

Towards an Open Data Infrastructure for Sciences with Photon and Neutron Sources

Frank Schlünzen
DESY



Science & Technology Facilities Council
ISIS



- The PaNdata Consortium
- Science at our Photon and Neutron facilities
- PaNdata user communities
- PaNdata objectives



- 11 Partners - lead STFC
- Projects:
 - PaNdata Europe (05/10-11/11)
 - PaNdata ODI (10/11-09/14)
- In operation
 - 8 Synchrotron Sources
 - 5 Neutron Sources
 - 1 Free Electron Laser (FEL)
 - Instruments / Beamlines >> 100
 - Users > 35.000
 - Investment > $3 \cdot 10^9$ €
 - Data Rate > 10PB / yr
- Under construction/commissioning
 - 2 Synchrotron Sources
 - 4 FELs
 - 1 Neutron Source (ESS)
 - Investment > $2 \cdot 10^9$ €
 - Data Rate >> 10PB / yr

Investigation of a new species of an early human ancestor by X-ray microtomography.

The skull and more than 40% of the entire body has been investigated with x-ray tomography providing detailed insights about the hominids life 1.98 million years ago.

The analysis of the data has only just started, but the preliminary visualisation of the complete skull shows fossilised insect eggs and an extended low density area that could point at remnants of the brain after its bacterial decay.



Rendering of the 3-D scan of the skull of *Australopithecus sediba* child.
Credits: P. Tafforeau



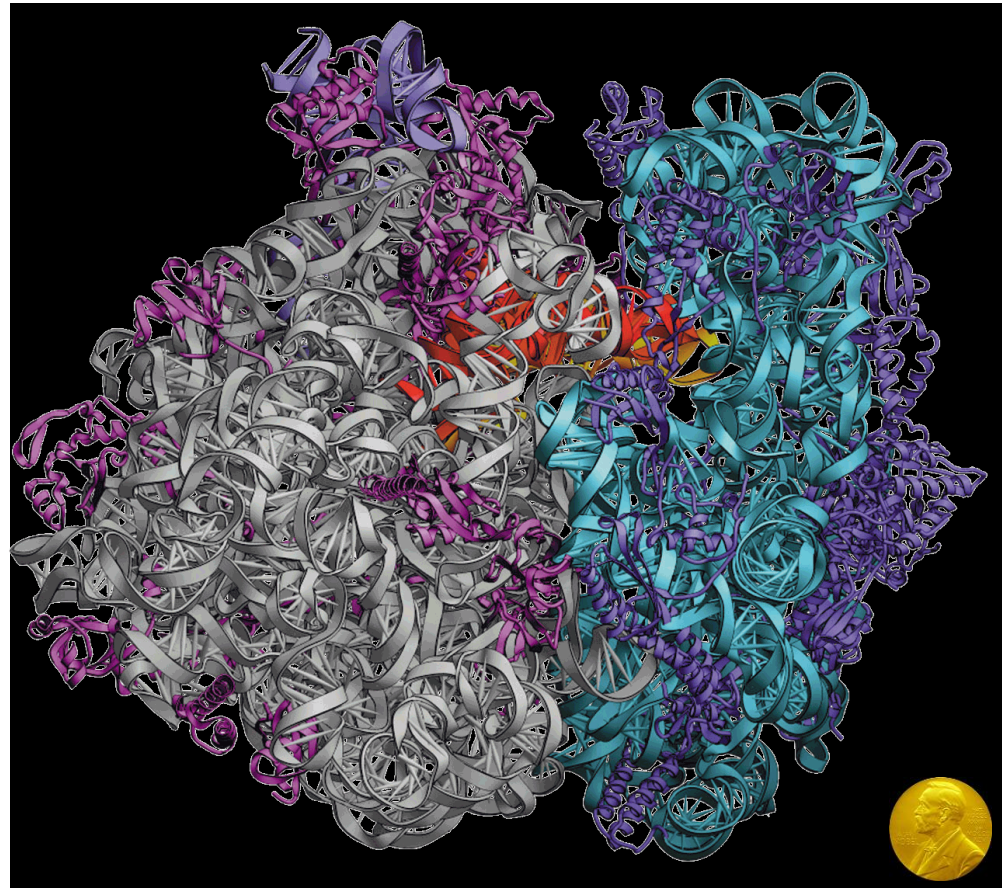
Berger et al., Science 328 (5975): 195-204
Australopithecus sediba: A New Species of *Homo*-Like Australopith from South Africa

