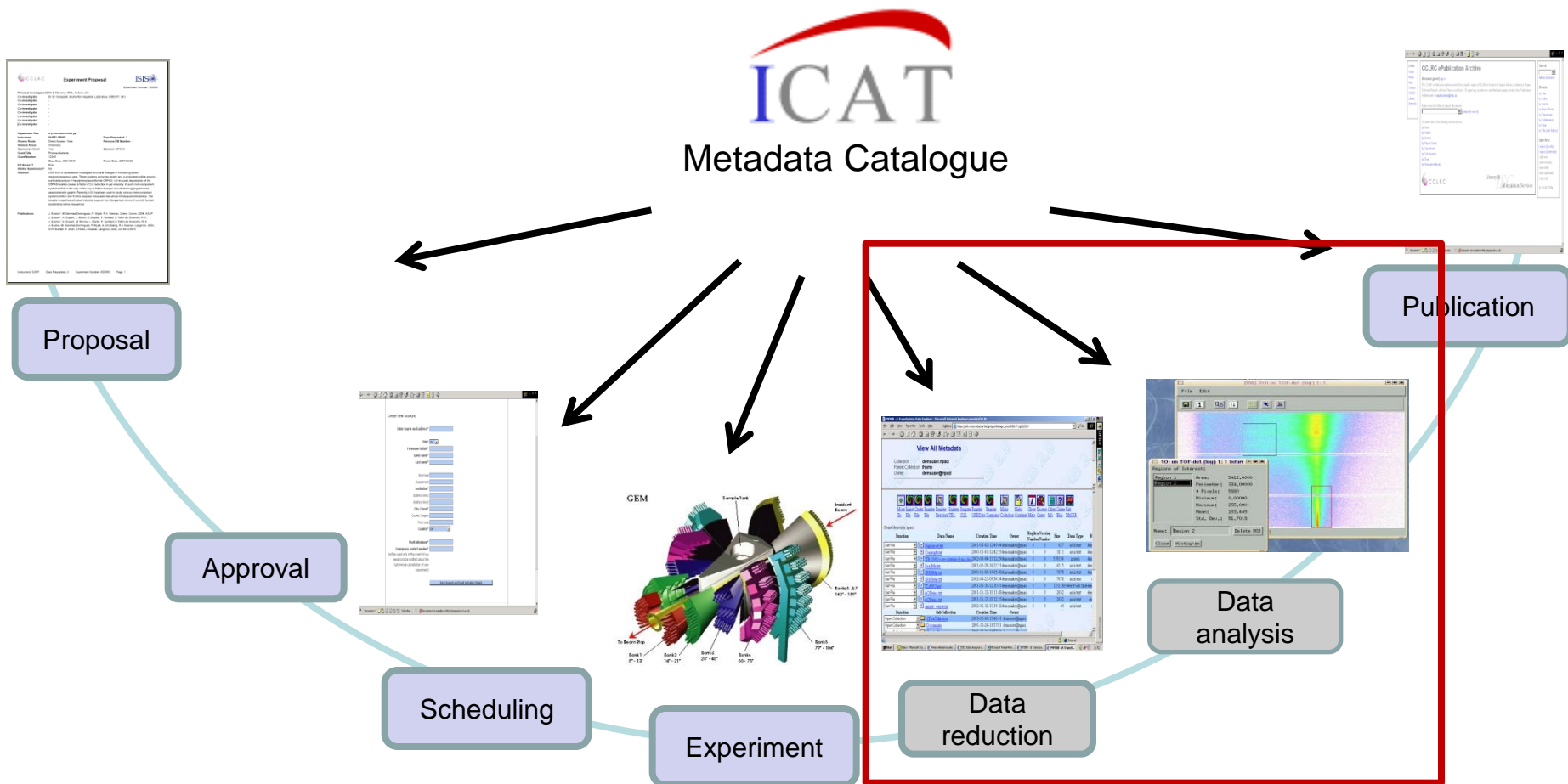
GLOSSARY : SITE-MAP : ACC

# Science context and provenance

- We need to preserve the understanding of the science
  - Information about instruments, sensors, samples, data sampling conditions, parameters measured, coverage, units and data rates.
  - Information on intention, methodology, and actors
  - Information on the data collection environment
  - Calibration information on the instruments, with errors and tolerances
- *Tacit knowledge* concerning the science,
- And how the data is processed to generate conclusions
  - The relationships between artefacts used in the scientific process.
  - Different types of digital artefacts, :
    - data, software, visualisation, documents, and workflows



