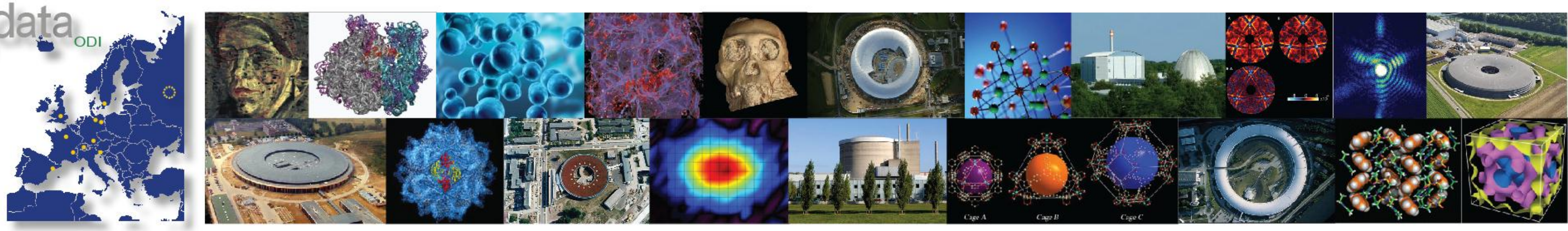


Towards an open data infrastructure for photon and neutron facilities

Brian Matthews and Juan Bicarregui

Scientific Computing Department (SCD)

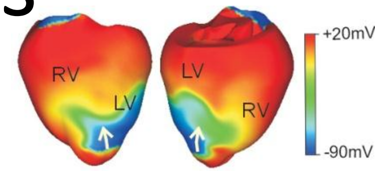
Rutherford Appleton Laboratory (RAL)
Science and Technology Facilities Council (STFC), U. K.

brian.matthews@stfc.ac.uk

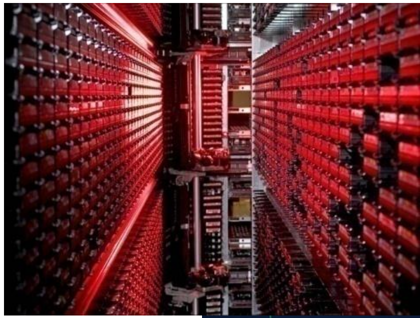
pandata ODI

Science & Technology Facilities Council

Science and Technology Facilities Council



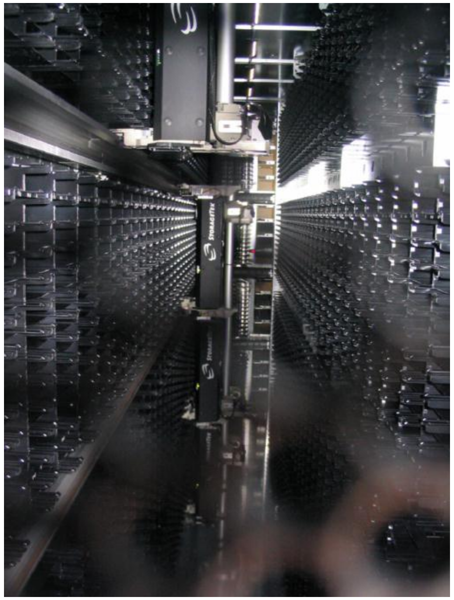
- Provide large-scale scientific facilities for UK Science
  - particularly in physics and astronomy
  - ISIS and Diamond Light Source facilities



- Scientific Computing Department
  - Provides advanced IT development and services to the STFC Science Programme
  - Strong role in management of our science data

