

Building an Open Data Infrastructure for Science: *Turning Policy into Practice*

Juan Bicarregui

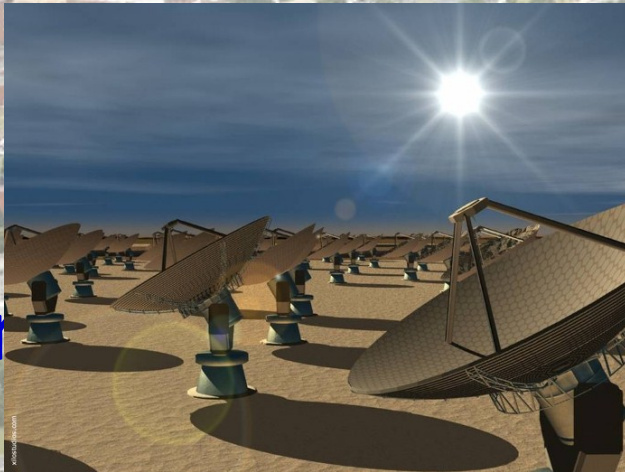
Head of Data Services Division

STFC Department of Scientific Computing

Overview

- Introduction
 - What is STFC?
 - What do we need from our data infrastructure?
- An example project
 - The PaNdata Collaboration
- Fostering Collaboration on a Global Scale
 - The Research Data Alliance

What is STFC?



Square Kilometre Array

- Synchrotron Radiation Source
- Lasers
- Space Science
- Particle Physics



Large Hadron Collider



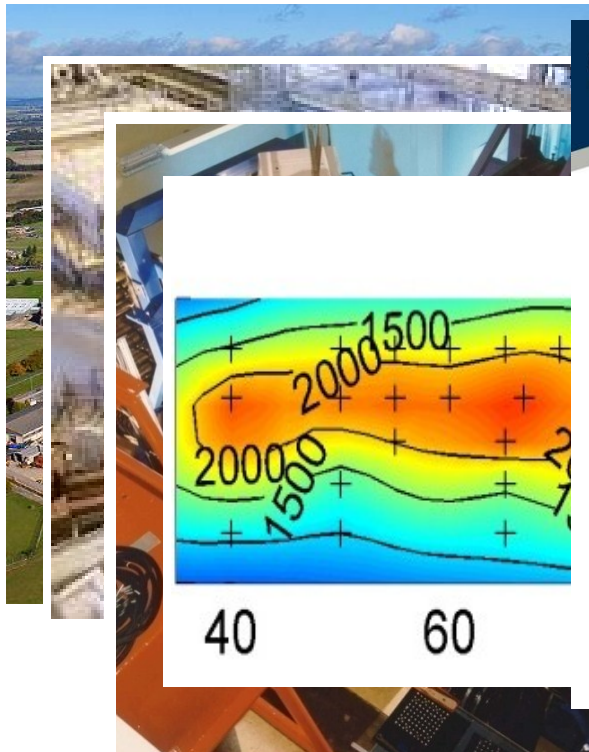
Daresbury Laboratory

Data Management

Communications



ESRF & ILL, Grenoble



Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

Science & Technology Facilities Council
ISIS

UNIVERSITY OF CAMBRIDGE

UK-SHEC

EPSRC

Centre for Ecology & Hydrology

University of Kent

CARDIFF UNIVERSITY

BIBI cymru

THE UNIVERSITY OF QUEENSLAND

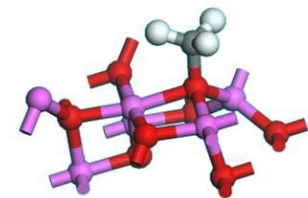
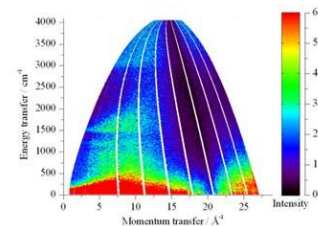
Schlumberger

Powerwave

University of Glasgow

INEOS ChlorVinyls

"Infrared spectroscopy only shows what is on the surface of the catalyst, but neutrons give the whole picture. We were able to modify the catalyst to give 50% less unwanted derivative."
-Dr Stewart Parker, ISIS and Professor David Lennon, University of Glasgow



Methyl chloride synthesis: Neutrons help industry

Putting Data Policy into Practice

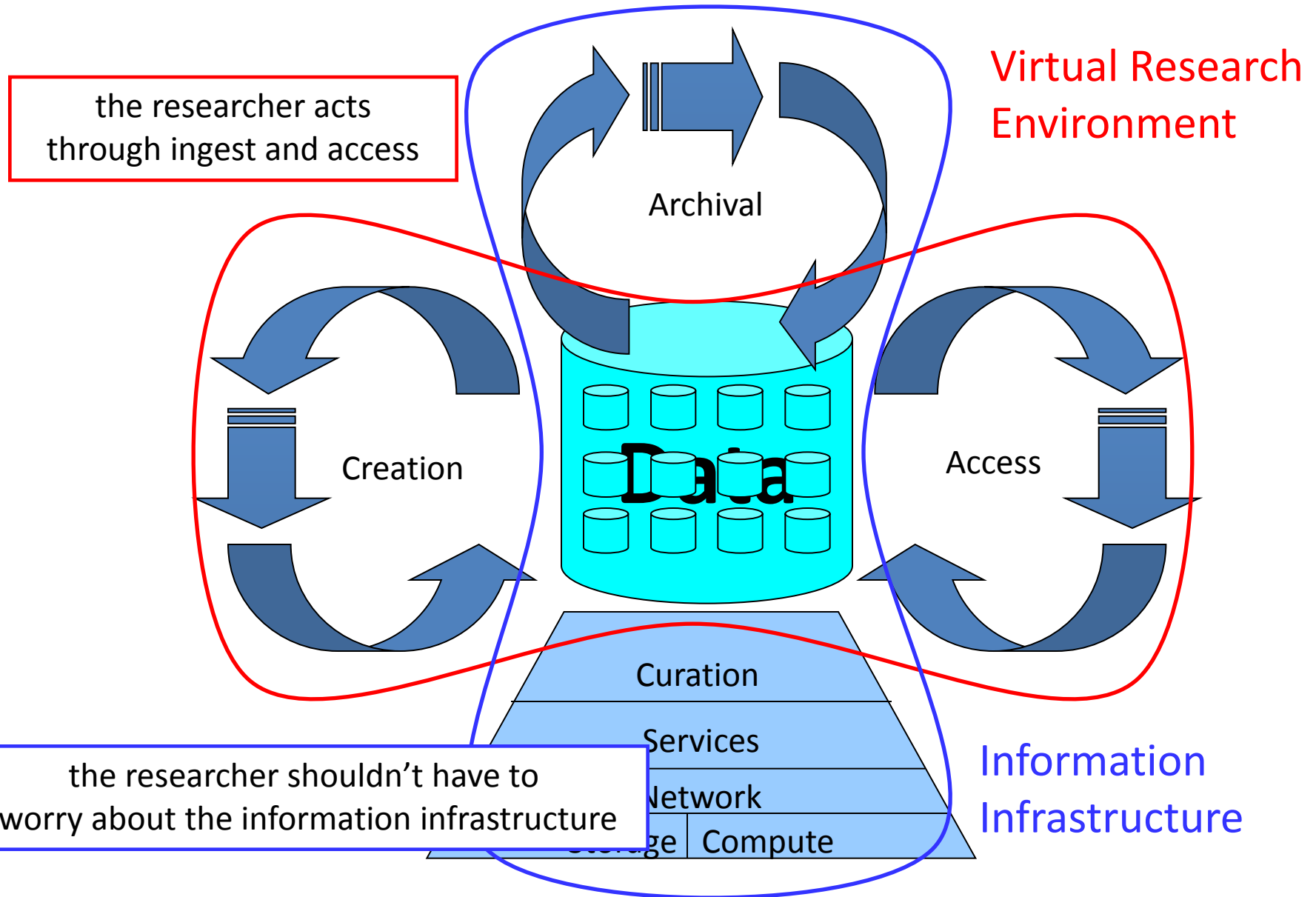
RCUK Principles on Data Policy

Seven (fairly) orthogonal principles:

