Towards an open data infrastructure for photon and neutron facilities

Brian Matthews and Juan Bicarregui

Scientific Computing Department (SCD)

Rutherford Appleton Laboratory (RAL)
Science and Technology Facilities Council (STFC), U. K.

brian.matthews@stfc.ac.uk
Science and Technology Facilities Council

• Provide large-scale scientific facilities for UK Science
  – particularly in physics and astronomy
  – ISIS and Diamond Light Source facilities

• Scientific Computing Department
  – Provides advanced IT development and services to the STFC Science Programme
  – Strong role in management of our science data
The computing centre at Rutherford Appleton Laboratory

The computing centre at RAL houses 20,000 computer processors and stores 10,000,000,000 Mbytes of data available on-line.

The centre is used for simulation and data analysis by researchers in all scientific disciplines from throughout the UK and their international collaborators.
STFC Rutherford Appleton Laboratory
Doing Science at Source Facilities

Neutrons and photons
Provide complementary views of matter:

Photons “see” electric charge – high atomic number nuclei

Neutrons “see” nucleons – especially hydrogen atoms

Visit facility on research campus

Place sample in beam

Diffraction pattern from sample

Fitting experimental data to model

Structure of cholesterol in crude oil
Facilities Science

- ~30,000 user visitors each year in Europe:
  - physics, chemistry, biology, medicine,
  - energy, environmental, materials, culture
  - pharmaceuticals, petrochemicals, microelectronics

- **Big Facilities for Small Science**

- Billions of € of investment
c. £400M for DLS+
running costs

- Over 5,000 high impact publications per year in Europe

- Each facility has dedicated data and computing infrastructure
  - But so far no integrated data repositories across facilities
  - Lacking sustainability & traceability