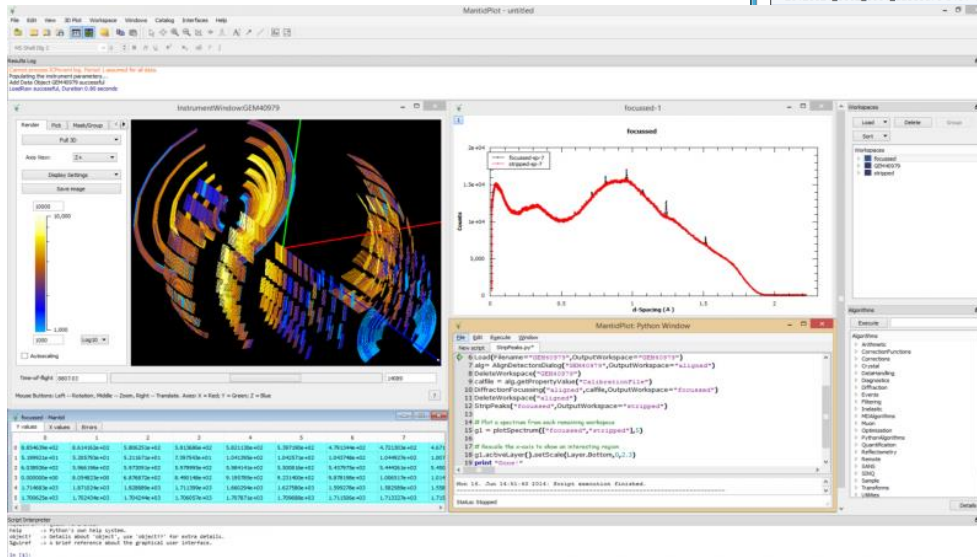
# Facility Data Lifecycle



Traditionally, these steps are decoupled from facilities. However, they are key to derive useful insights.

# Frameworks to capture provenance

Mantid



The screenshot shows the ICAT Job Portal web interface. At the top, there's a search bar and filters for project, user, instrument, experiment type, and number of channels. Below the filters, a table lists 7 datasets found. A dropdown menu is open over the table, showing options like 'Options...', 'Download', 'Show Download URL', and 'MSMM Viewer Project'. The table has columns for Name, Description, and Users.

Name	Description	Users
20120524_0002_0001_632c1ef9-9f32-4a39-a649-855ed592c27	coloc 3 Affibody 639 nm laser	
20120525_0004_0001_0bb36de-dd79-4c13-84ca-72a6a86de334	T47D 3 Affibody 639 nm laser	
20120524_0002_0001_e421cec3-d7eb-4e3f-baea-bf66fed31688	coloc 3 Affibody 639 nm laser	
20120525_0004_0001_e628e0b5-f699-45a4-93d7-61a952a35912	T47D 3 Affibody 639 nm laser	
20120524_0002_0001_c1b3dc65-0f05-4daf-be3f-e936291f812e	T47D 3 Affibody 639 nm laser	
20120524_0002_0001_da8e9d70-b461-4d0f-9e06-b3267809ed1d	T47D 3 Affibody 639 nm laser	
20120524_0002_0001_a599-6e62eb4f829e	T47D 3 Affibody 639 nm laser	