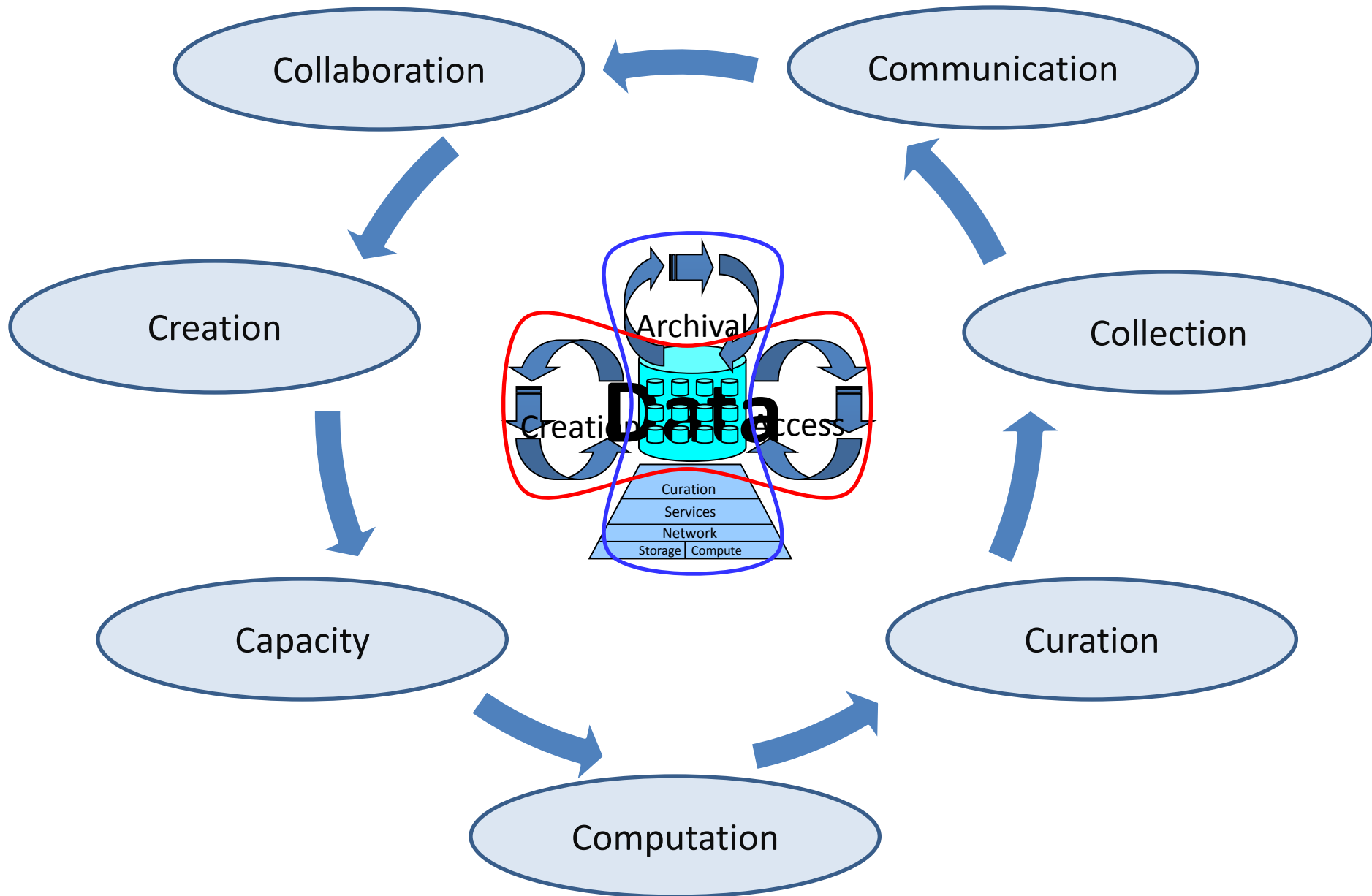
- Public good
 - Preservation
 - Discoverability
 - Confidentiality
 - First use
 - Recognition
 - Costs
- } Data
- } Access
- } Rights

Similar policy work going on at EU and Global levels

Data centric view of research



The 7 C's



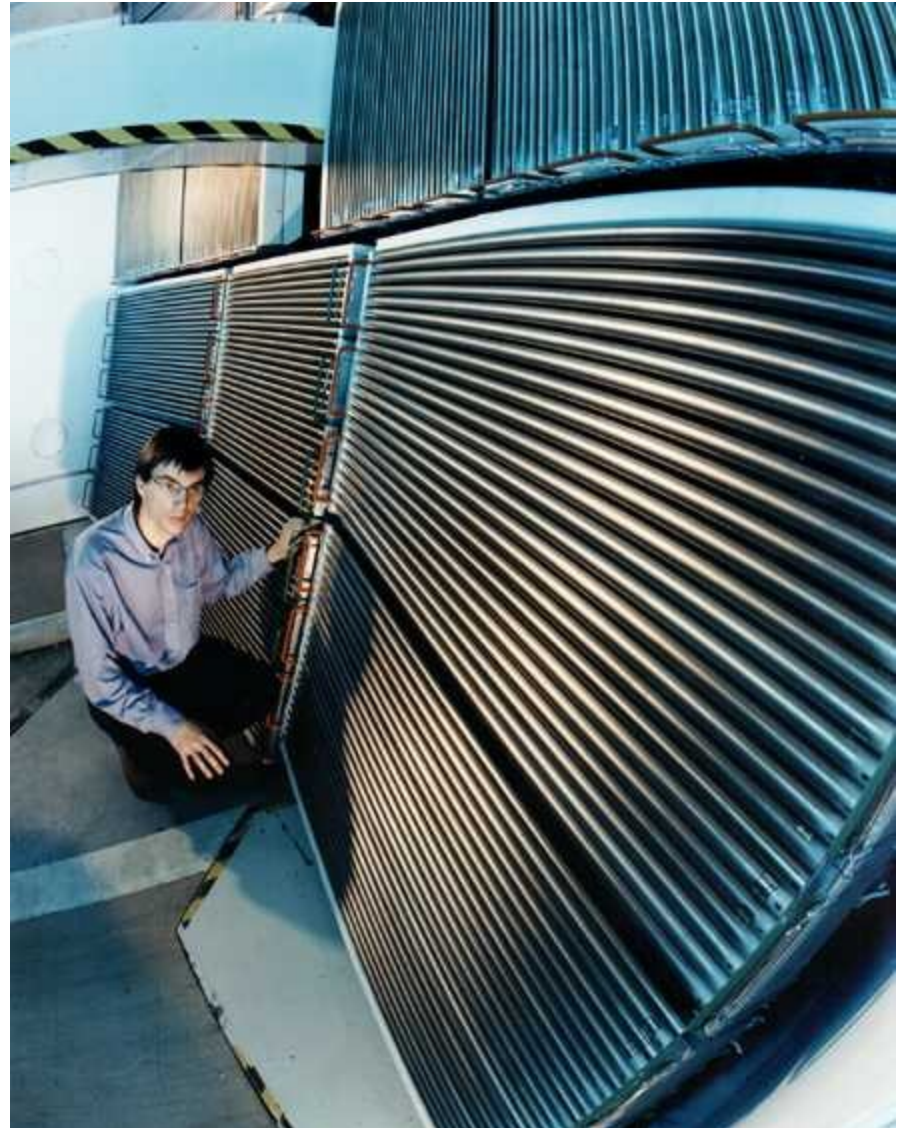
Creation

Linked systems for:

- Proposal submission
- User management
- Data acquisition

Metadata carried from each system to the next

Detectors moving from Hz to KHz, towards MHz,...

