PaNdata

Photon and Neutron Data Infrastructure

Zurich Airport Meeting

27 June 2011

Juan Bicarregui

Overview

The PaNdata Collaboration

The Vision

The PaNdata Europe Project

The PaNdata Open Data Infrastructure ~~Proposal~~ Project

Looking Forwards

The PaNdata Collaboration

- Established 2007 with ESRF, ILL, ISIS and Diamond
- Expanded since to 11 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 11 major European Research Infrastructures

STFC - SKA

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

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

STFC - SKA

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

CEA

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

INFM

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

STFC - LHC

PaN-data is coordinated by the e-Science Department at the Rutherford Appleton Laboratory, UK

PaN-data Applications

The partners operate hundreds of instruments used by over 30,000 scientists each year

These instruments support scientific fields as varied as:

- Physics, Chemistry, Biology, Material sciences, Energy technology, Environmental science, Medical technology and Cultural heritage

Applications include:

- crystallography that reveals the structures of viruses and proteins important for the development of new drugs
- neutron scattering that identifies stresses within engineering components such as turbine blades
- tomography that can image microscopic details of the 3D-structure of the brain

Industrial applications include pharmaceuticals, petrochemicals and microelectronics

Overview

The PaNdata Collaboration

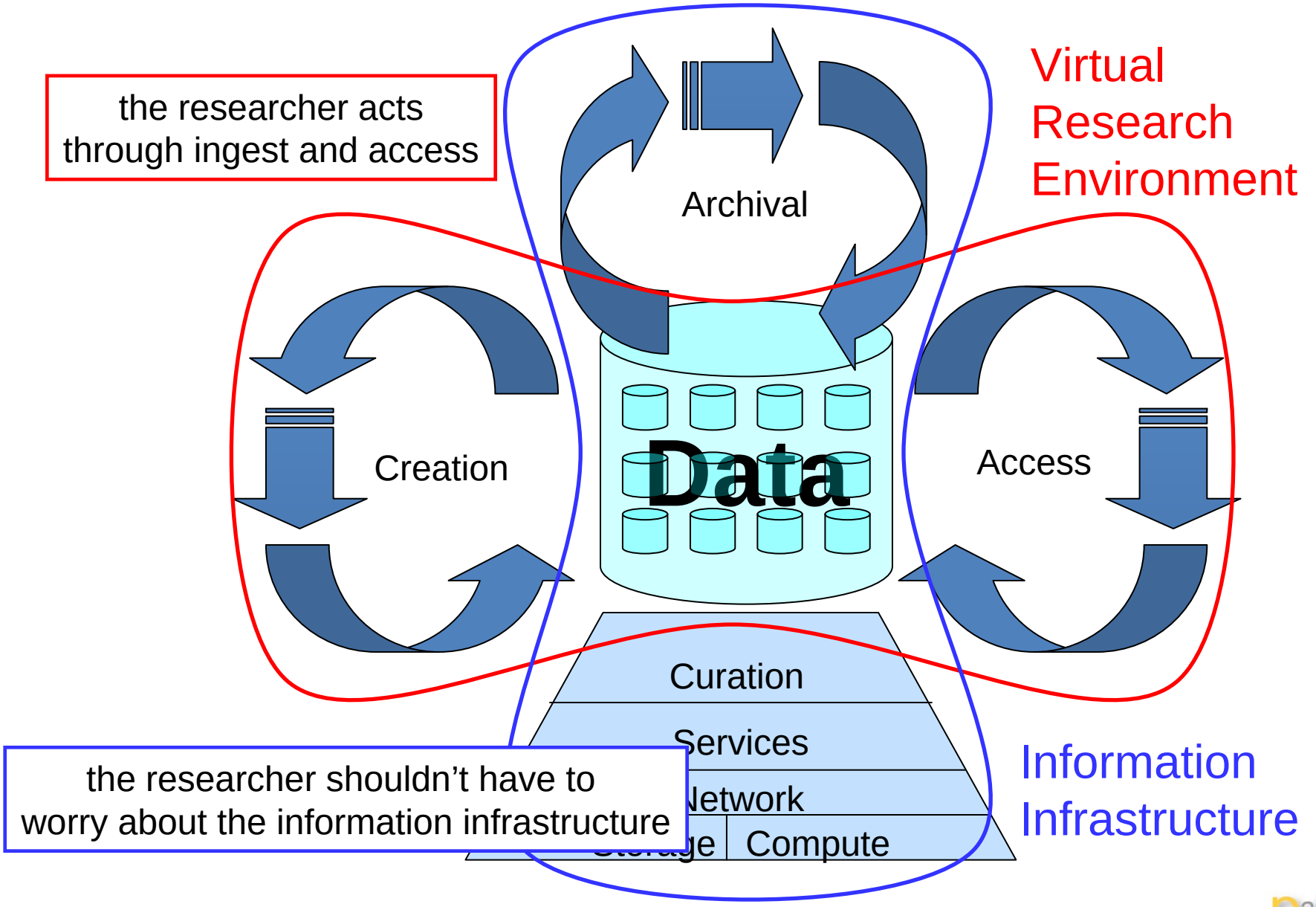
The Vision

The PaNdata Europe Project

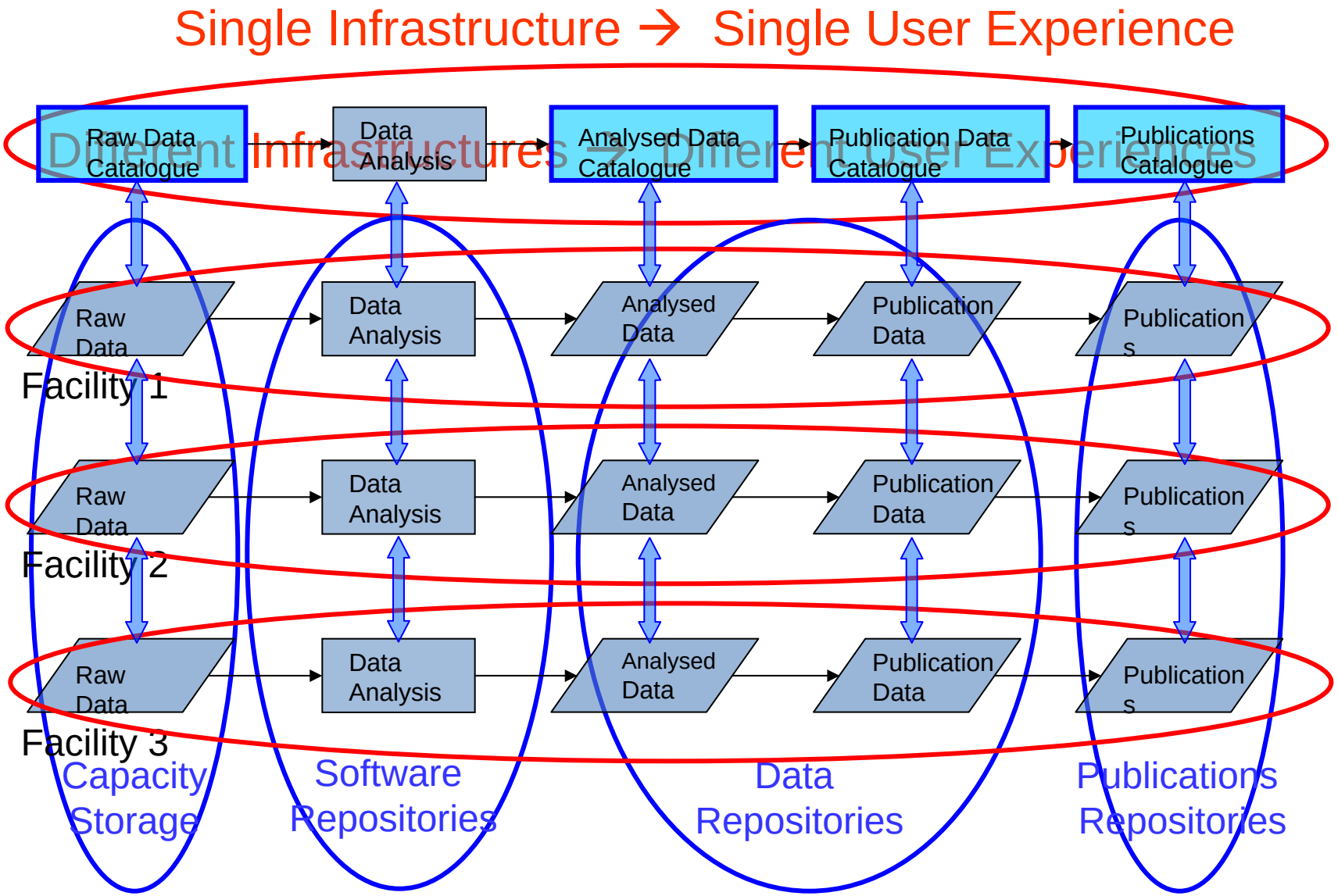
The PaNdata Open Data Infrastructure ~~Proposal~~ Project

Looking Forwards

What is e-Infrastructure?