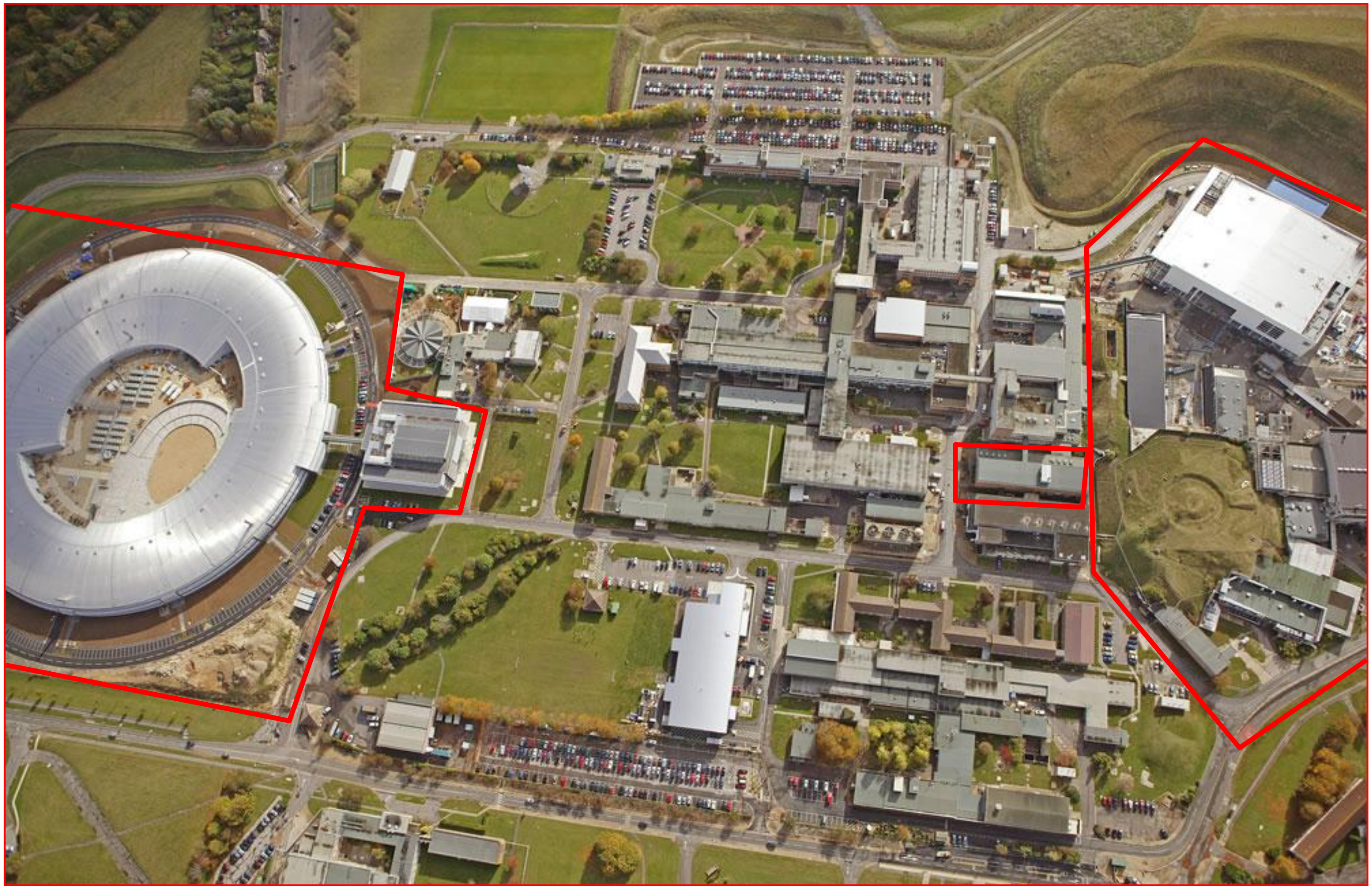
The computing centre at Rutherford Appleton Laboratory

The computing centre at RAL houses 20,000 computer processors and stores 10,000,000,000 Mbytes of data available on-line.

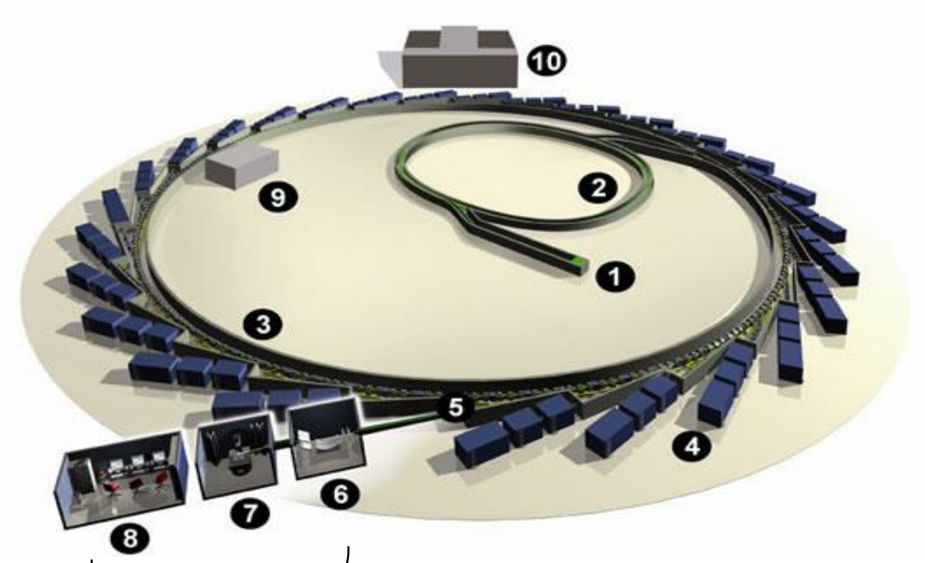


The centre is used for simulation and data analysis by researchers in all scientific disciplines from throughout the UK and their international collaborators.

