Overview of metadata and raw data cataloguing at Diamond.

Pierre Aller

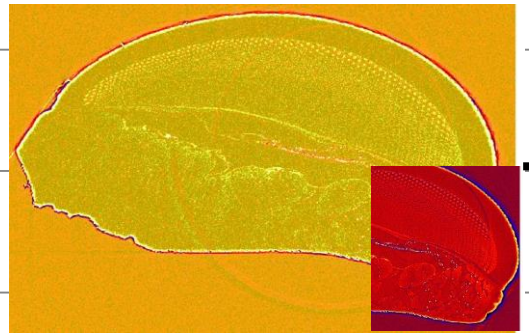
Alun Ashton

Further details from :
scientificsoftware@diamond.ac.uk

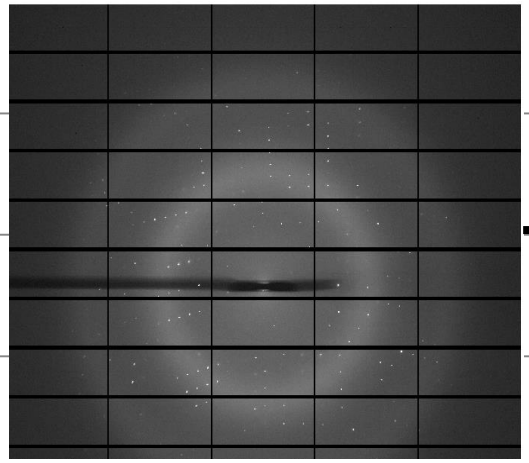
diamond

Diamond Data Archive

DLS Archived Data Volumes (GB)



MicroCT scan of a tropical bee's compound eye, imaged to understand how they navigate dense rainforest (G Taylor, E Baird, AJ Bodey, *et al*).



Diffraction pattern of an RNase 4 crystal at 1.6A, from Prof K. Ravi Acharya's group (at University of Bath)

Data Archived (GB)