PaNdata Vision



In words:

PANdata will provide our user communities with data repositories and data management tools to:

- deal with large sets and large data rates from the experiments,
- enable easy and standardised annotation of data,
- allow transparent and secure remote access to data,
- establish sustainable and compatible data catalogues, allow long-term preservation of data, and
- provide compatible open source data analysis software.

This will have a major impact on our scientific user community because it will offer:

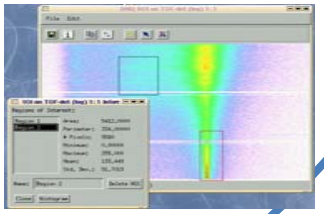
- cross facility and cross discipline data analysis,
- secure access to large data sets over the network instead of using portable media,
- maintaining the records of science by having properly annotated data,
- linking publications to data,
- allowing efficient software developments, and
- efficient scientific collaborations across Europe by providing compatible data formats and analysis software.

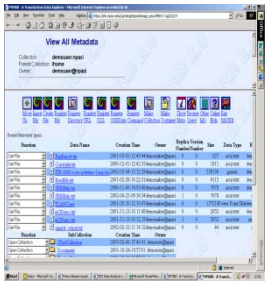
Metadata and Digital Curation

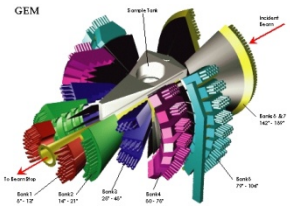
ICAT

Proposal

Record Publication







Approval

Data analysis

Subsequent publication registered with facility

Scientist submits application for beamtime

Scheduling

Experiment

Data cleansing

Facility committee approves application

Tools for processing made available

Facility registers, trains, and schedules scientist's visit

Scientists visits, facility run's experiment

Raw data filtered and cleansed

pandata

Overview

The PaNdata Collaboration

The Vision

The PaNdata Europe Project

The PaNdata Open Data Infrastructure ~~Proposal~~ Project

Looking Forwards

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**

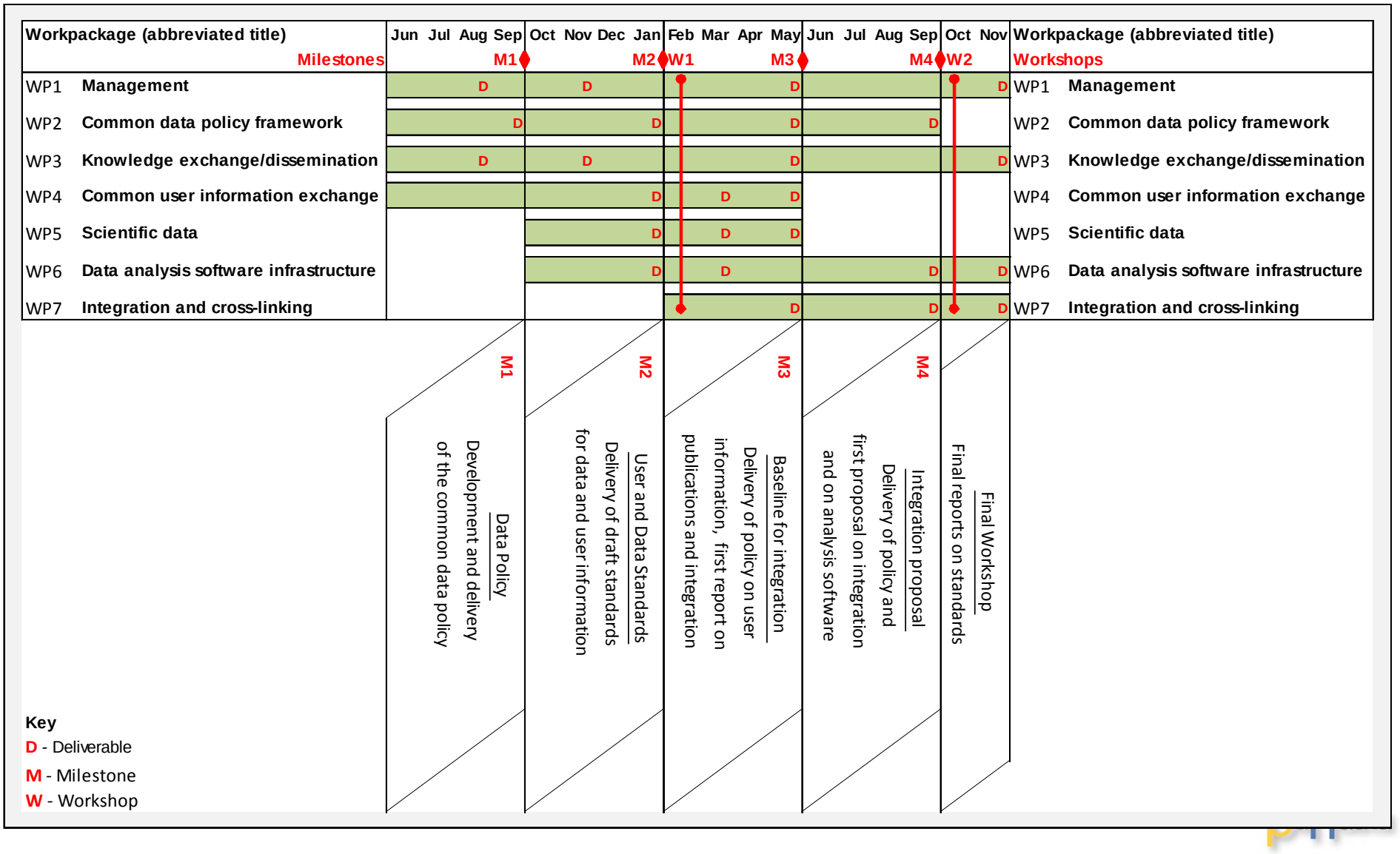
3. Definition of standards for common **scientific data formats**