STFC Rutherford Appleton Laboratory



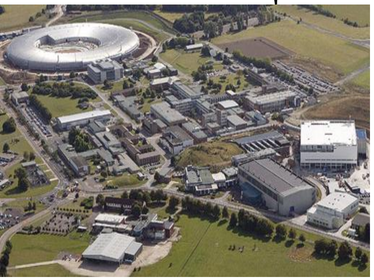
Doing Science at Source Facilities



Neutrons and photons
Provide complementary views of matter:

Photons “see” electric charge – high atomic number nuclei

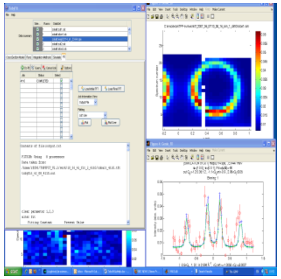
Neutrons “see” nucleons – especially hydrogen atoms



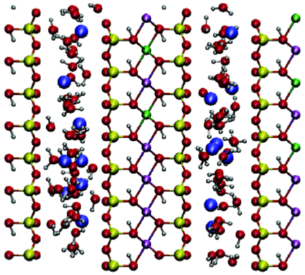
Visit facility on research campus

Place sample in beam

Diffraction pattern from sample



Fitting experimental data to model



Structure of cholesterol in crude oil

pandata ODI

Science & Technology Facilities Council

Facilities Science

- ~30,000 user visitors each year in Europe:
  - physics, chemistry, biology, medicine,
  - energy, environmental, materials, culture
  - pharmaceuticals, petrochemicals, microelectronics

- **Big Facilities for Small Science**

- Billions of € of investment

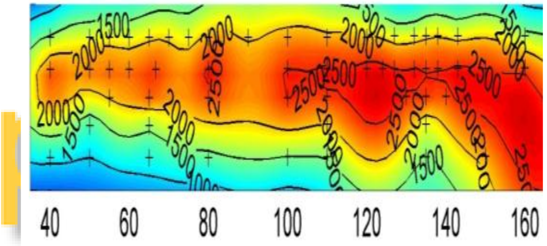
  c. £400M for DLS

  + running costs

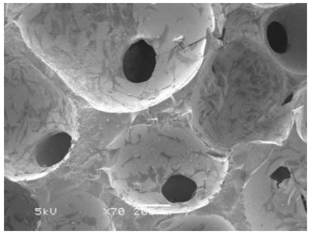
- Over 5.000 high impact publications per year in Europe

- Each facility has dedicated data and computing infrastructure
  - But so far no integrated data repositories across facilities
  - Lacking sustainability & traceability

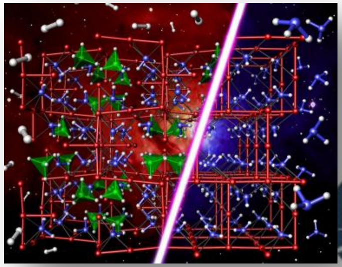
Longitudinal strain in aircraft wing



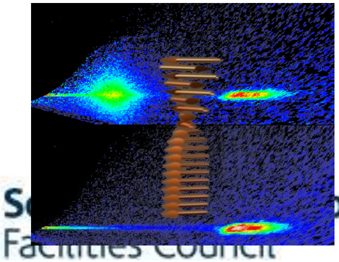
Bioactive glass for bone growth



Hydrogen storage for zero emission vehicles



Magnetic moments in electronic storage



ALBA DESY diamond elettra ESRF FRM II Forschungs-Neutronenquelle Heinz Maier-Leibnitz HELMHOLTZ ZENTRUM BERLIN für Materialien und Energie ILL NEUTRONS FOR SCIENCE Institut Laue-Langevin Science & Technology Facilities Council ISIS JCNS Jülich Centre for Neutron Science LLB Laboratoire Léon Brillouin MAX-lab PAUL SCHERRER INSTITUT PSI SOLEIL SYNCHROTRON

