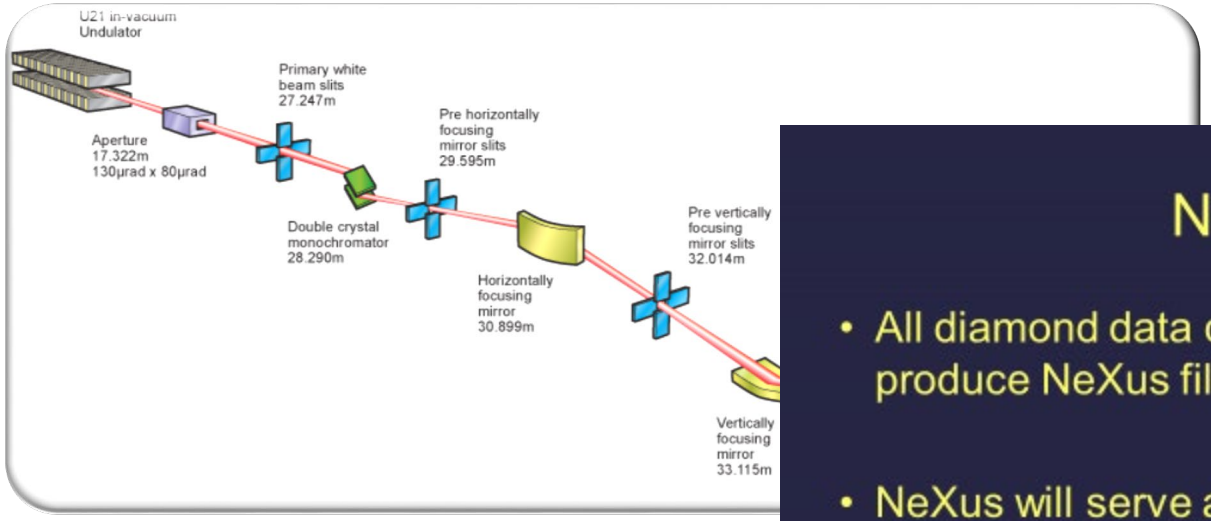
NeXus at DLS

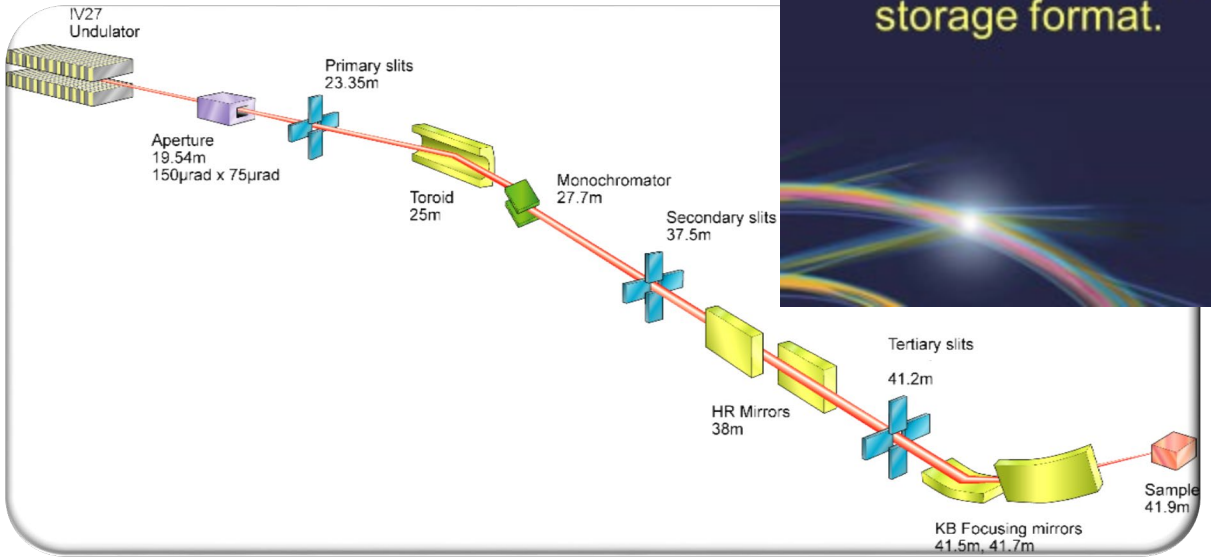
Alun Ashton

Mark Basham, Graeme Winter, Jake Filik, Peter Chang

diamond

Original Motivation

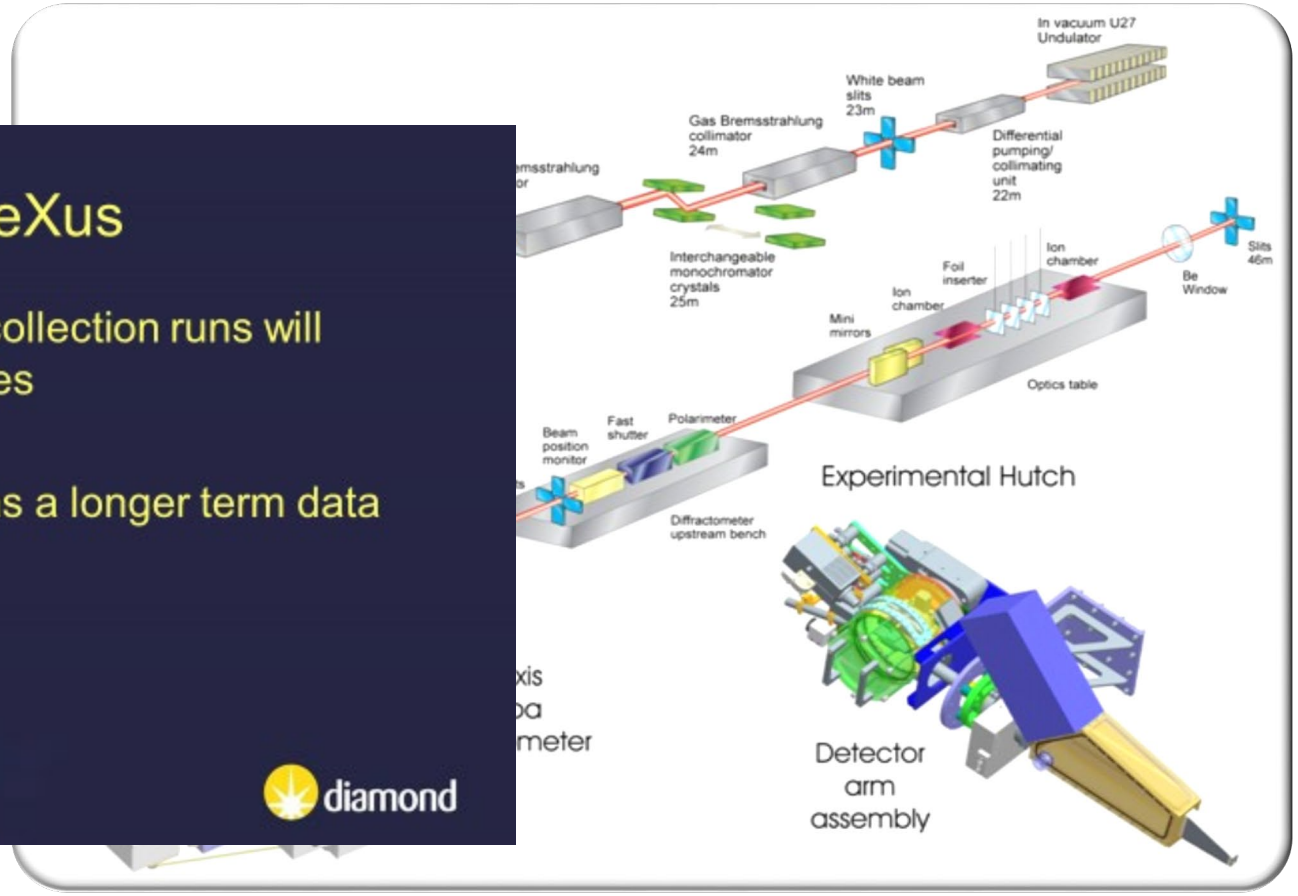




NeXus

- All diamond data collection runs will produce NeXus files
- NeXus will serve as a longer term data storage format.

