Linking raw data with scientific workflow and software repository: some early experience in PanData-ODI

Erica Yang, Brian Matthews

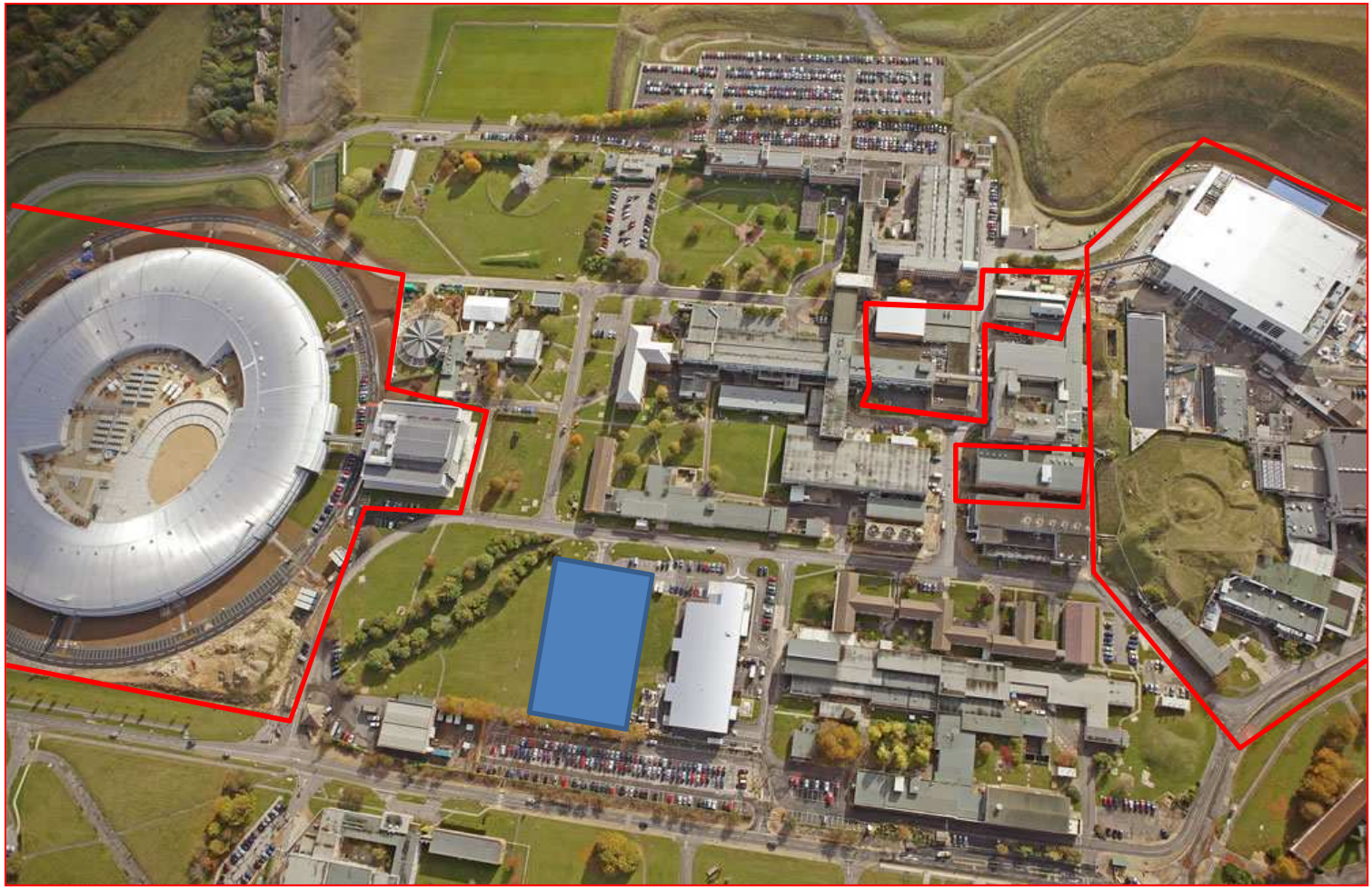
Scientific Computing Department (SCD)

Rutherford Appleton Laboratory (RAL)
Science and Technology Facilities Council (STFC), U. K.