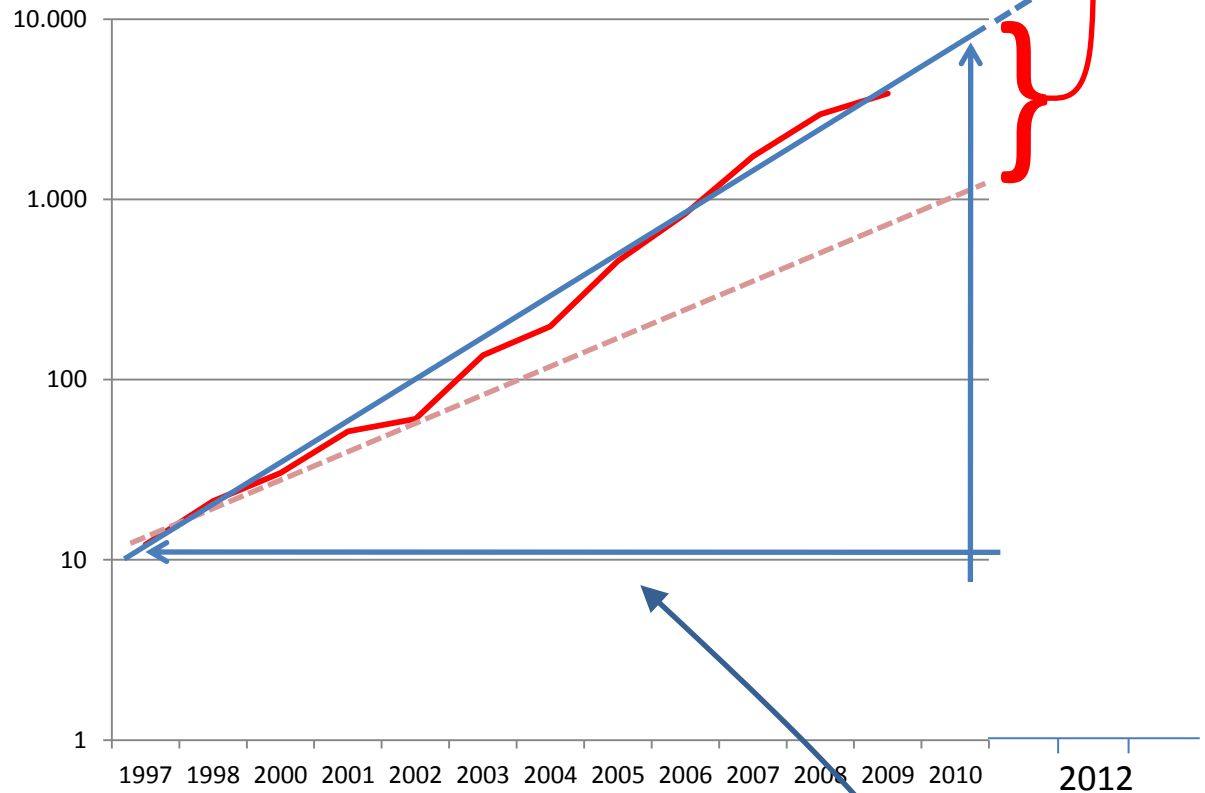


Examining the detectors on MAPS instrument on ISIS

Capacity

10 x Moore's Law (2 years)
2 x Moore's Law (1.5 years)

Total Data Stored (TeraBytes)

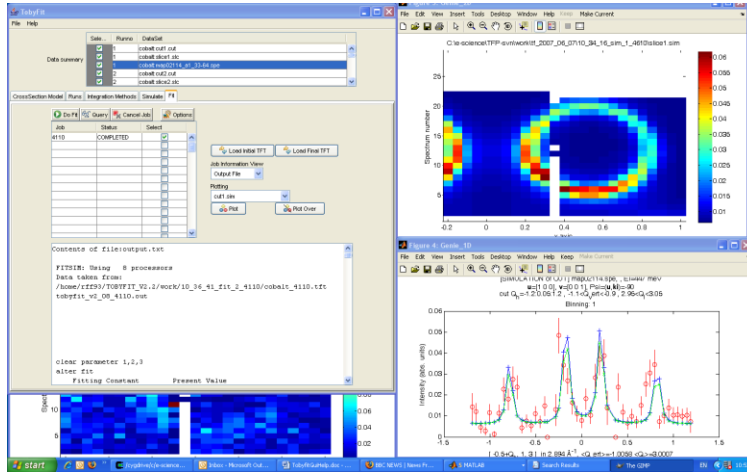


Moore's law for us
is about 15 months

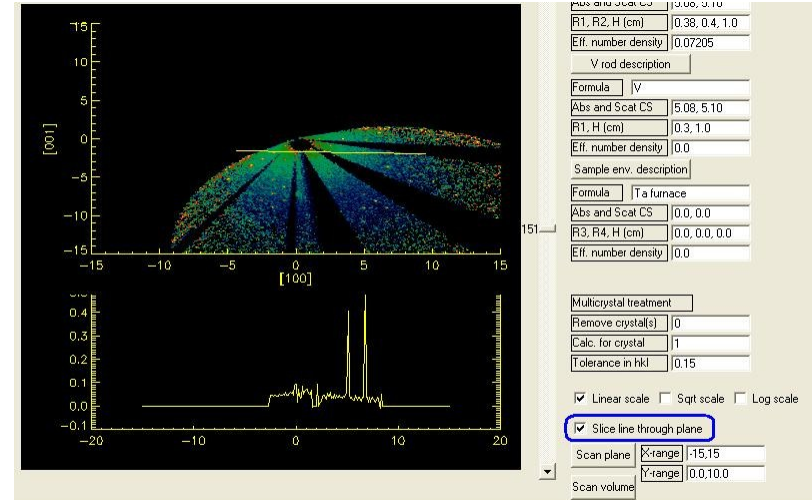
Currently store about
20 PetaBytes of data