4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0

3PB 1/15

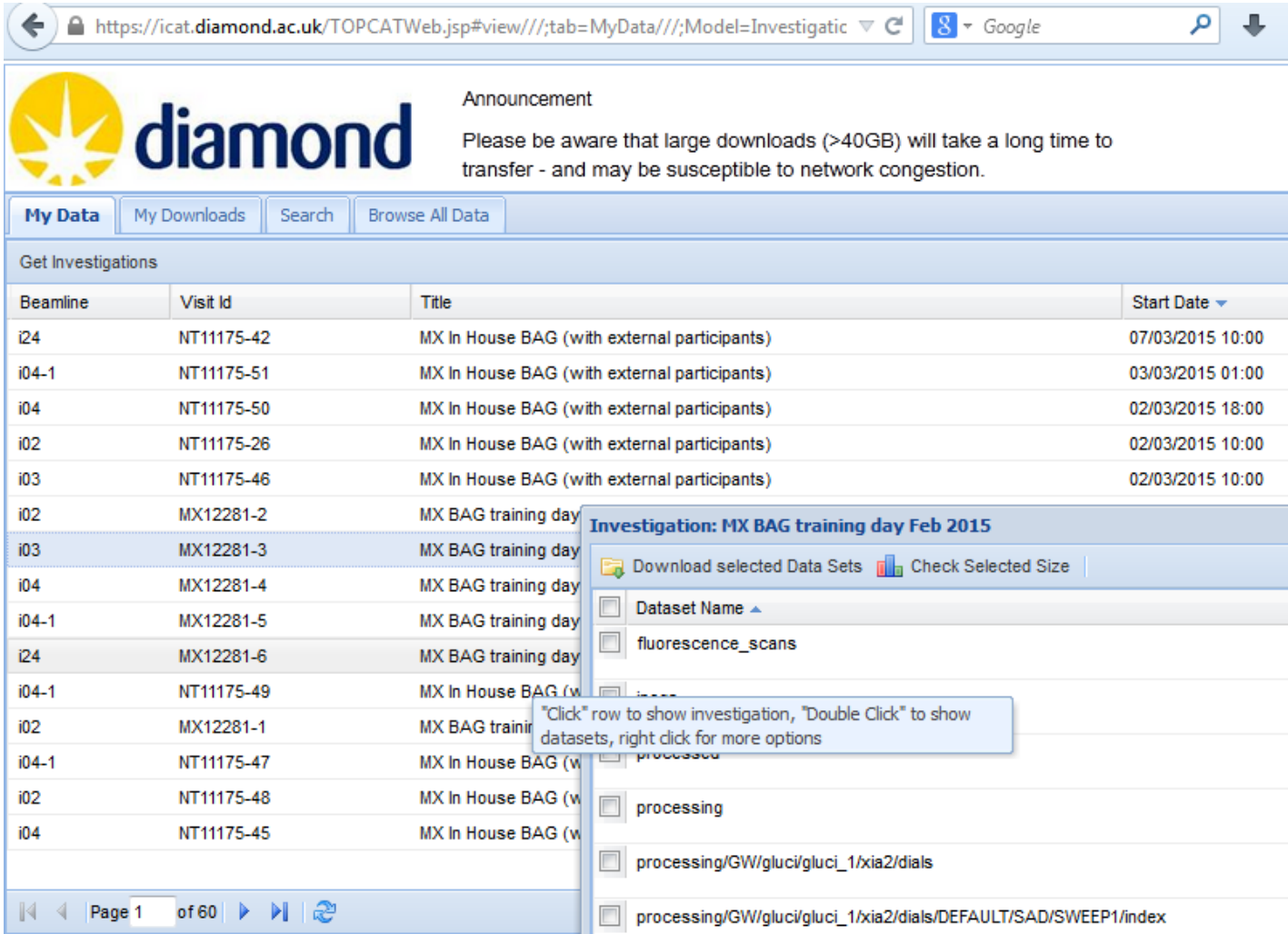
1PB 7/13

Jan-07 Jul-07 Jan-08 Jul-08 Jan-09 Jul-09 Jan-10 Jul-10 Jan-11 Jul-11 Jan-12 Jul-12 Jan-13 Jul-13 Jan-14 Jul-14 Jan-15 Jul-15