pandata ODI

Science & Technology Facilities Council

A Common Community

Number of Users shared between facilities

	BER II	BESSY II	DESY	DLS	ELETTRA	ESRF	ILL	ISIS	LLB	SINQ	SLS	SOLEIL	FRM-II	ANKA	neutron	photon	all
BER II	850	80	68	25	18	128	261	141	67	76	12	14	111	5	375	244	850
BESSY II	80	2306	238	45	134	399	67	33	31	26	149	93	42	31	175	758	2306
DESY	68	238	3563	88	121	735	194	91	55	44	155	130	103	43	356	1105	3563
DLS	25	45	88	3494	72	739	213	336	35	18	145	149	20	12	441	967	3494
ELETTRA	18	134	121	72	2731	455	85	43	23	4	66	316	9	20	145	839	2731
ESRF	128	399	735	739	455	10728	886	406	235	92	600	1069	144	80	1303	3256	10728
ILL	261	67	194	213	85	886	4338	741	343	229	69	176	349	10	1450	1246	4338
ISIS	141	33	91	336	43	406	741	2755	120	119	43	52	155	5	908	716	2755
LLB	67	31	55	35	23	235	343	120	1348	34	12	131	92	3	425	359	1348
SINQ	76	26	44	18	4	92	229	119	34	726	96	9	97	0	334	210	726
SLS	12	149	155	145	66	600	69	43	12	96	2424	182	18	18	169	923	2424
SOLEIL	14	93	130	149	316	1069	176	52	131	9	182	3656	14	26	299	1460	3656
FRM-II	111	42	103	20	9	144	349	155	92	97	18	14	1087	5	494	255	1087
ANKA	5	31	43	12	20	80	10	5	3	0	18	26	5	452	19	144	452
neutron	850	175	356	441	145	1303	4338	2755	1348	726	169	299	1087	19	7117	2350	8852
photon	244	2306	3563	3494	2731	10728	1246	716	359	210	2424	3656	255	452	4517	19902	24154
all	850	2306	3563	3494	2731	10728	4338	2755	1348	726	2424	3656	1087	452	8624	19242	30873

Details of how we count users: http://wiki.pan-data.eu/CountingUsers

PaN-Data facilities users

Total number of unique users:	**30873**	
Using only Neutrons:	6719	or 21.7% of all unique users
Using only Photons:	22021	or 71.3% of all unique users
Using Neutrons and Photons:	2133	or 6.9% of all unique users
Using more than one facility:	6863	or 22.2% of all users
Using more than one Photon source:	4252	or 17.6% of all photon users
Using more than one Neutron source:	1734	or 19.6% of all neutron users
All facilities have users in common with all other facilities, also across neutron and photon sources.		
Typically, **30-40%** of the users of any of the photon or any of the neutron sources also use at least one other facility		

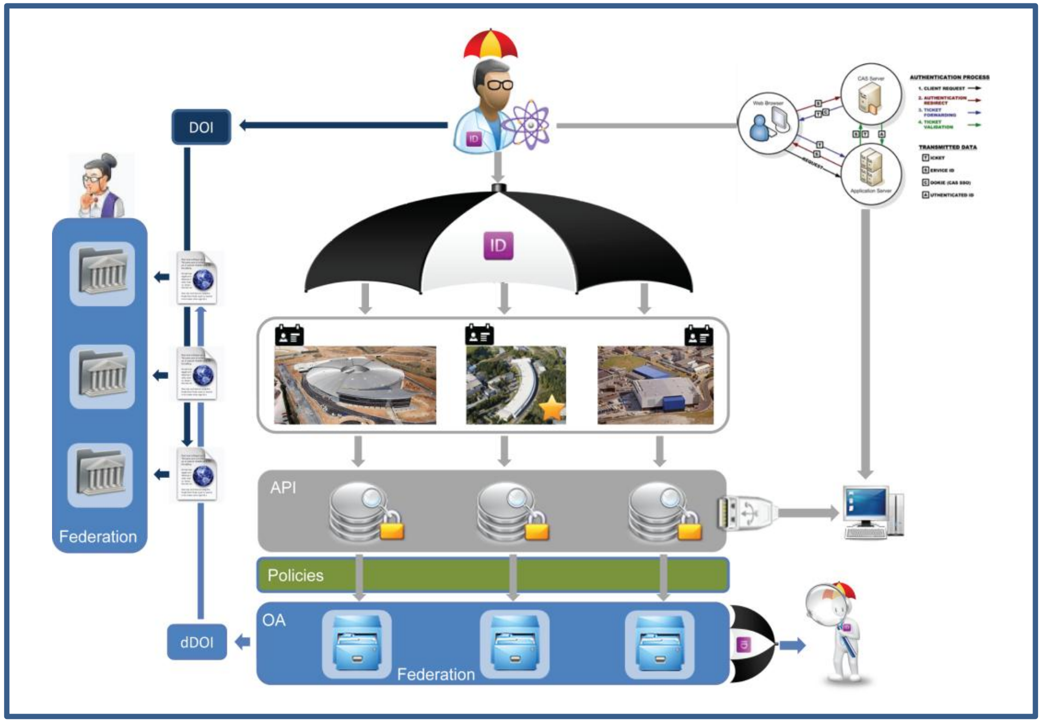
Benefit to be gained to the user community by coordinating the computing infrastructure