Moore's Law
x1000 in 13 years
Doubling every 1.3 years

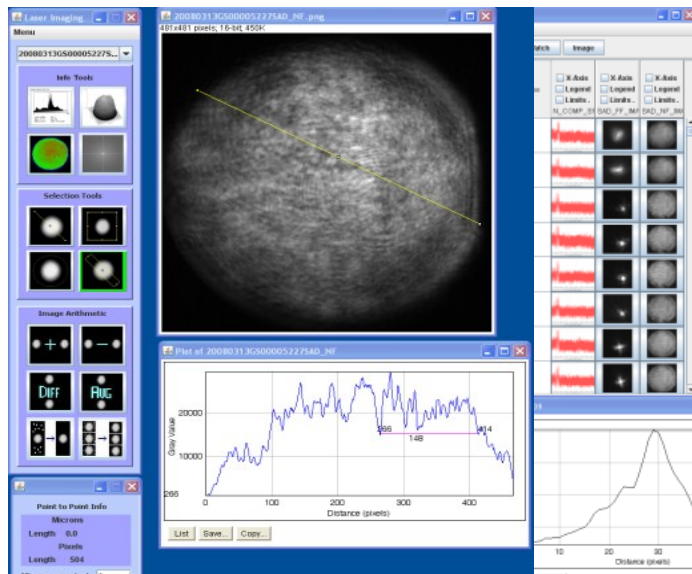
Computation



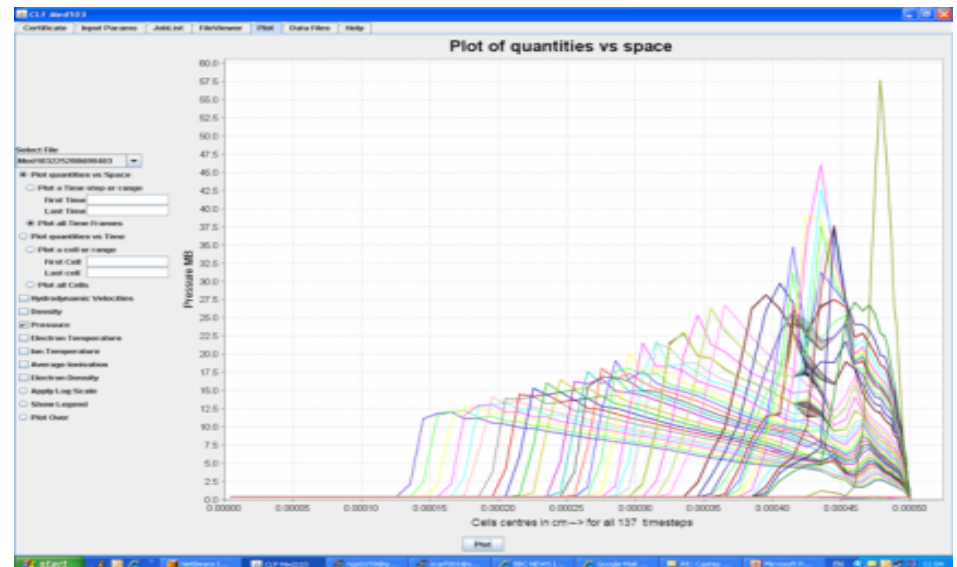
Fitting of experimental data to model



Compute intensive components on HPC
(Blue Gene/Q at DL, Emerald GPGPU cluster at RAL)



Real-time diagnostics of instrument
performance and data flow pipeline.



Computational derivation of properties from theory

Curation

Facility Archives

- All **ISIS** data (~25 years) > 3,000,000 files
- All **Diamond** Data (~5 years) > 100,000,000 files

LHC Tier 1

- UK hub for **LHC** data (11PB)

Other Research Councils

- **NERC** JASMIN+CEMS super data cluster (£4.5M)
- **BBSRC** Institutes data archive
- **MRC** Data Support service

JISC

- JISCmail (1Million users)
- National Grid Service

Universities

- **Imperial College** - National Service for Computational Chemistry Software - **EPSRC** funded
- **Oxford, UCL, Southampton Bristol**, Emerald GPU cluster (£1M) (**EPSRC**)

Others:

- The SCARF compute cluster for RAL Facilities, Users and Collaborators
- The STFC Publications Archive
- The CCPs (Collaborative Computational Projects)
- The Chemical Database Service
- The IBM Blue Gene/Q ICE-CSE

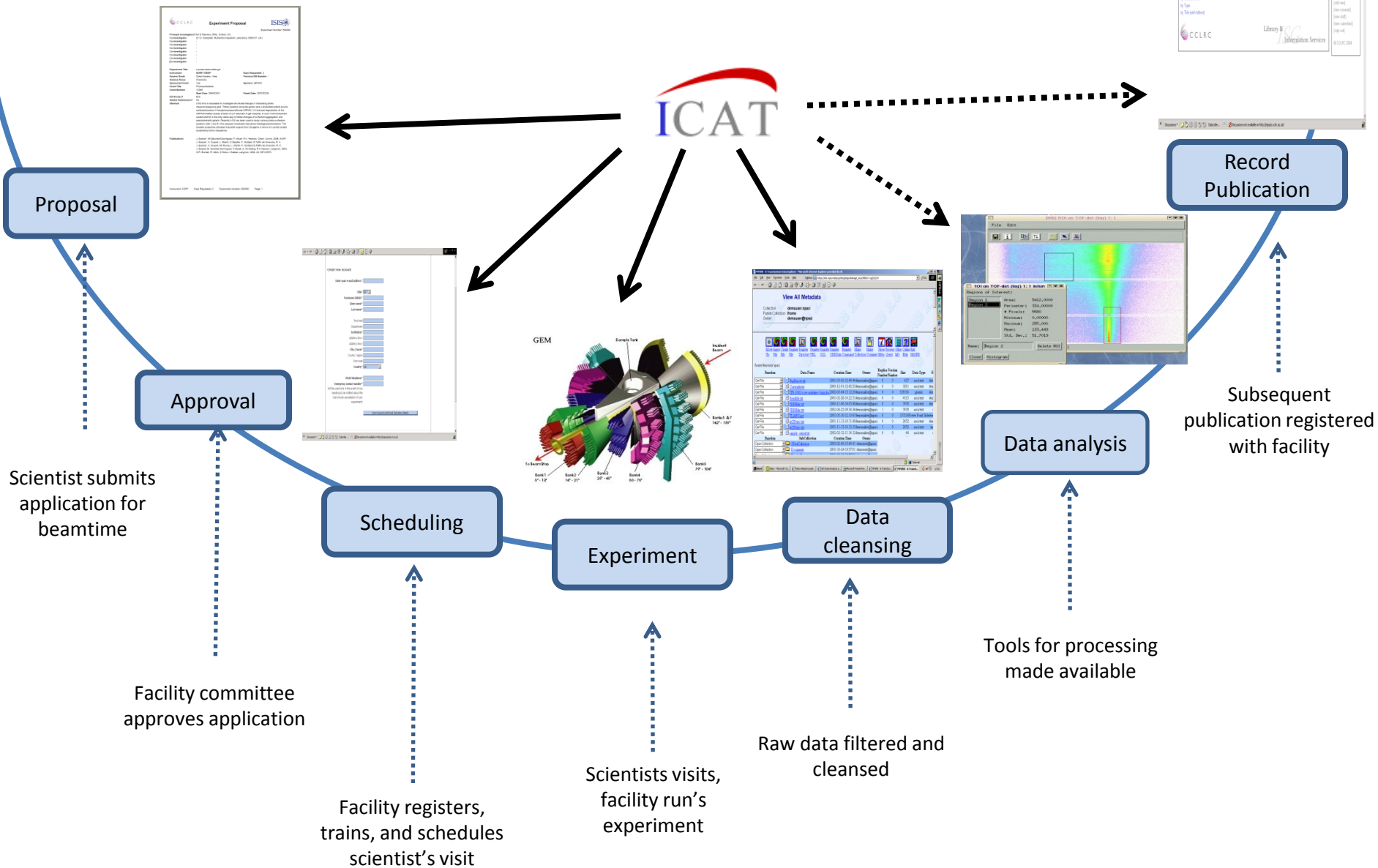


The **StorageTek**
tape robots
100PB Capacity



JASMIN & CEMS
4PB Parallel disk

Collection



Communication

“The web has changed everything...”

Immense Expectations !

