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


***Juan Bicarregui***

**Head of Data Services Division**

**STFC Department of Scientific Computing**



**IDCC 2013, International Digital Curation Conference, 14-17 January 2013, Amsterdam**

Science & Technology Facilities Council

# Overview

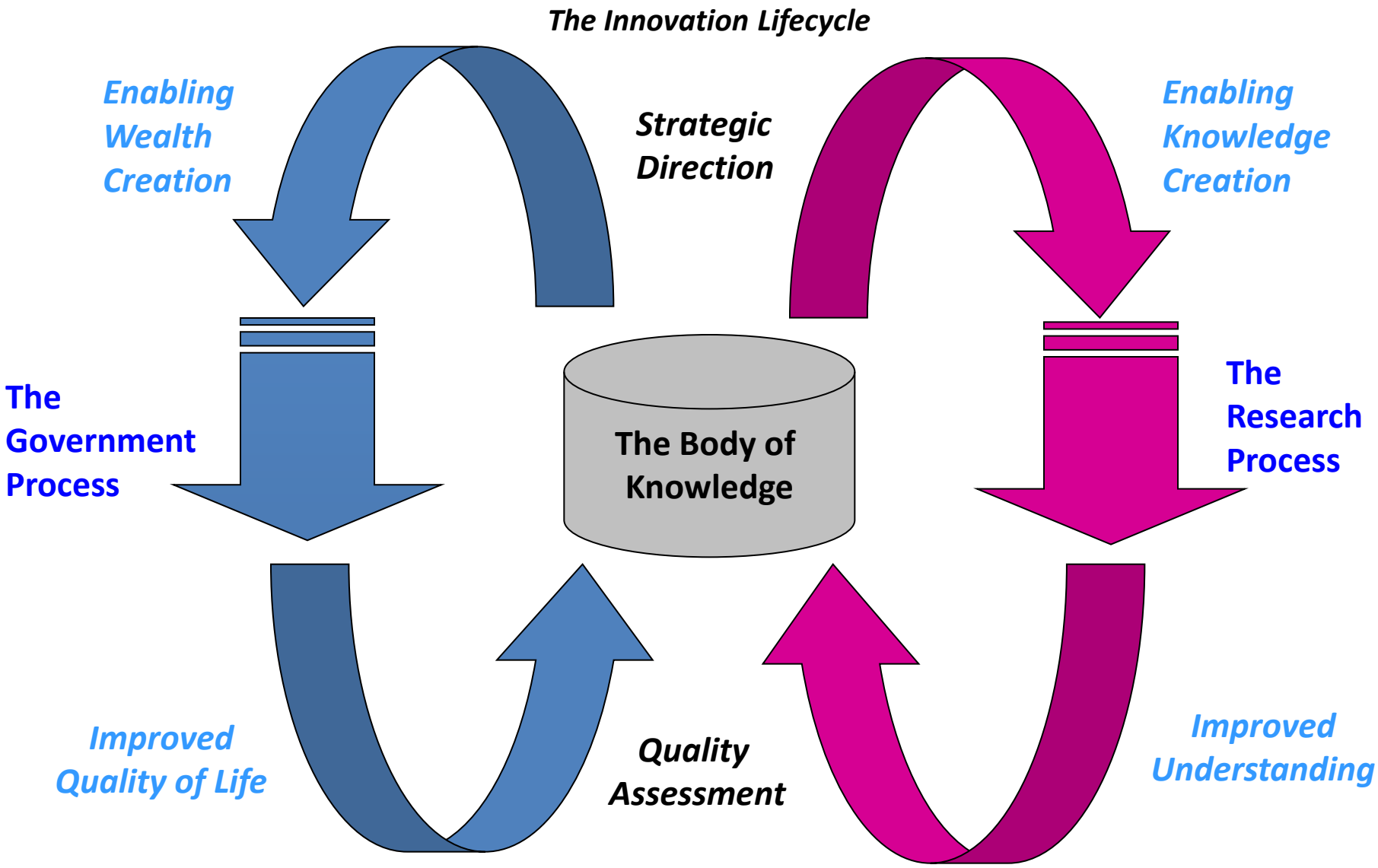
1. The Policy Context
   - OECD
   - EC/NSF/…
   - G8+5
   - RCUK
   - Royal Society
   - G8
2. PaNdata
   - Photon and Neutron Open Data Infrastructure
3. The Research Data Alliance
   - Fostering Collaboration on a global scale

# 1. The Policy Context

- OECD, 2004-2006
  - Principles and *Guidelines* for Access to Research *Data* from *Public* Funding
- EC, 2007-2012
  - Recommendation on access to and preservation of scientific information
- G8+5, 2011-2012
  - Global Research Infrastructure Sub Group on Data
- Research Councils UK, 2011
  - Joint Principles on Data
- Royal Society, 2011-2012
  - Science as an Open Exercise
- G8 Ministerial Statement, 2013
  - Grand Challenges, Global Research Infrastructures,
  - Open Scientific Research Data, Open Access

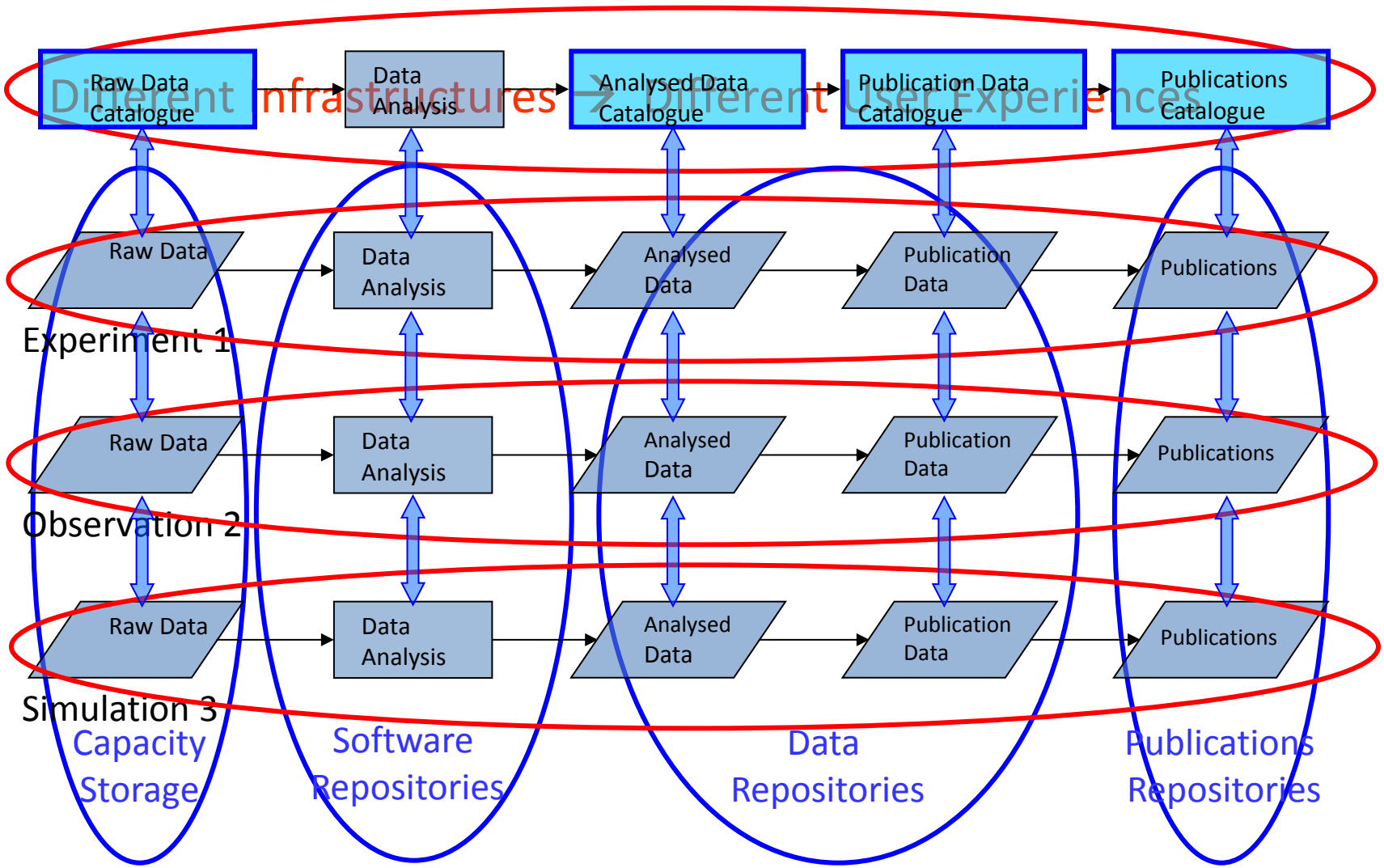
The views expressed herein are the personal views of the author and do not necessarily reflect the views of the policy makers

