

The Photon and Neutron Data Initiative – PaN-data



- Why?
- With whom?
- When?
- What?

IT is transforming the practice of science

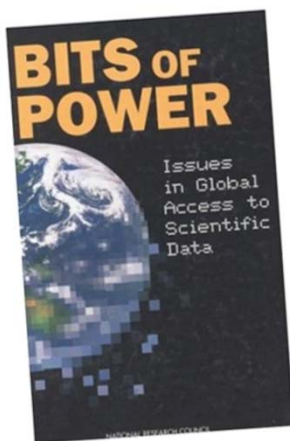
Science is increasingly computational, data-intensive, collaborative

Data management is now central to the scientific endeavor

Why PaN-data?...because in the P + N community:

1. scientific data is often considered private property
2. open access to scientific data is almost impossible
3. scientific data is not managed professionally
4. invaluable data sets are lost or inaccessible
5. barriers to interact with data must be lowered
6. scientists use several facilities for their research
7. the data deluge makes everything worse

1. Scientific data is often considered private property



US National Research Council, Study: “Bits of Power, Issues in Global Access to Scientific Data”, 1997

“The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for ... data derived from publicly funded research”

OECD Principles and Guidelines for Access to Research Data from Public Funding (2007):

“Sharing and open access to publicly funded research data not only helps to maximise the research potential but provides greater returns from the public investment in research”

ORGANISATION DE
COOPÉRATION ET
DE DÉVELOPPEMENT
ÉCONOMIQUES



1. Scientific data is often considered private property

ESFRI Position Paper on Digital Repositories:

“Research Infrastructures should guarantee that raw research data are made available through portals and databases.”

06/09/2007 – e-IRG ESFRI



Data's shameful neglect

“Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly”

Nature **461**, 145 (10 September 2009) | doi:10.1038/461145a

2. Open access to scientific data is almost impossible

- Data is not on-line
- Data is poorly or not described
- No search tools
- No persistent identifiers
- Authentication/authorisation for scientists is difficult
- Open access is not (yet) well accepted
- Institutions lack infrastructure



3. Scientific data is not managed professionally

Data management is left to the individual scientist

- Stored on unreliable support: USB disks/keys, laptop disks, PC disks
- Access/retrieval difficult
- High risk of data loss



3. Scientific data is not managed professionally

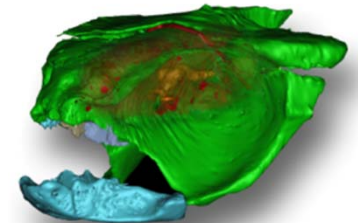
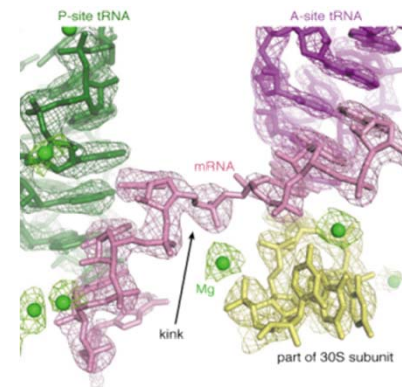
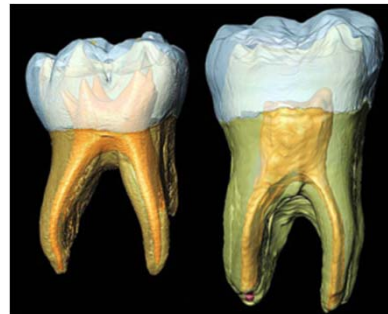
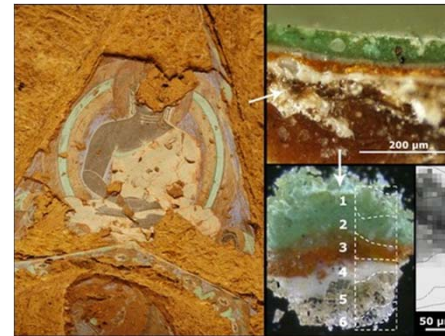
Data management is left to the individual scientist

- Stored on unreliable support: USB disks/keys, laptop disks, PC disks
- Access/retrieval difficult
- High risk of data loss



Some data simply goes up into smoke!