Crystallography of ribosomes

The ribosome, the protein-factory of the cell have been investigated for 30 years by X-ray crystallography cumulating into a high resolution characterization of protein bio-synthesis.

As an important side effect it permits to study the interaction and ways of inhibition of a wide spectrum of antibiotics, which has been taken up by a number of pharmaceutical companies to support design of new antibiotics.

In 2009 the Nobel prize has been awarded to A.Yonath, V.Ramakrishnan and T.Steitz for their work on structure and function of the ribosome.



http://rna.ucsc.edu/rnacenter/ribosome_images.html



Schlünzen et al., Nature 413 (6858): 814-821

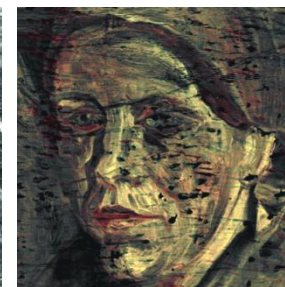
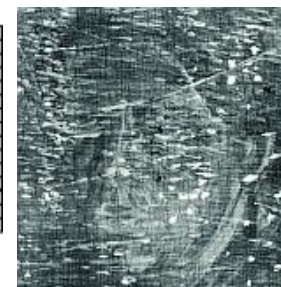
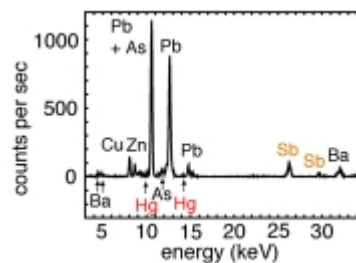
Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria

Investigation of van Gogh's paintings with x-ray fluorescence spectroscopy.



In the winter 1884/85 van Gogh painted a number of portraits in Nuenen/NL, which he later on sent to his brother Theo in Paris. Here the trace of some paintings got lost.

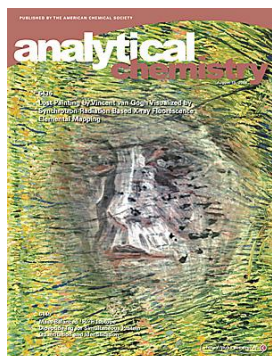
As the reconstruction revealed, van Gogh – always short of cash - had simply recycled the canvas 2 years later after moving to Paris. Mapping the chemical elements permitted now to recover the hidden painting in color and beauty.


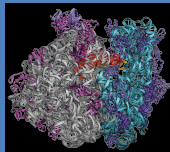



http://hasylab.desy.de/news_events/research_highlights/archive/visualizing_a_lost_painting_by_vincent_van_gogh/index_eng.html

Dik et al. *Anal. Chem.*, 80 (16), 6436–6442

Visualization of a Lost Painting by Vincent van Gogh Using Synchrotron Radiation Based X-ray Fluorescence Elemental Mapping



			
Technique	micro tomography	Crystallography	Fluorescence Spectroscopy
Community	Paleontology	Structural Biology	Arts
interest for other communities	Evolutionary Biology, Anthropology, Earth sciences	Medicine, Drug design, Life Science	History, Literature, Chemistry, Political / Geographical
Databases	Paleodb, Anthrobase, ...	PDB, NADB, EMDB, CSD, ...	few separated databases
Data rates	100MB's / sec ~10MB / 2D-image	300MB / sec ~6MB / 2D-image	Small ~10kB / Spectrum
Data volumes	100's GB	Up to few TB	small
Data formats	Tiff	cbf, cif, ascii, pdb, ...	Ascii, Excel, ...
Archived	Yes	Certainly not	Probably not
Reproducible?	Not at all	To some extend	Not at all
Accessible?	No	No	No