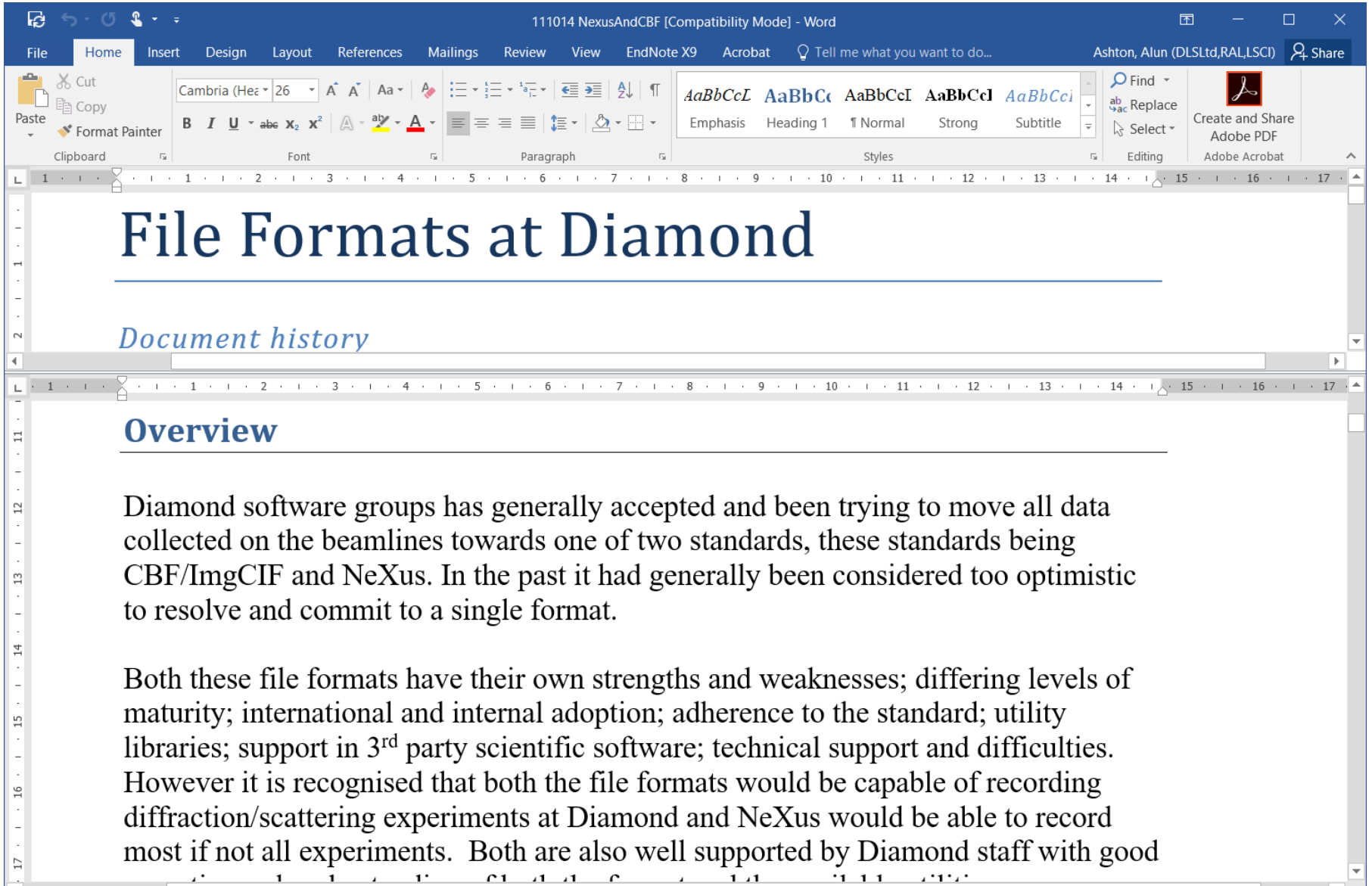
diamond



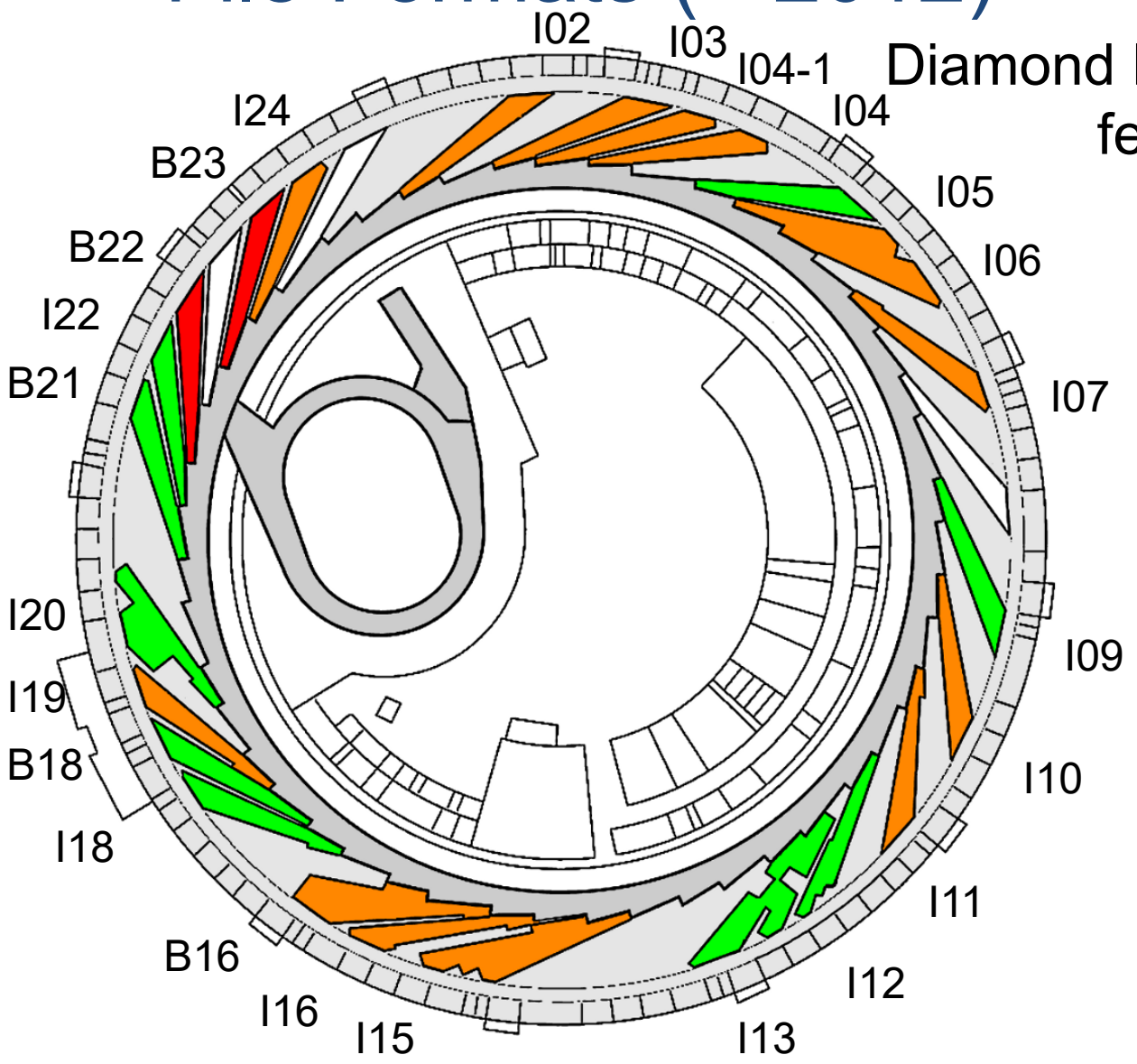
- Well described measurements
- Clean beamline agnostic reduced data

diamond

2011 File Formats at Diamond Report



File Formats (< 2012)



Diamond has a policy of, where feasible, to standardise on file formats, the choice being NeXus/HDF5

Green: predominantly using NeXus.