4. Strategy for the interoperation of **data analysis software** enabling the most appropriate software to be used independently of where the data is collected

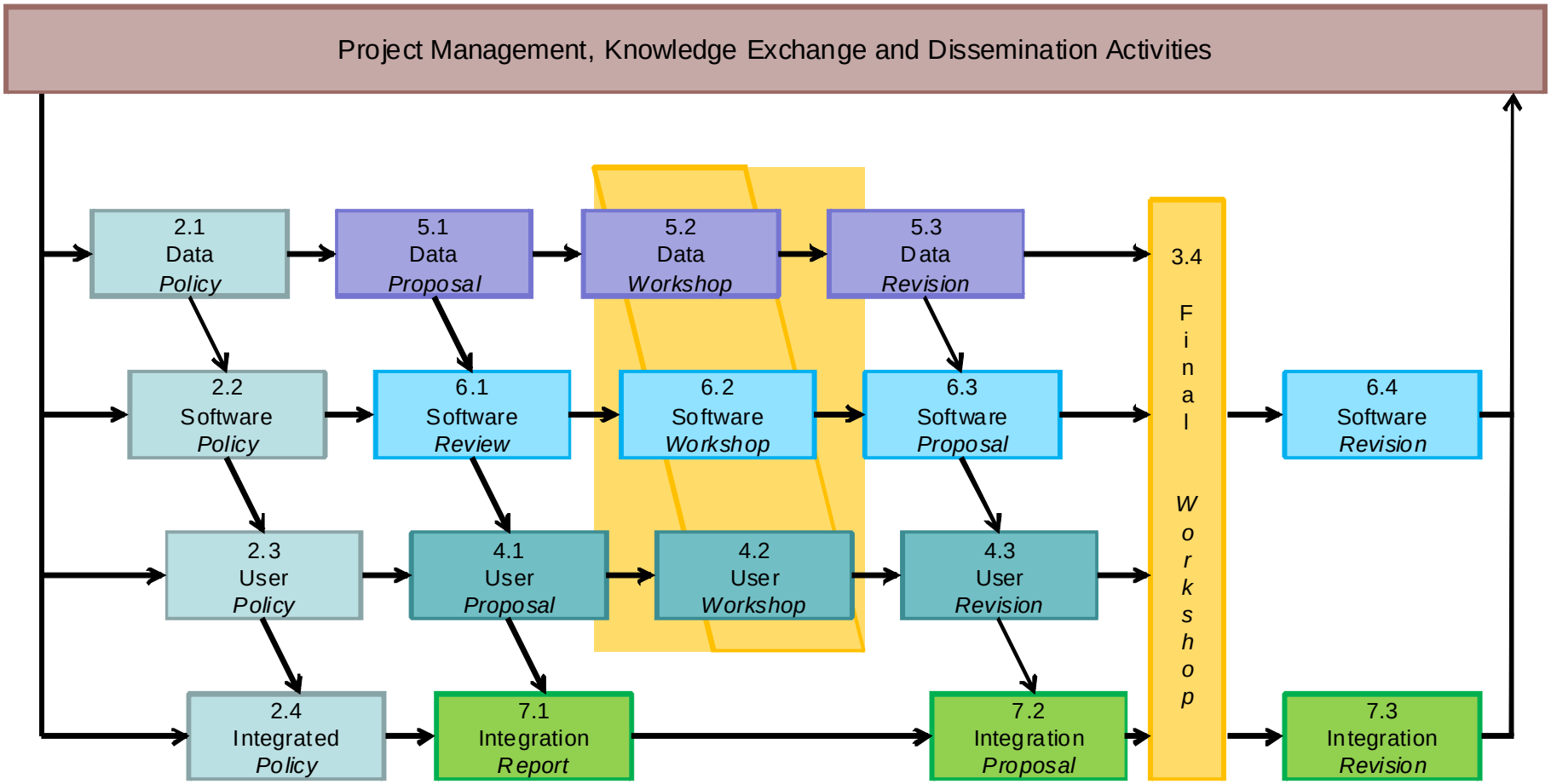
5. **Integration and cross-linking** of research outputs completing the lifecycle of research, linking all information underpinning publications, and supporting the long-term preservation of the research outputs

PaN-data Europe Timeline

PaN-data Europe runs from June 2010 until December 2011 with workshops in Spring and Autumn 2011.