# Economic Impact

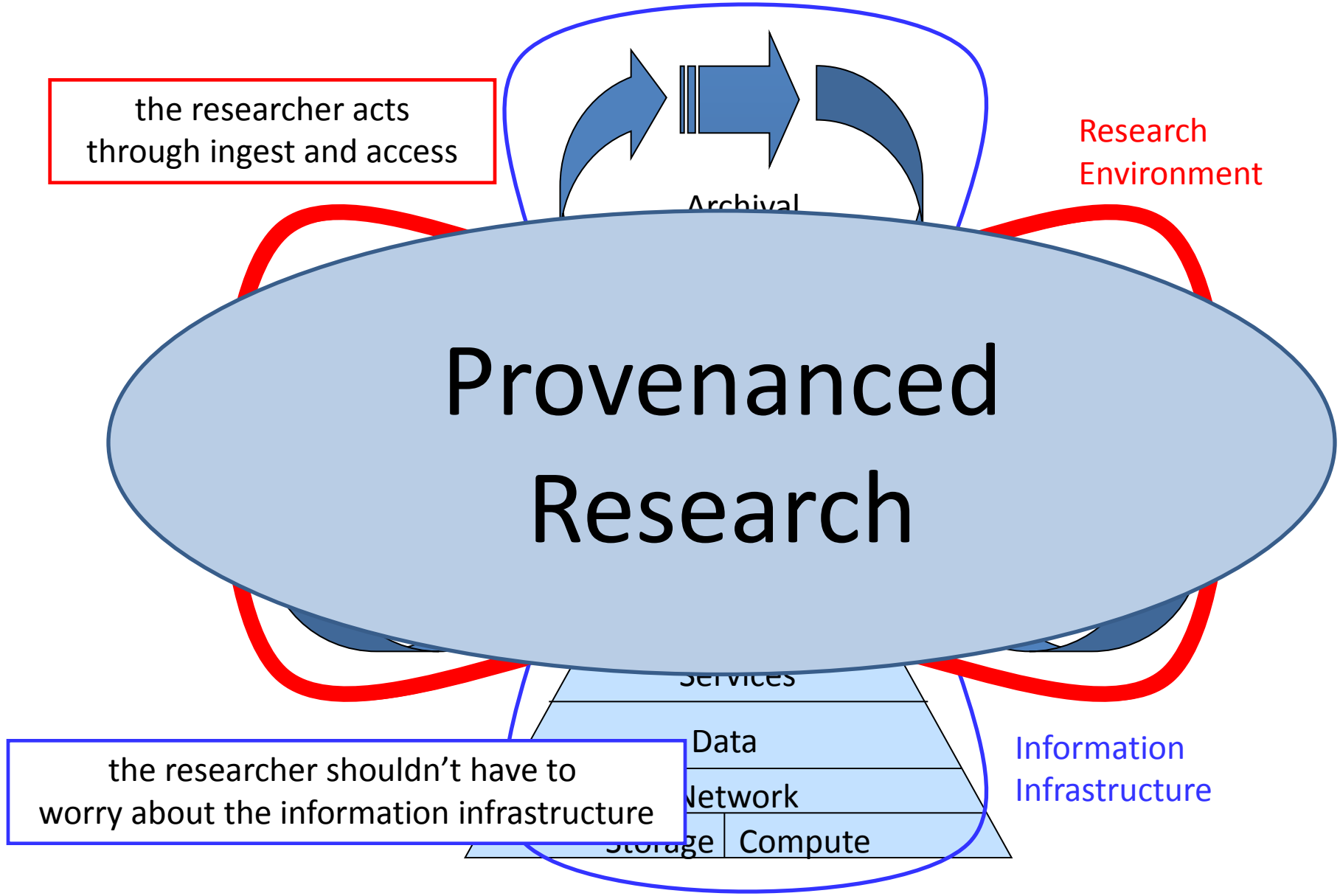



Aggregation of Knowledge lies at the heart of the innovation lifecycle

# Technology Sharing



# Open Science



# RCUK principles: Data are a Public Good

*Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.*

***Public good*** – is nonrival and non-excludable [wikipedia]

consumption by one does not reduce availability for others

no one can be effectively excluded from using

***Research Data***

recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings

***As few restrictions as possible***

Later (distinguish registration from restriction)

***Timely***

Later (discipline specific)

***Responsible***

Later (maximising access does not necessarily maximising research benefit)

