

5-WAY MEETING



DATA SESSION



NICHOLAS SCHWARZ, APS
STUART CAMPBELL, NSLS-II
DULA PARKINSON, ALS
AMEDEO PERAZZO, SLAC

9 October 2017

PURPOSE

- Most of us are anticipating a multiple order-of-magnitude increase in data rates and volumes over the coming years.
- Advanced computing and the ability to cope with data is critical to the success of our facilities.
- Needs are driven by new scientific opportunities, which are enabled by new sources, new measurement techniques, advances in detectors, multi-modal data utilization, and advances in data analysis algorithms.

Today's Topics:

- (Common) Data Policies & Practices
- Using External Computing & Storage Resources
- What are we doing now and where can we go together? What are our most pressing needs?

AGENDA

- Introduction
- Policies & Practices at European Facilities
- ALS, Dula Parkinson
- APS, Nicholas Schwarz
- NSLS-II, Stuart Campbell
- SLAC, Amedeo Perazzo
- Discussion

EUROPEAN FACILITIES

EUROPEAN FACILITIES

PanData Data Policy

- **Photon and Neutron data (PaNdata)** infrastructure initiative developed a template data policy
- Many European facilities have or are in the process of adopting this policy
- Data interoperability between facilities

- The facility is the custodian of raw data and metadata
- Raw data and metadata will be stored for ten years
- The experimental team will have sole access to the data during a three-year embargo period, renewable if necessary
- After the embargo, the data will be released under an open access license

EUROPEAN FACILITIES

Facility Status

ESRF

- Continues to roll-out their data management tools to more beamlines, with the goal of archiving all facility data and retaining that data for ten years by 2020.
- On-site redundant tape archival system.
- Working toward rich metadata in ICAT and HDF5 NeXus files.
- PaNdata Software Catalog (<https://software.pan-data.eu/>) - catalog and disseminate software used at European synchrotron and neutron sources.
- <http://www.esrf.eu/datapolicy>

Beamline	Status	Techniques	Data Policy Implementation			
			Metadata Collection*	Data archiving*	Raw Data HDF5*	Open access to data*
ID01	KMAP		Implemented	In progress		
BM01A						
BM01B						
ID02						
BM02						
ID03						
BM05	Tomography		In progress	In progress		
ID06-LVP						
ID06						
BM08						
ID09						
ID10						
ID11	Tomography		In progress	In progress		
ID12						
ID13						
BM14						
BM15A						
BM15B						
ID16A	Fluo, Tomo		Implemented	In progress	Implemented	
ID16B	Tomo		In progress	In progress		
ID17	MRI, Tomography		Implemented	In progress		
ID18						
ID19	Tomography		In progress	In progress		
ID20 et			In progress	In progress		
ID21	Microscopy		Implemented	In progress		
ID22						
ID23-1	MX		Implemented	In progress		
ID23-2	MX		Implemented	In progress		
BM23						
ID24						
BM25A						
BM25B						
ID26						
BM26A						
BM26B						
ID27						
ID28						
BM28						
ID29	MX		Implemented	In progress		
BM29	BIOSAXS		Implemented	In progress		
ID30A-1	MX		Implemented	In progress		
ID30A-3	MX		Implemented	In progress	Implemented	
ID30B	MX		Implemented	In progress		
BM30A						
BM30B						
ID31						
ID32			In progress	In progress		
BM32						
CryoEM						

EUROPEAN FACILITIES

Facility Status

ALBA

- Eventually want to follow PaNdata data retention policies
- Planning to prototype ISPyB and ICAT metadata catalogs

DESY

- Centralized storage system used by most beamlines
- Embargo period of 3 years
- Data access via web portal

DLS

- STFC hosts DLS's tape archive system off-site
- In 2014, DLS archived approximately 1 PB of data; in 2017, DLS archived a total of 7.2 PB of data
- Data is transferred to tape after 40 days on spinning disk
- Uses ICAT as the facility metadata catalog

EUROPEAN FACILITIES

Facility Status

ESS

- Very similar policy to PaNdata
- Two copies of the raw data and meta data will be kept in physically different locations; 10 years on spinning disk before moving to tape archival

ISIS

- Now has an explicit data policy following the PaNdata template
- Saves all data in NeXus as well as legacy RAW format

MAX-IV

- Experimenting with metadata catalogs, ISPyB and ICAT, but may very likely adopt the same solution as ESS
- Intends to use off-site computing and storage facility

SwissFEL

- Will adopt a policy that follows PaNdata standard
- Computing and storage resources will be provided by the nearby Innovation Park
- SwissFEL and the ESS will co-develop a new metadata catalog instead of adopting an existing catalog



www.anl.gov