Dependencies



Dependencies between the major project tasks

Overview

The PaNdata Collaboration

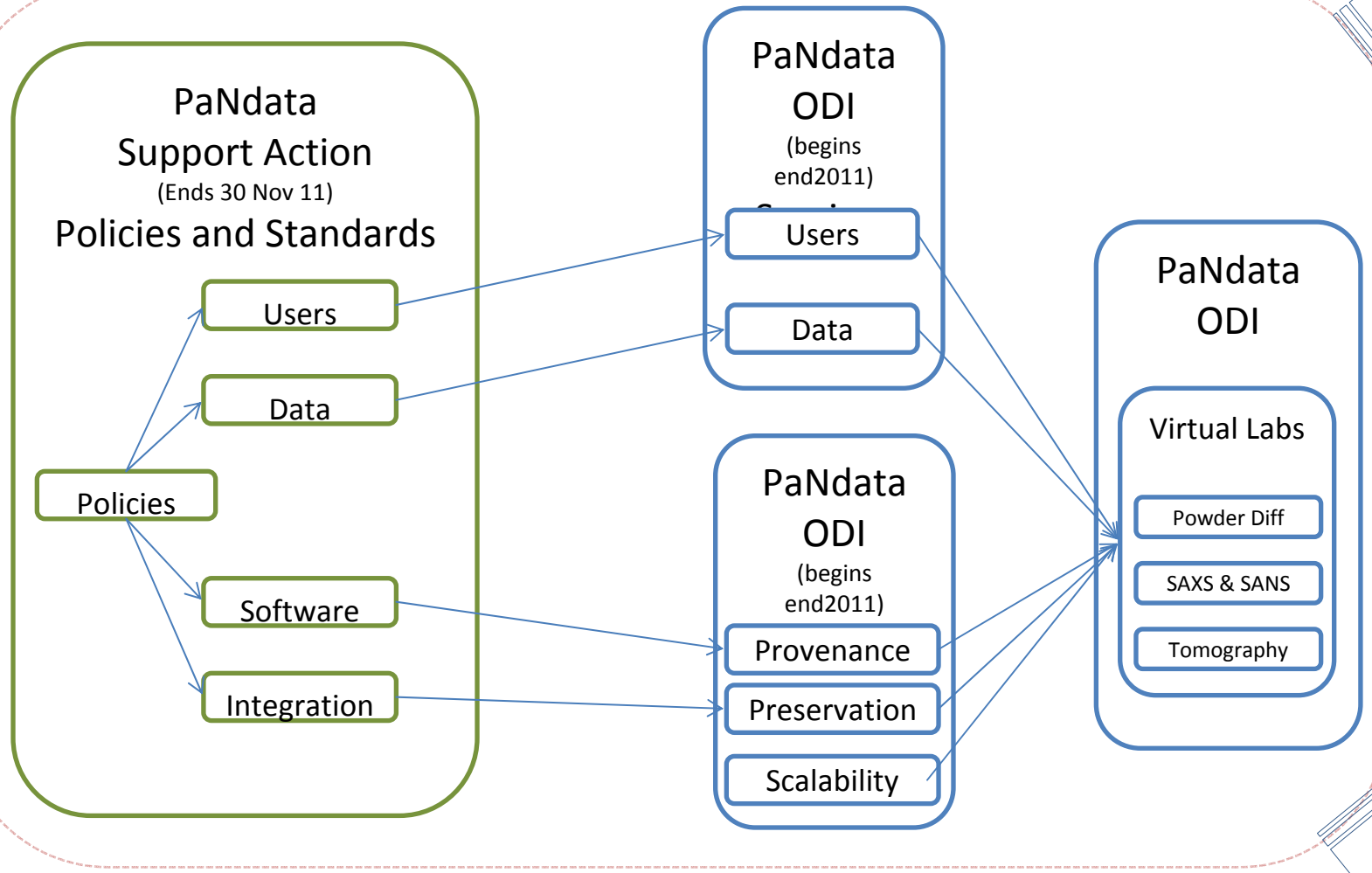
The Vision

The PaNdata Europe Project

The PaNdata Open Data Infrastructure ~~Proposal~~ Project

Looking Forwards

ERA Open Access Sharing Initiatives (examples, etc)



ERA Infrastructure Platform Initiatives (EGI, etc)

Objectives

Objective 2 – Users

To deploy, operate and evaluate a system for pan-European **user identification** across the participating facilities and implement common processes for the joint maintenance of that system.

Objective 3 – Data

To deploy, operate and evaluate a generic **catalogue of scientific data** across the participating facilities and promote its integration with other catalogues beyond the project.

Objective 4 – Provenance

To research and develop a conceptual framework, defined as a **metadata model, which can record the analysis process**, and to provide a software infrastructure which implements that model to **record analysis steps** hence enabling the **tracing of the derivation of analysed data outputs**.

Objective 5 – Preservation

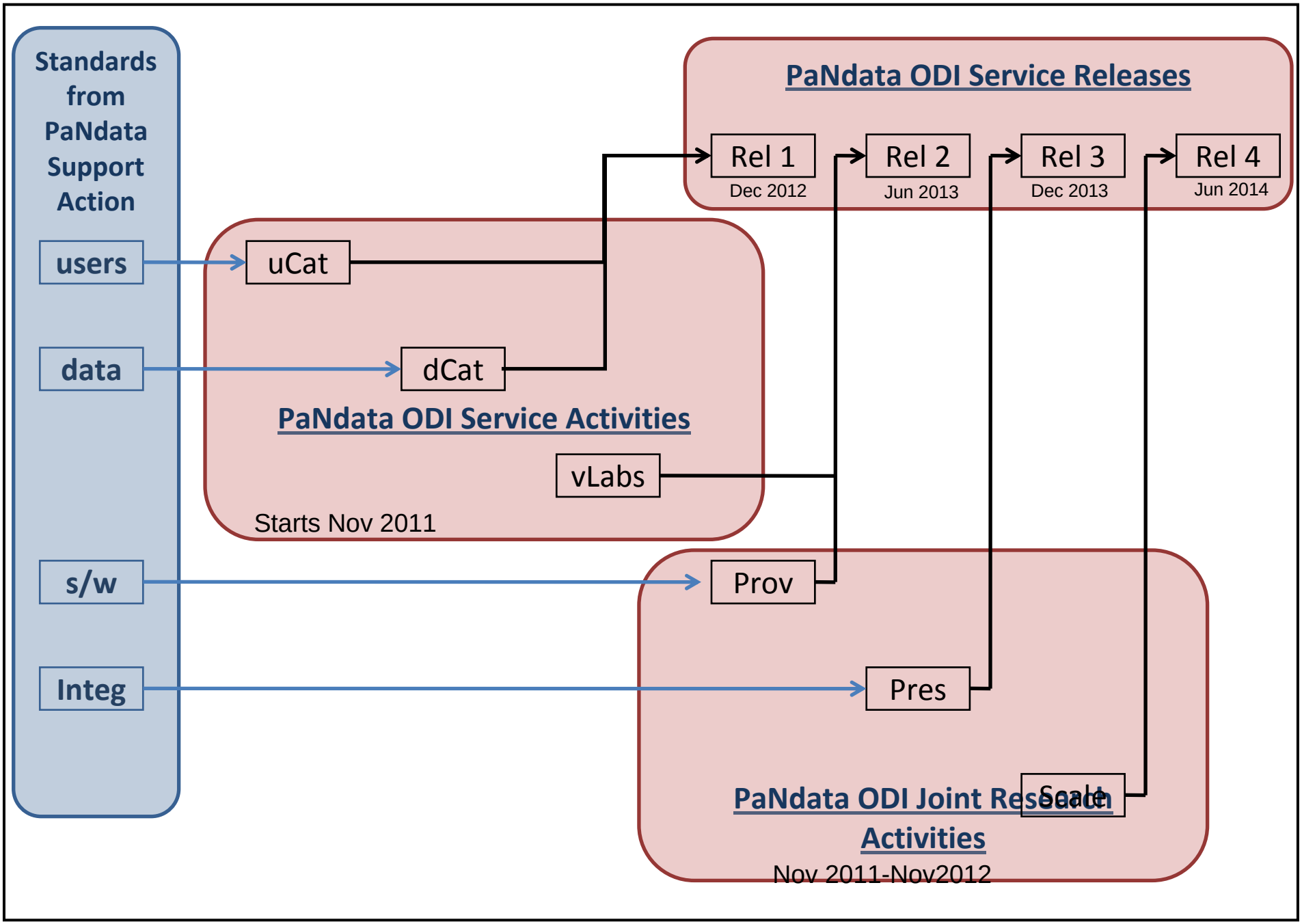
To add to the PaNdata infrastructure extra capabilities oriented towards **long-term preservation** and to integrate these within selected virtual laboratories of the project to demonstrate benefits. These capabilities should, as for the developments in the provenance JRA, be integrated into the normal scientific lifecycle as far as possible. The conceptual foundations will be the **OAIS** standard and the **NeXus** file format.