erica.yang@stfc.ac.uk

BACKGROUND

STFC Rutherford Appleton Laboratory



STFC RAL: CLF, Diamond, ISIS, SCD

pandata europe

Science & Technology Facilities Council

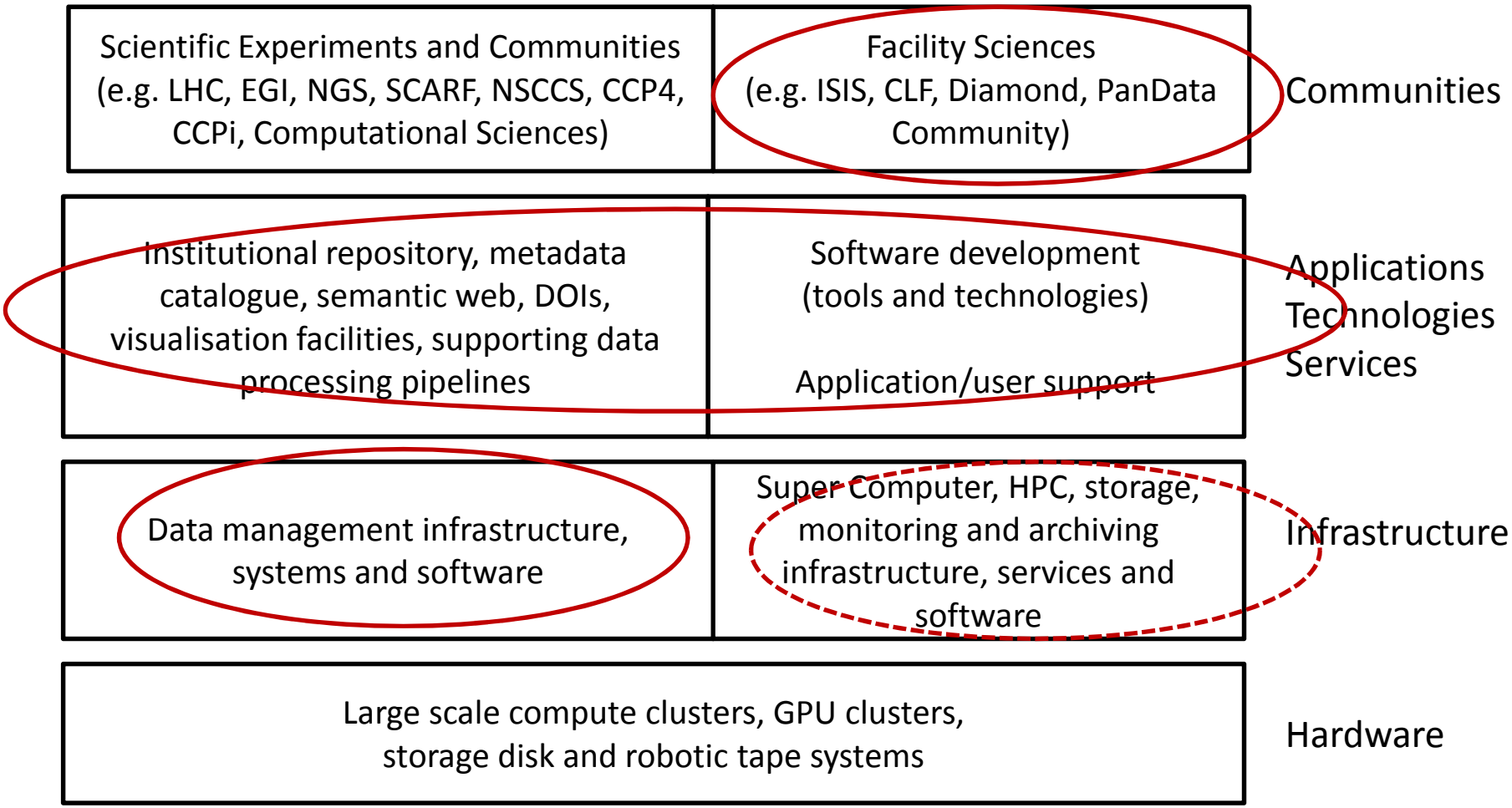
What we do

Scientific Computing Department ...

Scientific Experiments and Communities (e.g. LHC, EGI, NGS, SCARF, NSCCS, CCP4, CCPi, Computational Sciences)	Facility Sciences (e.g. ISIS, CLF, Diamond, PanData Community)	Communities
Institutional repository, metadata catalogue, semantic web, DOIs, visualisation facilities, supporting data processing pipelines	Software development (tools and technologies) Application/user support	Applications Technologies Services
Data management infrastructure, systems and software	Super Computer, HPC, storage, monitoring and archiving infrastructure, services and software	Infrastructure
Large scale compute clusters, GPU clusters, storage disk and robotic tape systems		Hardware

What we do

Scientific Computing Department ...



I am a computer scientist ...

I2S2, SRF, PanData-ODI, SCAPE, PinPoinT, TomoDM

pandata europe

Science & Technology Facilities Council

PANDATA-ODI

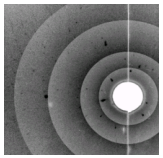
pandata europe

Science & Technology Facilities Council

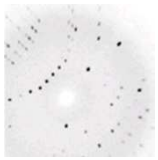
PaN-data ODI– an Open Data Infrastructure for European Photon and Neutron laboratories

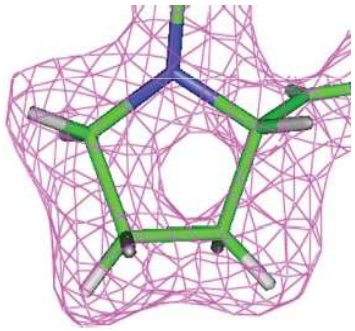
Federated data catalogues supporting cross-facility, cross-discipline interaction at the scale of atoms and molecules

Neutron diffraction



X-ray diffraction





High-quality structure refinement

- Unification of data management policies
- Shared protocols for exchange of user information
- Common scientific data formats
- Interoperation of data analysis software
- Data Provenance WP: Linking Data and Publications
- Digital Preservation: supporting the long-term preservation of the research outputs