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**

Provide common tools and user experience across facilities

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



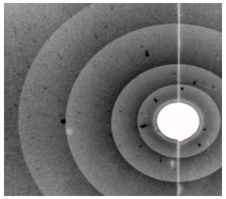
pandata ODI

Science & Technology Facilities Council

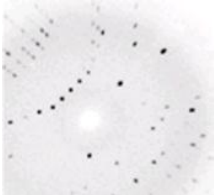
Sharing and combining data

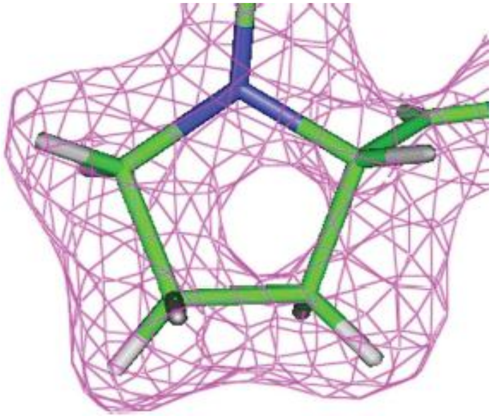
- Users move around to collect different views on their samples
  - Different instruments have different characteristics
- Combining data can give additional insights
- Needs
  - Common data formats
  - Common data access
  - Common metadata
- Context
  - Providing better provenance information