Orange: Mixed NeXus and other formats or considering NeXus in the next 12 months.

Files can be generated by Detector, EPICS or Data Acquisition

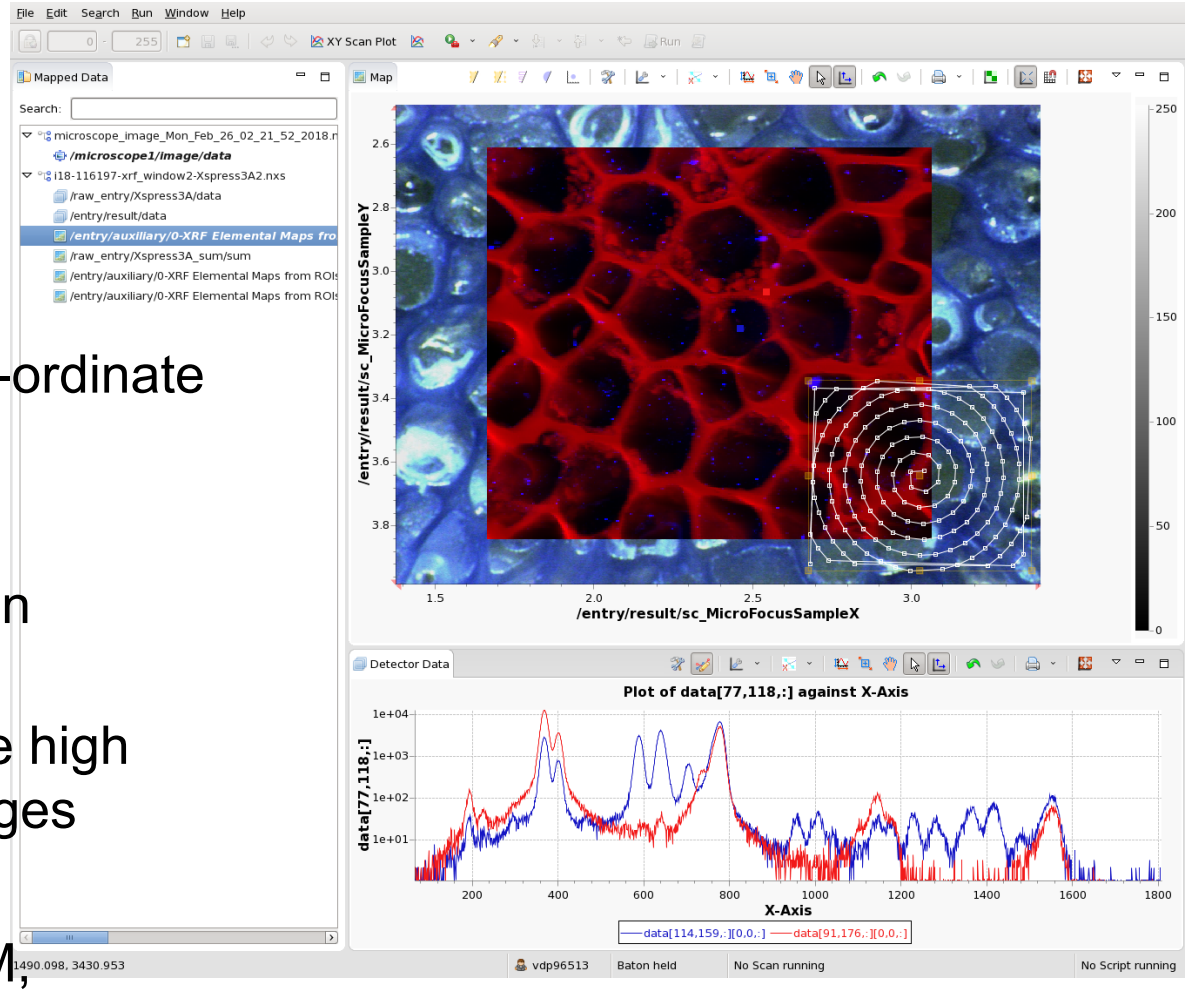
Recent Motivation/Drivers/Priority

- Increased data volumes
- Live visualisation
- Increased automation

diamond

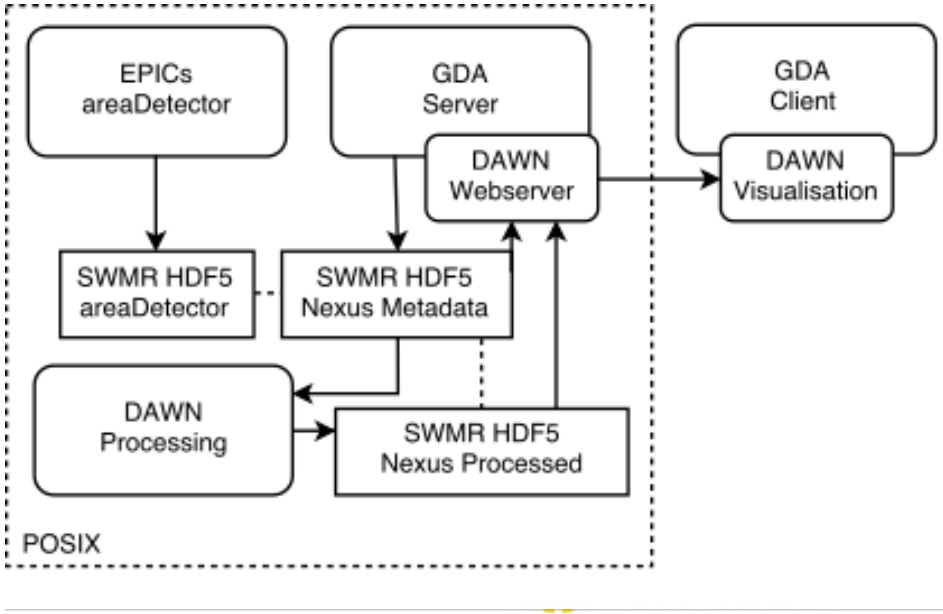
Live Grid Scan Visualisation

- HDF5 SWMR allows visualisation during scan
  - Raw and processed data
- Nexus tagging used to identify:
  - Appropriate data for visualisation
  - Dimensions of data to visualise in sample co-ordinate
  - Complex scan trajectories – spirals
  - RGB images from optical microscopes
- Since sample stage co-ordinate used visualisation independent of scan resolution
  - Overlay – Coarse sample location scans, fine high resolution scans and optical microscope images
- Required tags independent of experiment
  - Same UI used for – XRF, XRD, SAXS, STXM, Ptychography, ARPES, FTIR…
  - Consistent sample stage coordinates.
  - No application definition.