Science & Technology Facilities Council ISIS

NEUTRONS FOR SCIENCE Institut Laue-Langevin

DESY

PAUL SCHERRER INSTITUT PSI

ALBA

HELMHOLTZ ZENTRUM BERLIN für Materialien und Energie

ESRF

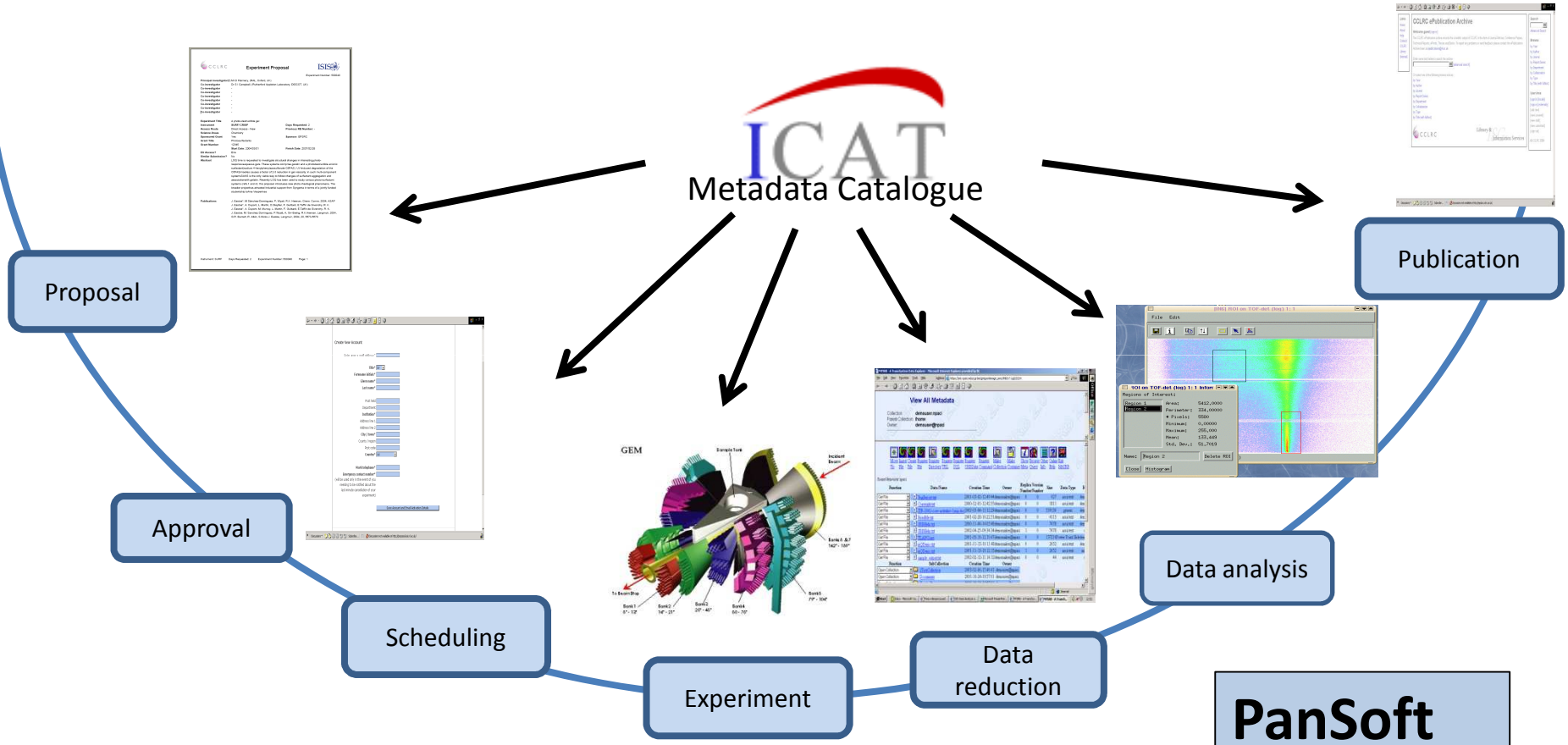
diamond

elettra

SOLEIL SYNCHROTRON

Laboratoire Léon Brillouin

(Facility) Data Continuum

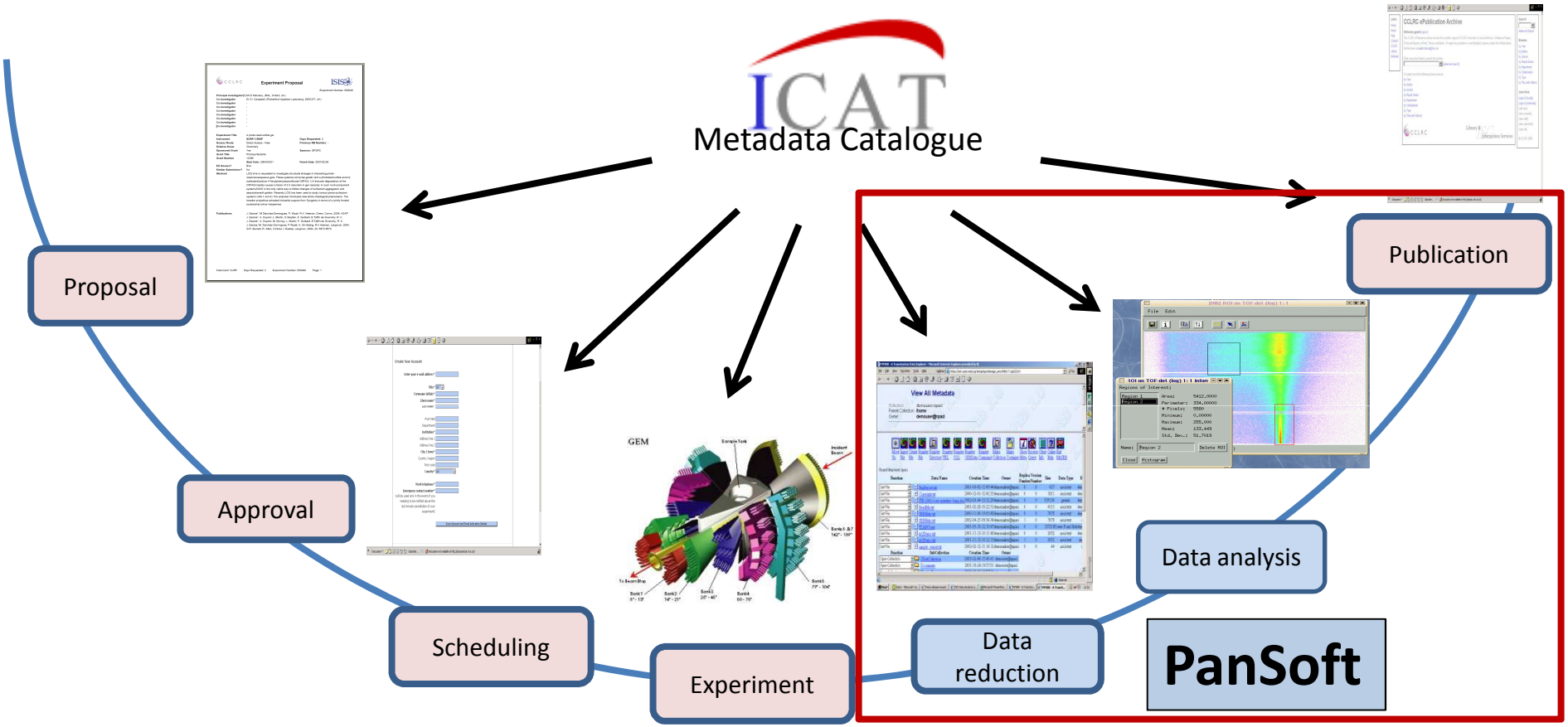


Services on top of ICAT:
DOI, TopCAT, Eclipse ICAT Explorer ...