- Web enables:
 - access to *everything*
 - Everything on-line
- Interlinking enables:
 - Validation of results
 - Repetition of experiment
- Discovery enables:
 - new knowledge from old



STFC's “e-pubs” Institutional Repository has records of 30,000 publications spanning 25 years

Collaboration

Technology integration facilitates scientific collaboration

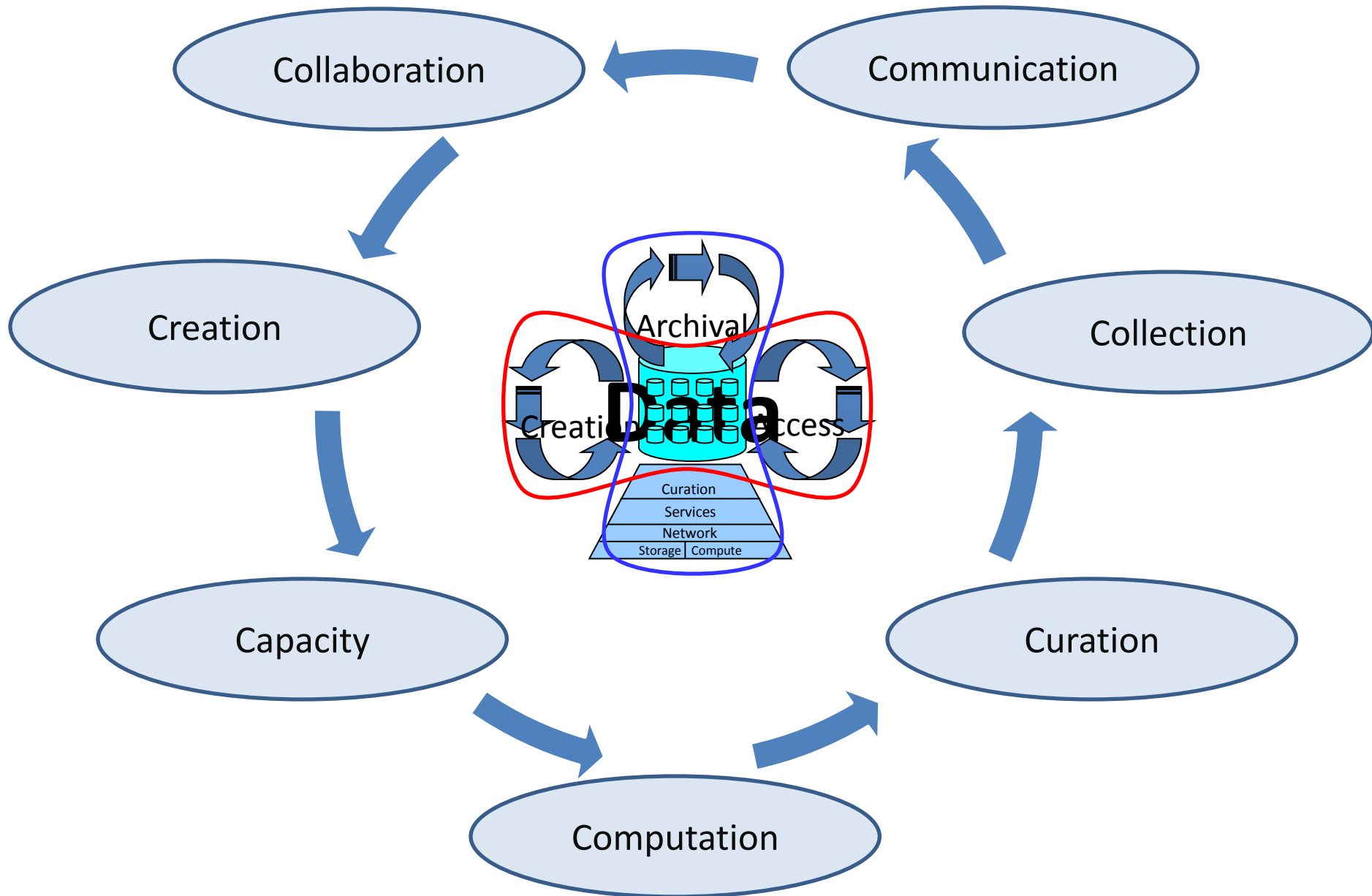
- Cross facility/beamline
- Cross disciplinary

Technology integration improves facility efficiency

- PaN-data –Photon and Neutron Data infrastructure
 - ICAT also used in Australian Synchrotron and Oak Ridge National Lab
- Research Data Alliance



The 7 C's



Overview

- Introduction
 - What is STFC?
 - What do we need from our data infrastructure?
- An example project - PaNdata
- Fostering Collaboration on a Global Scale - RDA

The PaNdata Collaboration

- Established 2007 with 4 partners
- Expanded since to 13 organisations
(see next slide)
- Aims:
 - “...to construct and operate a shared data infrastructure for Neutron and Photon laboratories...”

2007	2008	2009	2010	2011	2012	2013	2014
EDNS (4)							
			EDNP (10)				
				PaNdataEurope(11)			
					Pandata ODI(11)		

PaN-data Partners

PaN-data bring together 13 major European Research Infrastructures

ISIS is the world's leading pulsed spallation neutron source

ILL operates the most intense slow neutron source in the world

PSI operates the Swiss Light Source, SLS, and Neutron Spallation Source, SINQ, and is developing the SwissFEL Free Electron Laser

HZB operates the BER II research reactor the BESSY II synchrotron

CEA/LLB operates neutron scattering spectrometers from the Orphée fission reactor

JCNS Juelich Centre for Neutron Science

ESRF is a third generation synchrotron light source jointly funded by 19 European countries

Diamond is new 3rd generation synchrotron funded by the UK and the Wellcome Trust

DESY operates two synchrotrons, Doris III and Petra III, and the FLASH free electron laser

Soleil is a 2.75 GeV synchrotron radiation facility in operation since 2007

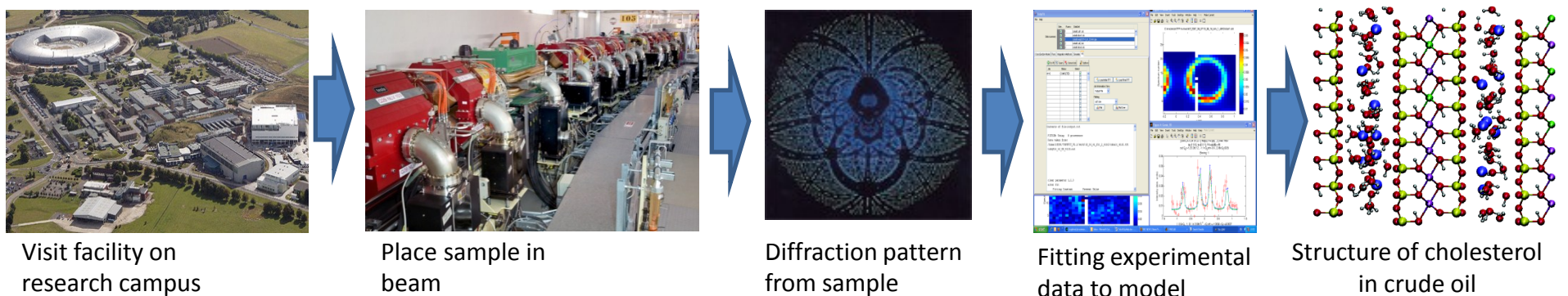
ELETTRA operates a 2-2.4 GeV synchrotron and is building the FERMI Free Electron Laser

ALBA is a new 3 GeV synchrotron facility due to become operational in 2010

MaxLab, Max IV Synchrotron

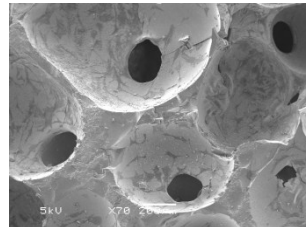
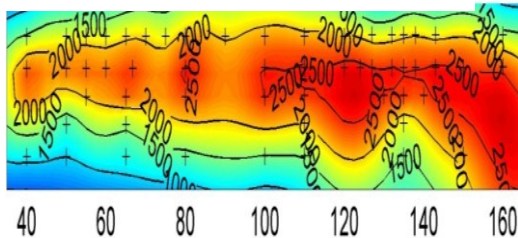
PaN-data is coordinated by the STFC Department of Scientific Computing

