Data Analysis Exercise of User Crossover Between Facilities

Shelly Ren
ORNL/NScD

April 22, 2015
Background

- At the RDA Third Plenary Meeting in March 2014, the EU pan-data project team presented a user survey of users who performed experiments both at neutron and at photon sources in Europe. SNS was invited to join the exercise during the discussion.

- In August 2014, Susan White-Depace kindly and enthusiastically collected user data from APS, ALS, and NSLS to enable a study of crossover users between synchrotron facilities in the US, and between APS and SNS/HFIR.

- Today, I will present the results of the latest EU survey including SNS/HFIR & LCLS users, and the study of user crossover between several facilities in the US.
User Facility Survey

- How many users do the facilities have in total?
- How many users are using just a single or multiple different facilities?
- How many users perform experiments both at neutron and at photon sources?
- How many users are using the facilities outside the US?
- What types of institutions do our users come from?
- How many repeat users show up at our facilities?
- How critical is it for facilities to provide an integrated federated identity management service and a common data infrastructure?
Count, Hash, and Compare

• We used a 2 year time span to avoid effects from proposal submission deadlines, to reduce the influence of facility downtime occurring at different time of the years, and to also include users who come on a regular basis, but not every year.

• Survey time span and criteria:
  
  o US: 10/01/2012-07/31/2014, anyone who is on the proposal which was run during the date range.
  
  o EU: 08/01/2012-07/31/2014, on a proposal submitted during that time, or on a proposal submitted prior to 08/01/2012, but had an experiment during the time span.
  
  o EU: 06/01/2010-05/31/2012.
Count, Hash, and Compare

• We used a Secure Hash Algorithm (SHA-256) to convert users’ email addresses to almost-unique, fixed size 256-bit (32-byte) irreversible hashes, and saved hash results in a facility data file. (A hash function is any function that can be used to map digital data of arbitrary size to digital data of fixed size).

• Identical email addresses result in identical sha256 hashes regardless where and how it's being calculated, which allows us to match the hashes and extract the basic information without ever knowing other facilities users email addresses.

• We ran a script which could pull hashed data from different facility hash data files, count for matches, and produce some basic statistics.
Shared Users – US Survey Results

Counting users who worked on an experiment which was run between 10/01/2012 to 07/31/2014 at APS, and SNS/HFIR

<table>
<thead>
<tr>
<th></th>
<th>APS</th>
<th>SNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS</td>
<td>7200</td>
<td>420</td>
</tr>
<tr>
<td>SNS</td>
<td>420</td>
<td>2588</td>
</tr>
</tbody>
</table>

In other words, 5.8% of the 7200 APS users did experiments at SNS/HFIR, and 16% of the 2588 SNS/HFIR users did experiments at APS.
Shared Users – US Survey Results

Counting users who worked on an experiment between 10/01/2012 to 07/31/2014 at APS, ALS, and NSLS

<table>
<thead>
<tr>
<th></th>
<th>APS</th>
<th>ALS</th>
<th>NSLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS</td>
<td>7200</td>
<td>321</td>
<td>513</td>
</tr>
<tr>
<td>ALS</td>
<td>321</td>
<td>3052</td>
<td>108</td>
</tr>
<tr>
<td>NSLS</td>
<td>513</td>
<td>108</td>
<td>3130</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>APS</th>
<th>ALS</th>
<th>NSLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS</td>
<td>4.46%</td>
<td>7.13%</td>
<td></td>
</tr>
<tr>
<td>ALS</td>
<td>10.52%</td>
<td>3.54%</td>
<td></td>
</tr>
<tr>
<td>NSLS</td>
<td>16.39%</td>
<td>3.45%</td>
<td></td>
</tr>
</tbody>
</table>
## Shared Users -- EU Survey Results

For period from 06/01/2010 through 05/31/2012:

### Photon Sources

<table>
<thead>
<tr>
<th>Facility</th>
<th># Users</th>
<th>Users using at least one other Photon Source</th>
<th>Users using at least one Neutron Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBA</td>
<td>773</td>
<td>412 (53.2%)</td>
<td>69 (8.9%)</td>
</tr>
<tr>
<td>ANKA</td>
<td>452</td>
<td>152 (33.6%)</td>
<td>19 (4.2%)</td>
</tr>
<tr>
<td>BESSY II</td>
<td>2306</td>
<td>774 (33.5%)</td>
<td>183 (7.9%)</td>
</tr>
<tr>
<td>DESY</td>
<td>4197</td>
<td>1423 (33.9%)</td>
<td>489 (11.6%)</td>
</tr>
<tr>
<td>DLS</td>
<td>4407</td>
<td>1153 (26.1%)</td>
<td>554 (12.5%)</td>
</tr>
<tr>
<td>ELETTRA</td>
<td>3167</td>
<td>951 (30.0%)</td>
<td>150 (4.7%)</td>
</tr>
<tr>
<td>ESRF</td>
<td>10287</td>
<td>3789 (36.8%)</td>
<td>1361 (13.2%)</td>
</tr>
<tr>
<td>SLS</td>
<td>3827</td>
<td>1556 (40.6%)</td>
<td>392 (10.2%)</td>
</tr>
<tr>
<td>SOLEIL</td>
<td>4568</td>
<td>1847 (40.4%)</td>
<td>411 (8.9%)</td>
</tr>
</tbody>
</table>

