Towards the Preservation of the Scientific Memory

Brian Matthews

Shirley Crompton, Catherine Jones, Simon Lambert

Scientific Computing Department
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:
  - 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\textsuperscript{nd} Pb and into their 3\textsuperscript{rd}
  - Diamond catalogue containing over 600 million files
  - Cataloguing 12000 Files per minute

- **EU-XFEL** 5000 frame/sec : ~ 50 GB/s
## A special case?

### Number of Users shared between facilities

<table>
<thead>
<tr>
<th></th>
<th>ALBA</th>
<th>BER II</th>
<th>DESY</th>
<th>DLS</th>
<th>ELETTRA</th>
<th>ESRF</th>
<th>FRM-II</th>
<th>ILL</th>
<th>ISIS</th>
<th>LLB</th>
<th>SINQ</th>
<th>SLS</th>
<th>SOLEIL</th>
<th>neutron</th>
<th>photon</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBA</td>
<td>773</td>
<td>7</td>
<td>61</td>
<td>58</td>
<td>51</td>
<td>281</td>
<td>2</td>
<td>51</td>
<td>13</td>
<td>5</td>
<td>10</td>
<td>77</td>
<td>105</td>
<td>69</td>
<td>400</td>
<td>773</td>
</tr>
<tr>
<td>BER II</td>
<td>7</td>
<td>1563</td>
<td>115</td>
<td>46</td>
<td>27</td>
<td>179</td>
<td>157</td>
<td>383</td>
<td>198</td>
<td>98</td>
<td>191</td>
<td>62</td>
<td>36</td>
<td>580</td>
<td>329</td>
<td>1563</td>
</tr>
<tr>
<td>DESY</td>
<td>61</td>
<td>115</td>
<td>4197</td>
<td>137</td>
<td>222</td>
<td>851</td>
<td>116</td>
<td>255</td>
<td>113</td>
<td>62</td>
<td>95</td>
<td>315</td>
<td>188</td>
<td>469</td>
<td>1294</td>
<td>4197</td>
</tr>
<tr>
<td>DLS</td>
<td>58</td>
<td>46</td>
<td>137</td>
<td>4407</td>
<td>102</td>
<td>810</td>
<td>30</td>
<td>267</td>
<td>399</td>
<td>33</td>
<td>52</td>
<td>229</td>
<td>192</td>
<td>546</td>
<td>1130</td>
<td>4407</td>
</tr>
<tr>
<td>ELETTRA</td>
<td>51</td>
<td>27</td>
<td>222</td>
<td>102</td>
<td>3167</td>
<td>433</td>
<td>11</td>
<td>77</td>
<td>35</td>
<td>20</td>
<td>18</td>
<td>179</td>
<td>367</td>
<td>141</td>
<td>900</td>
<td>3167</td>
</tr>
<tr>
<td>ESRF</td>
<td>281</td>
<td>179</td>
<td>851</td>
<td>810</td>
<td>433</td>
<td>10287</td>
<td>139</td>
<td>900</td>
<td>369</td>
<td>190</td>
<td>174</td>
<td>963</td>
<td>1286</td>
<td>1313</td>
<td>3586</td>
<td>10287</td>
</tr>
<tr>
<td>FRM-II</td>
<td>2</td>
<td>157</td>
<td>116</td>
<td>30</td>
<td>11</td>
<td>139</td>
<td>1095</td>
<td>347</td>
<td>137</td>
<td>89</td>
<td>161</td>
<td>33</td>
<td>29</td>
<td>509</td>
<td>259</td>
<td>1095</td>
</tr>
<tr>
<td>ILL</td>
<td>51</td>
<td>383</td>
<td>255</td>
<td>267</td>
<td>77</td>
<td>900</td>
<td>347</td>
<td>4649</td>
<td>731</td>
<td>301</td>
<td>395</td>
<td>156</td>
<td>222</td>
<td>1518</td>
<td>1347</td>
<td>4649</td>
</tr>
<tr>
<td>ISIS</td>
<td>13</td>
<td>198</td>
<td>113</td>
<td>399</td>
<td>35</td>
<td>369</td>
<td>137</td>
<td>731</td>
<td>355</td>
<td>180</td>
<td>2880</td>
<td>89</td>
<td>233</td>
<td>94</td>
<td>56</td>
<td>745</td>
</tr>
<tr>
<td>LLB</td>
<td>5</td>
<td>98</td>
<td>62</td>
<td>33</td>
<td>20</td>
<td>190</td>
<td>89</td>
<td>301</td>
<td>1235</td>
<td>74</td>
<td>39</td>
<td>151</td>
<td>391</td>
<td>323</td>
<td>1235</td>
<td></td>
</tr>
<tr>
<td>SINOQ</td>
<td>10</td>
<td>191</td>
<td>95</td>
<td>52</td>
<td>18</td>
<td>174</td>
<td>161</td>
<td>395</td>
<td>233</td>
<td>74</td>
<td>1219</td>
<td>224</td>
<td>31</td>
<td>590</td>
<td>415</td>
<td>1219</td>
</tr>
<tr>
<td>SLS</td>
<td>77</td>
<td>62</td>
<td>315</td>
<td>229</td>
<td>179</td>
<td>963</td>
<td>33</td>
<td>156</td>
<td>94</td>
<td>39</td>
<td>224</td>
<td>3827</td>
<td>399</td>
<td>371</td>
<td>1470</td>
<td>3827</td>
</tr>
<tr>
<td>SOLEIL</td>
<td>105</td>
<td>36</td>
<td>188</td>
<td>192</td>
<td>367</td>
<td>1286</td>
<td>29</td>
<td>222</td>
<td>56</td>
<td>151</td>
<td>31</td>
<td>399</td>
<td>4568</td>
<td>394</td>
<td>1817</td>
<td>4568</td>
</tr>
</tbody>
</table>