pandata europe

Science & Technology Facilities Council

Data Continuum



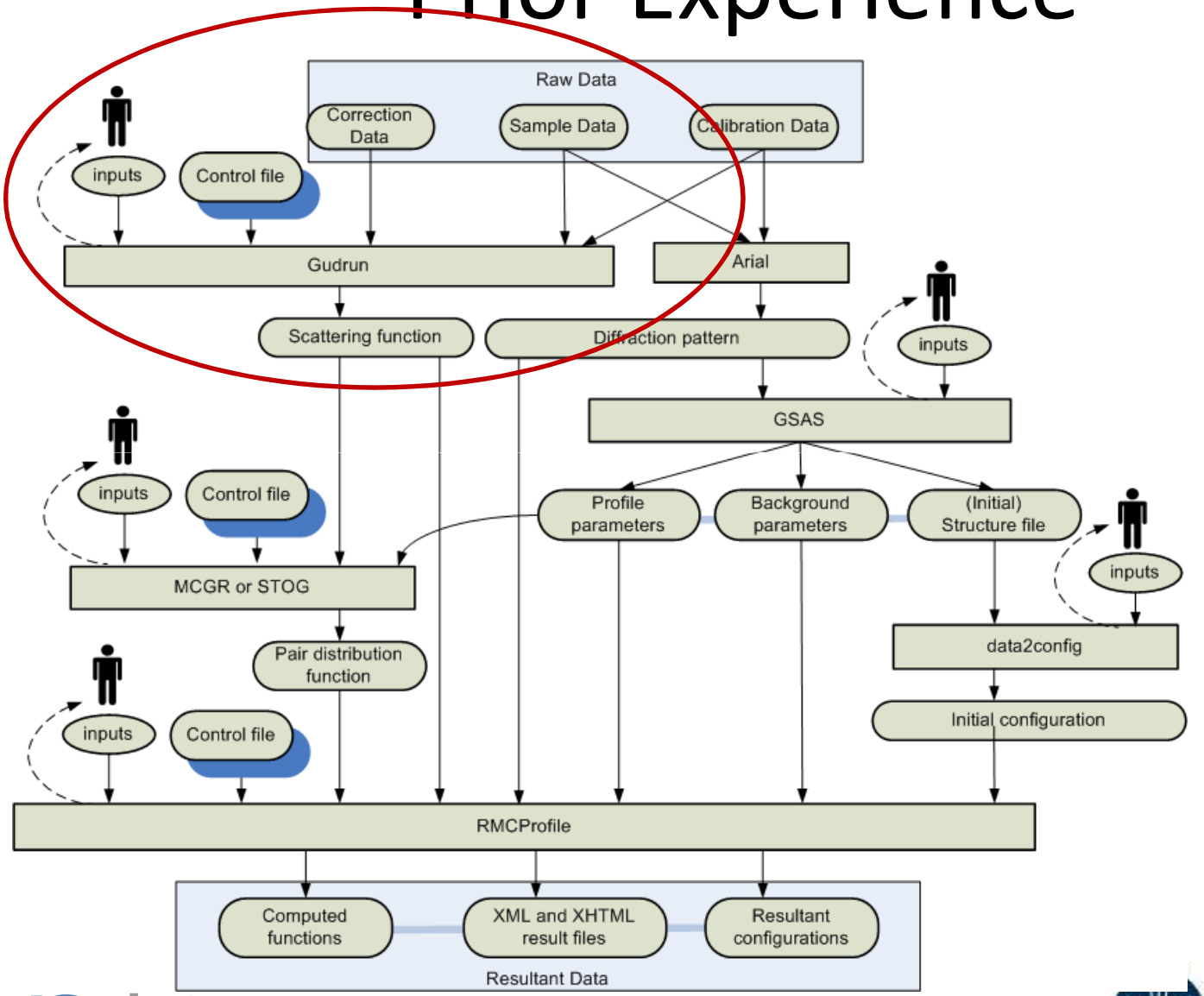
Well developed and supported

- These are with users.
- Traditionally, these, although very useful for data *citation*, *reuse and sharing*, are very difficult to capture!