The Science we do - Structure of materials



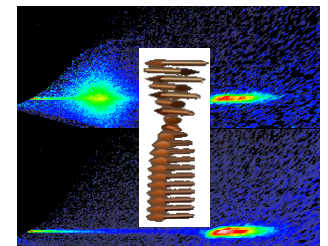
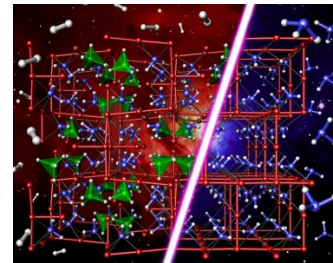
- Over 30,000 user visitors each year:
 - physics, chemistry, biology, medicine,
 - energy, environmental, materials, culture
 - pharmaceuticals, petrochemicals, microelectronics
- Over 5.000 high impact publications per year
 - But so far no integrated data repositories
 - Lacking sustainability & traceability

Longitudinal strain in aircraft wing



Bioactive glass for bone growth

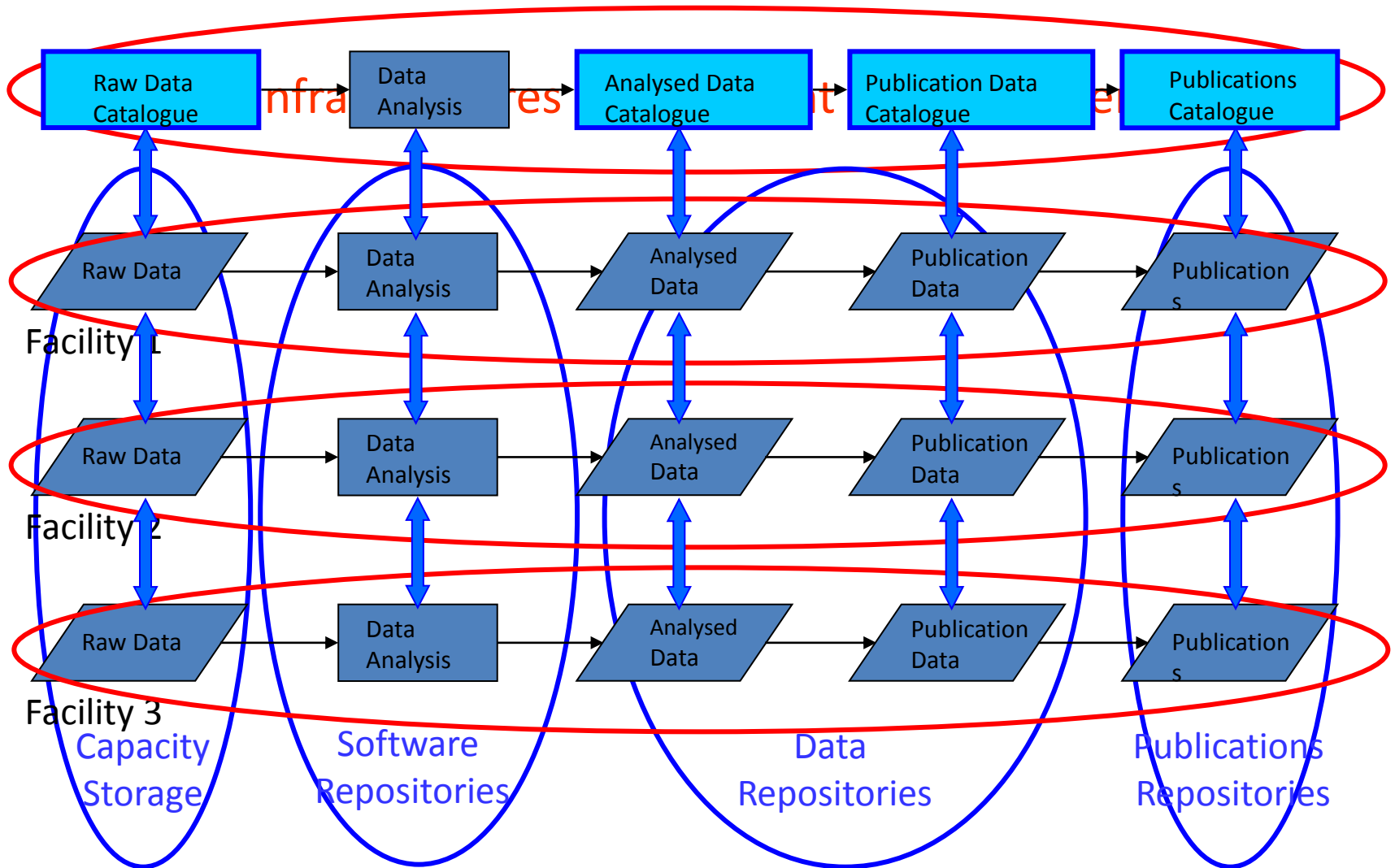
Hydrogen storage for zero emission vehicles



Magnetic moments in electronic storage

PaNdata Vision

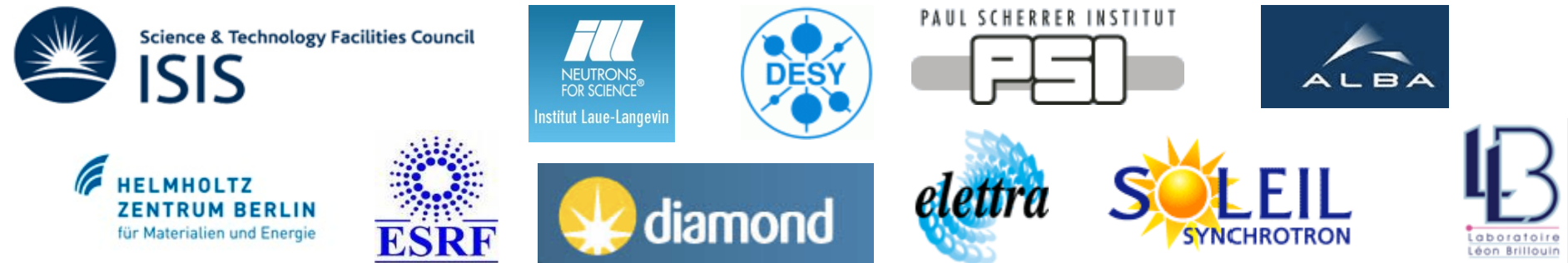
Single Infrastructure → Single User Experience