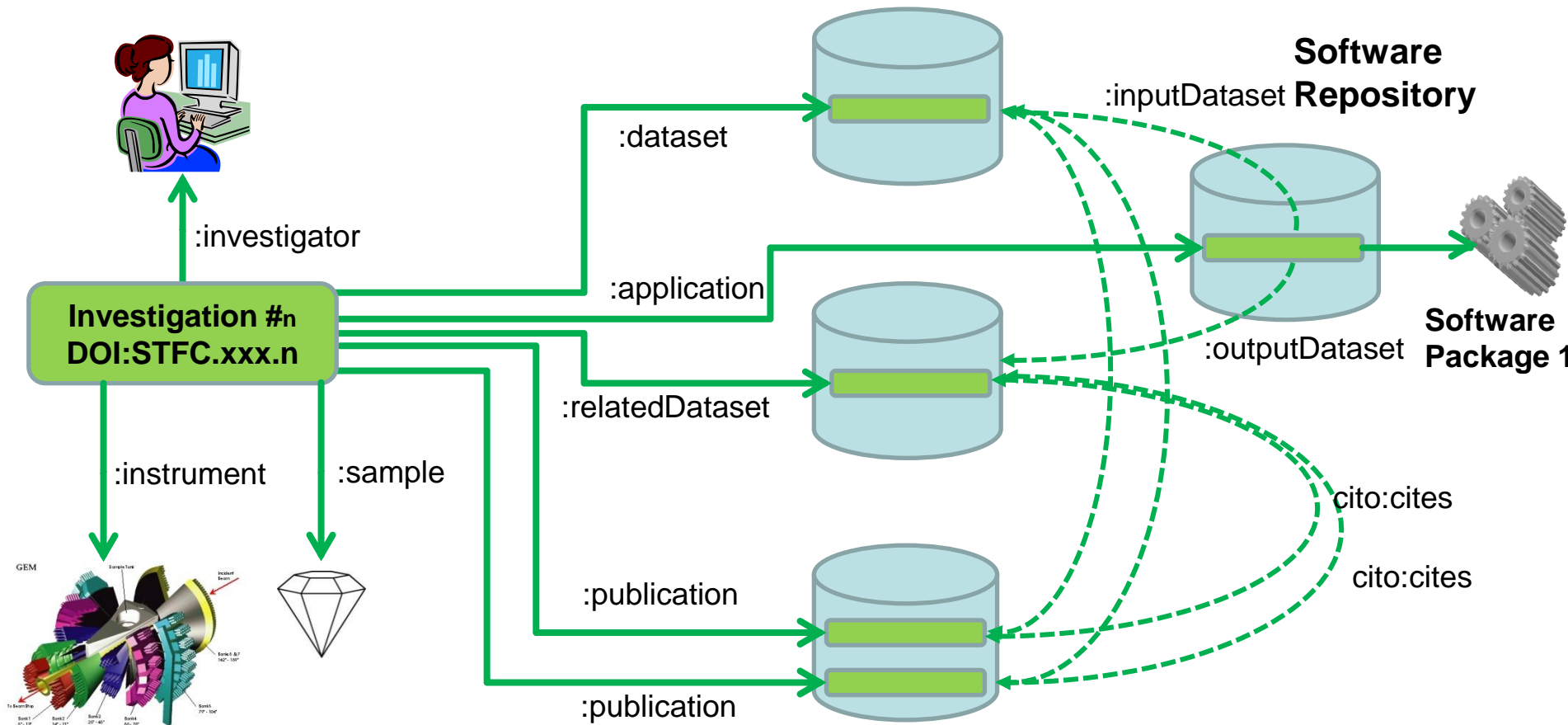
ICAT Job Portal



Science & Technology  
Facilities Council




# Investigation Research Objects: Record Experiments not Data



- Own metadata format (CSMD)
- OAI-ORE
- W3C Prov ontology
- Assume that the software is in a repository