### Neutron Sources

<table>
<thead>
<tr>
<th>Facility</th>
<th># Users</th>
<th>Users using at least one other Neutron Source</th>
<th>Users using at least one Photon Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>BER II</td>
<td>1563</td>
<td>619 (39.6%)</td>
<td>381 (24.3%)</td>
</tr>
<tr>
<td>FRM-II</td>
<td>1095</td>
<td>534 (48.7%)</td>
<td>281 (25.6%)</td>
</tr>
<tr>
<td>ILL</td>
<td>4549</td>
<td>1671 (35.9%)</td>
<td>1367 (29.4%)</td>
</tr>
<tr>
<td>ISIS</td>
<td>2380</td>
<td>1079 (37.4%)</td>
<td>754 (26.1%)</td>
</tr>
<tr>
<td>LLB</td>
<td>1235</td>
<td>408 (33.0%)</td>
<td>330 (26.7%)</td>
</tr>
<tr>
<td>SINQ</td>
<td>1219</td>
<td>610 (50.0%)</td>
<td>424 (34.7%)</td>
</tr>
<tr>
<td>SNS</td>
<td>4311</td>
<td>666 (15.4%)</td>
<td>252 (5.8%)</td>
</tr>
</tbody>
</table>

[http://pan-data.eu/node/99](http://pan-data.eu/node/99)
# Shared Users -- EU Survey Results

For period from 08/01/2012 through 07/31/2014:

## Photon Sources

<table>
<thead>
<tr>
<th>Facility</th>
<th># Users</th>
<th>Users using at least one other Photon Source</th>
<th>Users using at least one Neutron Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBA</td>
<td>1303</td>
<td>679 (52.1%)</td>
<td>122 (9.3%)</td>
</tr>
<tr>
<td>BESSY II</td>
<td>838</td>
<td>418 (49.8%)</td>
<td>76 (9.0%)</td>
</tr>
<tr>
<td>DESY</td>
<td>3680</td>
<td>1579 (42.9%)</td>
<td>343 (9.3%)</td>
</tr>
<tr>
<td>DLS</td>
<td>10445</td>
<td>2598 (24.8%)</td>
<td>1136 (10.8%)</td>
</tr>
<tr>
<td>ELETTRA</td>
<td>3422</td>
<td>1171 (34.2%)</td>
<td>149 (4.3%)</td>
</tr>
<tr>
<td>ESRF</td>
<td>10786</td>
<td>4242 (39.3%)</td>
<td>1165 (10.8%)</td>
</tr>
<tr>
<td>LCLS</td>
<td>1123</td>
<td>329 (29.2%)</td>
<td>17 (1.5%)</td>
</tr>
<tr>
<td>SLS</td>
<td>3981</td>
<td>1637 (41.1%)</td>
<td>365 (9.1%)</td>
</tr>
<tr>
<td>SOLEIL</td>
<td>5134</td>
<td>2145 (41.7%)</td>
<td>349 (6.7%)</td>
</tr>
</tbody>
</table>

## Neutron Sources

<table>
<thead>
<tr>
<th>Facility</th>
<th># Users</th>
<th>Users using at least one other Neutron Source</th>
<th>Users using at least one Photon Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>BER II</td>
<td>237</td>
<td>162 (68.3%)</td>
<td>86 (36.2%)</td>
</tr>
<tr>
<td>MLZ</td>
<td>1430</td>
<td>601 (42.0%)</td>
<td>353 (24.6%)</td>
</tr>
<tr>
<td>ILL</td>
<td>4138</td>
<td>1252 (30.2%)</td>
<td>1304 (31.5%)</td>
</tr>
<tr>
<td>ISIS</td>
<td>3406</td>
<td>891 (26.1%)</td>
<td>1052 (30.8%)</td>
</tr>
<tr>
<td>SINQ</td>
<td>1424</td>
<td>614 (43.1%)</td>
<td>501 (35.1%)</td>
</tr>
<tr>
<td>SNS</td>
<td>3723</td>
<td>581 (15.6%)</td>
<td>327 (8.7%)</td>
</tr>
</tbody>
</table>

http://pan-data.eu/Users2014-Results
## Shared Users -- EU Survey Results

For period from 08/01/2012 through 07/31/2014:

