
ESRF
The European Synchrotron

Data Management for Photon Science

H. Reichert – reichert@esrf.fr

THE ESRF IN A NUTSHELL

- Partnership between 21 countries
- World's most productive synchrotron laboratory
- Research in all areas involving condensed matter, materials, and living matter
- ~30 public beamlines (instruments); 14 CRG beamlines (national teams)
- 600 Staff: 500 with a technical background, 60 post-docs, 40 PhD students
- > 8000 user visits for ~1500 projects
- ~1800 refereed publications / year
- Annual budget: ~100 M€ including the Upgrade Programme

ESRF UPGRADE PROGRAMME

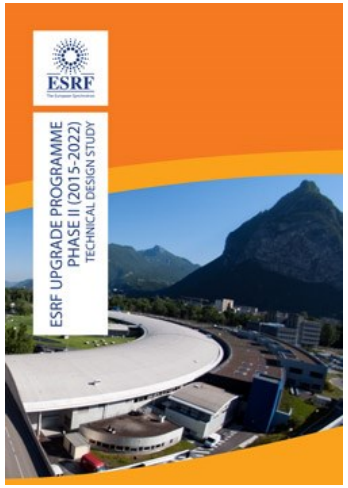
Phase I

19 upgraded or deeply refurbished BLs
Accelerator and source upgrade
Construction programme

2009 → 2015

Purple Book
2008

Orange Book
January 2015



2015 → 2022

Phase II

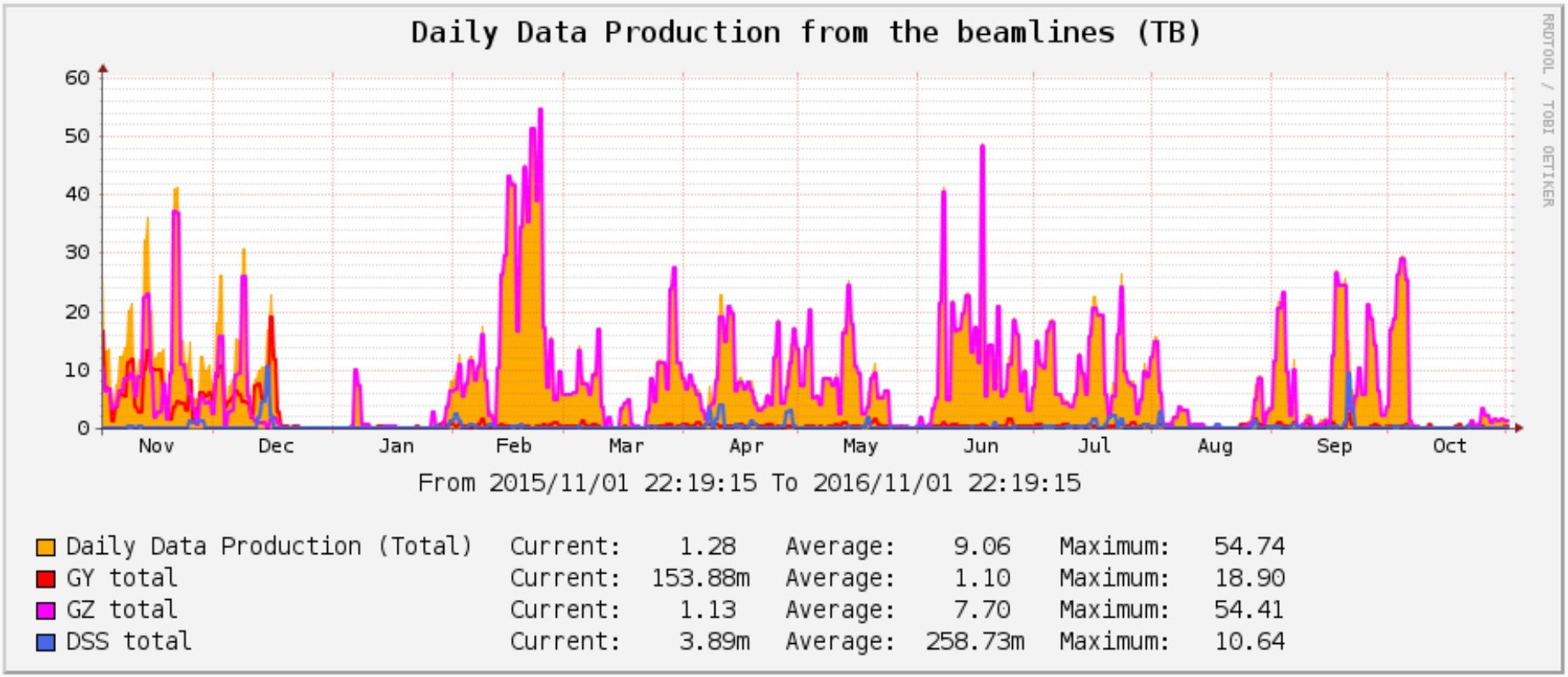
New storage ring
4 new BLs
Technology programme