4. Invaluable data sets are lost or inaccessible



5. Barriers to interact with data must be lowered

6. Scientists use several facilities for their research

- EU wide authentication and authorisation of scientists
- Common data format
- Federated data bases
- Search tools, visualisation tools

7. The data deluge makes everything worse

- New detectors → high data rates (GB/s), high frame rates (<1ms/frame)
- Very large number of files per experiment
- Need for automatic meta-data capture
- Need for on-line data analysis
- Computing infrastructure under pressure

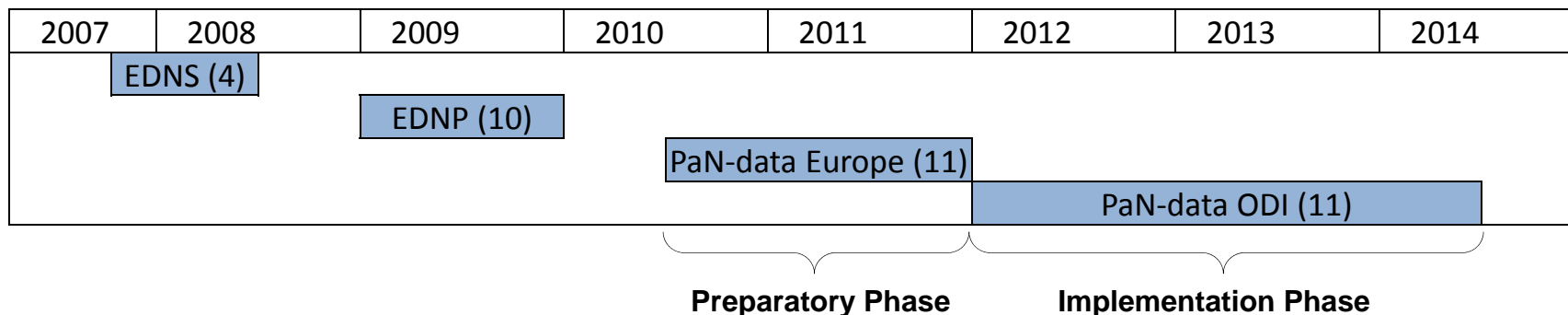


pandata

Established 2007 with 4 facilities

Expanded since to 11 facilities

Goal: *“...to construct and operate a shared data infrastructure for Neutron and Photon laboratories...”*



The PaN-data initiative

- Photons and Neutrons are complementary investigation tools
- Cross discipline experiments are increasing in number
- Neutron labs have built up data catalogues
- Synergy is essential for the project

Five P+N sites in Europe are in PaN-Data:

- ISIS + DIAMOND
- SINQ + SLS
- ILL + ESRF
- HMI + BESSY, now the HZB
- LLB + SOLEIL
- (+ DESY, ELETTRA, and ALBA)



The PaN-data RIs



Active Users of our Facilities

Over a period of two years (01/06/2009 – 31/05/2011):