- Practices vary from individuals to individuals, and from institutions to institutions

pandata europe

Science & Technology Facilities Council

Prior Experience



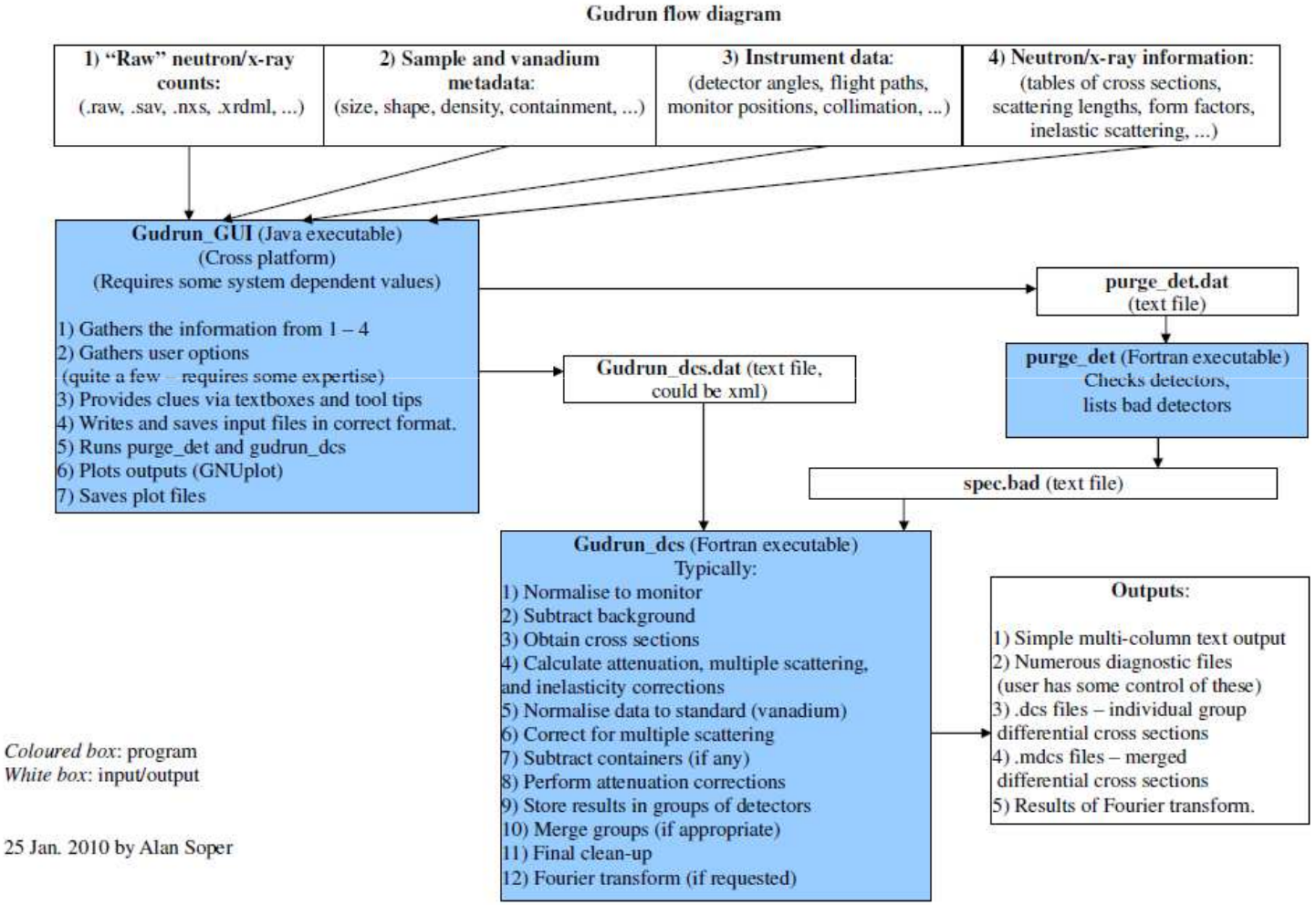
Raw data

Derived data

Resultant data

Credits: Martin Dove, Erica Yang (Nov. 2009)