~4 orders of magnitude
Improvement in performance!
(on average)

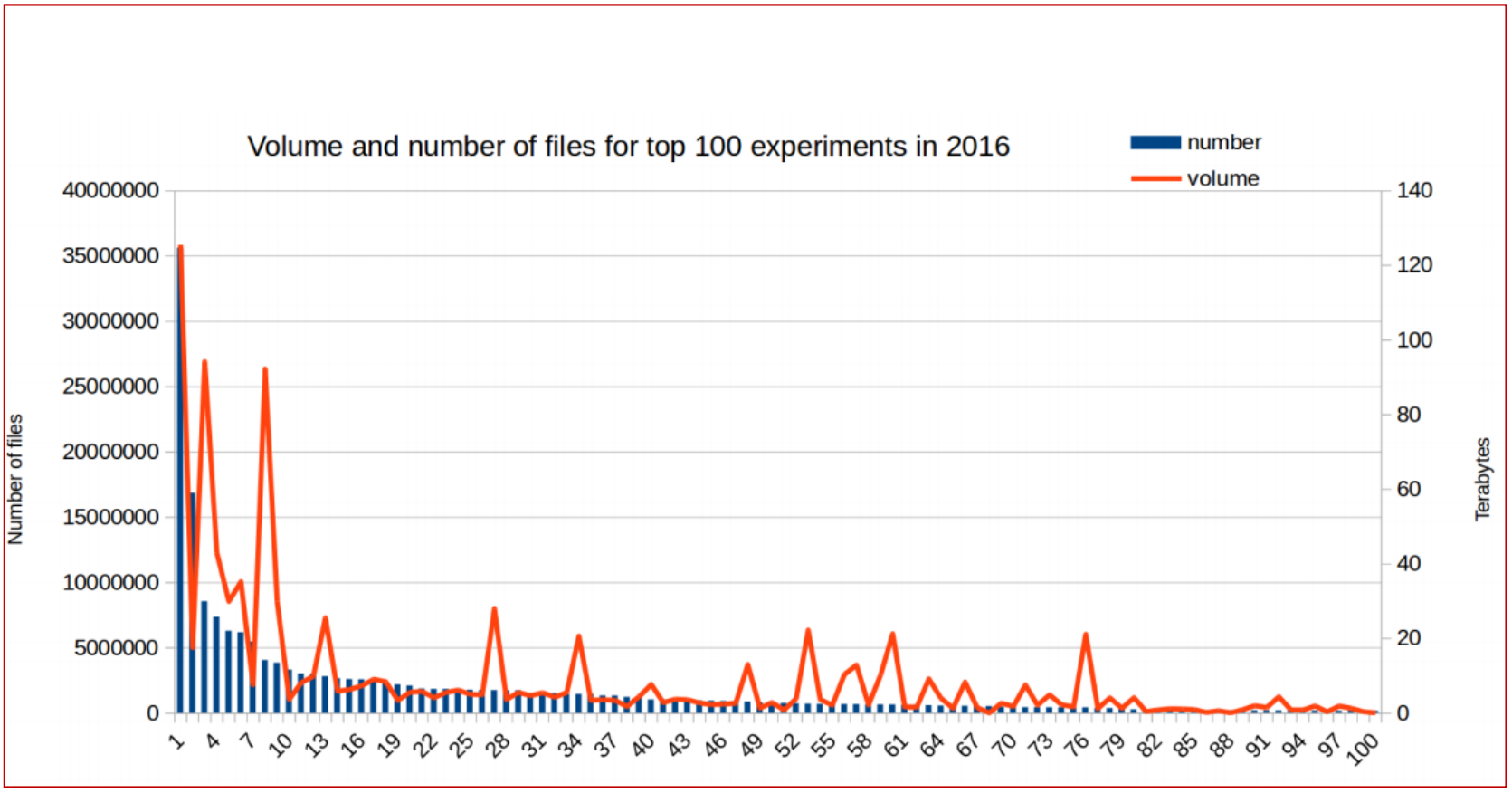
DATA PRODUCTION TODAY

- The ESRF beamlines produce currently 4 PB/year
- The data rate is increasing steadily
- In 2021 (first full year of operation with the new storage ring) we expect at least 10-15 PB/year
- The number of files per experiment is dramatically increasing