<table>
<thead>
<tr>
<th>Number of Users shared between facilities</th>
<th>ALBA</th>
<th>BERII</th>
<th>BESSYII</th>
<th>DESY</th>
<th>DLS</th>
<th>ELETTRA</th>
<th>ESRF</th>
<th>MLZ</th>
<th>ILL</th>
<th>ISIS</th>
<th>LCLS</th>
<th>SINC</th>
<th>SLS</th>
<th>SOLEIL</th>
<th>SNS</th>
<th>neutron</th>
<th>photon</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBA</td>
<td>1303</td>
<td>5</td>
<td>43</td>
<td>90</td>
<td>274</td>
<td>128</td>
<td>356</td>
<td>8</td>
<td>83</td>
<td>36</td>
<td>2</td>
<td>23</td>
<td>124</td>
<td>161</td>
<td>7</td>
<td>122</td>
<td>679</td>
<td>1303</td>
</tr>
<tr>
<td>BERII</td>
<td>5</td>
<td>237</td>
<td>27</td>
<td>28</td>
<td>20</td>
<td>0</td>
<td>42</td>
<td>60</td>
<td>104</td>
<td>50</td>
<td>0</td>
<td>75</td>
<td>22</td>
<td>7</td>
<td>37</td>
<td>102</td>
<td>80</td>
<td>237</td>
</tr>
<tr>
<td>BESSYII</td>
<td>43</td>
<td>27</td>
<td>838</td>
<td>128</td>
<td>96</td>
<td>95</td>
<td>143</td>
<td>16</td>
<td>31</td>
<td>16</td>
<td>28</td>
<td>29</td>
<td>119</td>
<td>100</td>
<td>11</td>
<td>76</td>
<td>418</td>
<td>838</td>
</tr>
<tr>
<td>DESY</td>
<td>90</td>
<td>28</td>
<td>128</td>
<td>3680</td>
<td>396</td>
<td>255</td>
<td>901</td>
<td>110</td>
<td>157</td>
<td>92</td>
<td>151</td>
<td>82</td>
<td>326</td>
<td>246</td>
<td>63</td>
<td>343</td>
<td>1579</td>
<td>3680</td>
</tr>
<tr>
<td>DLS</td>
<td>274</td>
<td>20</td>
<td>96</td>
<td>396</td>
<td>10445</td>
<td>297</td>
<td>1606</td>
<td>82</td>
<td>485</td>
<td>763</td>
<td>70</td>
<td>144</td>
<td>559</td>
<td>526</td>
<td>124</td>
<td>1136</td>
<td>2598</td>
<td>10445</td>
</tr>
<tr>
<td>ELETTRA</td>
<td>128</td>
<td>0</td>
<td>95</td>
<td>255</td>
<td>297</td>
<td>3422</td>
<td>480</td>
<td>21</td>
<td>99</td>
<td>41</td>
<td>68</td>
<td>14</td>
<td>218</td>
<td>379</td>
<td>12</td>
<td>149</td>
<td>1171</td>
<td>3422</td>
</tr>
<tr>
<td>ESRF</td>
<td>355</td>
<td>42</td>
<td>143</td>
<td>901</td>
<td>1606</td>
<td>480</td>
<td>10766</td>
<td>203</td>
<td>731</td>
<td>356</td>
<td>102</td>
<td>203</td>
<td>899</td>
<td>1390</td>
<td>155</td>
<td>1165</td>
<td>4242</td>
<td>10785</td>
</tr>
<tr>
<td>MLZ</td>
<td>8</td>
<td>66</td>
<td>16</td>
<td>110</td>
<td>82</td>
<td>21</td>
<td>203</td>
<td>1430</td>
<td>409</td>
<td>167</td>
<td>3</td>
<td>222</td>
<td>52</td>
<td>46</td>
<td>158</td>
<td>601</td>
<td>353</td>
<td>1430</td>
</tr>
<tr>
<td>ILL</td>
<td>83</td>
<td>104</td>
<td>31</td>
<td>167</td>
<td>485</td>
<td>99</td>
<td>731</td>
<td>409</td>
<td>4138</td>
<td>606</td>
<td>3</td>
<td>364</td>
<td>130</td>
<td>239</td>
<td>316</td>
<td>1252</td>
<td>1304</td>
<td>4138</td>
</tr>
<tr>
<td>ISIS</td>
<td>36</td>
<td>50</td>
<td>16</td>
<td>92</td>
<td>763</td>
<td>41</td>
<td>356</td>
<td>157</td>
<td>606</td>
<td>3406</td>
<td>9</td>
<td>236</td>
<td>101</td>
<td>84</td>
<td>267</td>
<td>891</td>
<td>1052</td>
<td>3406</td>
</tr>
<tr>
<td>LCLS</td>
<td>2</td>
<td>0</td>
<td>28</td>
<td>151</td>
<td>70</td>
<td>68</td>
<td>102</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>17</td>
<td>9</td>
<td>44</td>
<td>6</td>
<td>17</td>
<td>329</td>
<td>1123</td>
</tr>
<tr>
<td>SINC</td>
<td>23</td>
<td>75</td>
<td>29</td>
<td>82</td>
<td>144</td>
<td>34</td>
<td>203</td>
<td>222</td>
<td>384</td>
<td>236</td>
<td>1</td>
<td>1424</td>
<td>250</td>
<td>65</td>
<td>185</td>
<td>614</td>
<td>501</td>
<td>1424</td>
</tr>
<tr>
<td>SLS</td>
<td>124</td>
<td>22</td>
<td>119</td>
<td>326</td>
<td>559</td>
<td>218</td>
<td>899</td>
<td>52</td>
<td>130</td>
<td>101</td>
<td>79</td>
<td>250</td>
<td>3981</td>
<td>396</td>
<td>64</td>
<td>365</td>
<td>1637</td>
<td>3981</td>
</tr>
<tr>
<td>SOLEIL</td>
<td>161</td>
<td>7</td>
<td>100</td>
<td>246</td>
<td>526</td>
<td>379</td>
<td>1390</td>
<td>45</td>
<td>239</td>
<td>84</td>
<td>44</td>
<td>65</td>
<td>366</td>
<td>5134</td>
<td>40</td>
<td>349</td>
<td>2145</td>
<td>5134</td>
</tr>
<tr>
<td>SNS</td>
<td>7</td>
<td>37</td>
<td>11</td>
<td>63</td>
<td>124</td>
<td>12</td>
<td>155</td>
<td>156</td>
<td>316</td>
<td>267</td>
<td>6</td>
<td>185</td>
<td>64</td>
<td>40</td>
<td>3723</td>
<td>581</td>
<td>327</td>
<td>3723</td>
</tr>
</tbody>
</table>