# Data Journal Mock up

**Science & Technology  
Facilities Council**

**ISIS**

Cycles

Investigation

Edit

Archived Versions

Previous Investigation

Next Investigation

**Investigation title:** Reversible B-H Bo

**Release Date:** 27-07-2013

**Creator:** Dr Amber Thompson

**DOI:** 10.5286/ISIS.E.24079414

**Date of Experiment:** 30-06-2010 - 27-07-20

**Facility:** ISIS Pulsed Neutron & Muon Source

**Publisher:** ISIS Data Journal, STFC

**Data format:** RAW/Nexus


Select the data format above to find out more

**Data Citation**

The recommended format for citing this data is [author], [date], [title], [publisher], [doi]

For Example:

Dr Amber Thompson; (2010): Reversible B-H Bo  
doi:10.5286/ISIS.E.24079414

**Science & Technology  
Facilities Council**

**PROTOTYPE of the ISIS Data Journal**  
The archive for ISIS research data

ISIS

Cycles

Investigation

Edit

Archived Versions

Edit

RB1010274

Techniques

Please select the techniques used

Technique Name

☒ Neutron Diffraction

☐ Single Crystal Diffraction

Links to External Resources

► Click here for help with Links to External Resources

Relationship Type	External Resource's Name (displayed on landing page)	External Resource's Type	External Resource's URL/DOI	Note (optional)	
uses method in	Mantid 3.2	computer application	doi:10.5286/software/mantid3.2		Remove
provides data for	Experimental Crystal Structure Determination	model	doi:10.5517/ccykjnn	Structure	Remove
is cited as data source by	CY Tang; N Phillips; JJ Bates; AL Thompson; M J Gutmann;	journal article	doi:10.1039/C2CC33361A		Remove
is cited as data source by	CY Tang; N Phillips; JJ Bates; AL Thompson; M J Gutmann; ;	journal article	http://publ.org/net/epubs/work/63093	this is the ePubs version	Remove
obtains support from	EP/F019181/1. Small molecule functionalization by metal-m	grant application	http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/		Remove