- **The data rate drives the ESRF IT infrastructure**



DATA PRODUCTION TODAY



Limit for long-term archiving:

- **1000 files / 8hours / experiment**
- **~20 000 files / experiment / week → HDF5 on all beamlines**

PANDATA USER SURVEY

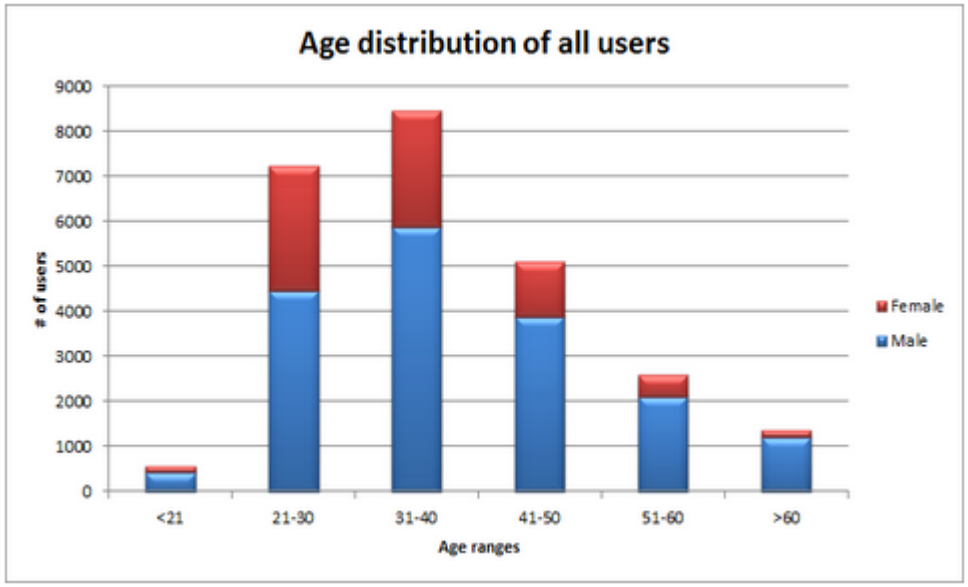
Counting Users of 17 RIs via the User Offices:
http://pan-data.eu/Users2014-Results