Objective 6 – Scalability

To develop **a scalable data processing** framework, combining **parallel filesystems** with a parallelized standard data formats (pNexus pHDF5) to permit applications to make most efficient use of dedicated multi-core environments and to permit simultaneous ingest of data from various sources, while maintaining the possibility for real-time data processing.

Objective 7 – Demonstration

To deploy and operate the services and technology developed in the project in **virtual laboratories** for three specific techniques providing a set of integrated end-to-end data services.


Standards from PaNdata Support Action
users
data
s/w
Integ
PaNdata ODI Service Releases
Rel 1
Dec 2012
Rel 2
Jun 2013
Rel 3
Dec 2013
Rel 4
Jun 2014
uCat
dCat
PaNdata ODI Service Activities
vLabs
Starts Nov 2011
Prov
Pres
Scale
PaNdata ODI Joint Research Activities
Nov 2011-Nov2012

WP2 Dissemination

Objectives

Engagement with other initiatives and dissemination of project results, in particular to other research infrastructures.

Task 2.1. Establish an external web site as an extension to the existing website for the PaNdata collaboration (www.pandata.eu).

Task 2.2. Establish an interest group for project news items via community channels, informing them of project progress.

Task 2.3. Presentations to relevant international audiences at conferences, symposia, other project meetings etc.

Task 2.4. Provision of the open source software and appropriate documentation to potential partner bodies.

Task 2.5. Workshops to present the integrated systems to user and facility communities.

D2.1 : Project Website (M1) – November 2011

D2.2 : Dissemination plan (M3)

D2.3 : First Open Workshop (M15) – January 2013

D2.4 : Open Source software distribution procedure (M21)

D2.5 : Second Open Workshop (M27) - January 2014

WP3 User Catalogue and AAA **Service**

Objectives

To deploy, operate and evaluate a protocol for pan-European user identification across the participating facilities and implement common processes for the joint maintenance of that system.

Task1: Consultation on existing software components → recommendations for technologies to be implemented.

Task 2: Set up team includes representatives from the user office and/or IT staff of the partners.

Task 3: Specify an architecture which ... builds on the IRUVX "umbrella" concept.

Task 4: Implement ... the necessary local modifications (including trust management).

Task 5: Implement a standard affiliation database which is accessible for update and use by the participating facilities ...

Introduce a central affiliation database according to the PaNdata de-facto standard.

Provide an interface of the local WUO systems to this standard.

Organise and support the migration of the local WUOs to this new affiliation database.

Task 6: Deploy the user management system at all participating facilities.

A major factor will be the integration with the facility's bespoke user administration systems.

The deployment will include setting up of an administration authority for the system.

Task 7: Evaluate the system within a subset of the collaborating facilities.

Task 8: Operate and report on the AAA trust system for the remainder of the project.

Task 9: Maintain communication with other user authentication systems (through Workpackage 2) ...

D3.1 : Specification of AAA infrastructure (M6) Apr 2012

D3.2 : Pilot deployment of initial AAA service infrastructure (M12) Nov 2012

D3.3 : Production deployment of AAA service infrastructure (M18) Apr 2013

D3.4 : Evaluation of initial AAA service infrastructure (M24) Nov 2014

WP5 Virtual Laboratories (Service)

Objectives

To deploy a set of integrated end-to-end user and data services supporting three specific techniques:

- Structural 'joint refinement' against X-ray & neutron powder diffraction data
- Simultaneous analysis of SAXS and SANS data for large scale structures
- Access to tomography data exemplified through paleontological samples

D5.1: Specific requirements for the virtual laboratories (M6) Apr 2012

D5.2: Deployment of Specification of the three virtual laboratories (incorporating any specific requirements software to support them) (M18) Apr 2013

D5.3: Report on the implementation of the three virtual laboratories (M30) Apr 2014

WP6 Provenance (JRA)

Objectives

To develop a conceptual framework, which can record and recall the data continuum, and especially the analysis process, and to provide a software infrastructure which implements that model to record analysis steps hence enabling the tracing of the derivation of analysed data outputs

Task 1: Requirements for Provenance

Task 2: Modelling the data continuum

Task 3: Ontologies for specific instruments/techniques

Task 4: Tool Support for the Data Continuum

Task 5: Tracing the Data Continuum

Task 6: Evaluation

D6.1: Model of the data continuum in Photon and Neutron Facilities (M12) Nov 2012

D6.2: Common ontology definition and definition of tools to support the use of provenance for Photon and Neutron Facilities (M18) Apr 2012

D6.3: Tools for building research objects in Photon and Neutron Facilities (M24) Nov 2013

D6.5: Evaluation report on provenance management in Photon and Neutron Facilities (M30) Apr 2014

WP7 Preservation (JRA)

Objectives

To incorporate models and tools oriented towards long-term data preservation into the PaNdata infrastructure, focussing on several aspects considered of benefit: an OAIS-based infrastructure; persistent identifiers; and certification of authenticity and integrity

Task 1. Baseline and OAIS application

Task 2. Persistent identifiers (for datasets)

Task 3. Representation information and archiving

RI for datasets, and AIPs (Archival Information Packages)

This will include software as a kind of representation information, and the need to preserve the software itself.