Preliminary Findings – US

• 420 shared users performed experiments at both APS and SNS/HFIR from 10/01/2012 to 07/31/2014.

• APS and ALS shared 321 users, APS and NSLS shared 513 users, and ALS and NSLS shared 108 users from 10/01/2012 to 07/31/2014.

• Users conducted experiments across the synchrotron facilities, as well as at both synchrotron and neutron facilities in the US.
Preliminary Findings -- EU

• Two US and thirteen EU facilities participated the second user survey for the time from 08/01/2012 to 07/31/2014, which featured 4,840 users from US facilities, and 35,000 users from EU facilities.

• It is noteworthy that 1,091 out of 4,840 users from US facilities, or 22.5% of the US facility users were also active users at the European photon and/or neutron sources.

• Amazingly, there is not a single pair of facilities around the world without common users.
Conclusions

• Scientists are visiting both photon and neutron facilities within the US and around the world.

• A common comprehensive data infrastructure and user management would be greatly beneficial to both users as well as facilities to enhance the productivity of research.
  o Standardization of data policies
  o Federation of user authentication
  o Federation of data catalogue
  o Common data access
  o Data provenance, preservation, and scalability
  o Common user platforms
Conclusions

• “Virtual Data Facility” demo at SC 2014:
  o Integrated Federated Identity Management
  o Digital Object Identification Service
  o Portable Data Services Environment
  o Data Replication
  o Mobile Data Services
  o User Environment

• The journey from research and prototype to production is never easy, but we are at this turning point and could make an inordinate positive impact on our communities through collaboration and integration.
Going Forward

- Github and Google group are set up for science facilities in EU, a way to "Share the Experience" and let others see:
  - Facility overview, technology, integration, project, team, contact, links to software and services.
  - What have we done, what are we doing, what is next?
  - Diamond Light Source, Elettra, ESRF, and ISIS.
  - Schedule, reports, stats and others.
Going Forward

- Benefit from DOE projects/experiments database.
- Share best practice and build common infrastructure among user facilities.
- Deliver data and tool to users’ finger tip.
- Expand the scope of collaboration across user facilities including computing, neutron and light sources, nano centers, and others.
Acknowledgement

• Thank you to Susan White-Depace at NUFO for energetically collecting the user survey data and to the staff at the APS, ALS, NSLS facilities for providing the user data.

• Thank you to Tom Griffin at STFC in UK for inviting us to join the user survey of the EU pan-data project, and to Frank Schluenzen at DESY in Germany for welcoming and counting SNS and LCLS users in the EU user survey.

• Thank you to Crystal Schrof and Thomas Proffen for their inspiration, advice and full support on this project!
Thank you for your attention!