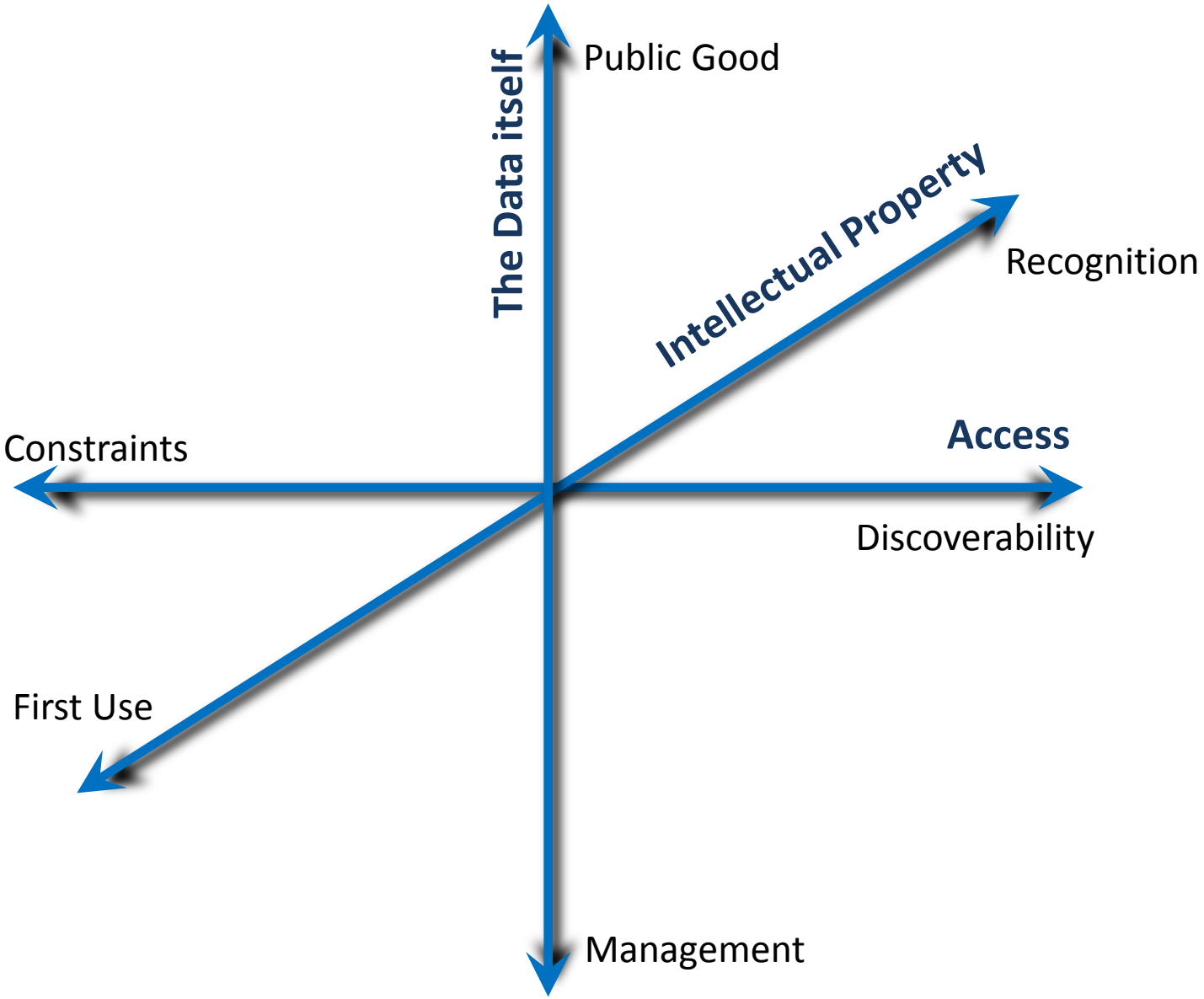
***Intellectual Property***

Later (balance contribution from sharing and from primary research)

# RCUK Principles on Data Policy

2) Data should be managed
3) Data should be discoverable
4) There may be constraints
5) Originators may have first use
6) Reusers have responsibilities
7) Data sharing is not free

# 3 Dimensions of policy



Science & Technology Facilities Council

# Overview

1. The Policy Context
   - OECD
   - EC/NSF/…
   - G8+5
   - RCUK
   - Royal Society
   - G8
2. PaNdata
   - Photon and Neutron Open Data Infrastructure
3. The Research Data Alliance
   - Fostering Collaboration on a global scale

# What is STFC?

Square Kilometre Array

Large Hadron Collider

Pr des:

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

Data Management

ications

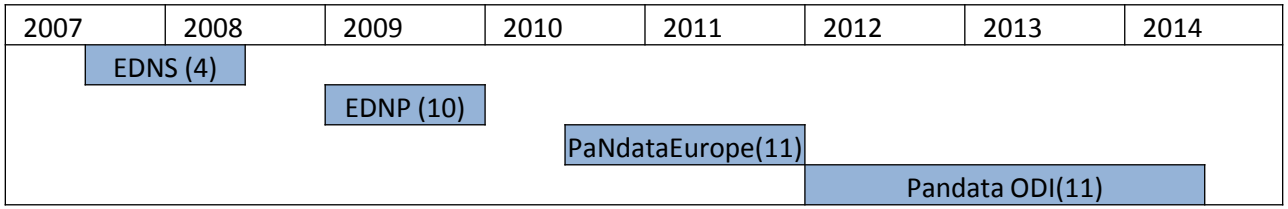
Daresbury Laboratory

ESRF & ILL, Grenoble

# 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..."*



# 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

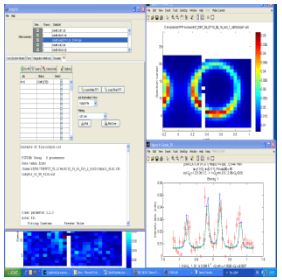
PaN-data is coordinated by the STFC Department of Scientific Computing