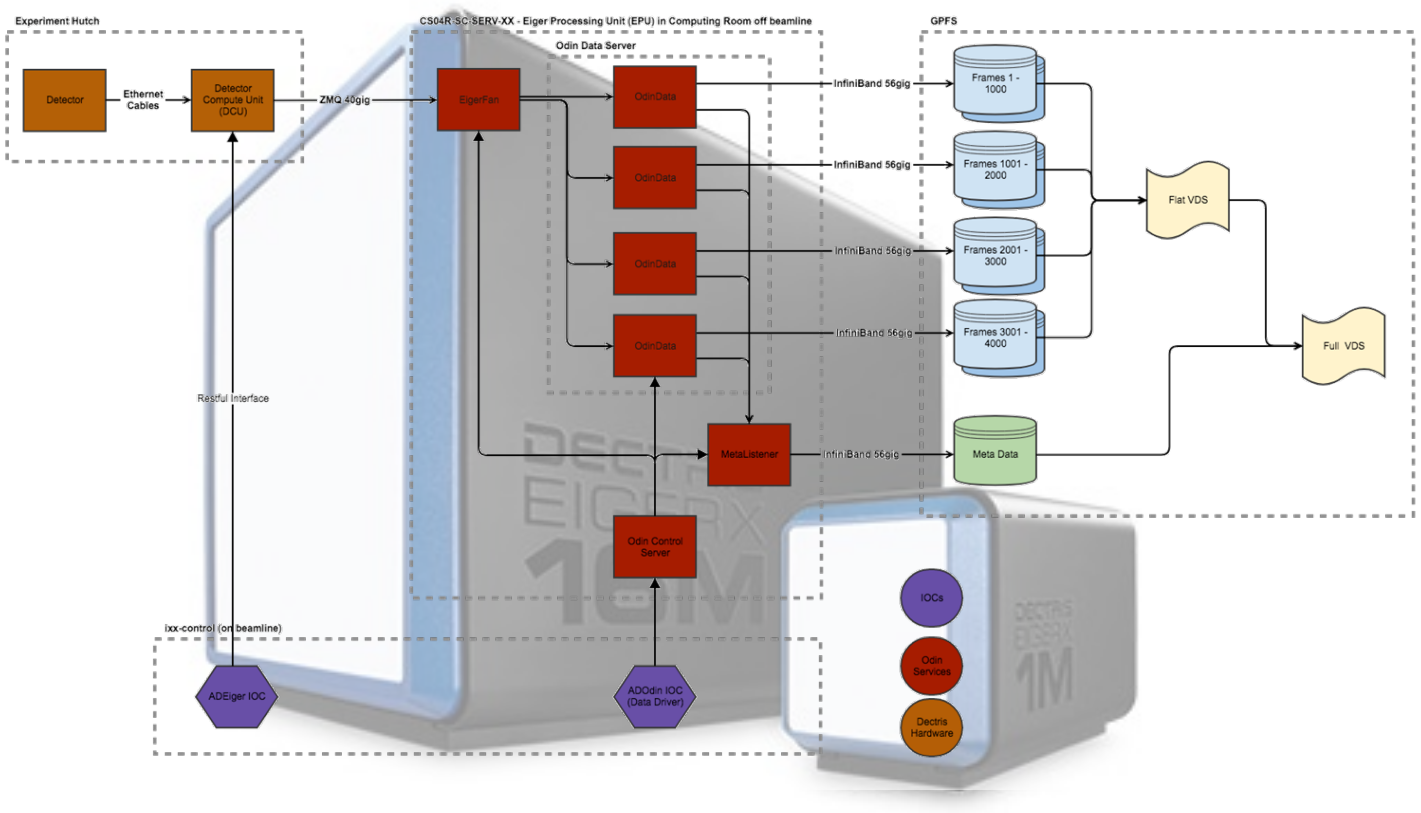
Live Grid Scan Visualisation - SWMR

- Not so simple as “SWMR lets you read the data as its being written”
- Has to be across POSIX compliant mounts
- Cluster and control machines all GPFS – SWMR works
- Visualisation clients mount file system as NFS – SWMR doesn’t work
- Webserver on control machine sends requested visualisation datasets (small) back to clients
- Currently deployed on 5 beamlines (first in 2016)
- XRF, XRD, SAXS, STXM, Ptychography…



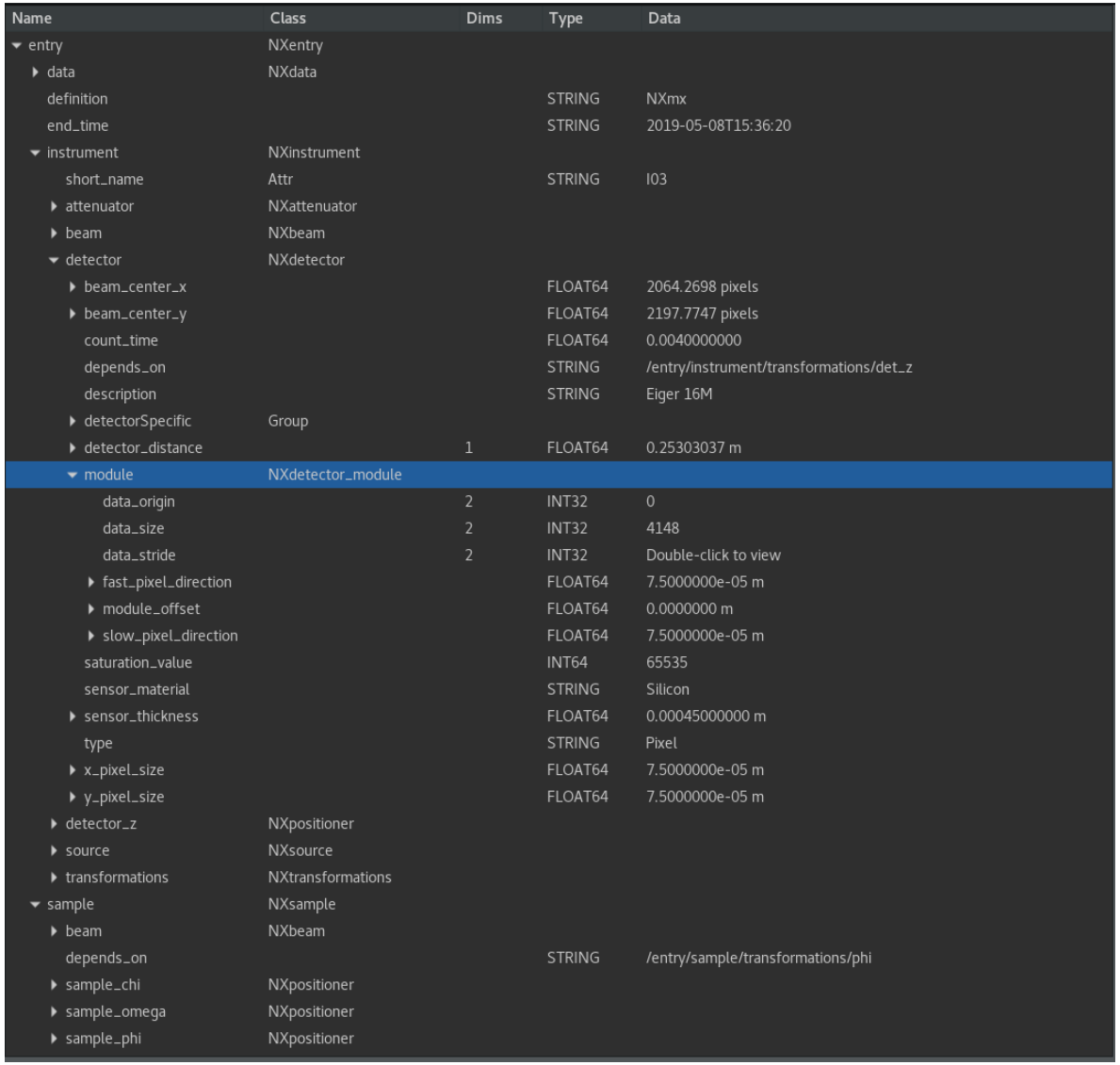
“Moving” from ImgCIF to NeXus

- Writing both NXmx standard and manufacturer NeXus pointing to data files.
- NXmx.nxs has the depends_on attribute to allow reconstruction of a full model of the beamline => comparable to using full imgCIF format.
- Work needed for VDS inverse beam and 3D masks for static and dynamic shadows.