35 968 users; P – 28 073, N – 10 324

7 895 (21.8%) use only Neutrons

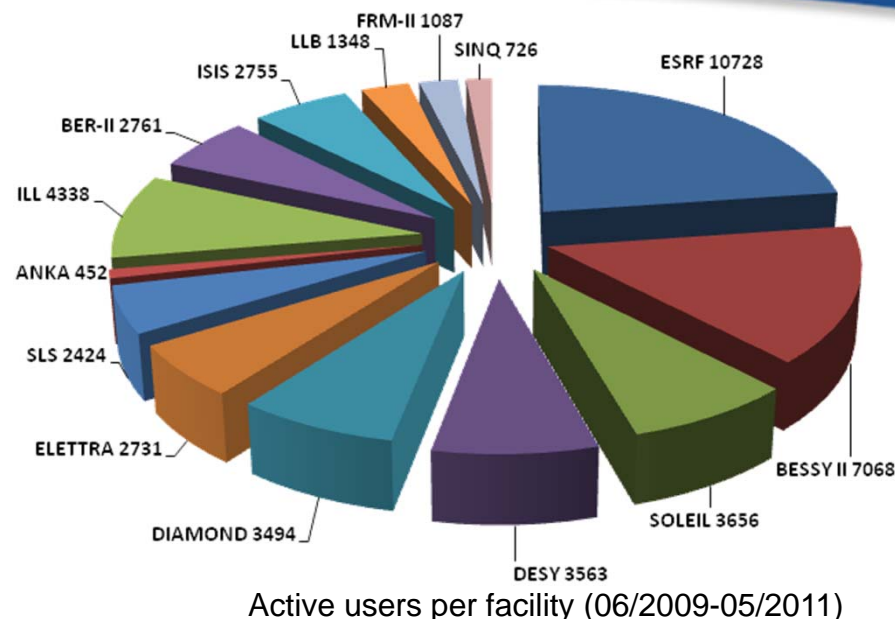
25 644 (71.2%) use only Photons

2 429 (6.7%) use Neutrons **and** Photons

7 757 (21.6% of all users) use more than one facility

4 830 (17.2% of all photon users) use more than Photon source

1 983 (19.2% of all neutron users) use more than one Neutron source



pandata

Goal: a shared data infrastructure for Photon and Neutron RIs

- Harmonise data policies in laboratories
- Harmonise authentication and authorisation
- Standardise data formats and annotation of data
- Allow transparent and secure remote access to data
- Establish sustainable and compatible distributed data catalogues
- Allow long term preservation of data
- Provide tools/interfaces for curating data
- Provide compatible open source data analysis software
- **A very ambitious work programme!**

Our data policy stipulates:

1. The facility shall act as a **custodian** for the data.
2. **All raw data will be curated** in a well-defined format with a unique ID.
3. **Metadata** is captured automatically and resides either within the raw data files, and/or in an associated on-line catalogue.
4. **Access to raw data** and the associated metadata obtained from an experiment **is restricted to the experimental team for a maximum period of 3 years**. Thereafter, it will become publicly accessible.
5. The embargo period can be extended on requests.
6. The on-line catalogue will link the data to the proposal and to the publication.
7. Ownership of all results (intellectual property) derived from the analysis of the raw data is determined by the contractual obligations of the person(s) performing the analysis.
8. Analysis of openly accessible data must acknowledge the source of the data and cite its unique identifier and any publication linked to the same raw data.



A Common Data Format

- Reduce need for local expertise
- Reduce number of conversion utilities
- Reduce redundant software development
- Increase metadata description – a prerequisite for data archiving/curation
- Increase cooperation in software development
- Increase functionality of generic software
- Create a community standard
- Create critical mass to influence detector manufactures
- Ease cross facility experiments