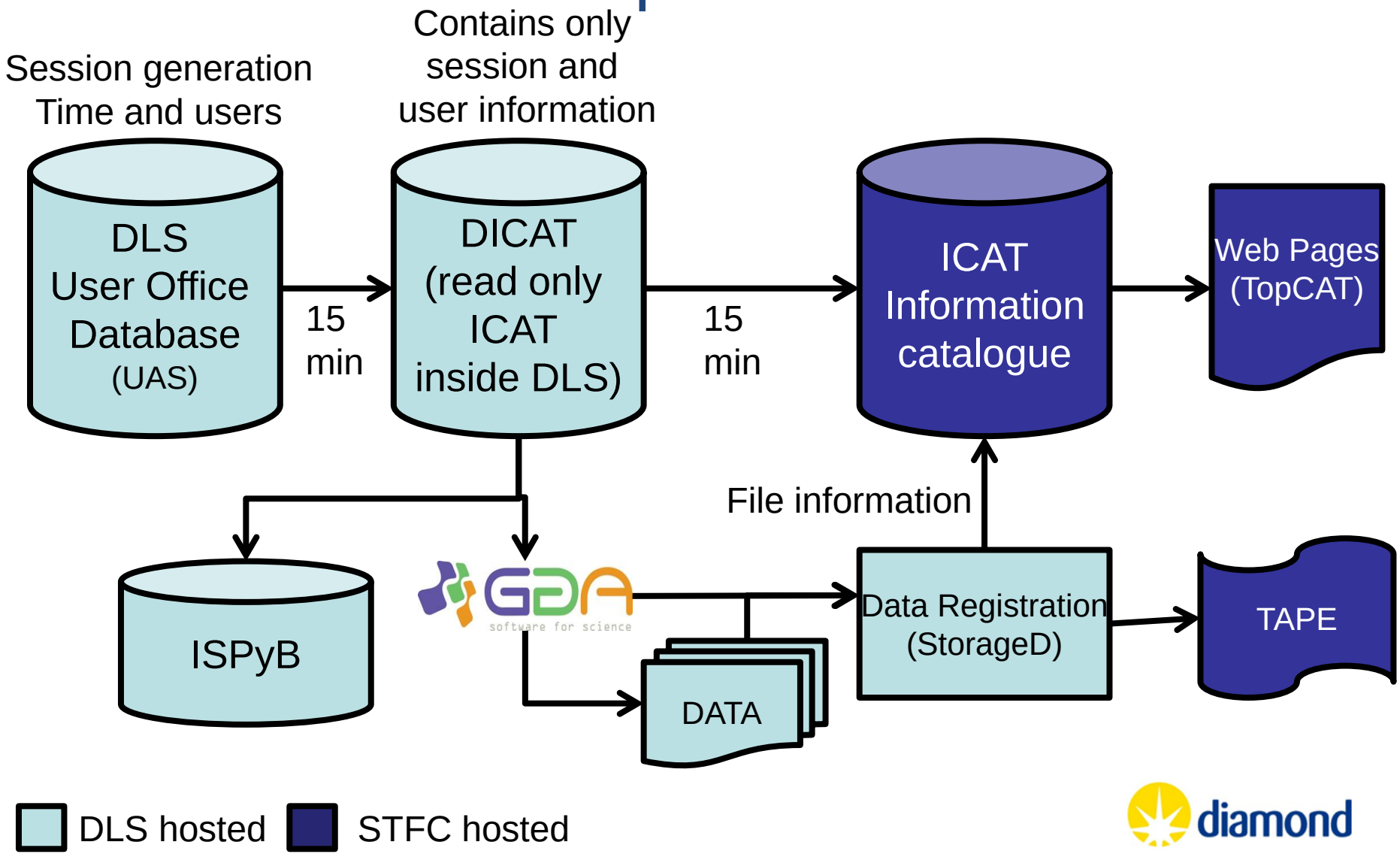
Access to Diamond Data Archive

http://icat.diamond.ac.uk

Ongoing work to