Science Applications

- Detector position/orientation specification in NXmx used in MX, XRD and SAXS/WAXS allowing common tools to be used across techniques.
- NXtomo used on all tomography beamlines allowing automatic reconstruction.
- Nxcxi_ptycho Coming soon, and will allow significantly reduced user input.



Name	Class	Dims	Type	Data
entry	NXentry			
data	NXdata			
definition			STRING	NXmx
end_time			STRING	2019-05-08T15:36:20
instrument	NXinstrument			
short_name	Attr		STRING	I03
attenuator	NXattenuator			
beam	NXbeam			
detector	NXdetector			
beam_center_x			FLOAT64	2064.2698 pixels
beam_center_y			FLOAT64	2197.7747 pixels
count_time			FLOAT64	0.0040000000
depends_on			STRING	/entry/instrument/transformations/det_z
description			STRING	Eiger 16M
detectorSpecific	Group			
detector_distance		1	FLOAT64	0.25303037 m
module	NXdetector_module			
data_origin		2	INT32	0
data_size		2	INT32	4148
data_stride		2	INT32	Double-click to view
fast_pixel_direction			FLOAT64	7.5000000e-05 m
module_offset			FLOAT64	0.0000000 m
slow_pixel_direction			FLOAT64	7.5000000e-05 m
saturation_value			INT64	65535
sensor_material			STRING	Silicon
sensor_thickness			FLOAT64	0.00045000000 m
type			STRING	Pixel
x_pixel_size			FLOAT64	7.5000000e-05 m
y_pixel_size			FLOAT64	7.5000000e-05 m
detector_z	NXpositioner			
source	NXsource			
transformations	NXtransformations			
sample	NXsample			
beam	NXbeam			
depends_on			STRING	/entry/sample/transformations/phi
sample_chi	NXpositioner			
sample_omega	NXpositioner			
sample_phi	NXpositioner			

NXProcess – used to persist sequences of processing steps in DAWN and SAVU

NeXus/HDF5 Tree

Name	Class	Dims	Type	Data
entry1	NXentry			
processed	NXentry			
process	NXprocess			
0	NXnote			
data	SDS		STRING	{"filePath":"/dls/i22/data/2019/cm22951-1/processing/SAXS_calibration.nxs"}
id	SDS		STRING	uk.ac.diamond.scisoft.analysis.processing.operations.DiffractionMetadataImportOperati
name	SDS		STRING	Import Detector Calibration
passed	SDS	1	INT32	0
saved	SDS	1	INT32	0
type	SDS		STRING	application/json
1	NXnote			
2	NXnote			
3	NXnote			
4	NXnote			
5	NXnote			
6	NXnote			
7	NXnote			
data	SDS		STRING	{"pixelSplitting":false,"
id	SDS		STRING	uk.ac.diamond.scisoft.
name	SDS		STRING	Azimuthal Integration
passed	SDS	1	INT32	0
saved	SDS	1	INT32	0
type	SDS		STRING	application/json
8	NXnote			
9	NXnote			
date	SDS		STRING	2019-01-30T17:58
origin	NXnote			
program	SDS		STRING	DAWN
version	SDS		STRING	2.11.0.v20181121-082
result	NXdata			

NeXus/HDF5 Tree

Name	Class	Dims	Type	Data
entry	NXentry			
final_result_tomo	NXdata			
framework_citations	NXcollection			
input_data	NXcollection			
intermediate	NXcollection			
plugin	NXprocess			
1	NXnote			
2	NXnote			
3	NXnote			
4	NXnote			
active	SDS	1	INT8	1
citation	NXcite			
data	SDS	1	STRING	{"in_datasets": ["tomo"], "reg_par": 0.0, "ratio": 0.95, "log": true, "algorithm": "gridrec", "out_d...
desc	SDS	1	STRING	{"in_datasets": "Create a list of the dataset(s) to process.", "reg_par": "Regularization parameter ...
hide	SDS	1	STRING	[]
id	SDS	1	STRING	savu.plugins.reconstructions.tomopy_recon
name	SDS	1	STRING	TomopyRecon
user	SDS	1	STRING	["reg_par", "log", "algorithm", "filter_name", "preview", "centre_of_rotation"]

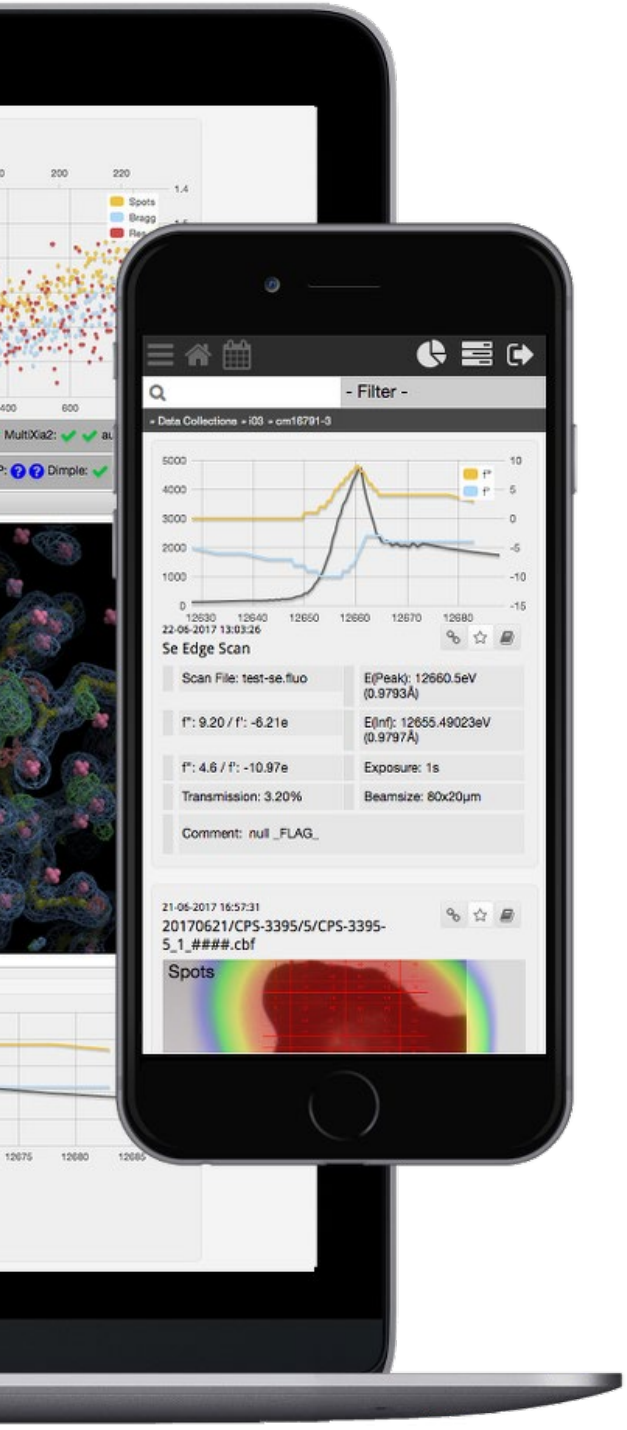
What tools are we using?

- DAWN (I) Visualisation and Automated processing for SAXS, WAXS, XRF, XRD, XANES, reciprocal space remapping….
- Savu (I/E) Tomography (absorption, diffraction, florescence)
  - PyFAI, PyMCA, TomoPy, Astra……
- Ptypy (E) Ptychography
- Dials / XIA2 (I/E)
- HDF5 support with Mantis, Igor, Matlab, PyMCA…..