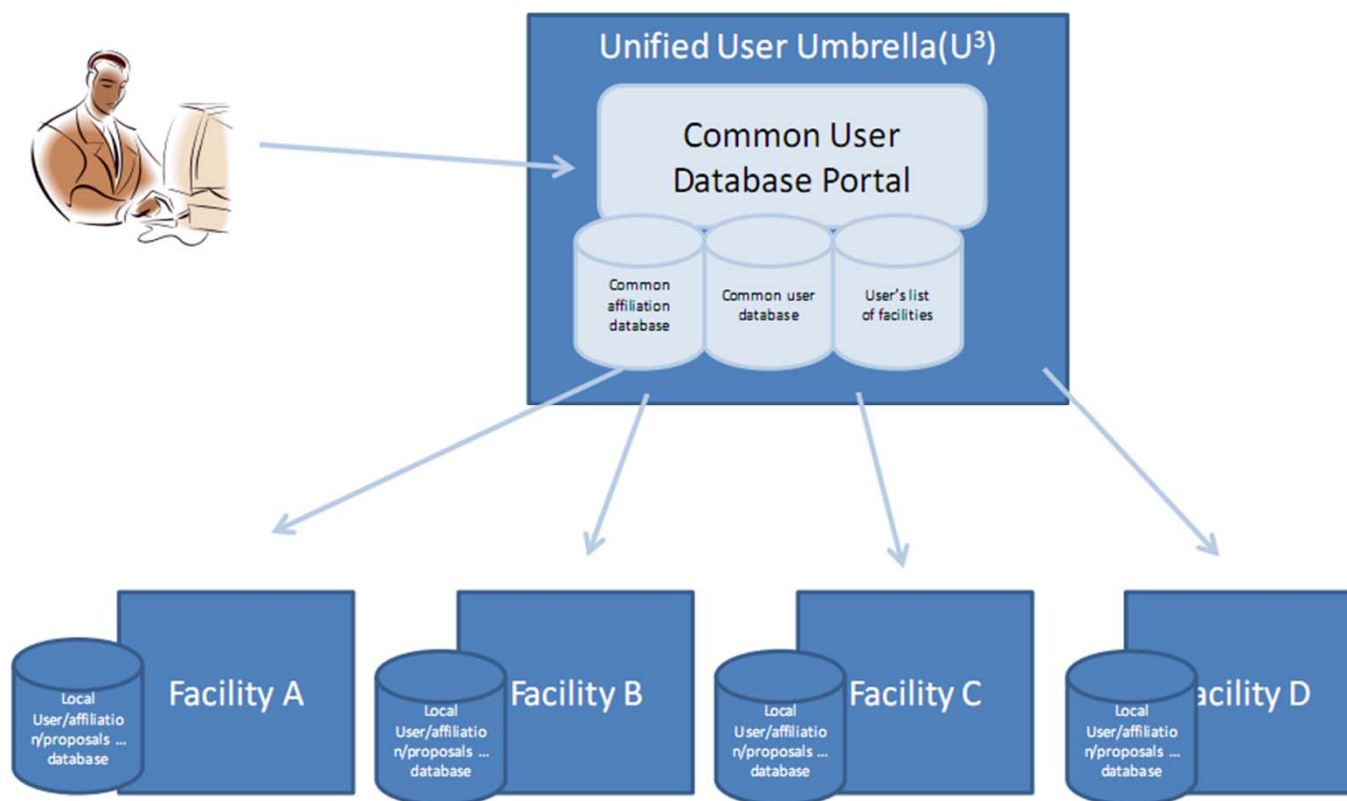
The PaN-data consortium
has decided to adopt
HDF5/Nexus

Build on the positive impact of standards!



A Unique ID for scientists

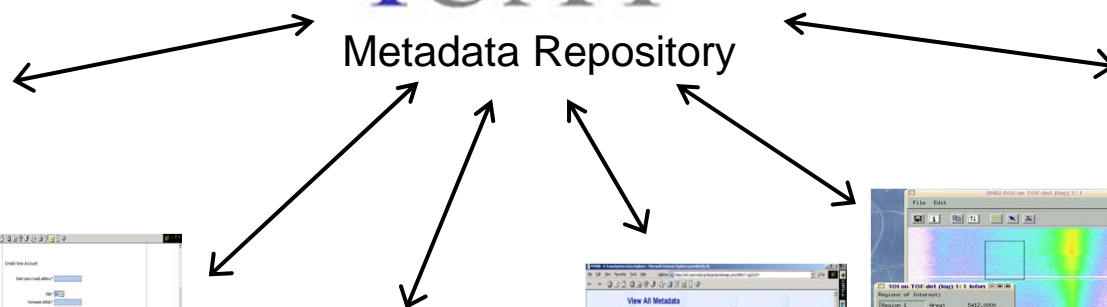
- A unique point to update user information (e.g. affiliation)
- A possible platform to manage proposals and facility events
- A prototype implementation (based on Shibboleth) is operational