Task 4. Integrity of datasets

Mechanisms for maintaining and checking integrity of datasets. (for individual datasets (as preservation actions are performed) and for data holdings as a whole.

Task 5. Evaluation and reporting

D7.1 Implementation of persistent identifiers for PaNdata datasets (M15) Jan 2013

D7.2 Mechanisms and tools for representation information and archiving (M21) July 2013

D7.3 Mechanisms and tools for integrity of datasets(M27) Jan 2014

D7.4 Report on evaluation of preservation mechanisms (M30) Apr 2014

WP8 Scalability (JRA)

Objectives

To develop a **scalable data processing framework combining parallel filesystems with a parallelized standard data format** (pNexus pHDF5) to permit applications to make most efficient use of dedicated multi-core environments and to permit simultaneous ingest of data from various sources, while maintaining the possibility for real-time data processing.

Task 1: pNexus API (Develop a pHDF5 compliant Nexus API.)

Task 2: Investigate parallel file systems.

Task 3: Investigate implementations on specific file systems

MPI-I/O implementations and pHDF5/pNexus on an even smaller number of preselected file systems.

Task 4: Coupling of advanced (pre-)processing engines.

- Test the capability of the system to cope with multiple parallel data streams. This will contain for example explicit tests feeding a pHDF5-file consisting of a large number of individual images into a multi-core analysis engine.

Task 5: Demonstration.

D8.1: Definition of pHDF5 capable Nexus implementation (M9) - Software

D8.2: Evaluation of Parallel filesystems and MPI I/O implementations (M9) - Report

D8.3: Implementation of pNexus and MPI I/O on parallel filesystems (M21) - Prototype

D8.5: Examination of Distributed parallel filesystem (M21) - Report

D8.6: Demonstrate capabilities on selected applications (M21) - Demonstrator

D8.7: Evaluation of coupling of prototype to multi-core architectures (M30) - Report