# The Science we do - Structure of materials

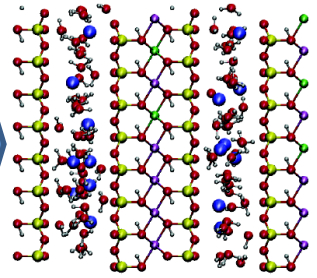
Visit facility on research campus

Place sample in beam

Diffraction pattern from sample



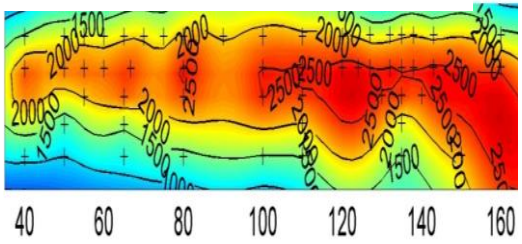
Fitting experimental data to model

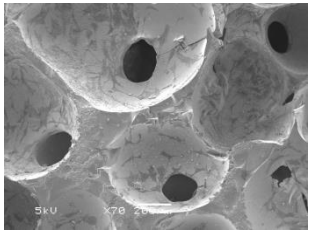


Structure of cholesterol in crude oil

- 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

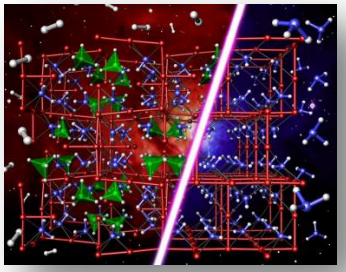
Longitudinal strain in aircraft wing

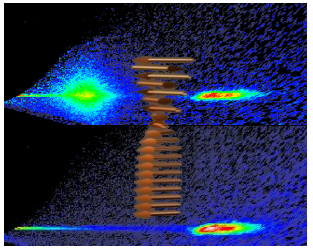


Bioactive glass for bone growth

Hydrogen storage for zero emission vehicles



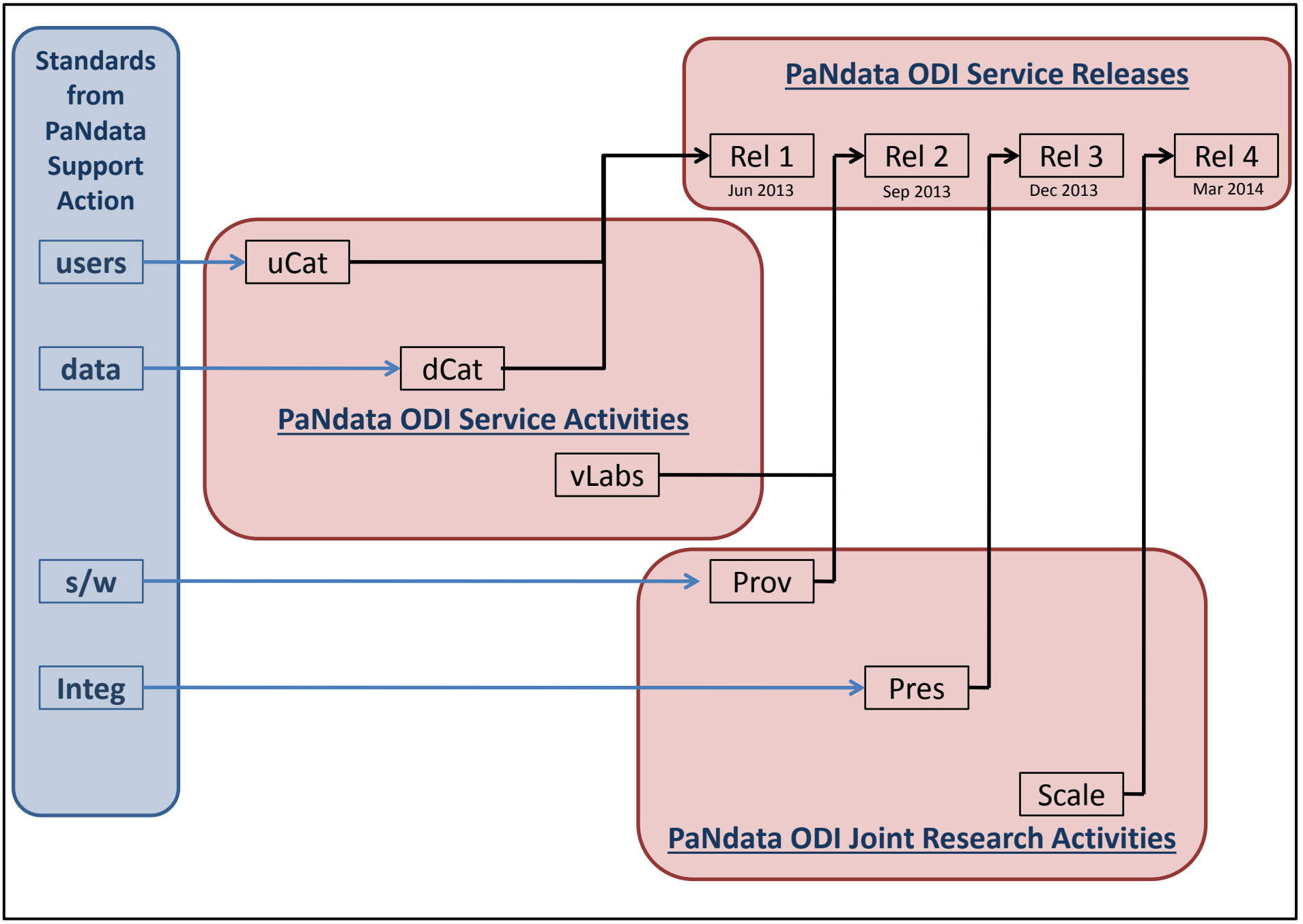


Magnetic moments in electronic storage

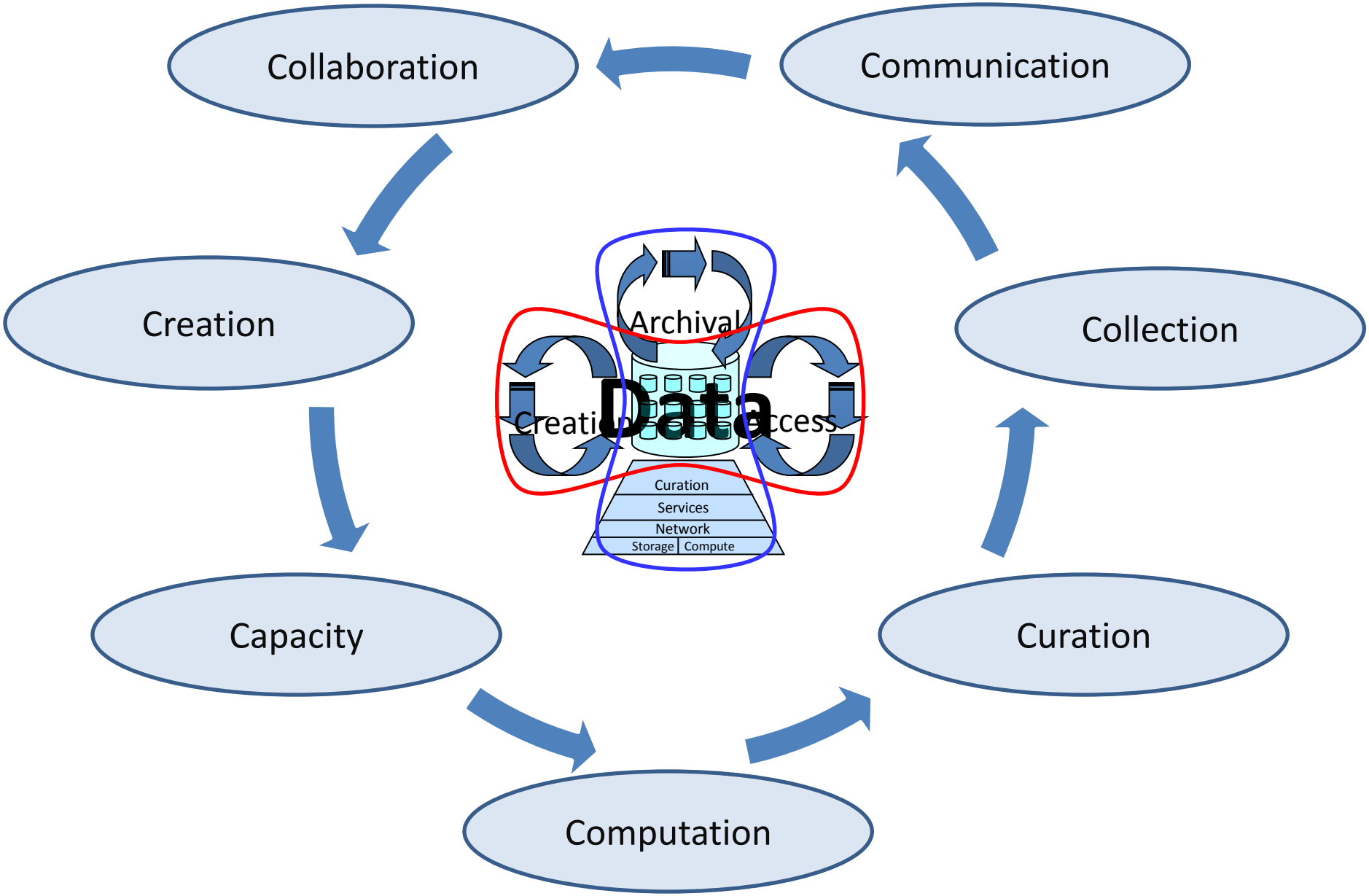
## PaN-data Standardisation

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


Standards from PaNdata Support Action
users
data
s/w
Integ
PaNdata ODI Service Releases
Rel 1
Jun 2013
Rel 2
Sep 2013
Rel 3
Dec 2013
Rel 4
Mar 2014
uCat
dCat
PaNdata ODI Service Activities
vLabs
Prov
Pres
Scale
PaNdata ODI Joint Research Activities

# The 7 C's


Collaboration
Communication
Creation
Collection
Capacity
Curation
Computation
Archival
Data
Curation
Services
Network
Storage
Compute