<http://www.icatproject.org/>



Metadata Repository



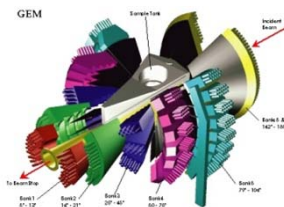
Proposal



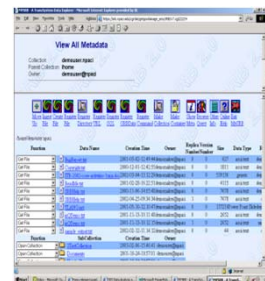
Approval



Scheduling

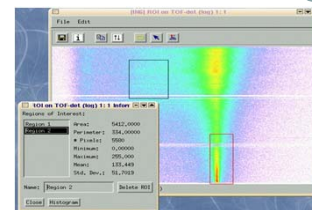


Experiment



Data cleansing

Data analysis



Record Publication

Subsequent publication registered with facility

Tools for processing made available

Raw data filtered and cleansed

Scientists visits, facility run's experiment

Facility registers, trains, and schedules scientist's visit

Facility committee approves application

Scientist submits application for beamtime

<http://paleo.esrf.eu>

ESRF paleontological microtomographic database

by ESRF

Categories

- [paleoanthropology \[11\]](#)
- [invertebrate paleontology \[26\]](#)
- [vertebrate paleontology \[8\]](#)

45 images

Specials

- [Most visited](#)
- [Best rated](#)
- [Random pictures](#)
- [Recent pictures](#)
- [Recent categories](#)
- [Calendar](#)

Menu

Quick search

- [Tags](#)
- [Search](#)
- [Comments](#)
- [About](#)
- [Notification](#)

Identification


- [Register](#)
- [Connection](#)

