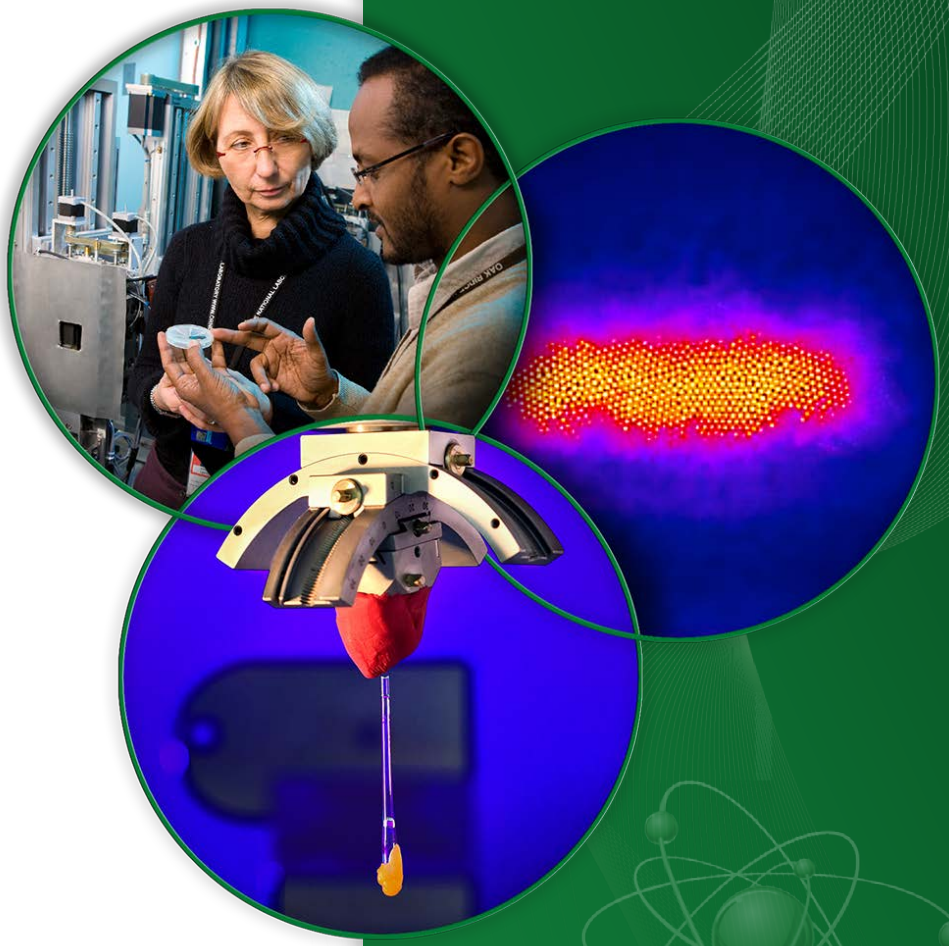
Data Analysis Exercise of User Crossover Between Facilities



Shelly Ren

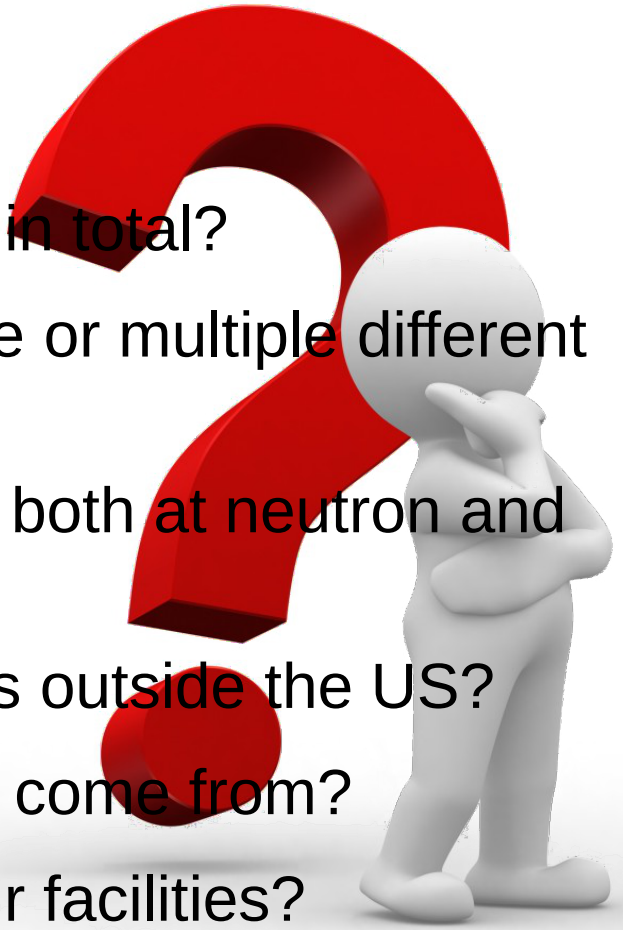
ORNL/NScD

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:
  - US: 10/01/2012-07/31/2014, anyone who is on the proposal which was run during the date range.
  - 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.
  - 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