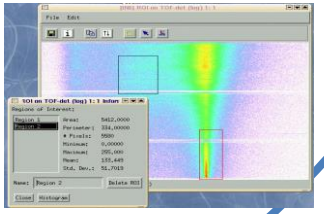
# Metadata Collection



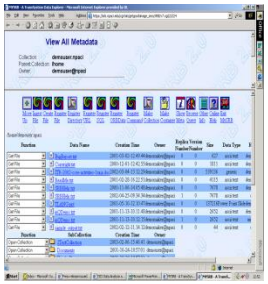
ICAT

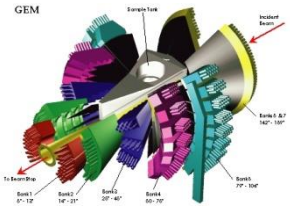
Proposal

Record Publication



Approval




Data analysis

Subsequent publication registered with facility

Scientist submits application for beamtime

Scheduling

Experiment

Data cleansing

Tools for processing made available

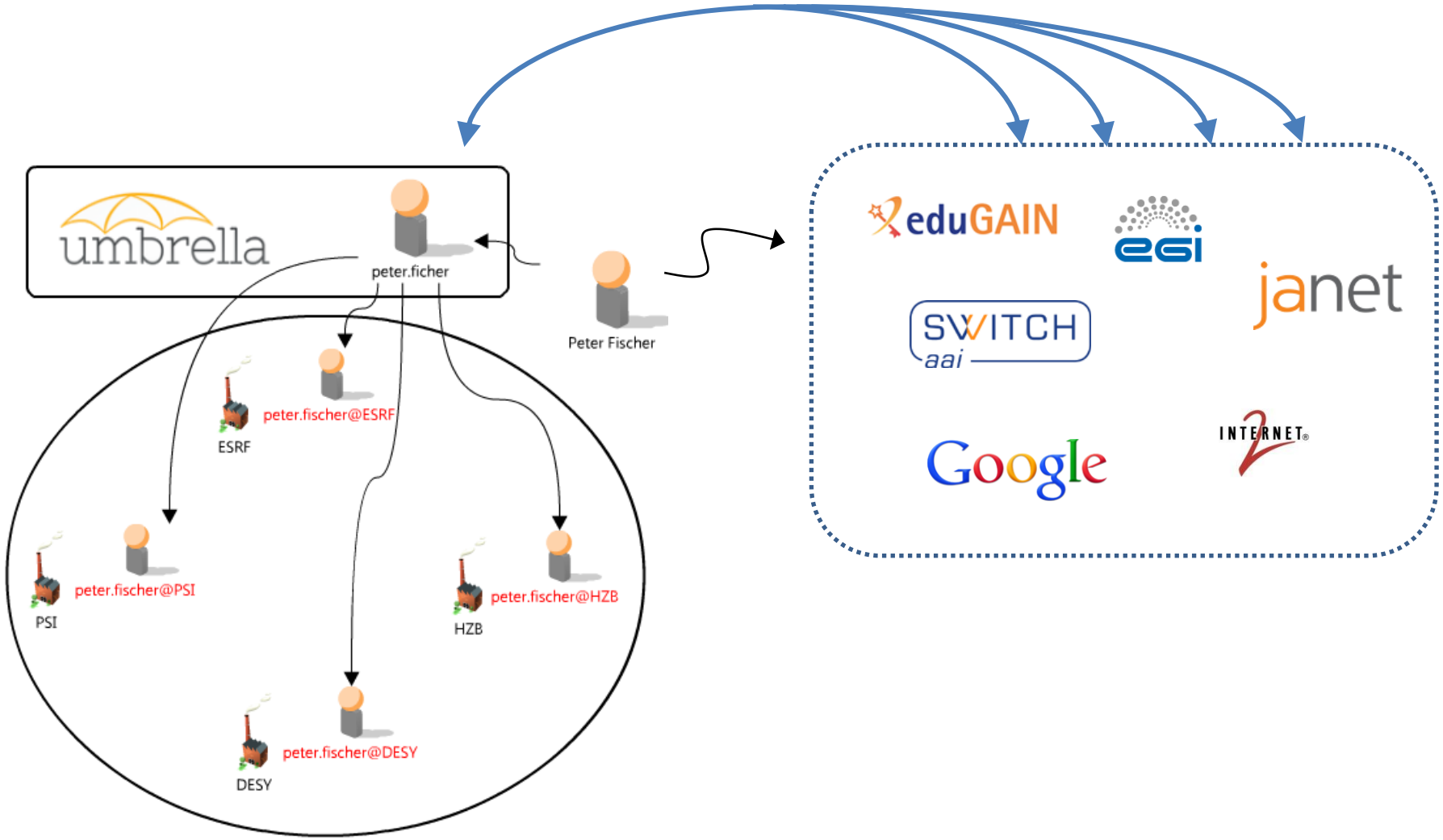
Facility committee approves application

Facility registers, trains, and schedules scientist's visit

Scientists visits, facility run's experiment

Raw data filtered and cleansed

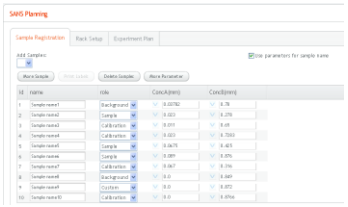
# Authentication




Credit: Bjorn Apt, PSI,

# Provenance:

## SANS2d: Experiment coordination



OpenGenie Script

**SampleTracks**

Sample Information

**Data Acquisition**

**Data Archive**

**Mantid**

**Data Processing**

raw data

**(Extended) ICAT Data Catalogue**

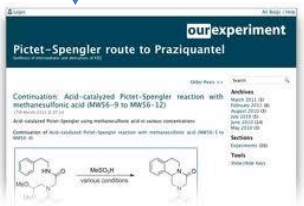
DOIs

**British Library DOI Server**

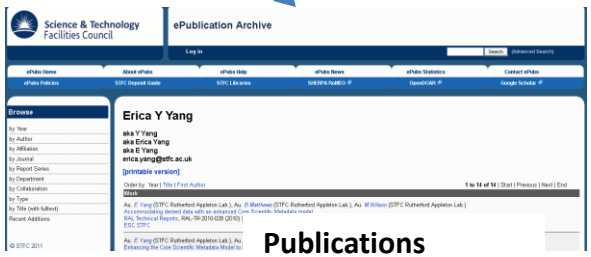
Outputs

derived data

New links



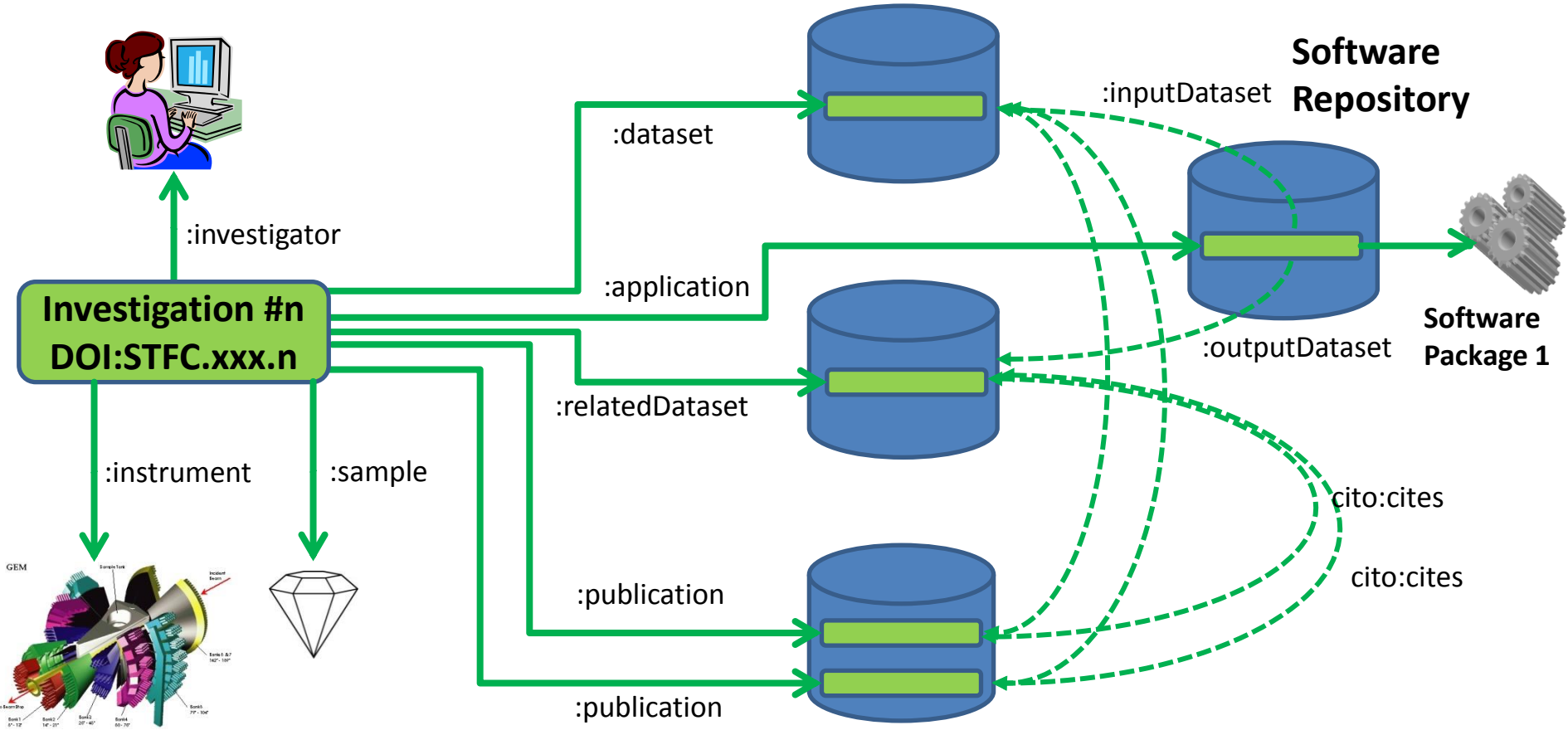

**ELN**




**Publications**

Credit: Brian Matthews, STFC,

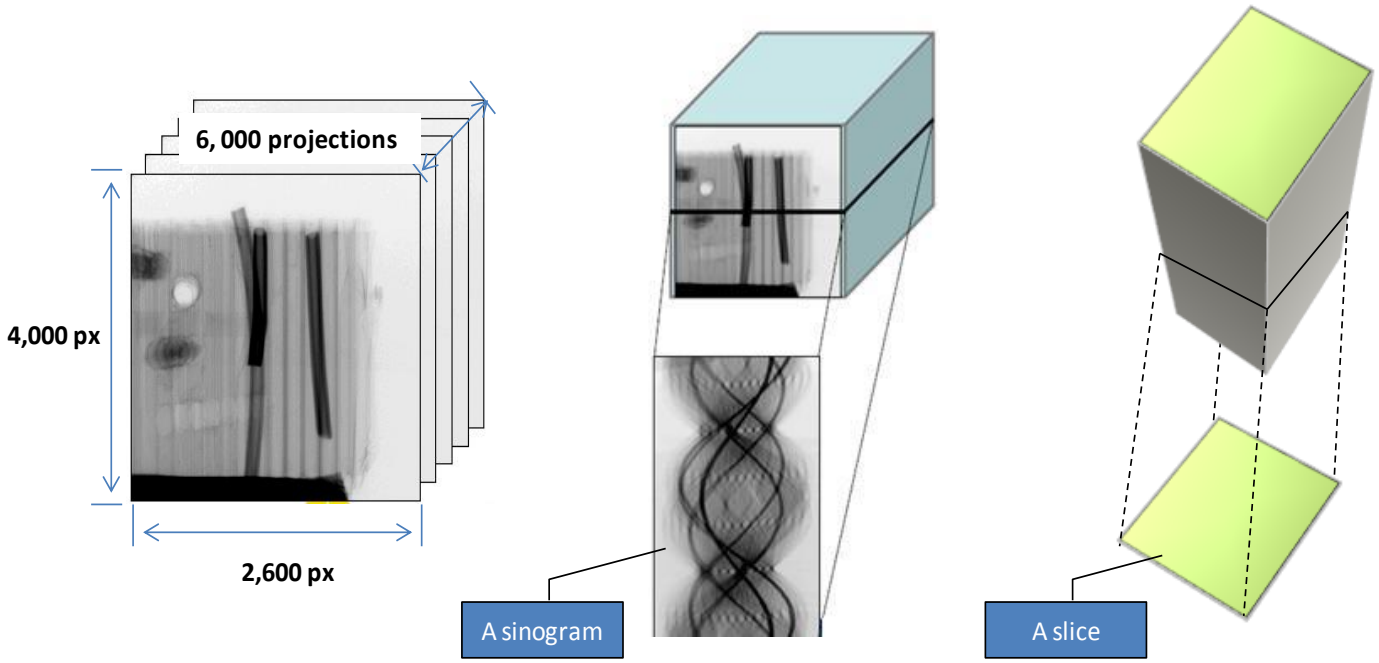
# Linking the software application into the research object




- Own metadata format (CSMD)
- OAI-ORE
- W3C Prov ontology
- Assume that the software is in a repository

Credit: Brian Matthews, STFC,

# Tomographic Reconstruction




Raw images