Quick connect


Username

Password


15 Most visited [15]



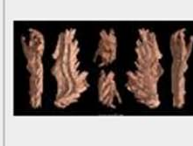
(171) Qafzeh 10 maxilla




(137) Engis 2 child



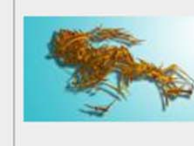
(102) Spider Orchestina




(80) Novispathodus cluster TQ84C30 PIMUZ 28766




(71) Trinil 11621 upper molar




(69) AMNH66253




(66) Qafzeh10 mandible




(57) Trichomya lengleti




(50) beetle Scolytidae



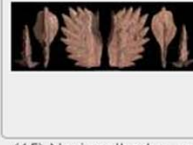
(47) 15529 000



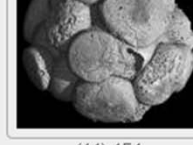
(47) Qafzeh 15 mandible



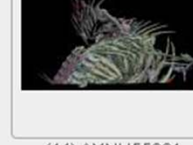
(45) Trinil 11620 upper molar



(45) Novispathodus sp TQ84C30 12



(44) 4F4



(44) AMNH55901





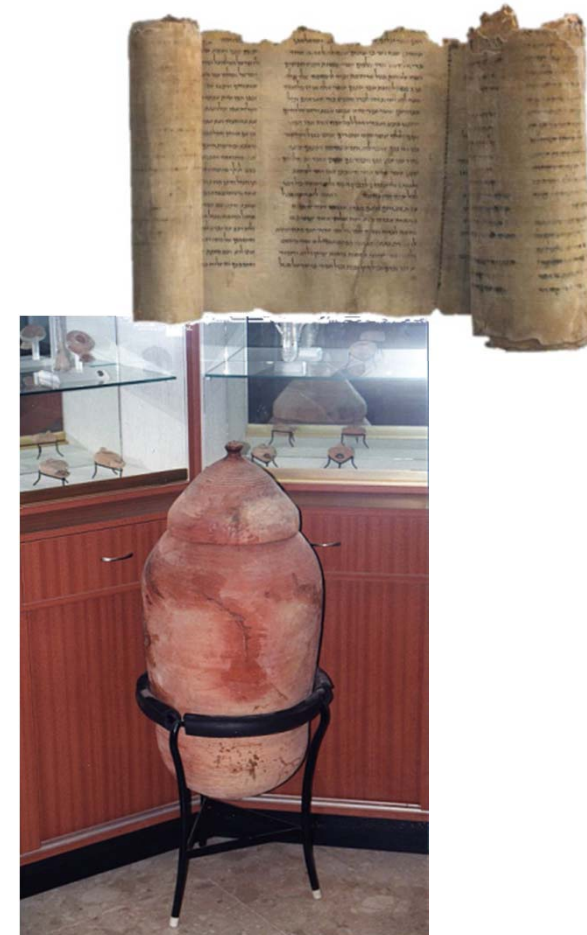
Data Archiving/Curation

Data Curation = preservation and maintenance of digital assets

Issues:

- Storage format evolution and obsolescence
- Persistence of the digital objects and their identifiers
- Rate of creation of new data and data sets
- Broad access and searching flexibility
- Obsolescence of data analysis code

*“Digital documents last forever -
or for five years, whichever comes first”
Jeff Rothenberg, 1997*



2100 years of sustainable data storage:
Qumran jar storing dead sea scrolls

Sharing analysis code

1. The software policy is expected to generate **significant savings** in the RIs.
2. Exchange information on purchasing conditions of commercial software packages.
3. RIs shall communicate their intention to develop new software.
4. No major DA-software package shall be developed by a single person.
5. Newly developed DA-software shall be **compatible with HDF5/Nexus**.
6. Newly developed software shall be **extensively documented and tested**.
7. No “branching” of software developments.
8. Software packages shall be preserved along with the computing environment for at least 3 years.
9. Source code shall be preserved ad infinitum.
10. Agree on the principle of “software package maintainers” and identify at least one software package per lab.
11. Compile a matrix of software packages maintained in each lab, including commercial software (version, OS, etc.), reflecting the current status.

An on-line prototype software catalogue is ready.



Photon and Neutron Software Catalogue

PaNsoft is a database of software used mainly for data analysis of neutron and photon experiments. PaNsoft is one element of a larger project, [PaNdata](#), which aims to provide a complete, shared data infrastructure for neutron and photon laboratories.

This database can be freely consulted. It gives an overview of software available for neutron and photon experiments and their use with respect to instruments at experimental facilities.

By [registering](#) and [logging-in](#) new software can be entered and it will appear in the database after moderation. Similarly, feedback can be given on the software presented herein and more generally via the forum hosted here.

[Browse software](#)

Software: [Recent Software](#) [Popular Software](#)

Recent Software

NAMD

NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems.