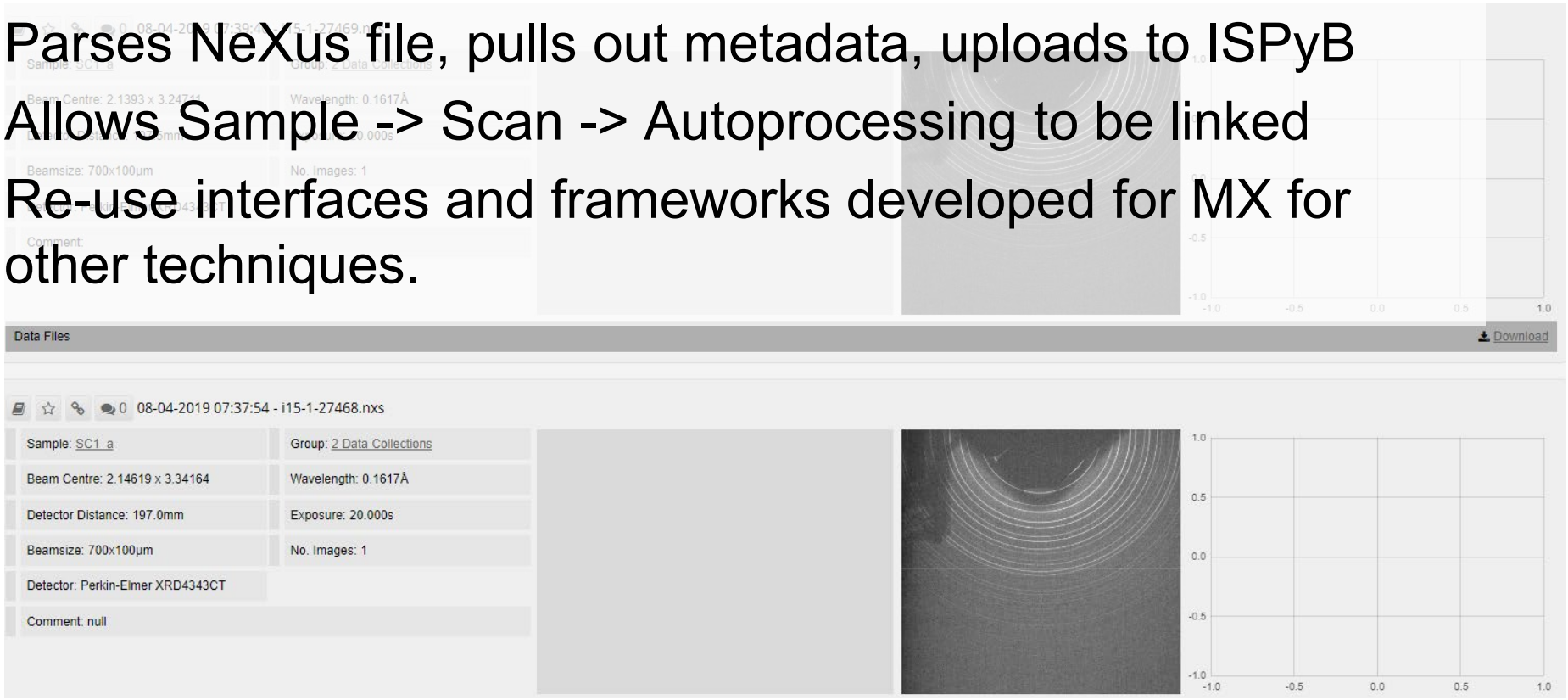
- A lot of these tools are automatically processing and ‘understanding the data’ without interacting with the users.

diamond

NeXus and ISPyB



- ISPyB traditionally database for MX experiments
- DataCollections, Sample, Detector etc. are quite generic
- ActiveMQ microservice subscribed to GDA scan topic
- Parses NeXus file, pulls out metadata, uploads to ISPyB
- Allows Sample -> Scan -> Autoprocessing to be linked
- Re-use interfaces and frameworks developed for MX for other techniques.



Lessons Learned from using HDF5 - NeXus

SWMR

- Required POSIX compliance can be expensive
  - One Scan -> 3 Detectors -> Three servers (+ control machine) -> all need to be GPFS
  - Debugging accidental NFS mount can be fun (doesn't work or works with delay…)
- Need careful tuning of flush rate vs chunk size vs compression for best performance

Virtual Datasets

- Powerful tool
- Optimise file/dataset structure for fastest possible data writing
- Use VDS to reshape and rearrange data to better reflect the experiment
- Again, need to be careful – can slow data read

Chunking

- Never use a chunk size of one
- Needs careful optimisation for special use cases e.g. tomography
- Reading data can be slow (NFS vs GPFS), even when data not large?

Versions

- SWMR, VDS need hdf5 1.10 – not backwards compatible, patchy uptake.

NXmx/Dectris

diamond

Diamond Overview – Current

- 5 Beamlines writing post 2014 axis tagging
- MX eiger + 2 + 1 writing application definitions from acquisition software
- 24 using core original NeXus writer
  - Pre-2014 tagging
- Only one scan per file -> file flagged for archive at end of scan
- Some beamlines also write legacy text/tiff formats (possibly only 2 or 3 without any NeXus).
- NeXus usefulness varies hugely per beamline
- More popular on beamlines that write multi-dimensional data, multiple datasets per file – i.e. all the benefits of HDF5
- Main issue against adoption is compatibility with common downstream analysis applications