(2,600 X 4,000 px each image/projection)
(Total: 6,000 images/projections)

Sinograms
from the projections

(Total: 4,000 sinograms
Each sinogram: 2,600 x 4,000 x 1 px)

Reconstructed slices
in a 3d volume

(Total: 4,000 slices)

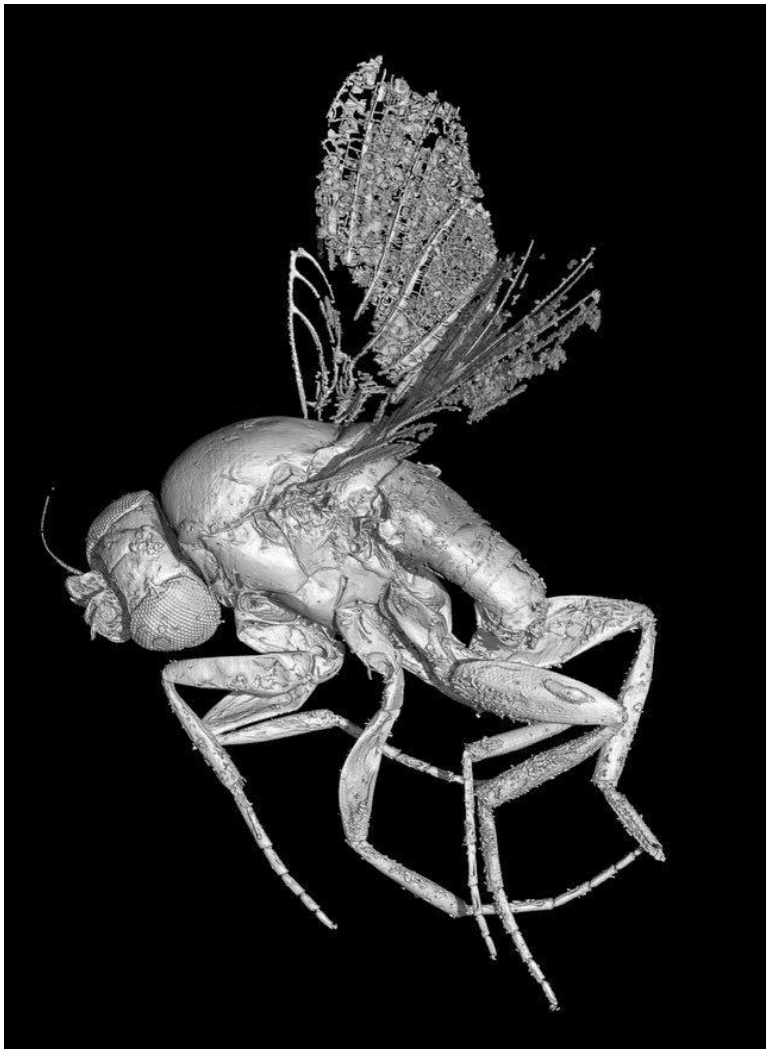
Credit: Mark Basham, Diamond,

~100Gb per 3D image - ~40 mins on 16 GPU cluster
~10 TB per experiment" - ~3 days on site
~ 1PB per year (per beamline)
Working on using the Emerald (376 GPUs)

# ESRF example: Amber inclusion

**Prioriphora schroederhohenwarthi**

- Xray imaging of 1mm *Prioriphora* (scuttle fly) from Cretaceous period
- found at Archingeay-Les Nouillers in opaque amber



Solorzano et al, 2011, Systematic Entomology (2011)

Science & Technology Facilities Council

# Overview

1. The Policy Context
   - OECD
   - EC/NSF/…
   - G8+5
   - RCUK
   - Royal Society
   - G8
2. PaNdata
   - Photon and Neutron Open Data Infrastructure
3. The Research Data Alliance
   - Fostering Collaboration on a global scale

# 3. The Research Data Alliance

New 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**

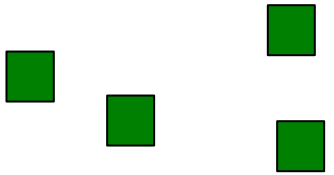
- The RDA is not a government organization or regulatory body and, instead, is a public body responsive to its members; and

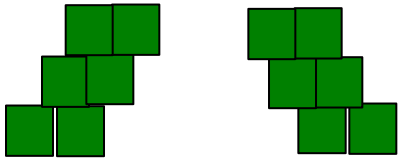
**Non-profit**

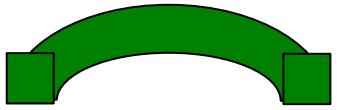
- RDA is not a commercial organization and will not design, promote, endorse, or sell commercial products, technologies, or services.