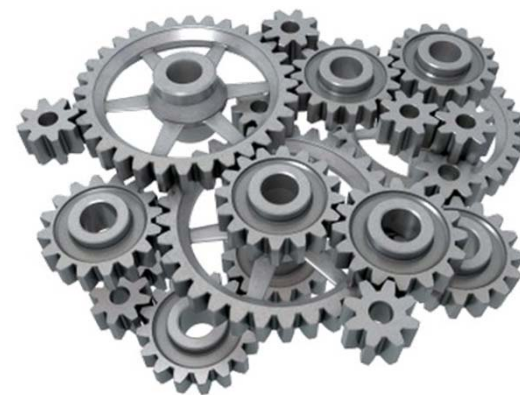
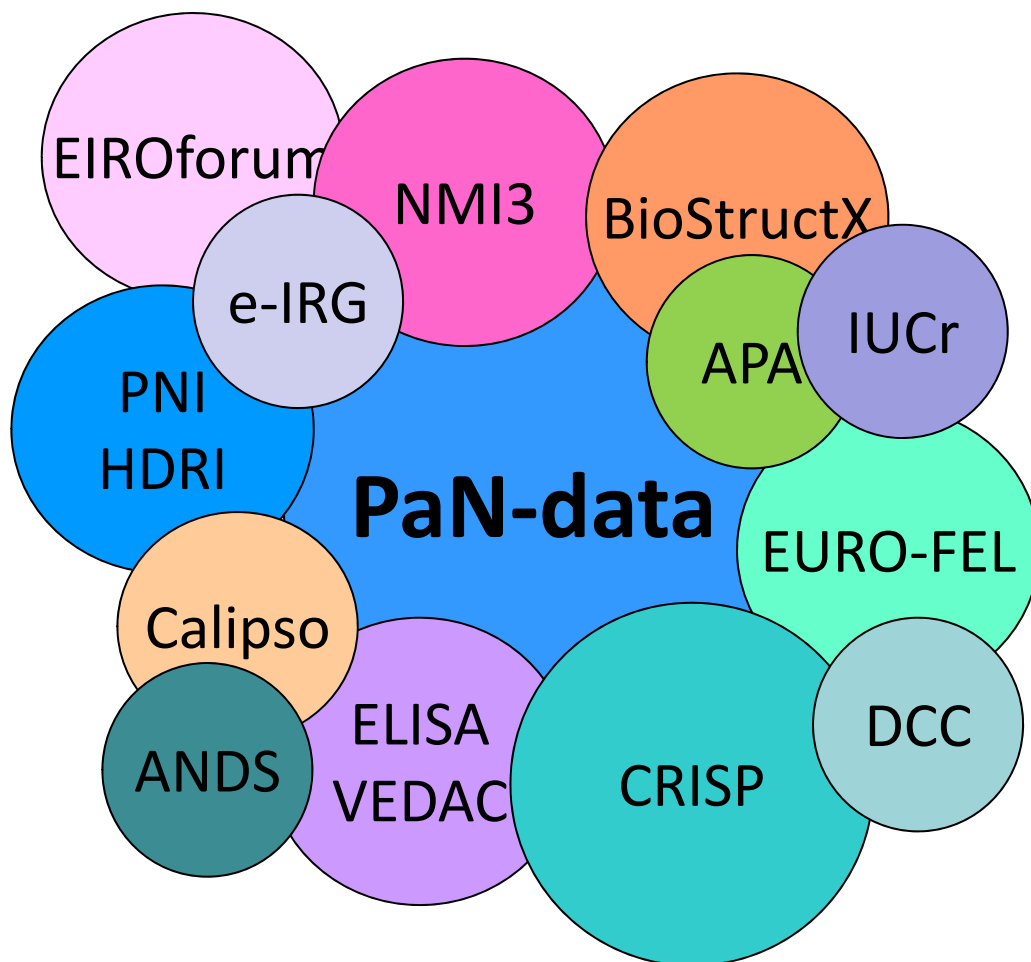
PORE3D

large software suite for filtering, segmentation and quantitative analysis of 3D images (CT, MRI, CLSM, ...).

Ominc

Logiciel d'acquisition des spectres infrarouge. Ce logiciel permet aussi un de créer des cartographies

PaN-data embedded in Europe



The work programme for the next 2.5 years

Building on the achievements of the preparatory work phase, and in strong collaboration with the CRISP consortium,

Deploy and operate:

A pan-European user ID system for scientists

A generic catalogue of scientific data across the RIs

Develop:

A conceptual framework for the tracing of data analysis steps

Tools for long-term data preservation

A scalable data processing framework



Conclusions

PaN-data has still a long way to go

We need the backing of policy makers, science managers, lab directors to implement the vision of PaN-data:

- ✓ **User authentication and authorisation,**
- ✓ **Data management and preservation,**
- ✓ **Data traceability.**

pandata

Thank you for your attention!