improve reliability, usability, scalability and downloading

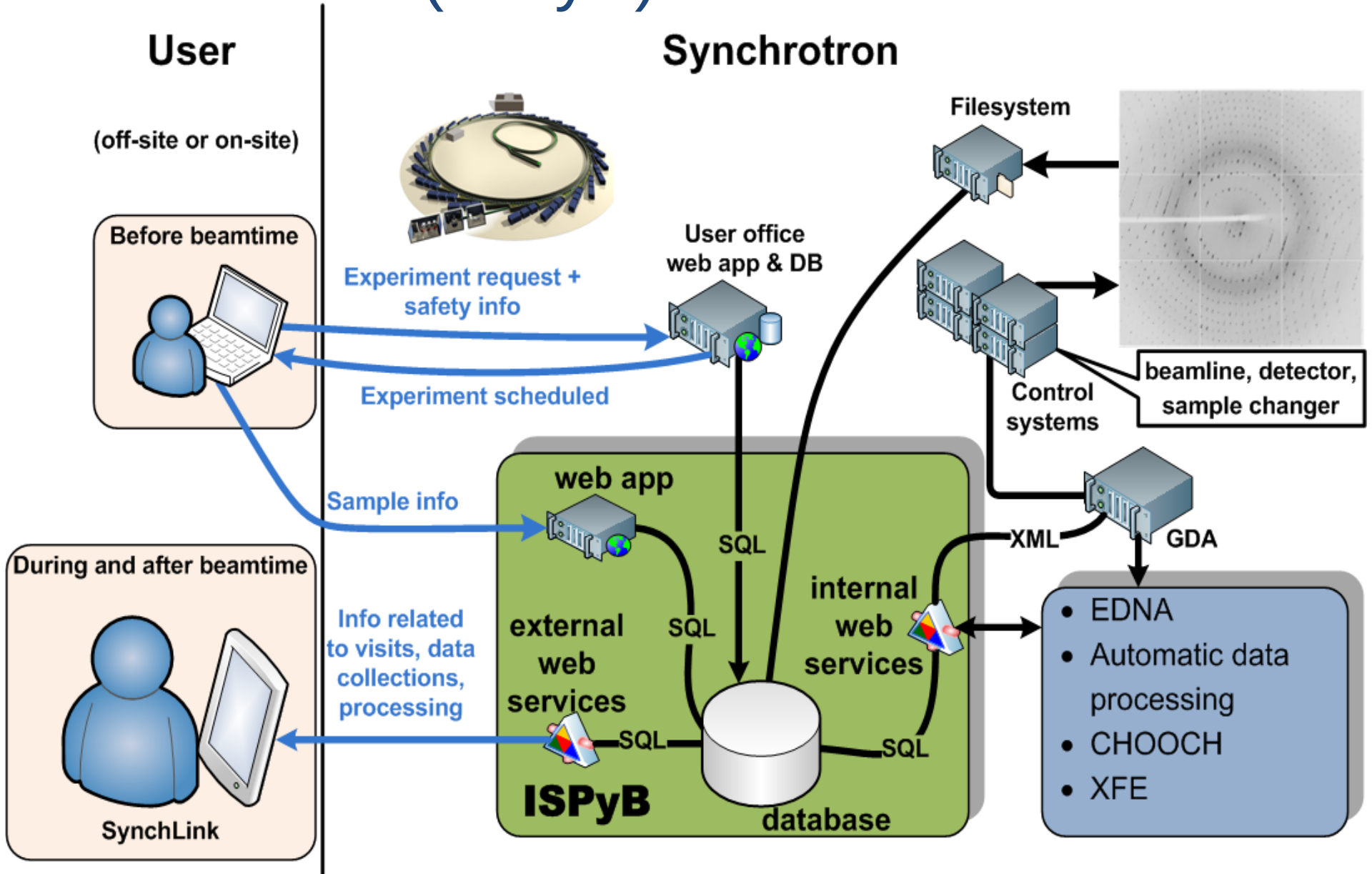


