PaNSIG (PaNDaData), and the interactions between SB-IG and PaNSIG

Erica Yang
erica.yang@stfc.ac.uk

Scientific Computing Department
STFC Rutherford Appleton Laboratory

Structural Biology IG
27 March 2014
RDA Plenary, Dublin
What is ICAT?

**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.

---

**Example ISIS Proposal**

**GEM** – High intensity, high resolution neutron diffractometer

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

**B-lactoglobulin protein interfacial structure**
PaNSIG: IG for Photon and Neutron Science

• Co-chairs
  • Amber Boehnlein/SLAC
  • Frank Schluenzen/DESY,
  • Brian Matthews/STFC

• A follow-on from the EU PaNData project
• Focusing on discussing coordinated efforts to
  • Address rising challenges around research data
  • Identify opportunities for exploiting research data

• Participants (so far)
  • Computing people: facility operation, data analysis code development, data infrastructure and HPC operation, and system integration
  • “Friendly” scientists (need more!)
PaNData: Photon and Neutron Data Infrastructure

- Established in 2007 with 4 facilities
  - now standing at 13
  - With “friends” around the world
- Combined Number of Unique Users more than 35000 in 2011
- Combines Scientific and IT staff from the collaborating facilities
- European Framework 7 Projects
  - PaNdata-Europe: SA, 2009-11
  - PaNdata-Open Data Infrastructure, IP, 2011-14
- Guesstimates
  - Investment > €4.000.000.000
  - Running costs > €500.000.000/yr
  - Publications > 10.000/yr
  - RCots/Publication ~ €50.000
  - Data volume >>>> 10PB/yr

Credit: Brian Matthews
## Counting Users

### 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>SINOQ</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>1563</td>
<td>7</td>
<td></td>
<td>10</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>69</td>
<td>400</td>
<td>105</td>
<td></td>
<td>115</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>33</td>
<td>52</td>
<td>198</td>
<td>98</td>
<td>191</td>
</tr>
<tr>
<td>DLS</td>
<td>58</td>
<td>46</td>
<td>137</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>157</td>
<td>383</td>
<td>46</td>
<td>137</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>157</td>
<td>383</td>
<td>46</td>
<td>137</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>74</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</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>74</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</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>74</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</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>2880</td>
<td>89</td>
<td>233</td>
<td>74</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</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>89</td>
<td>1235</td>
<td>74</td>
<td>74</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</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>74</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</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>39</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</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>31</td>
<td>46</td>
<td>347</td>
<td>137</td>
<td>89</td>
</tr>
</tbody>
</table>

- **Multiple facilities, and multiple techniques**
- **Experimental, computational, and maybe, data-intensive (?)**

[http://pan-data.eu/Users2012-Results](http://pan-data.eu/Users2012-Results)
PaN-Data Developments

Common data environment, common user experience

Shared Data Policy Framework

Federated User Authentication

Federated Data Catalogue

Common Software Catalogue

Common Ontology

Common Data Format

PanSoft

PanKOS

NeXus
ICAT Tool Suite and Clients

- A database
- With well defined data model
- With well defined APIs to access, search, download
  - Experiment data
  - Proposal to Publication
- Being rolled out in PaNData-ODI

User interactions with ICAT

- ICAT Data Explorer (Eclipse Plugin)
- ICAT + Mantid (desktop client)

ICAT APIs + Web Services

Desktop app
Clusters/HPC
Disk
Tape

ICAT
TopCAT (Web Interface to ICATs)
ICAT Job Portal
IDS (ICAT Data Service)
ICAT Data Explorer (Eclipse Plugin)
ICAT + Mantid (desktop client)
Interfaces - TopCAT
Interfaces – Mantid and DAWN
ICAT: current deployments

In production:

1. ISIS/UK: neutron
2. DLS/UK: synchrotron
3. CLF/UK: high power lasers
4. SNS/US: neutron
5. ILL/France: neutron

In prototype deployment:

1. ALBA/Spain: synchrotron
2. ESRF/France: synchrotron
3. DESY/Germany: synchrotron
4. Elletra/Italy: synchrotron
...
Data monitoring

Network monitoring

Data Synchronisation

Data Cataloguing

Behind the ICATs (Diverse experiments at RAL)

Data archive
Nexus and ICAT

How does diverse experiment metadata get into ICAT

A facility can put any metadata into ICAT as long as it conforms to the above model. But, it doesn’t help searching and discovering data!

(Data model diagram credit: Andy Gotz et. al./ESRF)
PanKOS: Ontology for Facility Science

Facilities, instruments, and techniques using Linked Data
(applications: cataloguing, annotation, searching, and linking)

**Diffraction**
- Neutron Diffraction/Elastic Neutron Scattering
  - Powder Diffraction
  - Single Crystal Diffraction
- X-ray Diffraction
  - Grazing Incidence Diffraction
  - Powder Diffraction
  - Resonant Diffraction
  - Small Angle Diffraction
  - Single Crystal Diffraction
  - Soft Diffraction
  - Surface Diffraction

**Other**
- Coherent Diffraction Imaging
- Diffraction Imaging (Topography)
- Enhanced Diffraction Imaging

**Diffusive - Diffusive MRI**

**Imaging**
- Holography

**Microscopy**
- X-Ray Photoemission Microscopy
- X-Ray Scanning Microscopy
- Scanning Transmission X-Ray Microscopy
- Tomographic Microscopy With CRLs

**Tomography**
- Fluorescence Tomography
How does it relevant to SB-IG?
An integrated structural biology platform for the UK

- Cell Biology OPPF-UK MPL RC@H
- Diamond Beamlines: Macromolecular Crystallography, Scattering, X-ray spectroscopy, infrared
- Computational environment / CCP4
- UK XFEL HuB@Diamond
- Fluorescence microscopy (CLF/STFC & DLS)
- Cryo-EM/ET

Diagram Credit: Martin Walsh
Challenges facing SB user communities

• Data Volume
  • e.g. 25PB/75 days XFEL SFX instrument == LHC data/year
• Diversity of data (and techniques)
• Complex workflows (multiple techniques)
• Variety of data analysis software/frameworks
  • Data reduction and analysis may be facility specific
  • Users lack of capability (hardware, software and infrastructure expertise) to exploit third party software, HPC, and advanced distributed computing frameworks
• Variety of data formats (TRUE)
• Vast variety of metadata collected
  • Experiments
  • Simulations
  • Data analysis software
• Hard to keep track (and keep a record) of data analysis workflows
A Facility-based Post-Experiment Data Analysis Environment

An opportunity to go hand-in-hand with the SB community?
Infrastructure for managing data flows

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

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

Infrastructure + Software + Expertise!

Some image credit: Avizo, Visualization Sciences Group (VSG)
Between SB-IG and PanSIG

Starting points for collaborations?

PanSoft PanKOS NeXus