Add new blank row

Save

Cancel

Contact Us

Cookies/Privacy

Terms & conditions

Cymraeg






FOI

Copyright

Glossary

Sitemap

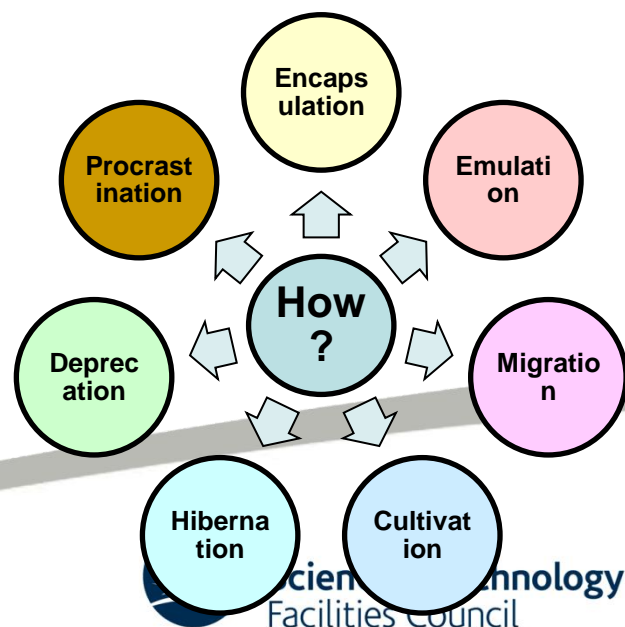
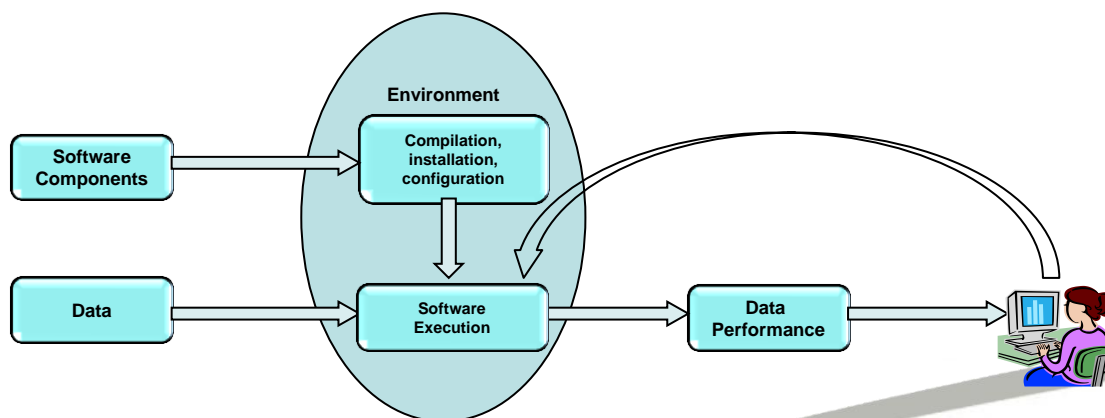
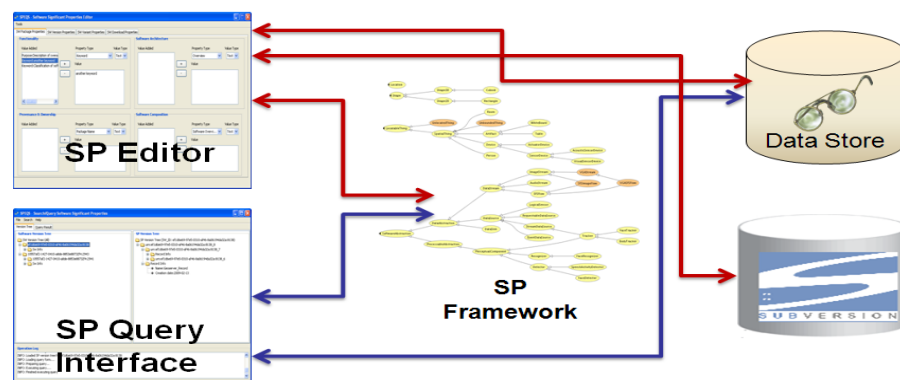
Accessibility





# Software preservation

- We need to tackle the issue of preserving s/w
- Exploratory study
  - A framework for software preservation
- SSI – practical steps
- Combines with s/w engineering approaches
- DOIs for software.



# Tacit Knowledge

- Capturing the human knowledge associated with science.
  - Blogs,
  - Electronic notebooks
  - Open science
  - Social media
- Business knowledge management
  - Communities of practice
  - After action review
  - Storytelling