Data is captured from every stage of the process

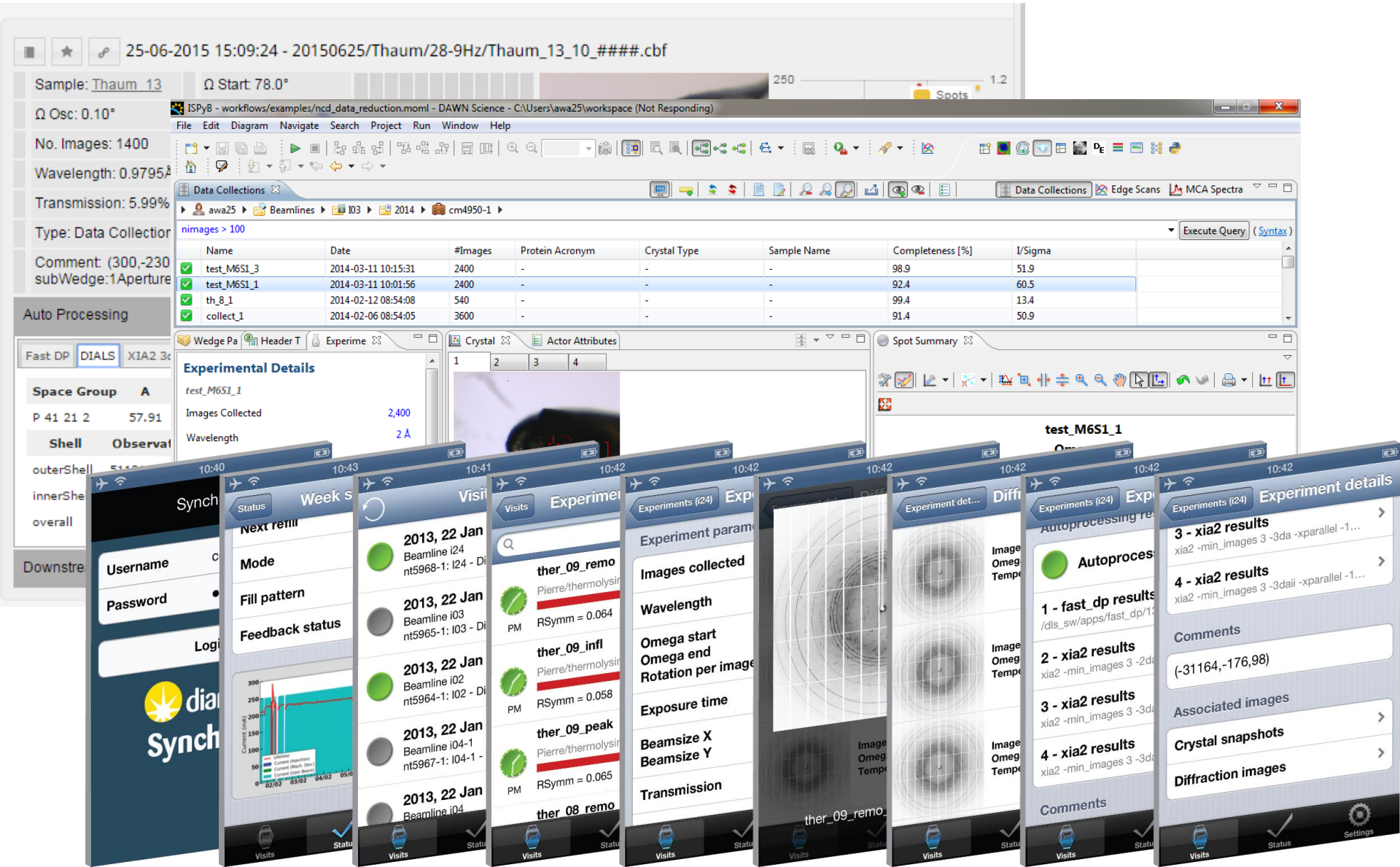


LIMS – Experiment Information Management.