Can be difficult to capture …



Credits: Alan Soper (Jan. 2010)

Data Provenance: Case Studies (so far)

	Description	Facility	Likely stages of continuum
1.	**Automated SANS2D reduction and analysis**	ISIS	- Experiment/Sample preparation - Raw data collection - Data reduction - Data Analysis
2.	**Tomography reconstruction and analysis**	Manchester Uni./DLS	- Raw data - Reconstructed data - Analysed data
3.	EXPRESS Services	ISIS	- Experiment preparation - Raw data collection - Data Analysis to standard final data product
4.	Recording publications arising from proposal	ISIS	- Proposal system - Raw Data collection - Publication recording
5.	DAWN + iSpyB	DLS, ESRF	- Experiment preparation - Data analysis steps

These case studies have given us unique insights into today's facilities ...

Looking for more case studies ...

TODAY'S FACILITIES – FACTS

Diamond and ISIS (the story so far ...)

- Diamond:
  - ~ 290TB and 120 million files [1]
  - In SCD data archive
  - Largest file: ~120GB – Tomography beamlines [3]
- ISIS [2]:
  - ~16TB and 11 million files
  - In ISIS data archive & SCD data archive (on going)
  - Largest file: ~16GB – the WISH instrument

Diamond Tomography

- Up to 120 GBs/file every 30 minutes
- 6,000 TIFF images/file
- Up to 200 GBs/hr
- ~5 TBs/day
- 1-3 days/experiment

How many images are there for each experiment?

ISIS

- Peak data copy rate (off instrument) 100MB/S
  - Expecting to reach 500MB-1GB/S (but not soon).
- Expect to grow to 10TB/cycle in 3-5 years
- Become interested in centrally hosted services (WISH)

WHAT DOES IT MEAN?

It means ...

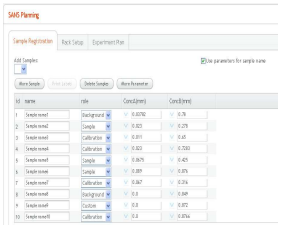
- Due to the volume, it is not cost effective to transfer the (raw + reconstructed) data back to the home institutions or elsewhere to process
  - The network bandwidth to universities mean that it will take a long time to transfer ...
  - So, users have to physically take data back home on storage drive ...
- It is impossible for users to do the reconstruction or analysis on their own computer/laptop.
  - How much RAM do you have on your laptop? And how big is the file from WISH? The Mantid story ...
- It is expensive to re-do the analysis back to the home institutions because of
  - Lack of hardware resources
  - Lack of metadata (a large number of files often mean that there is not much useful metadata)
  - Lack of expertise (e.g. parallel processing, GPU programming)
  - (Assuming software is open source ...)
- Facilities become interested, again, in centralised computing services, right next to the data
  - The ISIS WISH story ...
  - Diamond GPU cluster vs. SCD GPU cluster (directly linked to the data archive)

Users now ask for remote services

- Users are interested in remote data analysis services
  - "... Of course this would mean a *step change in the facilities provided and the time users spend at the facility*. ... "
  - In response, we are developing: "TomoDM" ...
- Then, what benefits can remote services bring?
  - Systematic recording of data continuum, thus allowing the recording of scientific workflows, software, and data provenance (the first four categories data as defined in the "Living Publication" [4])
  - Drive data processing (reduction & analysis) with data provenance
  - It is not only possible to create bi-directional links between raw data and publications, it is also possible to systematically create pair-wise bi-directional links between raw, derived, resultant data and publications.