	APS	SNS
APS	7200	420
SNS	420	2588

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

	APS	ALS	NSLS
APS	7200	321	513
ALS	321	3052	108
NSLS	513	108	3130

	APS	ALS	NSLS
APS		4.46%	7.13%
ALS	10.52%		3.54%
NSLS	16.39%	3.45%	

Shared Users -- EU Survey Results

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

Photon Sources			
Facility	# Users	Users using at least one other Photon Source	Users using at least one Neutron Source
ALBA	773	412 (53.2%)	69 (8.9%)
ANKA	452	152 (33.6%)	19 (4.2%)
BESSY II	2306	774 (33.5%)	183 (7.9%)
DESY	4197	1423 (33.9%)	489 (11.6%)
DLS	4407	1153 (26.1%)	554 (12.5%)
ELETTRA	3167	951 (30.0%)	150 (4.7%)
ESRF	10287	3789 (36.8%)	1361 (13.2%)
SLS	3827	1556 (40.6%)	392 (10.2%)
SOLEIL	4568	1847 (40.4%)	411 (8.9%)
Neutron Sources			
Facility	# Users	Users using at least one other Neutron Source	Users using at least one Photon Source
BER II	1563	619 (39.6%)	381 (24.3%)
FRM-II	1095	534 (48.7%)	281 (25.6%)
ILL	4649	1671 (35.9%)	1367 (29.4%)
ISIS	2880	1079 (37.4%)	754 (26.1%)
LLB	1235	408 (33.0%)	330 (26.7%)
SINQ	1219	610 (50.0%)	424 (34.7%)
SNS	4311	666 (15.4%)	252 (5.8%)

http://pan-data.eu/node/99

Shared Users -- EU Survey Results

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

Photon Sources			
Facility	# Users	Users using at least one other Photon Source	Users using at least one Neutron Source
ALBA	1303	679 (52.1%)	122 (9.3%)
BESSY II	838	418 (49.8%)	76 (9.0%)
DESY	3680	1579 (42.9%)	343 (9.3%)
DLS	10445	2598 (24.8%)	1136 (10.8%)
ELETTRA	3422	1171 (34.2%)	149 (4.3%)
ESRF	10786	4242 (39.3%)	1165 (10.8%)
LCLS	1123	329 (29.2%)	17 (1.5%)
SLS	3981	1637 (41.1%)	365 (9.1%)
SOLEIL	5134	2145 (41.7%)	349 (6.7%)
Neutron Sources			
Facility	# Users	Users using at least one other Neutron Source	Users using at least one Photon Source
BER II	237	162 (68.3%)	86 (36.2%)
MLZ	1430	601 (42.0%)	353 (24.6%)
ILL	4138	1252 (30.2%)	1304 (31.5%)
ISIS	3406	891 (26.1%)	1052 (30.8%)
SINQ	1424	614 (43.1%)	501 (35.1%)
SNS	3723	581 (15.6%)	327 (8.7%)

http://pan-data.eu/Users2014-Results

Shared Users -- EU Survey Results

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

Number of Users shared between facilities

	ALBA	BERII	BESSYII	DESY	DLS	ELETTRA	ESRF	MLZ	ILL	ISIS	LCLS	SINQ	SLS	SOLEIL	SNS	neutron	photon	all
ALBA	1303	5	43	90	274	128	356	8	83	36	2	23	124	161	7	122	679	1303
BER II	5	237	27	28	20	0	42	66	104	50	0	75	22	7	37	162	86	237
BESSY II	43	27	838	128	96	95	143	16	31	16	28	29	119	100	11	76	418	838
DESY	90	28	128	3680	396	255	901	110	167	92	151	82	326	246	63	343	1579	3680
DLS	274	20	96	396	10445	297	1606	82	485	763	70	144	559	526	124	1136	2598	10445
ELETTRA	128	0	95	255	297	3422	480	21	99	41	68	14	218	379	12	149	1171	3422
ESRF	356	42	143	901	1606	480	10786	203	731	356	102	203	899	1390	155	1165	4242	10786
MLZ	8	66	16	110	82	21	203	1430	409	167	3	222	52	46	158	601	353	1430
ILL	83	104	31	167	485	99	731	409	4138	606	3	384	130	239	316	1252	1304	4138
ISIS	36	50	16	92	763	41	356	167	606	3406	9	236	101	84	267	891	1052	3406
LCLS	2	0	28	151	70	68	102	3	3	9	1123	1	79	44	6	17	329	1123
SINQ	23	75	29	82	144	14	203	222	384	236	1	1424	250	65	185	614	501	1424
SLS	124	22	119	326	559	218	899	52	130	101	79	250	3981	366	64	365	1637	3981
SOLEIL	161	7	100	246	526	379	1390	46	239	84	44	65	366	5134	40	349	2145	5134
SNS	7	37	11	63	124	12	155	158	316	267	6	185	64	40	3723	581	327	3723
neutron	122	237	76	343	1136	149	1165	1430	4138	3406	17	1424	365	349	3723	10257	2606	11976
photon	1303	86	838	3680	10445	3422	10786	353	1304	1052	1123	501	3981	5134	327	5910	25914	32257
all	1303	237	838	3680	10445	3422	10786	1430	4138	3406	1123	1424	3981	5134	3723	12570	24923	41665

http://pan-data.eu/Users2014-Results

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.
  - Standardization of data policies
  - Federation of user authentication
  - Federation of data catalogue
  - Common data access
  - Data provenance, preservation, and scalability
  - Common user platforms

Conclusions

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