Neutron diffraction



X-ray diffraction



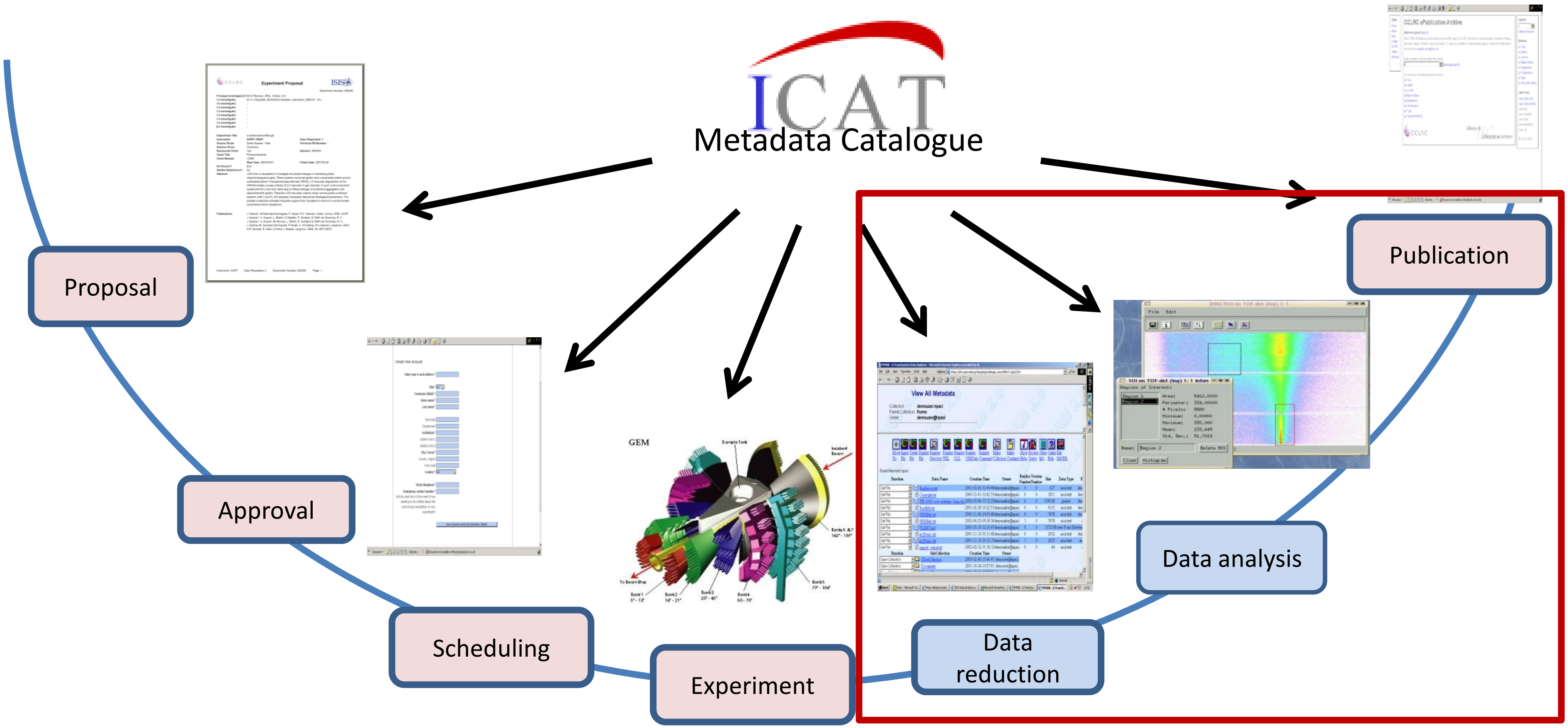


High-quality structure refinement

pandata ODI

Science & Technology Facilities Council

Data Continuum



▭ Well developed and supported in facility

- 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 ODI

Science & Technology Facilities Council

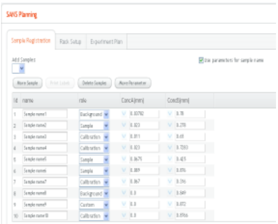
Managing the Data Continuum

- Provide better support for the data continuum
  - Recording of provenance
  - Linking raw and derived data, publications and software to the experiment
- Improve service to the user
  - Accurate record keeping of science results and context
  - Validation of science results
  - Publishing and reuse of “research objects”
  - Linking to other research objects

- However, capturing provenance of analysed data is hard and expensive
  - Lots of variation and input from users
  - Lots of blind alleys and retracing of steps
  - Mostly undertaken in the user’s institution

- So why bother ?
  - Good use cases where managing provenance well gives benefits.

Smart Research Framework: Automated Data Processing Pipeline for ISIS



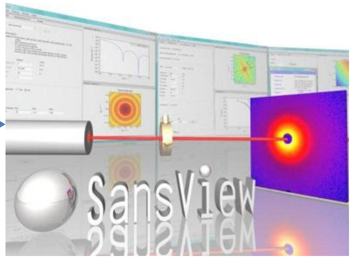
Id	name	role	ConcA(mm)	ConcB(mm)
1	Sample name1	Background	0.00702	0.78
2	Sample name2	Sample	0.023	0.278
3	Sample name3	Calibration	0.011	0.68
4	Sample name4	Calibration	0.023	0.7283
5	Sample name5	Sample	0.0675	0.425
6	Sample name6	Sample	0.089	0.876
7	Sample name7	Calibration	0.067	0.396
8	Sample name8	Background	0.0	0.849
9	Sample name9	Custom	0.0	0.872
10	Sample name10	Calibration	0.0	0.8766

OpenGenie Script

Raw data

Mantid

Reduced data



SampleTracks

Data Acquisition

Data Reduction

Model Fitting

Samples,
Experiment setup

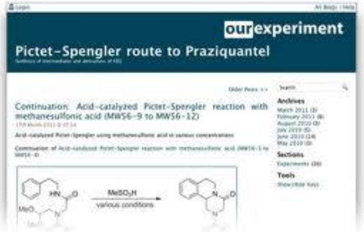
Raw data

Reduced data

Analysed data

ICAT

ICAT Data Catalogue



Pictet-Spengler route to Praziquantel

Continuation: Acid-catalyzed Pictet-Spengler reaction with methanesulfonic acid (MWS6-9 to MWS6-12)

Blog Posts in LabTrove

pandata ODI

Cameron Neylon (ISIS) and Erica Yang

Science & Technology Facilities Council

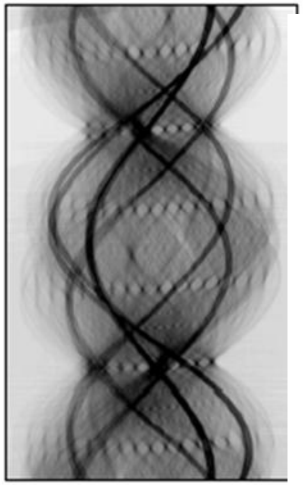
What does this mean?