August 2012 to July 2014 - 2 years

10 Photon Sources – ALBA, PETRA-III+FLASH, DLS, ELETTRA, ESRF, SLS, SOLEIL, BESSY-II, LCLS

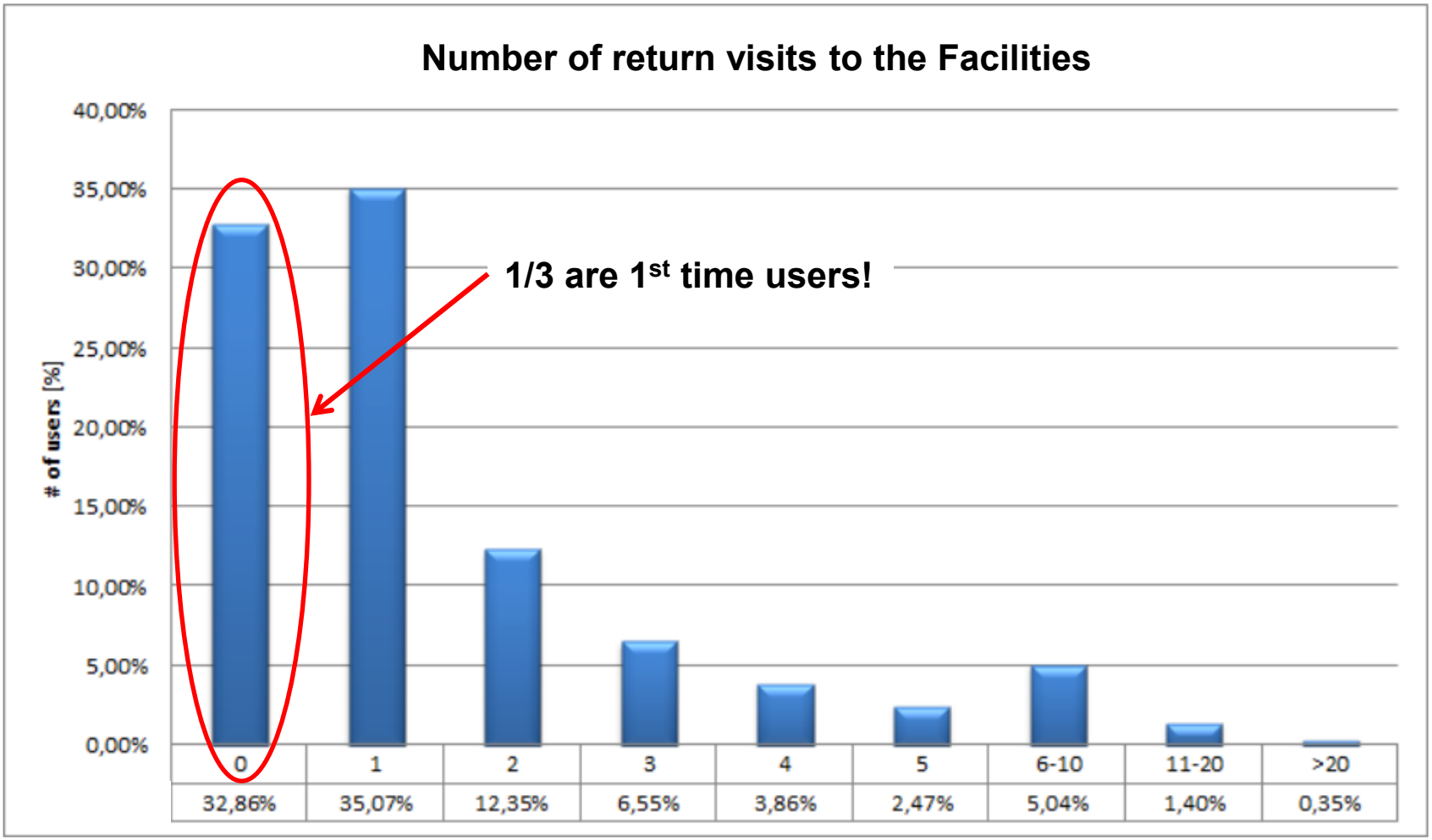
7 Neutron Sources – FRM-II, ILL LLB, ISIS, SINQ, BER-II, SNS

41 665 unique users from 95 countries (USA – 4 840)