# “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







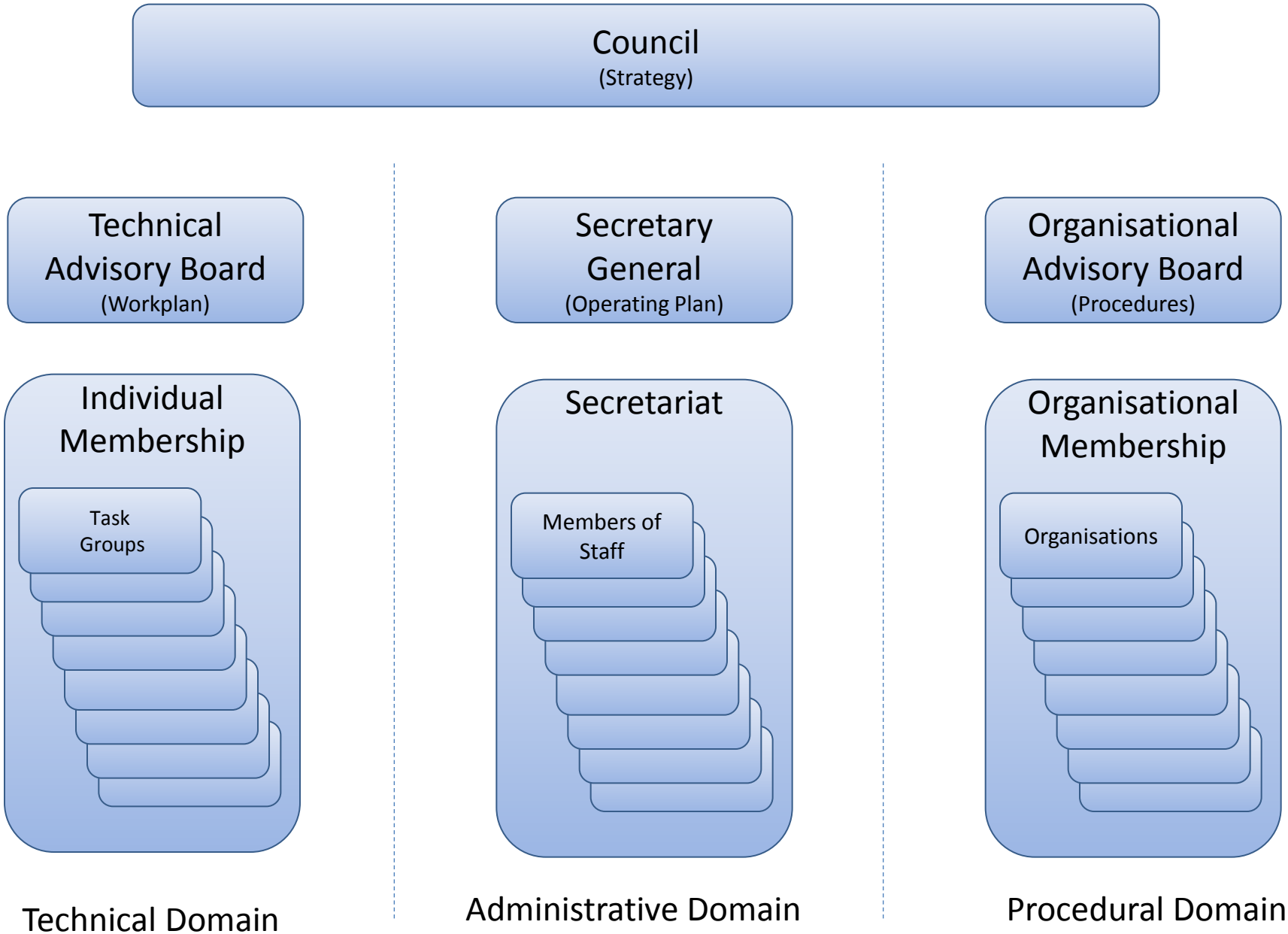
# RDA role

Two bridges we can build:

- Connecting Data
- Connecting People

What kind of organisation do we need to do this?

# RDA Bodies


Council
(Strategy)
Technical Advisory Board
(Workplan)
Secretary General
(Operating Plan)
Organisational Advisory Board
(Procedures)
Individual Membership
Task Groups
Secretariat
Members of Staff
Organisational Membership
Organisations
Technical Domain
Administrative Domain
Procedural Domain

## Data Practitioners Domain

## Administrative Domain

**Working Groups and Interest Groups**

- Carry out work of RDA
- Reach consensus on outputs
- May suggest BoFs about new topics
- Open to all but…
  - some commitment expected

**Plenary**

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

**Technical Advice Committee**

- advise on WG work activities
- Interacting directly with working groups
- advise on new WGs and new BoFs
- Give implementation suggestions to strategic direction from council

**Online Open Interaction Fora**

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

**Admistration 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

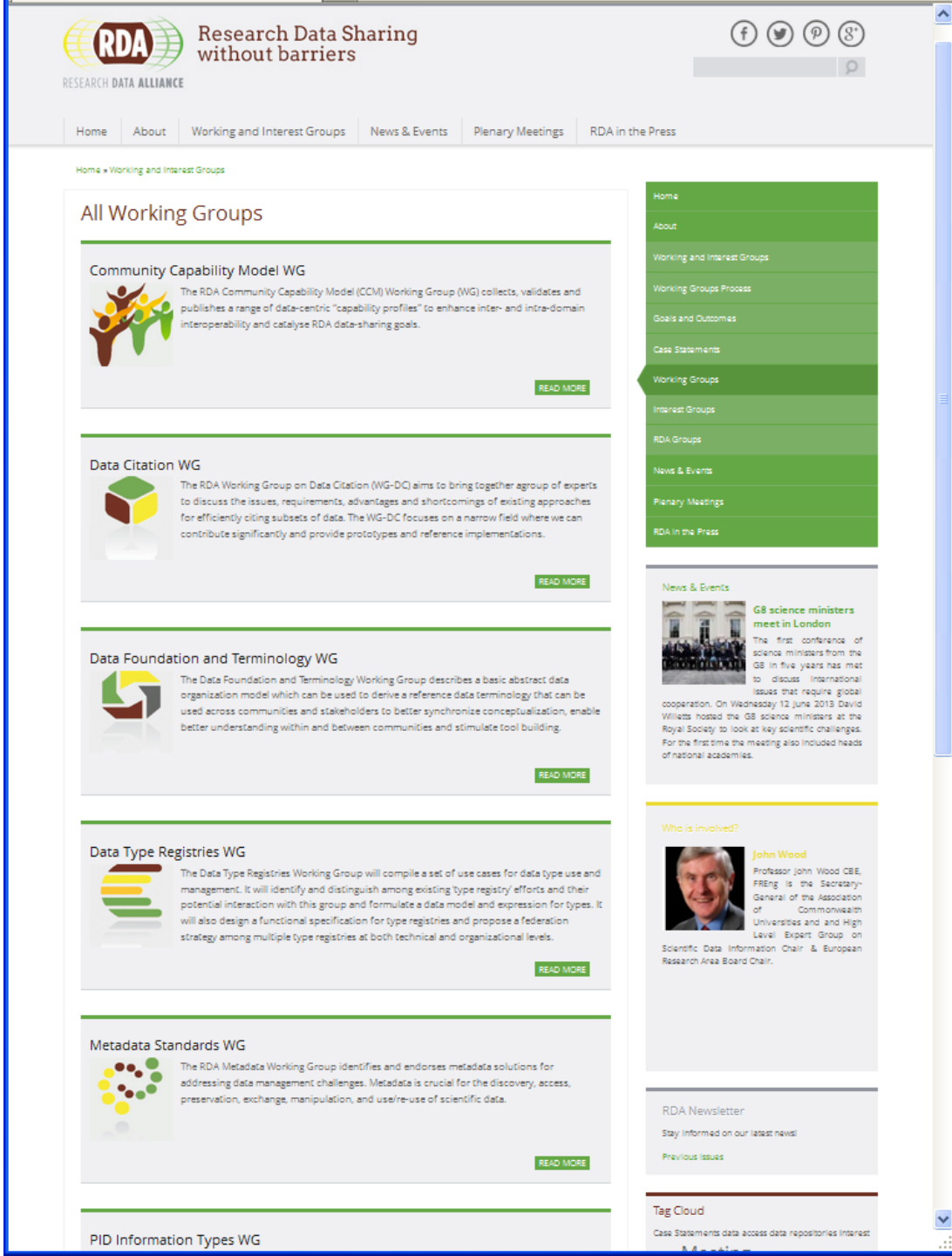
- - Council, funders,….