---

Longitudinal strain in aircraft wing

Bioactive glass for bone growth

Hydrogen storage for zero emission vehicles

Magnetic moments in electronic storage
## A Common Community

### Number of Users shared between facilities

<table>
<thead>
<tr>
<th>BER II</th>
<th>BESSY II</th>
<th>DESY</th>
<th>DLS</th>
<th>ELETTRA</th>
<th>ESRF</th>
<th>ILL</th>
<th>ISIS</th>
<th>LLB</th>
<th>SINQ</th>
<th>SLS</th>
<th>SOLEIL</th>
<th>FRM-II</th>
<th>ANKA</th>
<th>neutron</th>
<th>photon</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>850</td>
<td>80</td>
<td>68</td>
<td>25</td>
<td>18</td>
<td>128</td>
<td>261</td>
<td>141</td>
<td>67</td>
<td>76</td>
<td>12</td>
<td>14</td>
<td>111</td>
<td>5</td>
<td>375</td>
<td>244</td>
</tr>
<tr>
<td>BER II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BESSY II</td>
<td>80</td>
<td>2306</td>
<td>238</td>
<td>45</td>
<td>134</td>
<td>399</td>
<td>67</td>
<td>33</td>
<td>31</td>
<td>26</td>
<td>149</td>
<td>93</td>
<td>42</td>
<td>31</td>
<td>175</td>
<td>758</td>
</tr>
<tr>
<td>DESY</td>
<td>80</td>
<td>2306</td>
<td>238</td>
<td>45</td>
<td>134</td>
<td>399</td>
<td>67</td>
<td>33</td>
<td>31</td>
<td>26</td>
<td>149</td>
<td>93</td>
<td>42</td>
<td>31</td>
<td>175</td>
<td>758</td>
</tr>
<tr>
<td>DLS</td>
<td>25</td>
<td>45</td>
<td>88</td>
<td>3494</td>
<td>72</td>
<td>739</td>
<td>213</td>
<td>336</td>
<td>35</td>
<td>18</td>
<td>145</td>
<td>149</td>
<td>20</td>
<td>12</td>
<td>441</td>
<td>967</td>
</tr>
<tr>
<td>ELETTRA</td>
<td>18</td>
<td>134</td>
<td>121</td>
<td>72</td>
<td>2731</td>
<td>455</td>
<td>85</td>
<td>43</td>
<td>23</td>
<td>4</td>
<td>66</td>
<td>316</td>
<td>9</td>
<td>20</td>
<td>145</td>
<td>839</td>
</tr>
<tr>
<td>ESRF</td>
<td>128</td>
<td>399</td>
<td>735</td>
<td>739</td>
<td>455</td>
<td>10728</td>
<td>886</td>
<td>406</td>
<td>92</td>
<td>600</td>
<td>1069</td>
<td>144</td>
<td>80</td>
<td>1303</td>
<td>3256</td>
<td></td>
</tr>
<tr>
<td>ILL</td>
<td>261</td>
<td>67</td>
<td>194</td>
<td>213</td>
<td>85</td>
<td>886</td>
<td>4338</td>
<td>741</td>
<td>343</td>
<td>229</td>
<td>69</td>
<td>176</td>
<td>349</td>
<td>10</td>
<td>1450</td>
<td>1246</td>
</tr>
<tr>
<td>ISIS</td>
<td>141</td>
<td>33</td>
<td>91</td>
<td>336</td>
<td>43</td>
<td>406</td>
<td>741</td>
<td>2755</td>
<td>120</td>
<td>119</td>
<td>43</td>
<td>52</td>
<td>155</td>
<td>5</td>
<td>908</td>
<td></td>
</tr>
<tr>
<td>LLB</td>
<td>67</td>
<td>31</td>
<td>55</td>
<td>35</td>
<td>23</td>
<td>235</td>
<td>343</td>
<td>120</td>
<td>1348</td>
<td>34</td>
<td>12</td>
<td>131</td>
<td>92</td>
<td>3</td>
<td>425</td>
<td>359</td>
</tr>
<tr>
<td>SINQ</td>
<td>76</td>
<td>26</td>
<td>44</td>
<td>18</td>
<td>4</td>
<td>92</td>
<td>229</td>
<td>119</td>
<td>34</td>
<td>726</td>
<td>96</td>
<td>9</td>
<td>97</td>
<td>0</td>
<td>334</td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>12</td>
<td>149</td>
<td>155</td>
<td>145</td>
<td>66</td>
<td>600</td>
<td>69</td>
<td>43</td>
<td>12</td>
<td>96</td>
<td>2424</td>
<td>182</td>
<td>18</td>
<td>18</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>SOLEIL</td>
<td>14</td>
<td>93</td>
<td>130</td>
<td>149</td>
<td>316</td>
<td>1069</td>
<td>176</td>
<td>52</td>
<td>131</td>
<td>9</td>
<td>182</td>
<td>3656</td>
<td>14</td>
<td>26</td>
<td>299</td>
<td>1460</td>
</tr>
<tr>
<td>FRM-II</td>
<td>111</td>
<td>42</td>
<td>103</td>
<td>20</td>
<td>9</td>
<td>144</td>
<td>349</td>
<td>155</td>
<td>92</td>
<td>97</td>
<td>18</td>
<td>14</td>
<td>1087</td>
<td>5</td>
<td>494</td>
<td>255</td>
</tr>
<tr>
<td>ANKA</td>
<td>5</td>
<td>31</td>
<td>43</td>
<td>12</td>
<td>20</td>
<td>80</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>18</td>
<td>26</td>
<td>5</td>
<td>452</td>
<td>19</td>
<td>144</td>
</tr>
</tbody>
</table>

**Details of how we count users:** [http://wiki.pan-data.eu/CountingUsers](http://wiki.pan-data.eu/CountingUsers)
## PaN-Data facilities users

<table>
<thead>
<tr>
<th>Total number of unique users:</th>
<th>30873</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using only Neutrons:</td>
<td>6719  or 21.7% of all unique users</td>
</tr>
<tr>
<td>Using only Photons:</td>
<td>22021 or 71.3% of all unique users</td>
</tr>
<tr>
<td>Using Neutrons and Photons:</td>
<td>2133  or 6.9% of all unique users</td>
</tr>
<tr>
<td>Using more than one facility:</td>
<td>6863  or 22.2% of all users</td>
</tr>
<tr>
<td>Using more than one Photon source:</td>
<td>4252  or 17.6% of all photon users</td>
</tr>
<tr>
<td>Using more than one Neutron source:</td>
<td>1734  or 19.6% of all neutron users</td>
</tr>
</tbody>
</table>

*All* facilities have users in common with all other facilities, also across neutron and photon sources.