<http://www.multicians.org/>



Science & Technology  
Facilities Council

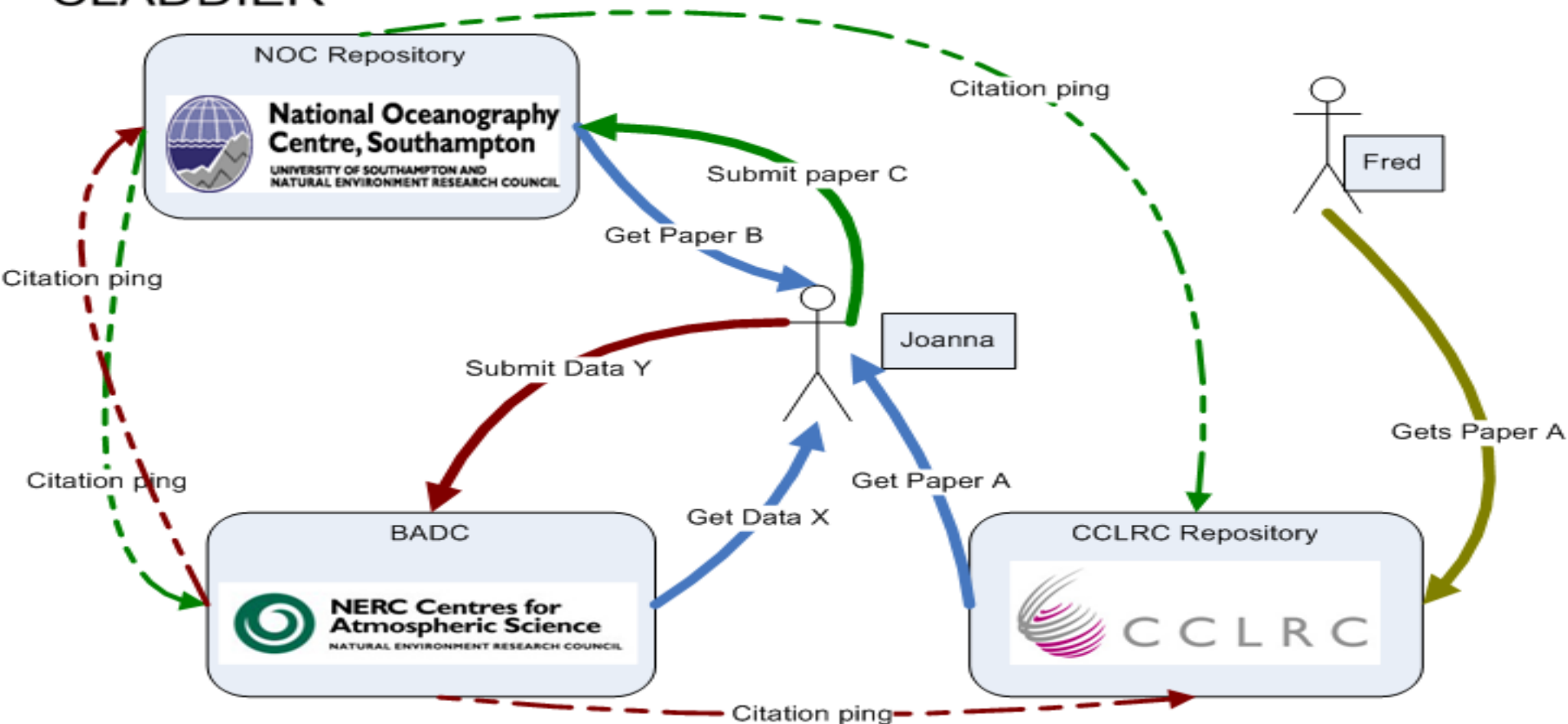
# Preserving the science memory in a distributed environment

- Research artefacts in different locations,
  - copies and versions in different places.
- Maintaining a link structure across repositories
  - under different jurisdictions
  - Different IPR and business models..
- Managing the trust relationships
  - guarantees on stability and quality.
- Attribution and rights management
  - credit can be properly assigned





# Citation, Location and Deposition in Discipline and Institutional Repositories



1  
Joanna gets data X and papers A and B for her research.

2  
Joanna submits a paper C to NOC. The repository automatically checks and notifies the cited repositories with a "citation ping"

3  
Joanna submits data Y to BADC. The data archive automatically checks and notifies the cited repositories with a "citation ping"

4  
Fred gets the paper A from CCLRC. The paper "knows" it is cited in paper C and data Y.



ology

# What's changed ?

- The “data deluge” has become true
  - Synchrotron and climate data will match LHC
- Sharing and publishing data recognised
  - High level data policy
  - Datacite
- Systematic data management become “standard”
  - Particularly in “big science”
  - But need to link to “bench science”



# Outstanding challenges

- Scaling
  - Vs of big data
- Better cases for preserving data
  - Especially benefits of preservation
- User-oriented preservation infrastructures
  - Based around linked data for dependencies
- Systematic collection of context and provenance
  - Automation
  - Research Objects
- Software preservation
- Preservation of tacit knowledge



# Preserving the scientific memory

Shift the focus to preserving the Science

[brian.matthews@stfc.ac.uk](mailto:brian.matthews@stfc.ac.uk)

[www.stfc.ac.uk/scd](http://www.stfc.ac.uk/scd)



Science & Technology  
Facilities Council