Overview

The PaNdata Collaboration

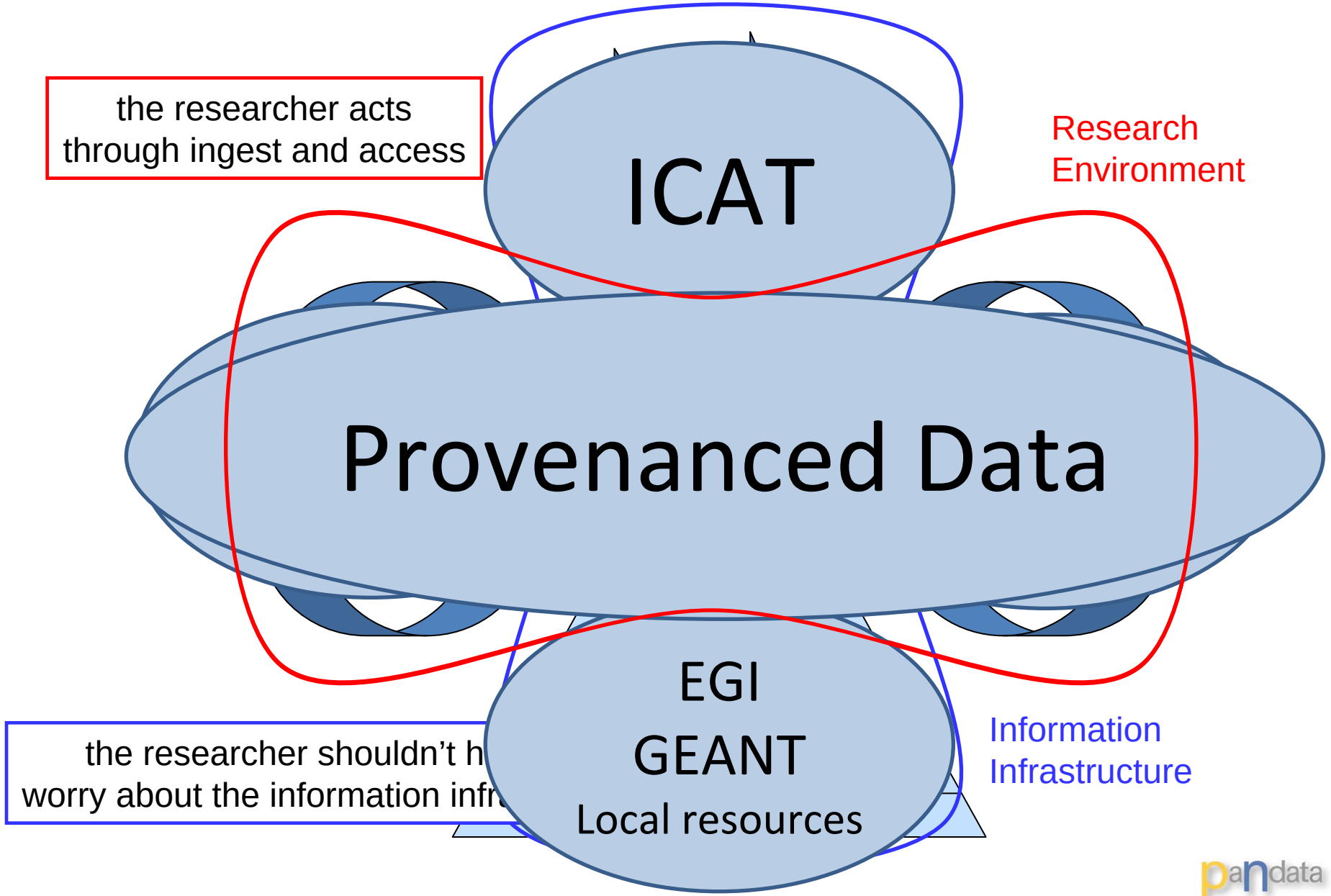
The Vision

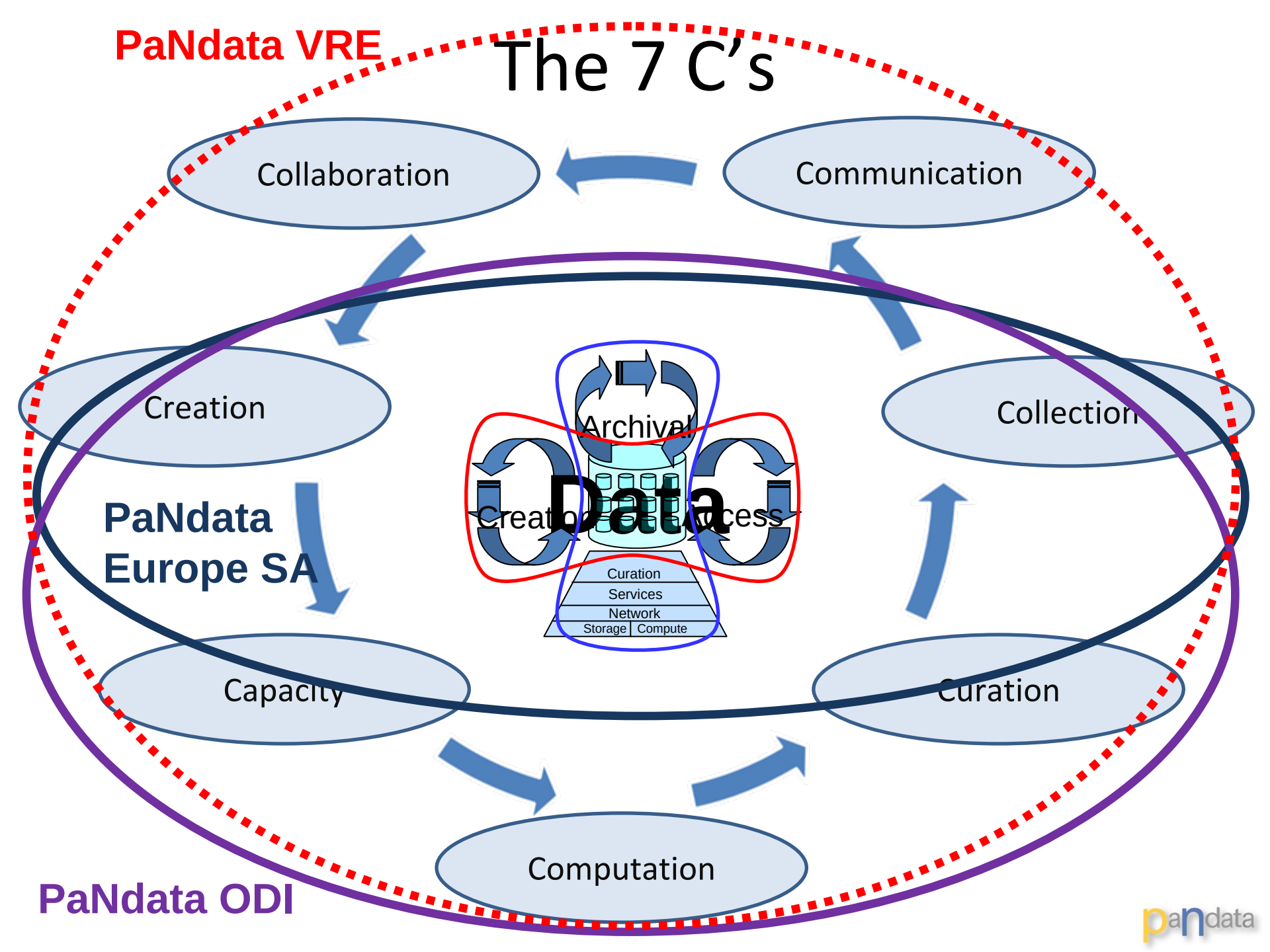
The PaNdata Europe Project

The PaNdata Open Data Infrastructure ~~Proposal~~ Project

Looking Forwards

The Research Lifecycle




PaNdata VRE
The 7 C's
Collaboration
Communication
Creation
Collection
Archival
Data
Curation
Services
Network
Storage
Compute
PaNdata
Europe SA
Capacity
Curation
Computation
PaNdata ODI

A Couple of Reflections

- **Seperate standardisation of:**
  - **Which user information to exchange**
    - **Start with standard affiliation names**
      - **(eg from National Federations)**
  - **Systems and Technology for exchange**

- **Very large overlap between CRISP and PaNdata**
  - **At least we should harmonise the meetings!**
  - **Final PaNdata Support Action in November**

Overview

The PaNdata Collaboration

The Vision

The PaNdata Europe Project

The PaNdata Open Data Infrastructure ~~Proposal~~ Project

Looking Forwards

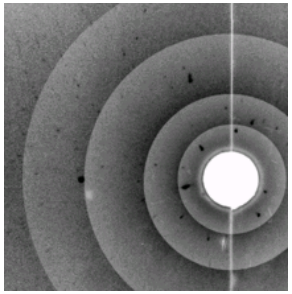
Thank You

www.pan-data.eu

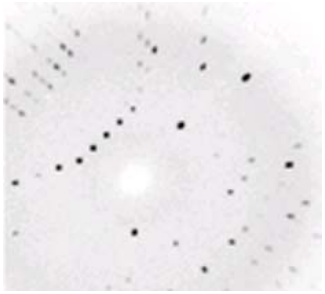
pandata

Science driver – Data Integration

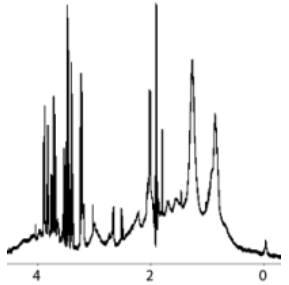
Neutron diffraction

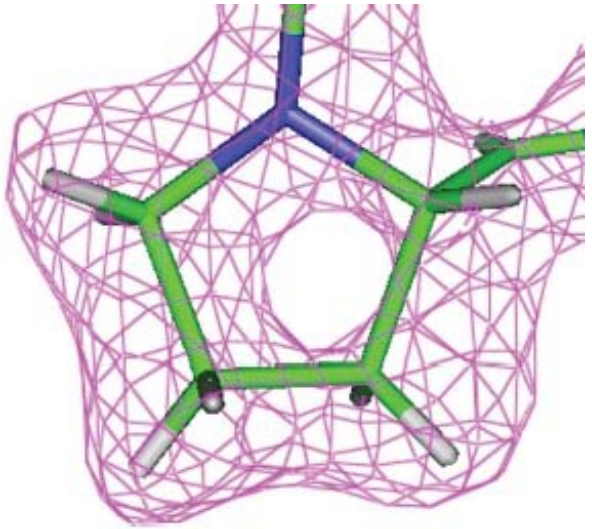


X-ray diffraction



NMR





High-quality structure refinement

pandata

WP leaders and Effort levels

Effort / Staff Months	Partner												Activity
	STFC	ESRF	ILL	Diamond	PSI	DESY	ELETTRA	Soleil	ALBA	HZB	CEA		
Workpackage	1	2	3	4	5	6	7	8	9	10	11		Type
Mgmt 1	10											10	COORD
Dissem 2	2	2	2	2	2	6	2	2	2	2	2	26	NA
uCat 3	3	18	3	3	18	3	3	3	3	3	3	63	SVC
dCat 4	6	3	6	3	6	3	18	3	3	3	3	57	SVC
vLabs 5	3	3	3	3	6	18	3	3	3	3	3	51	SVC
Prov 6	18		6				12					36	JRA
Pres 7	6	12	18									36	JRA
Scale 8				18	6	12						36	JRA
Total	48	38	38	29	38	42	38	11	11	11	11	315	