TWO EXAMPLES

SRF – Automated Data Processing Pipeline for ISIS

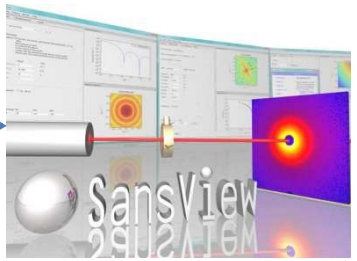


OpenGenie Script

Raw data

Mantid

Reduced data



SampleTracks | **Data Acquisition** | **Data Reduction** | **Model Fitting**

Samples, Experiment setup

Raw data

Reduced data

Analysed data

ICAT

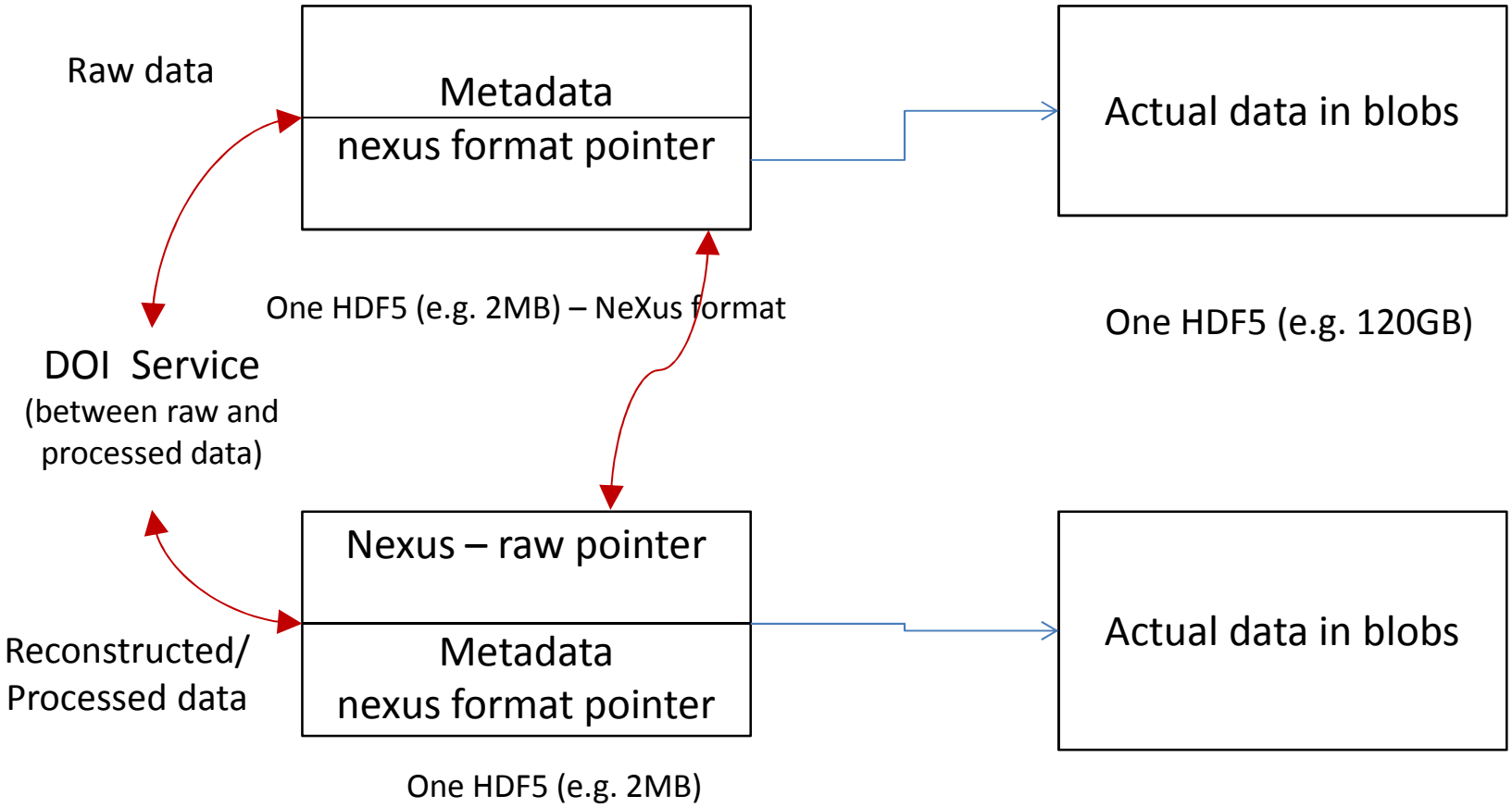
ICAT Data Catalogue

Blog Posts in LabTrove

pandata europe

Science & Technology Facilities Council

Links between metadata and files



But ...

- If we can capture all types of data
  - Samples
  - Raw data
  - Reduced data
  - Analysed data
  - Software
  - Data provenance (i.e. Relationships between datasets)
- Facilities operators do not normally
  - Perform validation and fixity checking of data files (volume vs. cost)
  - *Actually MUCH cheaper to simply back up all images without regard for quality, relevance or even duplication, than it is to build an infrastructure for automatically analyzing millions of image files to determine which are "worth keeping".* [1]
  - But lifetime of data on tape vs. Lifetime of tape
- Will DOIs good enough for data citation?