diamond

Motivation for the future

Diamond NeXus Working Group

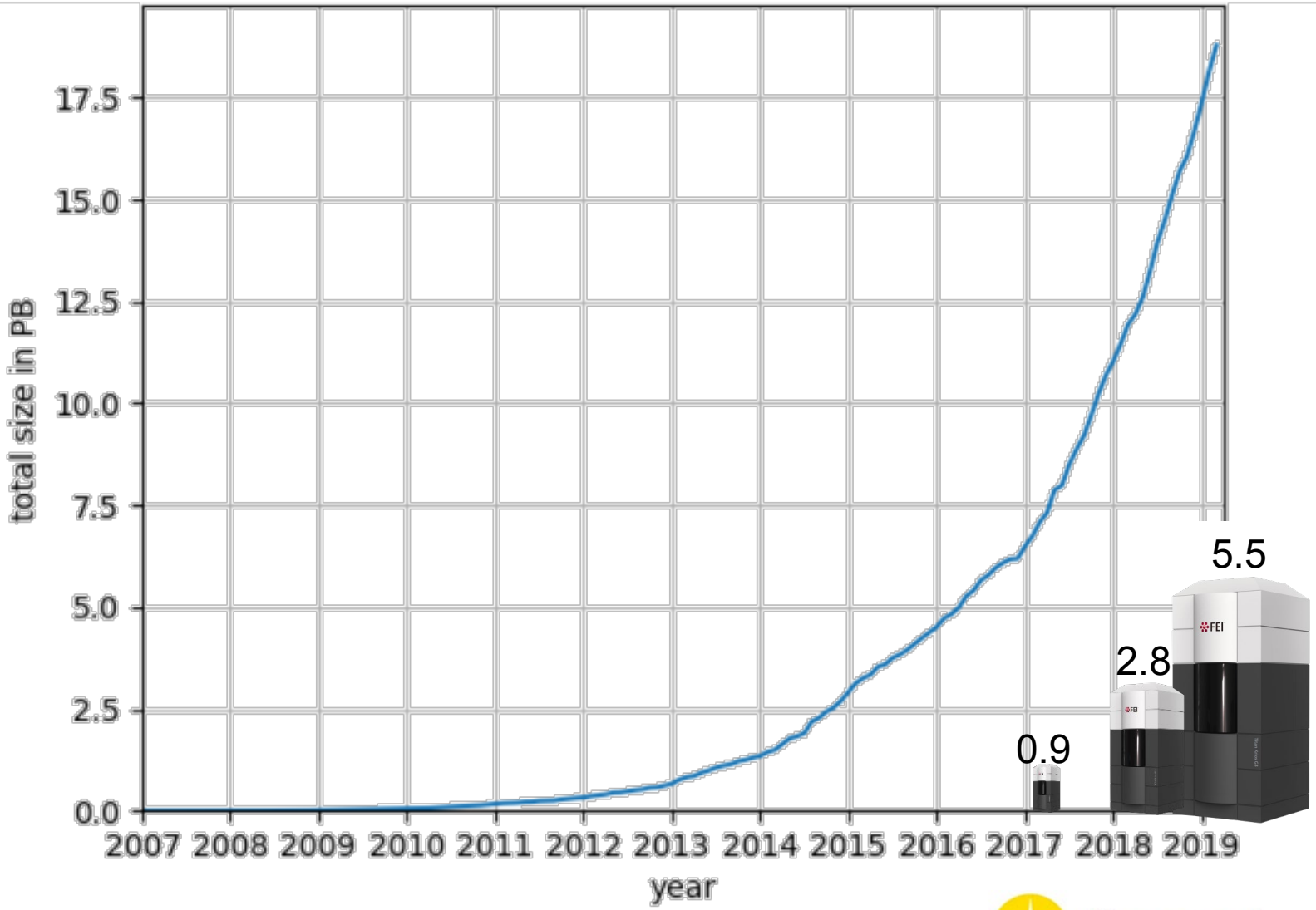
Chair: Steve Collins

To coordinate science input to Diamond NeXus development, support a coordinated approach with other facilities worldwide, and report and publicize Diamond's progress towards full adoption of NeXus files and philosophy.

- Coordinate universal adoption of a common file format for raw and processed data (hdf5)
- Coordinate universal adoption of a common metadata model (NeXus classes)
- Common approach to data processing workflows (including file validation for workflows)
- Promotion of FAIR principles for raw and processed data