How LIMS (ISPyB) is connected in DLS

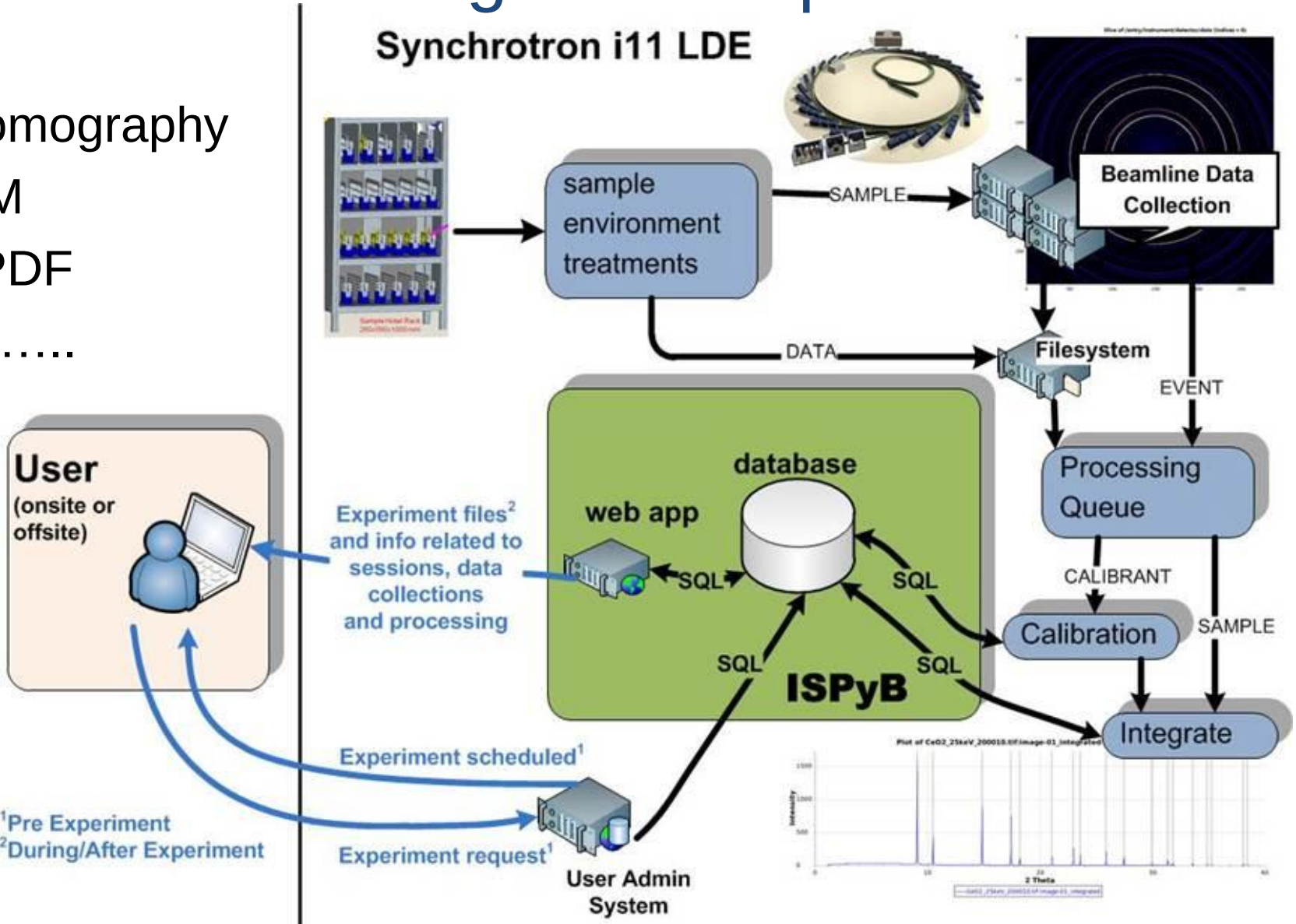


Interfaces


25-06-2015 15:09:24 - 20150625/Thaum/28-9Hz/Thaum_13_10_####.cbf
Sample: Thaum_13
Ω Start: 78.0°
Ω Osc: 0.10°
No. Images: 1400
Wavelength: 0.9795Å
Transmission: 5.99%
Type: Data Collection
Comment: (300,-230
subWedge:1Aperture
Auto Processing
Fast DP
DIALS
XIA2 3d
Space Group
P 41 21 2
57.91
Shell
outerShell
innerShe
overall
Downstre
ISPyB - workflows/examples/ncd_data_reduction.moml - DAWN Science - C:\Users\awa25\workspace (Not Responding)
File Edit Diagram Navigate Search Project Run Window Help
Data Collections
Edge Scans
MCA Spectra
awa25 Beamlines I03 2014 cm4950-1
nimages > 100
Execute Query
Syntax
Name
Date
#Images
Protein Acronym
Crystal Type
Sample Name
Completeness [%]
I/Sigma
test_M6S1_3 2014-03-11 10:15:31 2400 98.9 51.9
test_M6S1_1 2014-03-11 10:01:56 2400 92.4 60.5
th_8_1 2014-02-12 08:54:08 540 99.4 13.4
collect_1 2014-02-06 08:54:05 3600 91.4 50.9
Experimental Details
test_M6S1_1
Images Collected 2,400
Wavelength 2 Å
Crystal
Actor Attributes
Spot Summary
test_M6S1_1
Synch
Username
Password
Status
Week s
Next refill
Mode
Fill pattern
Feedback status
Visits
Experiments (i24)
Experiment details
Images collected
Wavelength
Omega start
Omega end
Rotation per image
Exposure time
Beamsize X
Beamsize Y
Transmission
3 - xia2 results
4 - xia2 results
Comments
(-31164,-176,98)
Associated images
Crystal snapshots
Diffraction images
Settings

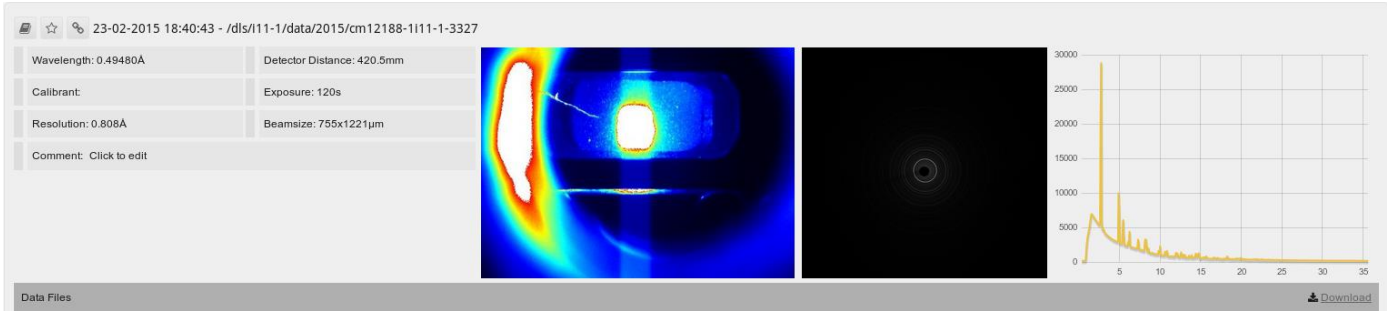
Accommodating more experiments

- Tomography
- EM
- xPDF
- ……..