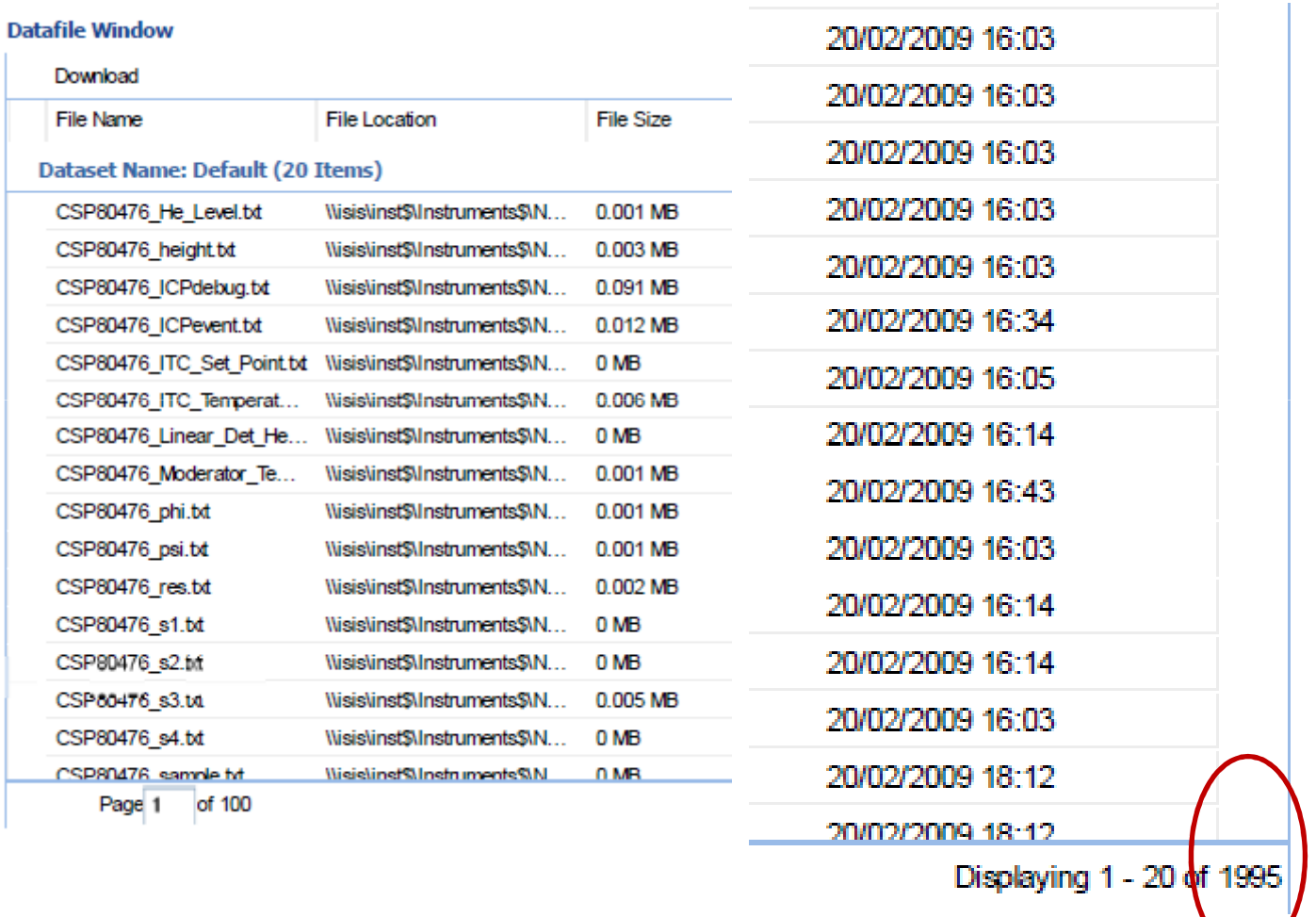
STFC EXPERIENCE OF USING DATACITE DOIS

pandata europe

Science & Technology Facilities Council

STFC DOI landing page

Behind the landing page



pandata europe

Science & Technology Facilities Council

Citation Issues

- At what granularity should data be made citable?
  - If single datasets are given identifiers, what about collections of datasets, or subsets of data?
  - What are we citing? Datasets or an aggregation of datasets
- STFC link DOI to experiments on large facilities which contain many data files
- Other organisations DOI usage policies use different granularities.
- Can there be a common, or discipline common DOI usage policy for data?
- Citation of different types of object: software, processes, workflows ...
- Private vs. Public datasets/DOIs

Summary

- Infrastructure services
  - Compute
  - Storage
  - Archiving + preservation (validation, integrity)
  - Remote data processing
  - DOIs
- It is already happening
  - DLS is already collecting processed data (e.g. iSpyB)
  - But, interesting issues remind to be resolved when lots of data become available ...

Acknowledgements

Project (s)	Institution(s)	People
PanData-ODI (case study)	Diamond, U.K.	Mark Basham, Kaz Wanelik
PanData-ODI: (case study)	Manchester University, U.K. Harwell Research Complex, RAL Diamond, U.K.	Prof. Philip Withers, Prof. Peter Lee
SCAPE, SRF, PanData-ODI	Scientific Computing Department, STFC	David Corney, Shaun De Witt, Chris Kruk, Michael, Wilson
I2S2, SRF	Chemistry, Southampton University, UK	Prof. Jeremy Frey, Dr. Simon Coles
I2S2	Cambridge University, UK	Prof. Martin Dove
I2S2, PanData-ODI, SRF, SCAPE	ISIS, STFC	Tom Griffin, Chris Morton-Smith, Alan Soper, Silvia Imberti, Cameron Neylon, Ann Terry, Matt Tucker

I2S2, SRF are funded by JISC, UK
PanData-ODI, SCAPE are funded by EC

pandata europe

Science & Technology Facilities Council

References

1. Colin Nave (Diamond), “Survey of Future Data Archiving Policies for Macromolecular Crystallography at Synchrotrons”, distributed via “dddwg-bounces@iucr.org”, July 2012.
2. Chris Morton-Smith (ISIS), “ISIS Data rates and sizes – up to March 2012”, May 2012.
3. Mark Basham (Diamond), “HDF5 Parallel Reading and Tomography Processing at DLS”, PanData-ODI DESY meeting, Hamburg, Germany, Feb. 2012.
4. John R Helliwell, Thomas C. Terwilliger, Brian McMahon, “The Living Publication”, April 2012.

Questions

pandata europe

Science & Technology Facilities Council