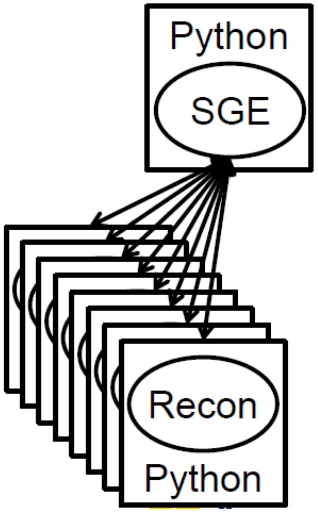
- We can automate the stages of the process “around the experiment”
  - Setting up the experiment
  - Auto-generating configuration and control scripts
  - Initial reduction of the data
- Capture and link (in ICAT)
  - Sample information
  - Experiment configuration and Control
  - Raw and reduced data,
  - Reduction software
- Auto-publish in the Blog the record (with links)
  - Accurate record of metadata for the user to refer to
  - Can be shared with research group or more widely
  - Auto-data publication.

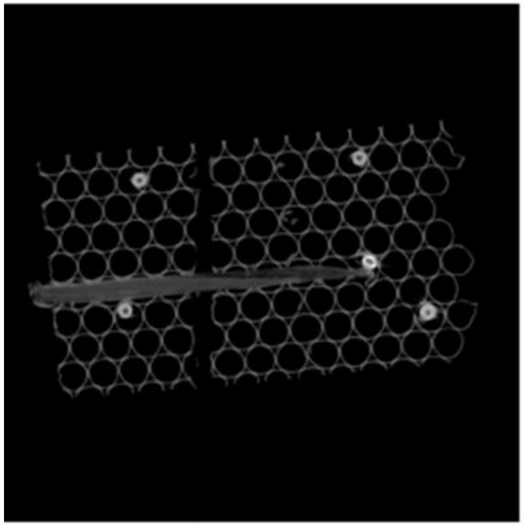
- Would expect to be able to transfer this to “Express Services”
  - A complete data package for the user.

Tomography - Reconstruction

- I13: high throughput tomography beamline
- Computationally heavy process
  - 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
- Each reconstruction
  - 15 individual runs on a GPU
  - Can take up to 45 mins



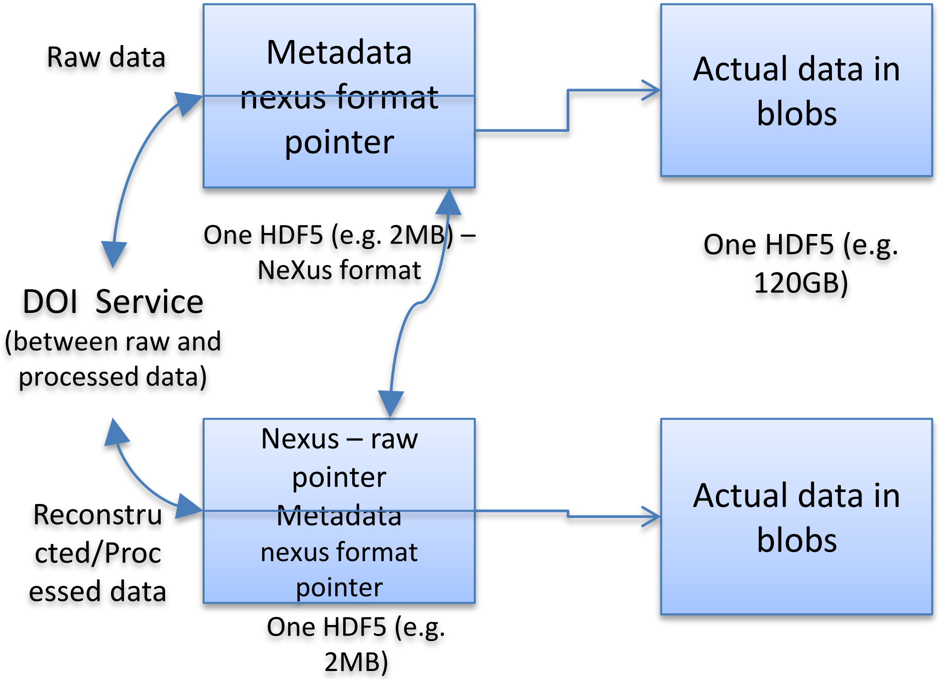




Mark Basham (DLS) and Erica Yang

Links between metadata and files

- It is not cost effective to transfer data to the home institutions
  - The network bandwidth
  - take data back home on storage drive
- It is expensive to do analysis at home institutions
  - It is impossible to process user's own computer
  - Lack of hardware resources
  - Lack of metadata
  - Lack of expertise (e.g. parallel processing, GPU programming)
- 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. ...* "
  - Capture Provenance of data products



Raw data

Metadata nexus format pointer

Actual data in blobs

One HDF5 (e.g. 2MB) – NeXus format

One HDF5 (e.g. 120GB)

DOI Service (between raw and processed data)