**Council**

- Set strategic direction
- Fin governance matters
- Approve new WGs (TAC advised)
- control balanced WG approach

# Example RDA Working Groups

- Data Citation
- Data Foundation and Terminology
- Data Type Registries
- Metadata Standards
- PID Information Types
- Practical Policy
- Standardisation of Data
- …



# Some Risks

- Standardisation is easy, I've done it a hundred times

  (apologies to Mark Twain)

- Two easy ways to standardise:
  - The Imperial model
  - The Esperanto model

- Justify need, define benefit, involve stakeholders
- Make a small steps and reassess

- "Never generalise from one example"

# Supporting Projects

Three projects supporting RDA through its first phase:

- RDA/Europe (previously iCordi) EC Project
- RDA/US NSF Project
- Support in Australia through ANDS

Steering Group setting it up:

- US – Fran Berman, Beth Plale
- EU – Leif Laaksonen, Peter Wittenburg, Juan Bicarregui
- Australia – Ross Wilkinson, Andrew Treloar

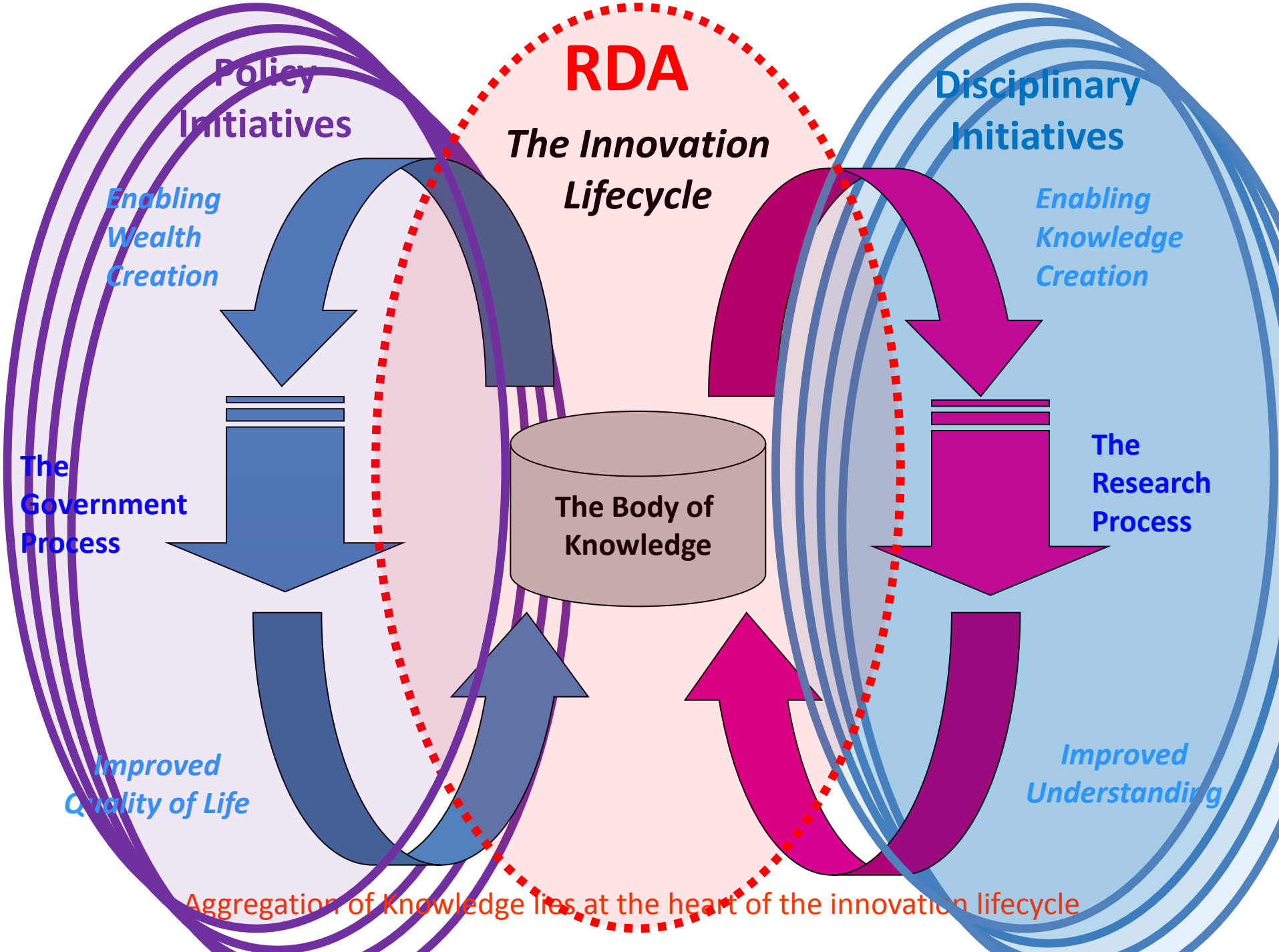
- TAB to be elected at 2nd Plenary
- First Oranisational Assembly at 2nd Plenary

# RDA Status in June 2013

- Pre-launch meetings in Munich and Washington September 2012,
  - ~200 Delegates
- Various Workshops eg through eIRG, IDCC, ….
- Launch and First Plenary, March 2013, Guttenberg,
  - ~250 participants
- Currently, 8 Working Groups and 14 Interest Groups
- **Second Plenary, September 16-18 2013, Washington**
- Third Plenary, March 26-28, 2014, Dublin
- Fourth Plenary, TBD

Please get involved by registering and participating in the discussions:

Website: rd-alliance.org/


RDA
The Innovation Lifecycle
Policy Initiatives
Enabling Wealth Creation
The Government Process
Improved Quality of Life
Disciplinary Initiatives
Enabling Knowledge Creation
The Research Process
Improved Understanding
The Body of Knowledge
Aggregation of Knowledge lies at the heart of the innovation lifecycle

Science & Technology Facilities Council

# Overview

1. The Policy Context
   - OECD
   - EC/NSF/…
   - G8+5
   - RCUK
   - Royal Society
   - G8

2. PaNdata
   - Photon and Neutron Open Data Infrastructure

3. The Research Data Alliance
   - Fostering Collaboration on a global scale


Thank You
www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
www.pan-data.eu
www.rd-alliance.org

# The End