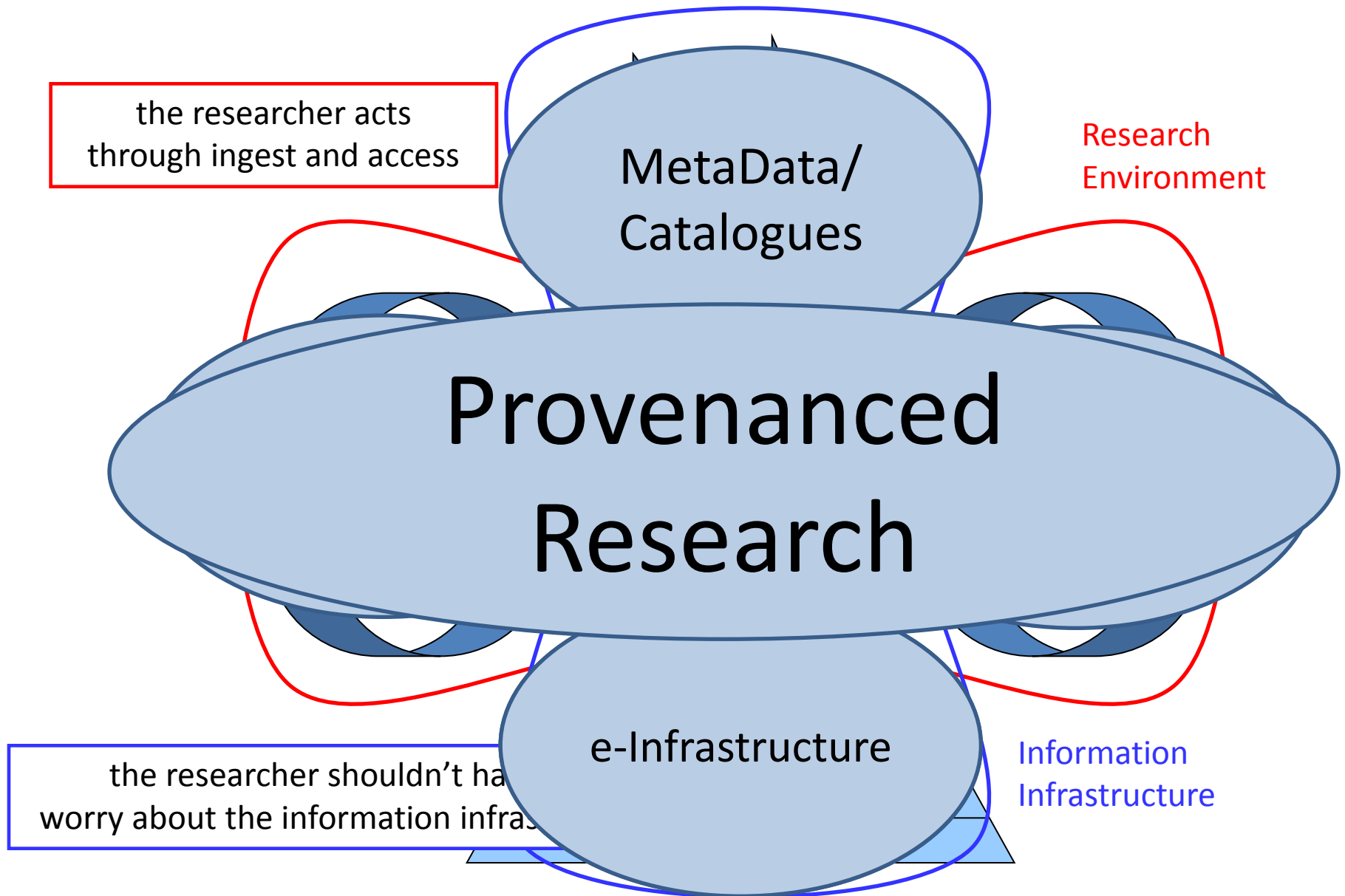
PaN-data Standardisation

PaN-data Europe is undertaking 5 standardisation activities:

1. Development of a **common data policy** framework
2. Agreement on protocols for shared **user information exchange**
3. Definition of standards for common **scientific data formats**
4. Strategy for the interoperation of **data analysis software** enabling the most appropriate software to be used independently of where the data is collected
5. **Integration and cross-linking** of research outputs completing the lifecycle of research, linking all information underpinning publications, and supporting the long-term preservation of the research outputs



The Research Lifecycle – a personal view



Overview

- Introduction
 - What is STFC?
 - What do we need from our data infrastructure?
- An example project – PaNdata
- Fostering Collaboration on a Global Scale - RDA

Research Data Alliance

Emerging international organization

Currently supported by:

EU

NSF

Australian National Data Service

To accelerate data-driven innovation
through research data sharing and exchange.

Infrastructure, Policy, Practice and Standards

Research Data Alliance

Vision and Purpose

Vision

*Researchers around the world
sharing and using research data without barriers.*

Purpose

*... to accelerate international
data-driven innovation and discovery
by facilitating research data
sharing and exchange,
use and re-use,
standards harmonization, and discoverability.
...through the development and adoption of
infrastructure, policy, practice, standards, and other
deliverables.*

RDA Principles

Openness

- Membership is open to all interested organizations,
- all meetings are public,
- RDA processes are transparent, and
- all RDA products are freely available to the public;

Consensus

- The RDA moves forward by achieving consensus and
- resolves disagreements through appropriate voting mechanisms;

Balance

- The RDA is organized on the principle of balanced representation for individual organizations and stakeholder communities;

Harmonization

- The RDA works to achieve harmonization across standards, policies, technologies, tools, and other data infrastructure elements;

Voluntary

- The RDA is not a government organization or regulatory body and, instead, is a public body responsive to its members; and

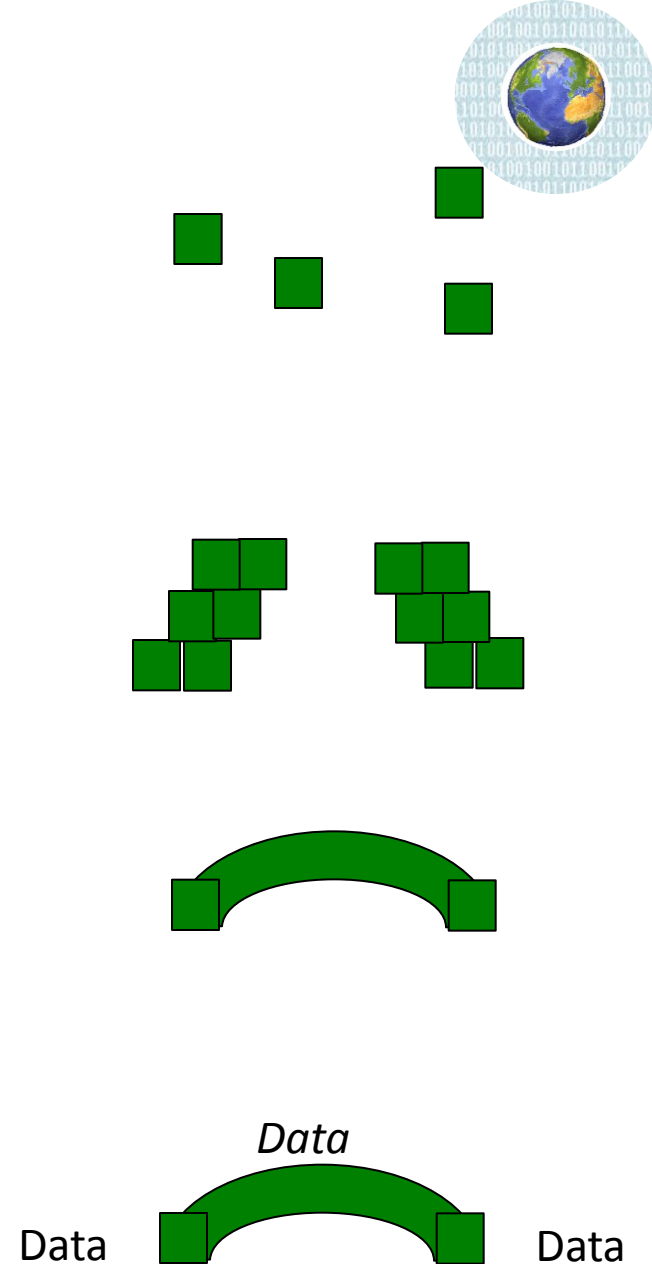
Non-profit

- RDA is not a commercial organization and will not design, promote, endorse, or sell commercial products, technologies, or services.