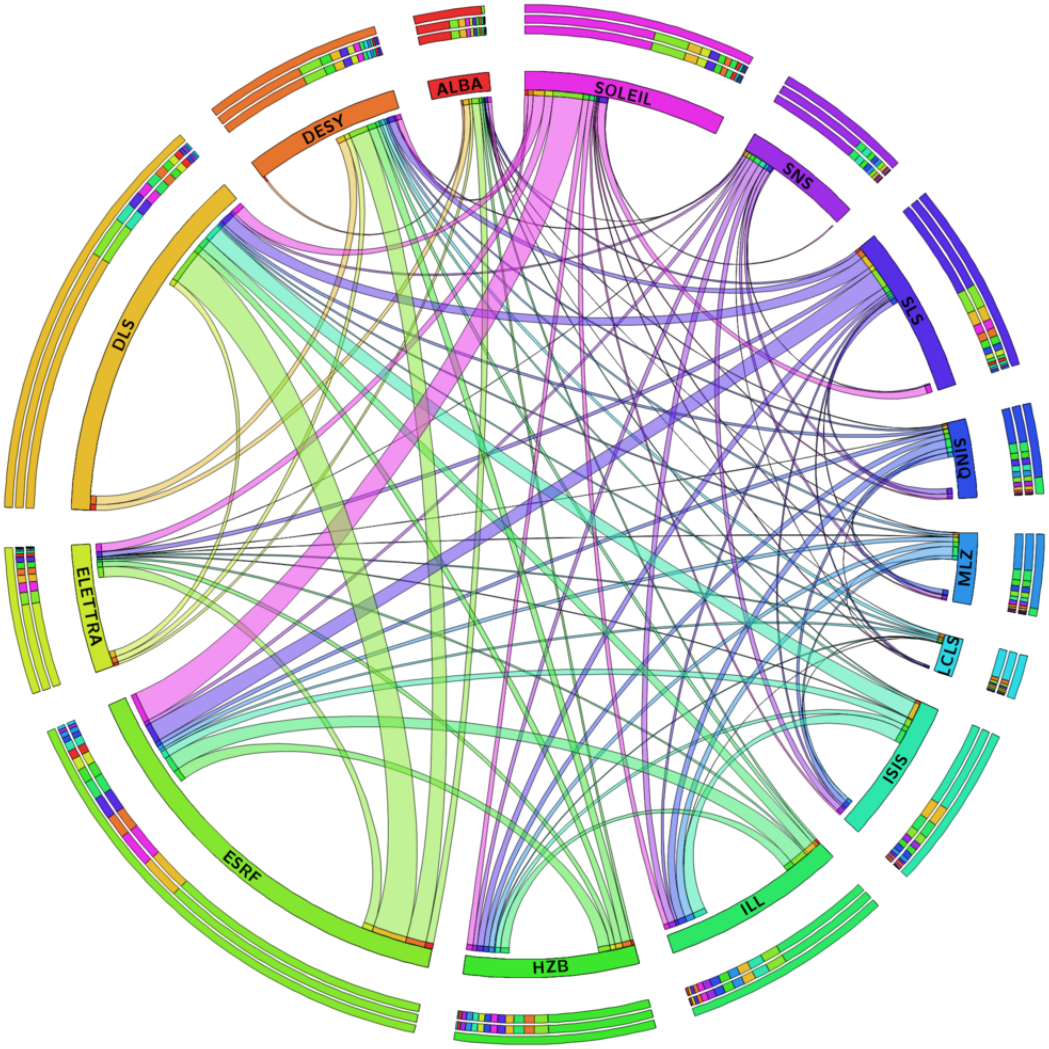
(Courtesy: F. Schlünzen, DESY)

PANDATA USER SURVEY



(Courtesy: F. Schlünzen, DESY)

PANDATA USER SURVEY



Chord chart on common users:

- Length of the facility arc represents the number of users of that facility
- Thickness of the arc between facilities relates to the number of common users

Conclusion

Users of Analytical Facilities are young and mobile

(Courtesy: F. Schlünzen, DESY)

PAN RIS WORK TOGETHER ON IT SINCE 2006

FP7+H2020 project/proposal

- VEDAC
- PaNData-Europe
- PaNData-ODI
- IRVUX
- CRISP
- EUCALL
- CALIPSO+
- NFFA
- PaNDaaS

Ouput

- User Statistics
- Data Policy
- Software Catalogue
- ICAT
- NeXus/HDF5 consensus
- Umbrella FIM
- ...

METADATA AND DATA POLICY

nature methods
Techniques for life scientists and chemists
Home | Current issue | Comment | Research | Archive ▼
Nature Methods **13**, 799 (2016) | doi:10.1038/nmeth.4026
Published online 29 September 2016

Subject terms: Publishing

As the research community embraces data sharing, academic journals can do their part to help. Starting this month, [redacted]

[redacted] Where applicable, they will include details about publicly archived data sets that have been analyzed or generated during the study. Where restrictions on access are in place—for example, in the case of privacy limitations or third-party control—authors will be expected to make this clear.