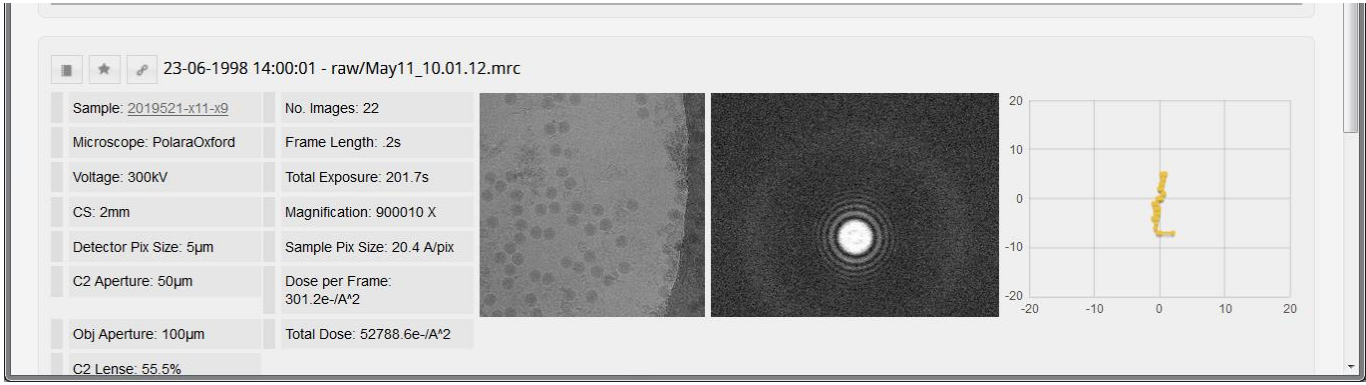
…With interfaces…

- Powder Diffraction



- Tomography

- EM



The RAW and processed data files

NeXus for RAW data

- Diamond Light Source is still committed to establishing NeXus as its primary file format for its facilities.
  - Key drivers are both scientific and technical.

- A combination of the Data Acquisition Software (GDA) and controls software EPICS available on all photon beamlines facilitates data collection in NeXus

diamond

HDF5 improvements for NeXus

- Diamond with other partners are funding improvement in HDF5 for
  - SWMR (Single Writer Multiple Reader)
  - Virtual Dataset, ability to link together multiple data sources/files.
- Diamond have contributed to h5py and Java data bindings to allow use of SWMR.

diamond

NeXus for experiments

- Diamond is working with communities to establish standards for all discipline areas e.g.
  - nxMX (investigating bridging imgCIF/CBF)
    - Andreas Förster: EIGER HDF5 data and NeXus format
    - Herbert J. Bernstein: Common diffraction image metadata specification in imgCIF, HDF5 and NeXus
  - Using nxTomo, nxARPES including data processing
  - Used for Non Crystalline Diffraction and soon for Pair Distribution Function experiments
  - nxSTXM, nxFLUO also being investigated

diamond

Future integration.

- Continue move towards full NeXus defined RAW and processed (including provenance) data
- Record more experiment and experiment types in a LIMS
- Harvest more information from raw data files and from LIMS to suitably enrich long term data archive
- Better links into external databases and repositories such as wwPDB, CCDC and other facility catologues.

diamond

Summary

- The LIMS systems at Diamond record all aspects of the experiment but not the raw data. This will eventually cover all experiments.
- The ‘files’ Diamond experiments produce should all eventually adhere to a descriptive standard
- The Diamond data archive should provide a link between the data and reasonable experiment parameters/results

Thanks

- NIAC: NeXus International Advisory Committee
- HDF5 Group
- ISPyB Collaborators
- STFC and ICAT Collaborators
- Diamond computing, software and science groups