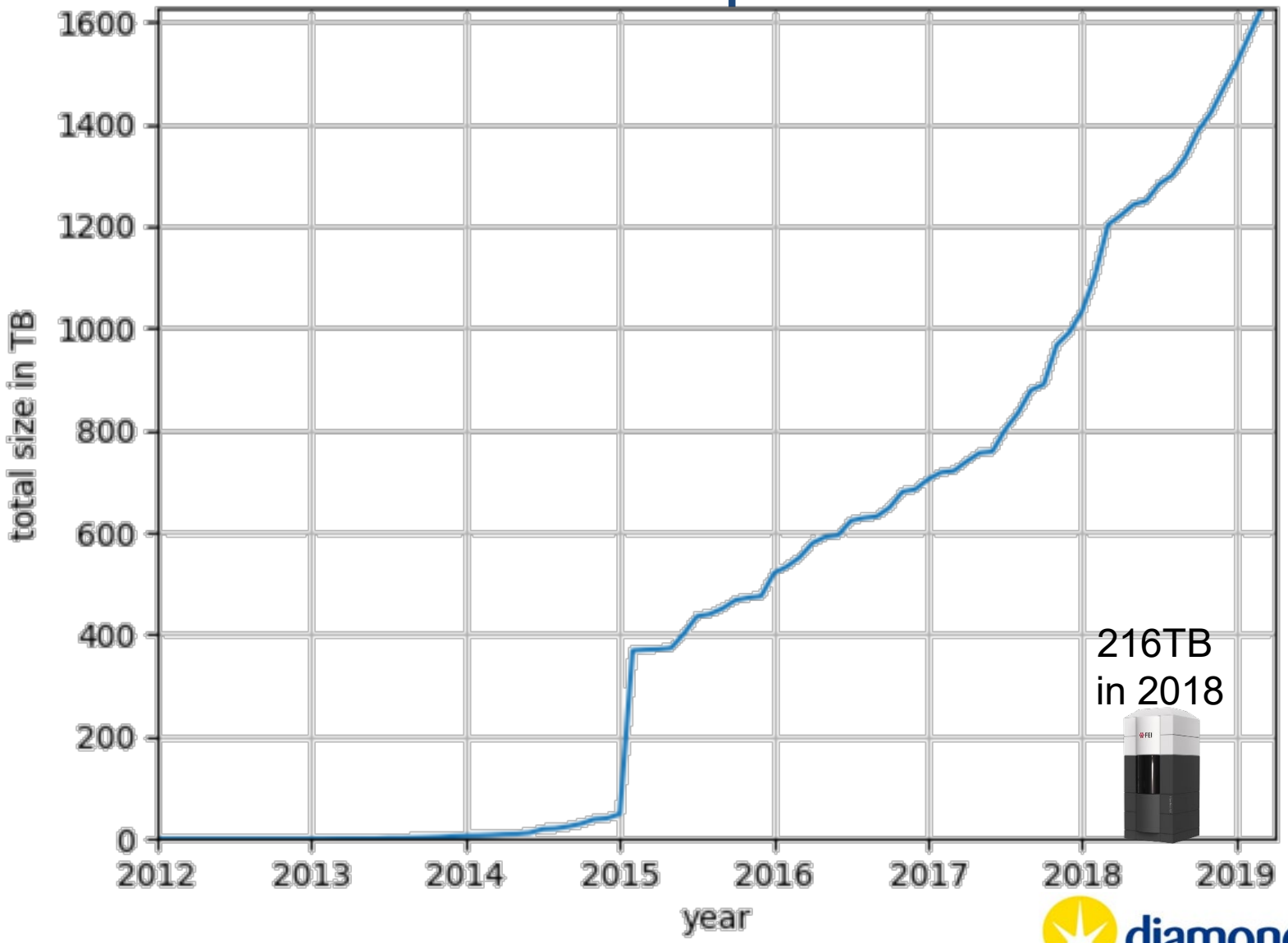
diamond

Total data archived from Diamond



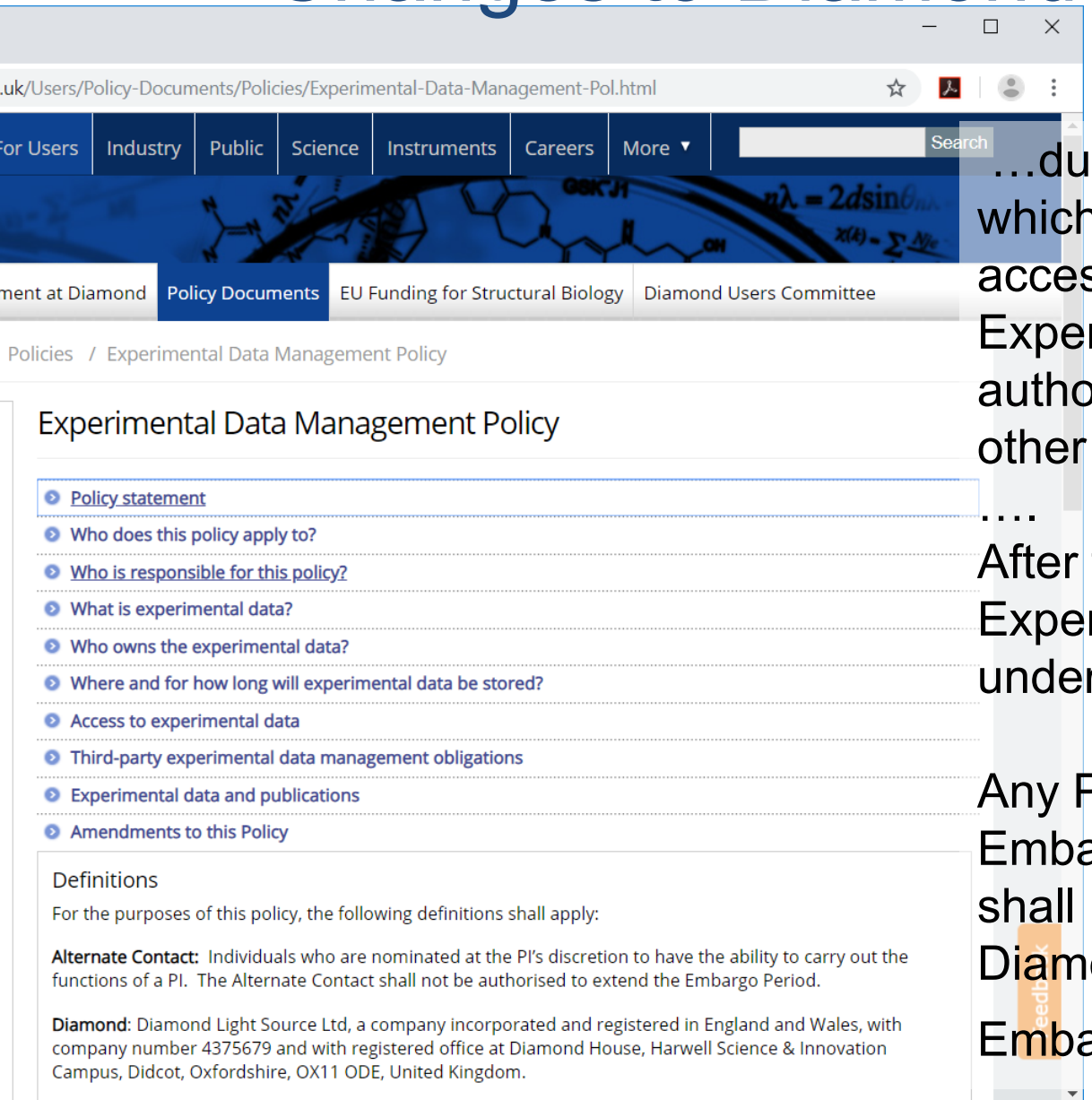
diamond

Retrieval plot



diamond

Changes to Diamond's experiment data policy



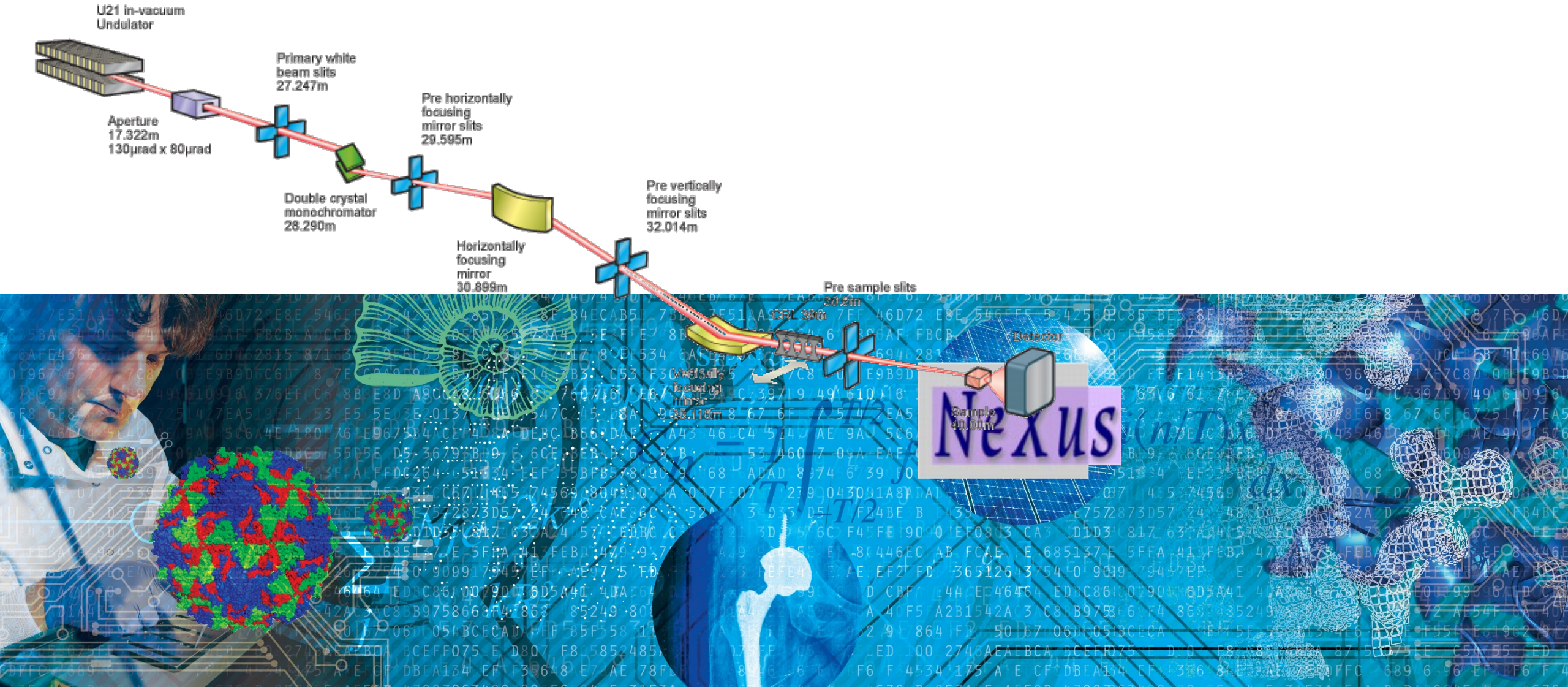
…during the Embargo Period (3 years from the date on which the User Generated Data is produced at Diamond), access to the Experimental Data will be restricted to the Experimental Team, and Diamond Employees and authorised Diamond service providers, for support and other facility related purposes.

….

After the Embargo Period, Diamond may make Experimental Data available on an open access basis under a CC-BY-4.0 licence.

Any PI or PI's Establishment that wishes to extend the Embargo Period for an additional 12 months at a time shall be entitled to submit an advance written request to Diamond's User Office within 3 months of the end of the Embargo Period…."

FAIR will need richer and contextual information



ada lovelace centre

An integrated, cross-disciplinary data intensive science centre, for better exploitation of research carried out at National Facilities including

- Diamond Light Source (DLS),
- ISIS Neutron and Muon Facility,
- Central Laser Facility (CLF),
- Culham Centre for Fusion Energy (CCFE).

Aim: to transform research at the facilities through a multi-disciplinary approach to data processing, computer simulation and data analytics. It will provide computing hardware, build software and provide computational and data analytics expertise that will spark a paradigm shift in the capability of scientists to design, analyse and interpret experiments.

diamond

Themes

The ALC will significantly enhance capability to support the Facilities' science programme

- Theme 1: Increases capacity in advanced software development for data analysis, interpretation, simulation and modelling
- Theme 2: Develop new generation of scientific data experts and scientific software engineers and research software engineers who can interact with science domain experts
- Theme 3: Provide significant compute infrastructure for managing, analysing and simulating the data generated by the facilities and for designing next generation Big-Science experiments

Focus is the science drivers and computational needs of Facilities

PaN community: PaNOSC, ExPaNDS

- P: WP2: Definition and adoption of common open standards for interoperability. Registering with and citing of these standards by standards bodies and publishers.
- P: WP3 (D3.5): Definition of standard metadata for scientific domains at the partner facilities to access to data beyond the generic search features of OpenAIRE, enabling new and interdisciplinary research leading to new insights and innovation for the society at large.

- E: WP2: Developing a common metadata framework to support FAIR data within the Photon and Neutron science community.
- E: WP3 (D3.2): Develop baseline standards for Metadata Catalogues based on harmonised policies (cf. WP2) to make Photon and Neutron data FAIR

Engaging User and Science Domains

- Instruct-ERIC, Instruct-ULTRA
  - Increase the quality and integration of the data being acquired at all installations…..
  - Increase the quality of reported information on data analysis processes.
  - Increase the quality and quantity of structural data deposited in public repositories, ….
  - Increase the openness and sharing of scientific data.

- EOSC-Life: *In EOSC-Life the 13 Biological and Medical Research Infrastructures in Europe join forces to create an open collaborative digital space for life science in the European Open Science Cloud. We will do this by publishing our data as FAIR Data Resources, ……*