Research Data Alliance:

“Building Bridges”

- Bridges to the future – data preservation
- Bridges to research partners
- Bridges across disciplines
- Bridges across regions
- Bridges to integration – to solve new problems
- Bridges across communities



RDA role

Two bridges we can build:

- Connecting Data
- Connecting People

What kind of organisation do we need to do this?









Data Practitioners Domain

Working Groups

- Carry out work of RDA
- Reach consensus on
- May suggest BoF topics
- Open to all RDA members
 - some commitment expected

Plenary

- Open to all persons involved in RDA
- Hears and comments on reports from WGs
- Suggests new BoFs
- Hears candidates for TAC

Technical Advisory Committee

- advise on WG work
- Interacting directly with WGs
- advise on new WGs and new topics
- Give implementation suggestions to strategic direction from council

Administrative Domain

Administration and Management Team

- Implement strategic direction set by council
- Supports the activities of the RDA
 - Arrange plenary meetings
 - Run the on-line for a
 - Manage documents
- Convene nominating committees for
 - Council and TAC
- Monitor and controls finances
- Prepare reports for
 - Council, funders,....

Council

- Set the vision
- Final approval of governance matters
- Approve new WGs (TAC advised)
- control balanced WG approach

- use for all kinds of activities, open to all RDA members

Online Open Interaction Fora

Research Data Alliance

Status in November 2012

- Initial meetings held in Munich and Washington
- ~200 Delegates
- ~12 vanguard Working Groups being established
- Council and Secretariat formed
- Launch planned for March 17 in Guttenberg

Website:

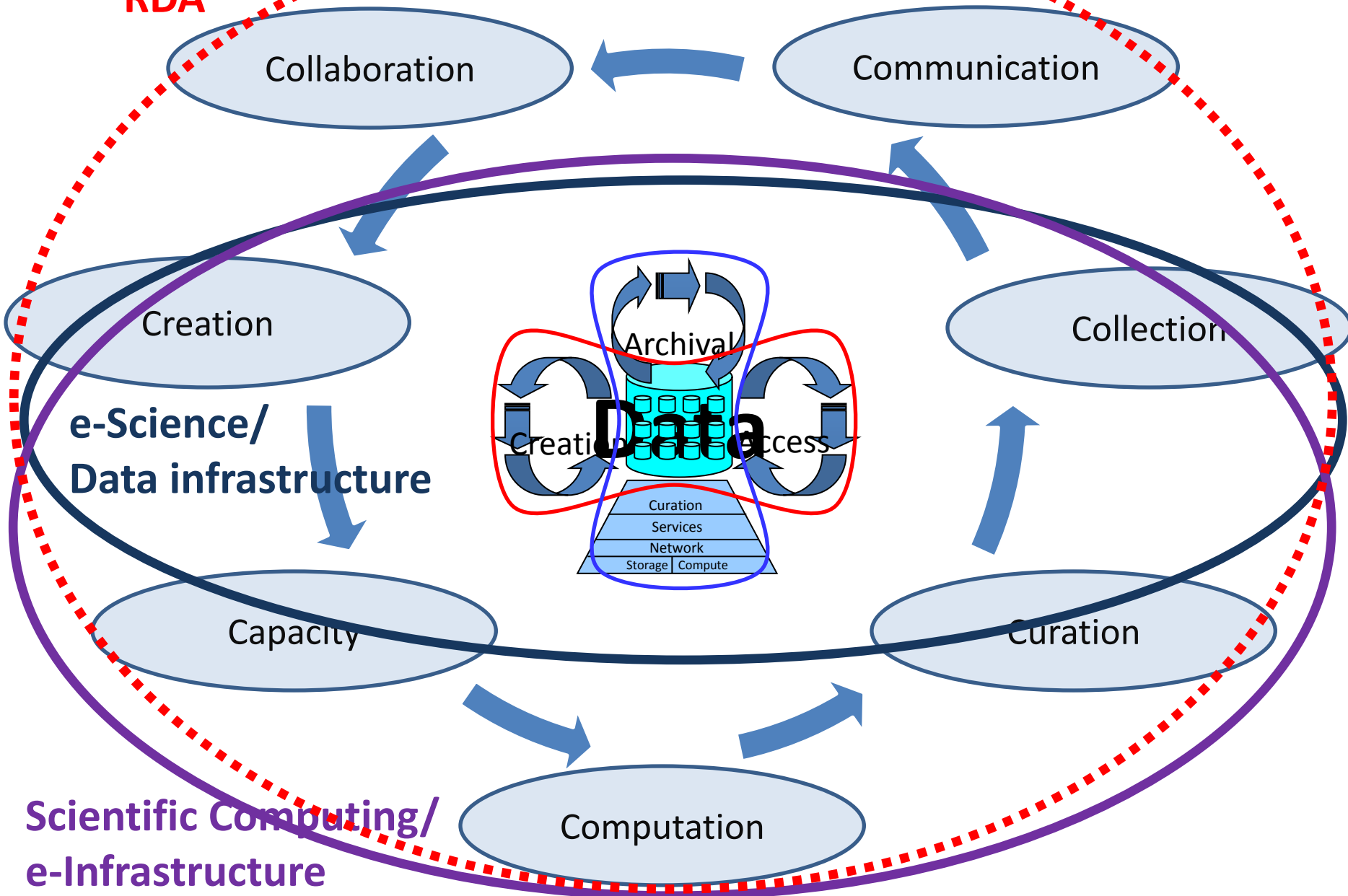
rd-alliance.org/

Washington Meeting:

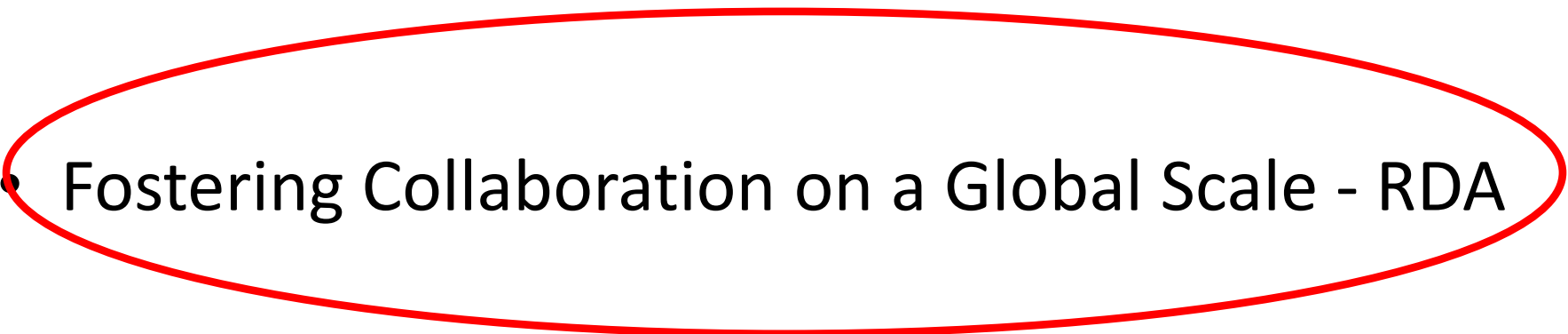
d2i.indiana.edu/data2012/ResearchDataAlliance

The Knowledge Lifecycle

RDA



Overview

- Introduction
 - What is STFC?
 - What do we need from our data infrastructure?
 - An example project – PaNdata
 - Fostering Collaboration on a Global Scale - RDA
- 

Thank You

www.stfc.ac.uk/SCD

www.pan-data.eu

www.rd-alliance.org/



The End