Project duration: 30 months

pandata

No.	Deliverable Name	WP	Nature	Diss.	Date
2.1	Project Website	2	O	PU	1
1.1	Project manag't structures, reporting, risk and quality procedures	1	R	CO	3
2.2	Dissemination plan	2	R	PU	3
3.1	Specification of AAA infrastructure	3	R	PU	6
5.1	Specific requirements for the virtual laboratories	5	R	PU	6
4.1	Requirements analysis for common data catalogue	4	R	PU	9
8.1	Definition of pHDF5 capable Nexus implementation	8	P	PU	9
8.2	Evaluation of Parallel filesystems and MPI I/O implementations	8	R	PU	9
1.2	First annual management report	1	R	CO	12
3.2	Pilot deployment of initial AAA service infrastructure	3	P	PU	12
6.1	Model of the data continuum in Photon and Neutron Facilities	6	R	PU	12
2.3	First Open Workshop	2	O	PU	15
4.2	Populated metadata catalogue with data from the virtual laboratories	4	R	PU	15
7.1	Implementation of persistent identifiers for PaNdata datasets	7	D	PU	15
3.3	Production deployment of AAA service infrastructure	3	D	PU	18
5.2	Deployment of Specification of the three virtual laboratories	5	R	PU	18
6.2	Common ontology def'n and def'n of tools to support provenance	6	R	PU	18
2.4	Open Source software distribution procedure	2	R	PU	21
4.3	Deployment of cross-facility metadata searching	4	D	PU	21
7.2	Mechanisms and tools for representation information and archiving	7	R	PU	21
8.3	Implementation of pNexus and MPI I/O on parallel filesystems	8	P	PU	21
8.4	Examination of Distributed parallel filesystem	8	R	PU	21
8.5	Demonstrate capabilities on selected applications	8	D	PU	21
1.3	Second annual management report	1	R	CO	24
3.4	Evaluation of initial AAA service infrastructure	3	R	PU	24
6.3	Tools for building research objects in Photon and Neutron Facilities	6	P	PU	24
2.5	Second Open Workshop	2	O	PU	27
4.4	Benchmark of performance of the metadata catalogue	4	R	PU	27
7.3	Mechanisms and tools for integrity of datasets	7	R	PU	27
8.6	Evaluation of coupling of prototype to multi-core architectures	8	R	PU	27
1.4	Final management report	1	R	CO	30
5.3	Report on the implementation of the three virtual laboratories	5	R	PU	30
6.4	Evaluation report on provenance management	6	R	PU	30
7.4	Report on evaluation of preservation mechanisms	7	R	PU	30

OECD Principles and Guidelines for Access to Research Data from Public Funding

13 principles

A – Openness

- **Openness means access on equal terms for the international research community at the lowest possible cost, ...**

B – Flexibility, C – Transparency, D – Legal conformity, E – Protection of intellectual property, F – Formal responsibility, G – Professionalism

H – Interoperability

- **Technological and semantic interoperability is a key consideration in enabling and promoting international and interdisciplinary access to and use of research data. ...**

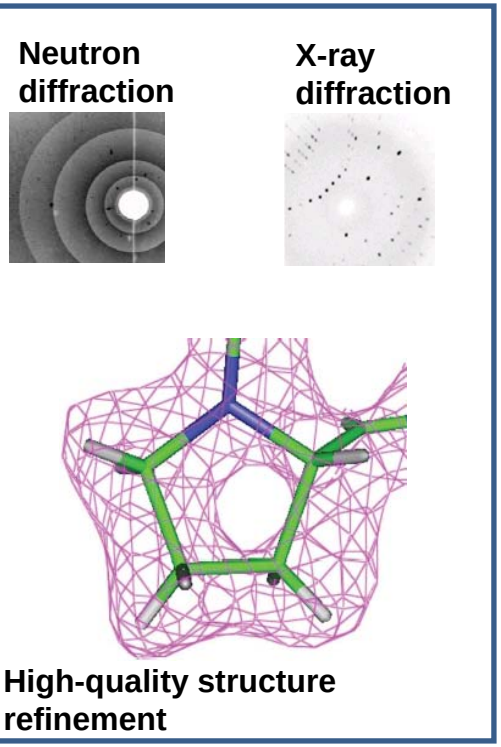
I – Quality, J – Security, K – Efficiency, L – Accountability

M – Sustainability

- **... taking administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention.**

[http://www.oecd.org/dataoecd/9/61/38500813.pdf]

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



- Unification of data management policies
- Shared protocols for exchange of user information
- Common scientific data formats
- Interoperation of data analysis software
- Linking Data and Publications and 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

Related projects

- IRUVX-PP Project
  - preparatory project to support the foundation of the EuroFEL Consortium
- ESRFUp
  - ESRF Upgrade program
- ILL 20/20
  - Upgrade program for ILL
- Open Source Development of the ICAT metadata catalogue
- NMI3
  - Integrated Infrastructure Initiative for Neutron Scattering and Muon Spectroscopy
- IA-SFS
  - Integrating Activity on Synchrotron and Free Electron Laser Science
- e-IRG - e-Infrastructure Reflection Group
  - e-IRG Report on Data Management
  - e-IRG Roadmap
- ESFRI - European Strategy Forum on Research Infrastructures

Digital Curation Projects

- **SCAPE : Scalable Preservation Environments**
  - automated, quality-assured workflows
  - policy-based planning and watch system
  - Case studies from: Library, Web Archiving, Scientific Research Data Sets
  - Partners include: Microsoft Research, The British Library
- **ODE: Opportunities for Data Exchange**
  - interoperability of data layers and data sharing
  - re-use and preservation
  - Partners include: Alliance for Permanent Access, CERN, Helmholtz Association
- **APARSEN – Alliance Permanent Access to the Records of Science Network**
  - To create a shared vision and framework for a sustainable digital information infrastructure providing permanent access to digitally encoded information
  - 30 Partners including: European Space Agency, Philips, Airbus, Microsoft Research
  - STFC leading
- **CASPAR and ENSURE - Enabling knowledge Sustainability Usability and Recovery**
  - scalable preservation solutions primarily for the commercial sector
  - Based on Open Archival Information System (OAIS) reference model (ISO:14721:2003) ,
  - 11 partners , IBM leading