Nexus – raw pointer Metadata nexus format pointer

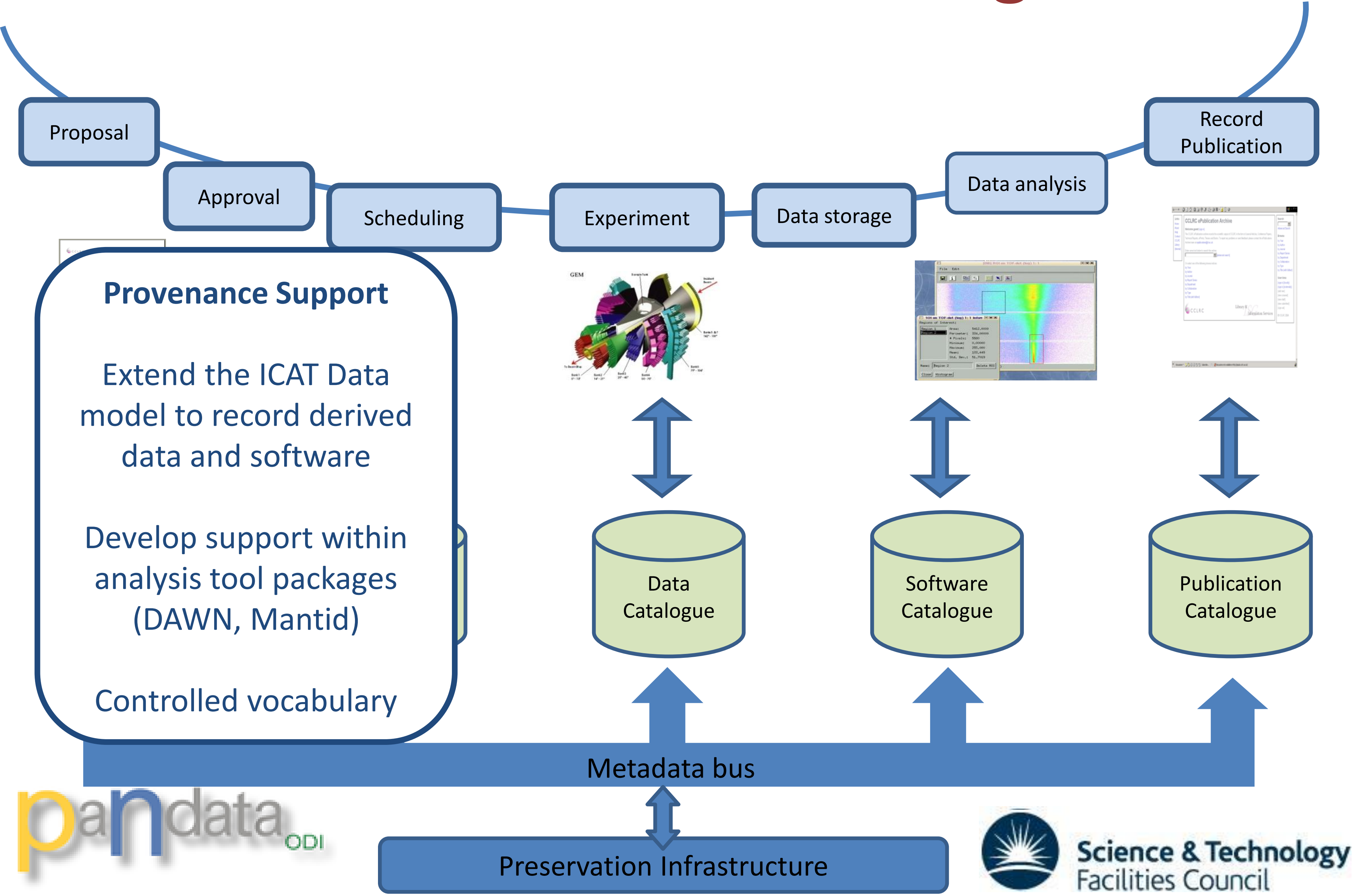
Actual data in blobs

Reconstructed/Processed data

One HDF5 (e.g. 2MB)

One HDF5 (e.g. 120GB)

PaN-data shared catalogues



Provenance Support

Extend the ICAT Data model to record derived data and software

Develop support within analysis tool packages (DAWN, Mantid)

Controlled vocabulary

Conclusions

- Developing a programme towards a common data infrastructure for facilities
  - Pooling limited resources
  - Common experience for users
  - Can transfer and share data more effectively
- Developing a common approach
  - Data and software catalogues
  - Provenance to capture the full context of the experiment
- Big facilities for small science

brian.matthews@stfc.ac.uk

pandata ODI

http://pan-data.eu/

Science & Technology Facilities Council