Typically, **30-40%** of the users of any of the photon or any of the neutron sources also use at least one other facility.

**Benefit to be gained to the user community by coordinating the computing infrastructure**
PaN-data ODI – an Open Data Infrastructure for European Photon and Neutron laboratories

**Federated data catalogues** supporting cross-facility, cross-discipline interaction at the scale of atoms and molecules

Provide common tools and user experience across facilities

- Unification of data management policies
- Shared protocols for exchange of user information
- Common scientific data formats
- Interoperation of data analysis software
- **Data Provenance: Linking Data and Publications**
- Digital Preservation: supporting the long-term preservation of the research outputs
Sharing and combining data

• Users move around to collect different views on their samples
  – Different instruments have different characteristics
• Combining data can give additional insights
• Needs
  – Common data formats
  – Common data access
  – Common metadata
• Context
  – Providing better provenance information
These are with users.
Traditionally, these, although very useful for data citation, reuse and sharing, are very difficult to capture!
Practices vary from individuals to individuals, and from institutions to institutions.
Managing the Data Continuum

• Provide better support for the data continuum
  – Recording of provenance
  – Linking raw and derived data, publications and software to the experiment

• Improve service to the user
  – Accurate record keeping of science results and context
  – Validation of science results
  – Publishing and reuse of “research objects”
  – Linking to other research objects

• However, capturing provenance of analysed data is hard and expensive
  – Lots of variation and input from users
  – Lots of blind alleys and retracing of steps
  – Mostly undertaken in the user’s institution

• So why bother?
  – Good use cases where managing provenance well gives benefits.
Smart Research Framework: Automated Data Processing Pipeline for ISIS

SampleTracks

OpenGenie Script

Data Acquisition

Raw data

Mantid

Data Reduction

Reduced data

Model Fitting

Analysed data

ICAT Data Catalogue

Blog Posts in LabTrove

Cameron Neylon (ISIS) and Erica Yang
What does this mean?

• We can automate the stages of the process “around the experiment”
  – Setting up the experiment
  – Auto-generating configuration and control scripts
  – Initial reduction of the data

• Capture and link (in ICAT)
  – Sample information
  – Experiment configuration and Control
  – Raw and reduced data,
  – Reduction software

• Auto-publish in the Blog the record (with links)
  – Accurate record of metadata for the user to refer to
  – Can be shared with research group or more widely
  – Auto-data publication.

• Would expect to be able to transfer this to “Express Services”
  – A complete data package for the user.
Tomography - Reconstruction

- I13: high throughput tomography beamline
- Computationally heavy process
  - Up to 120 GBs/file every 30 minutes
  - 6,000 TIFF images/file
  - Up to 200 GBs/hr
  - ~5 TBs/day
  - 1-3 days/experiment
- Each reconstruction
  - 15 individual runs on a GPU
  - Can take up to 45 mins

Mark Basham (DLS) and Erica Yang
Links between metadata and files

• It is not cost effective to transfer data to the home institutions
  – The network bandwidth
  – take data back home on storage drive

• It is expensive to do analysis at home institutions
  – It is impossible to process user’s own computer
  – Lack of hardware resources
  – Lack of metadata
  – Lack of expertise (e.g. parallel processing, GPU programming)

• Users are interested in remote data analysis services
  – “... Of course this would mean a step change in the facilities provided and the time users spend at the facility. ...”
  – Capture Provenance of data products
PaN-data shared catalogues

- Proposal
- Approval
- Scheduling
- Experiment
- Data storage
- Data analysis
- Record Publication

Provenance Support

Extend the ICAT Data model to record derived data and software

Develop support within analysis tool packages (DAWN, Mantid)

Controlled vocabulary
Conclusions

• Developing a programme towards a common data infrastructure for facilities
  – Pooling limited resources
  – Common experience for users
  – Can transfer and share data more effectively

• Developing a common approach
  – Data and software catalogues
  – Provenance to capture the full context of the experiment

• Big facilities for small science

brian.matthews@stfc.ac.uk

http://pan-data.eu/