The new policy (full details of which are available at www.go.nature.com/2bf4vqn) builds on our long-standing support for data availability as a condition of publication. It also extends our support for data citation, the practice of citing data sets in reference lists in a similar way to citing papers. [redacted]

ESRF OPEN DATA POLICY

- ESRF is custodian of data and metadata
- ESRF to collect high quality metadata to facilitate the use of data
- ESRF will keep metadata forever
- ESRF will keep raw (or reduced) data for 10 years
- Data will be registered in a data catalogue
- Data will be published with DOIs
- The experimental team has exclusive access to data during the embargo period (3 years which can be extended on request)
- Data will be made public after the embargo period under license CC-BY 4.0

…but the implementation is challenging

IMPLEMENTATION OF THE ESRF DATA POLICY (→ 2020)

METADATA

- Routinely used (Fast Tomography): ID11, ID16A, ID16B, ID17
- Under test: ID01(KMAP) ID21, ID31, ID30A-1, ID30A-3
- Coming soon: ID23-1, ID23-2, ID29, ID30B, BM29

Migration rate: ~ 10 beamlines per year

Data and metadata catalogued since June 2015

> 11.000.000 **FILES**	
17.000 **DATASETS**	
471.900 **PARAMETERS**	
418 **PROPOSALS**	
326 **USERS**	
2063 **SAMPLES**	

- Upgrade iCAT to version 4.7.0
- Development of IDS (iCAT Data Service) at ESRF for custom archival
- Customization and installation of the front-end TopCat

Joint efforts of the European accelerator-based light sources to tackle data management issues together

Common data policy implementation roadmap

- Metadata ontology
- Cross-facility searchable metadata catalogue
- Photon Science e-logbook
- Develop common tools to interact with archived data
- Link to public/commercial clouds (EOSC, AWS, …)

Make data Findable, Accessible, Interoperable and Reusable (FAIR)

http://ec.europa.eu/research/openscience

IMPLEMENTATION OF A COMMON OPEN DATA POLICY

Work on common on-line data reduction algorithms

- Data streaming, data format, compression algorithms
- Work hand-in-hand with main detector manufacturers
- Integrate algorithms into the lab specific framework

Share data analysis code developments

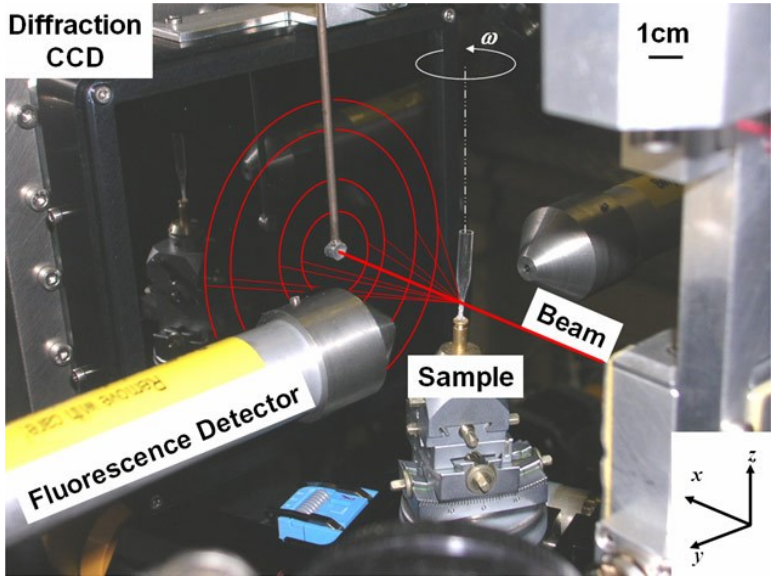
- Hire data scientists
- Each lab to champion the development/maintenance of a small set of analysis programs
- Make sure all programs are remote access compatible
- Make sure all programs are documented
- Organise code camps, eg. on parallel programming
- Set-up tutoring services for data analysis