### neutron

| neutron | 69     | 1563   | 469    | 546    | 141    | 1313   | 1095   | 4649   | 2880   | 1235   | 1219   | 371    | 394    | 10023  | 2334   | 10023  |

### photon

| photon  | 773    | 329    | 4197   | 4407   | 3167   | 10287  | 259    | 1347   | 745    | 323    | 415    | 3827   | 4568   | 2334   | 25336  | 25336  |

### all

| all     | 773    | 1563   | 4197   | 4407   | 3167   | 10287  | 1095   | 4649   | 2880   | 1235   | 1219   | 3827   | 4568   | 10023  | 25336  | 33025  |

[^1]: [http://pan-data.eu/Users2012-Results](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

Scan → Reconstruct → Segment + Quantify → 3D mesh + Image based Modelling → Predict + Compare

ICAT

Data Catalogue → Petabyte Data storage → Parallel File system → HPC CPU+GPU → Visualisation

Infrastructure + Software + Expertise!

ISIS:IMAT → DLS:I12/I13

Some image credit: Avizo, Visualization Sciences Group (VSG)
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

*Science & Technology Facilities Council*
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 practice
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
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

StorageD Client
Lustre file store

Data Storage

ICAT API
ICAT DB Metadata Catalogue
Metadata

TopCAT Web frontend
ICAT Data Service
Retrieved data

CASTOR Storage System
FUSE Data browser

Data Access

StorageD
Data (De-)
aggregator,
Metadata Ingestor

DB
Cache
Data

DB
Cache
Data

DB
Cache
Data

DB
Cache
Data
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.

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
DOI Data Access Process

About STFC
How we operate
Collaborate with STFC

ISIS Data

Data collected on the CRISP instrument at the ISIS facility

Investigation
DOI: 10.5286
Date of Exp:
Publisher: S
Data format:
Data Citatic

The recomm [author], [da
For Example
Griffin, et al

Science & Technology Facilities Council

Glossary : Site-Map : Acc

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

http://www.icatproject.org
Frameworks to capture provenance

Mantid

ICAT Job Portal
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
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

72 Multics Stories
http://www.multicians.org/
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
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.
Linking ISIS data

ISIS Beamtime Application: SXD Round: 2010 1
RB1010274 Reversible B-H Bond Activation at Catonic Rh(III) Centres: Structural Characterization of Key Hydrogen-containing Intermediates
PI: Aldridge Dr S simon.aldridge@chem.ox.ac.uk
University of Oxford
Department of Chemistry

Isis Data: persistent identifiers
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

www.stfc.ac.uk/scd