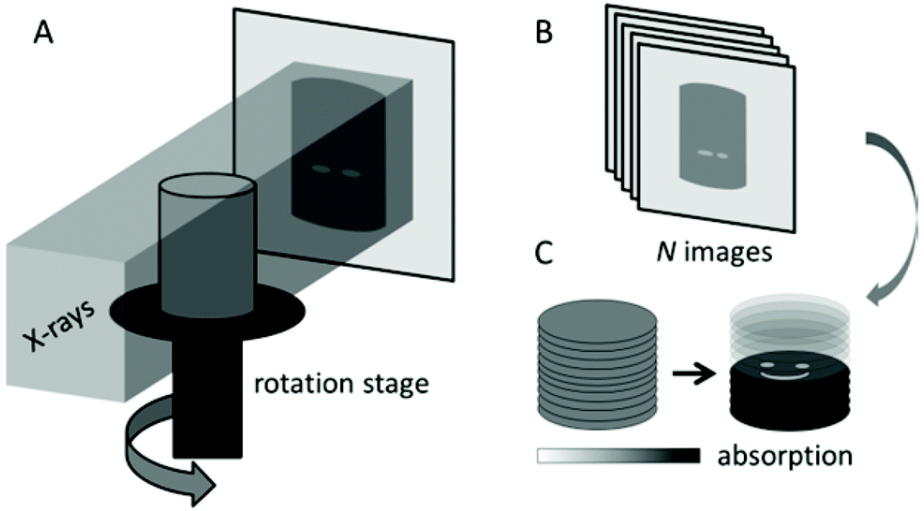
THE PANDAAS PROPOSAL

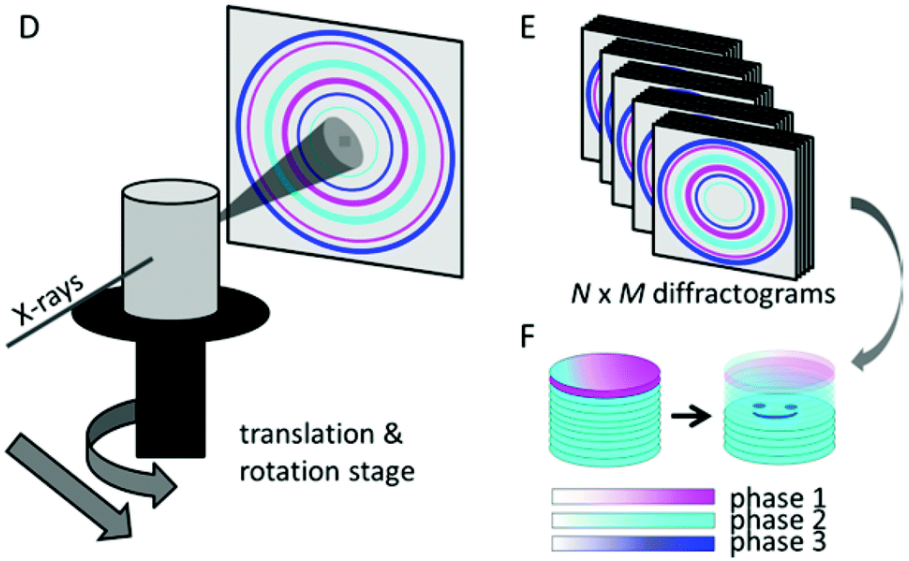
- ✓ Match data processing capabilities with advancements in detectors and sources
- ✓ Join forces to do this collaboratively in Europe
  - Share best practices and solutions
  - Share the workload
  - Create a homogeneous and compatible data reduction/analysis environment for our Users Community
  - Make the environment sustainable
- ✓ PaNDaaS H2020-Infradev-4 proposal (not approved)
  - ➢ "Big Data" private cloud infrastructure
  - ➢ Cloud federation
  - ➢ 21 Science use cases
  - ➢ 3.5 years
  - ➢ 21 partners (5 ESFRI projects)
  - ➢ 1553 PMs effort
  - ➢ 13.6 M€ (for human resources only)

EXAMPLE: X-RAY DIFFRACTION TOMOGRAPHY

- Data rate = 250Hz
- Experiment duration = 1 week
- Sample volume = 1000^3
- Data volume = 10PB
- To have reduced data on-line = <10TB
- Use PyFAI accelerated on GPU to reduce data in real-time by a factor of 1000







HARNESS THE DATA

Data volume is growing each year

Next generation experiments can produce PBs/week

<u>Multiple data bottlenecks:</u>

1. **Acquiring data from detectors to storage (Gigabytes/s)**
2. **→ Reducing data online fast enough (Gigabytes/min)**
3. **→ Analysing data fast enough (Terabytes/day)**
4. **→ Archiving data efficiently (Tera to Petabytes/day)**
5. **Getting results to users efficiently (Giga to Terabytes)**
6. **Helping Scientists analysing their data on-site (and off-site) (DaaS)**

Thank you for your attention