→ Good reasons to think about a data infrastructure

Survey to obtain reliable numbers of unique and common users:

- Combined user information from 14 facilities
 - Facilities are competitors!
 - Dislike plain (meaningless) cost/user comparisons

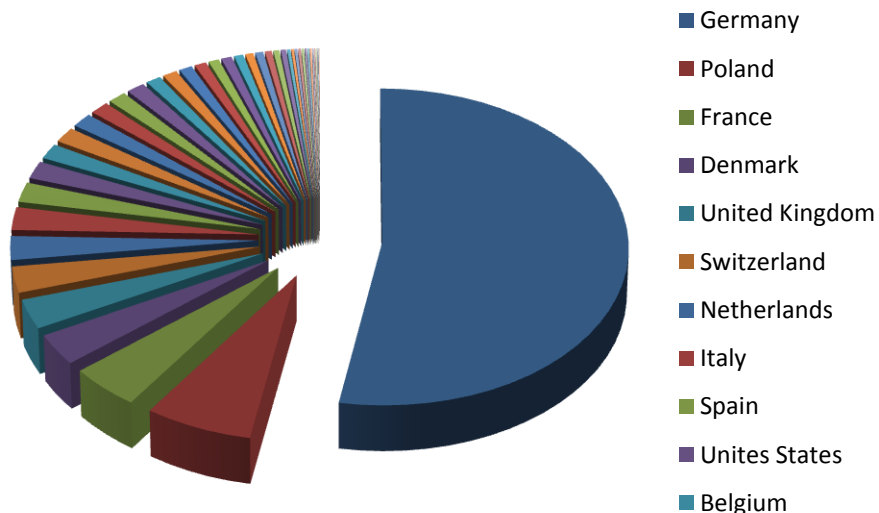
- Basic numbers:

- Total number of users entries: 47131 Photons: 34116 Neutrons: 13015
- **Total number of unique active users: 35968 Photons: 28073 Neutrons: 10324**
- 20-40% of a photon facilities users also perform experiments somewhere else
- 30-50% of a neutron facilities users also perform experiments somewhere else
- 20-30% of the neutron users also use photon sources

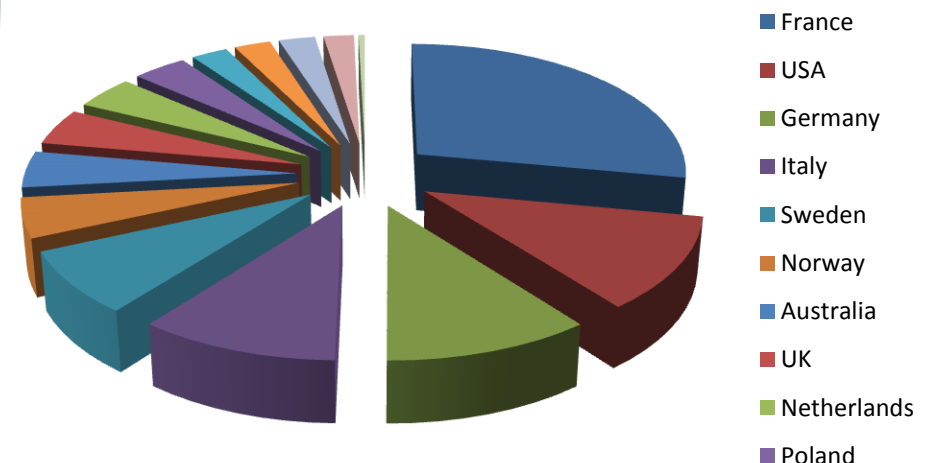
→ Good reasons for a **common** data infrastructure

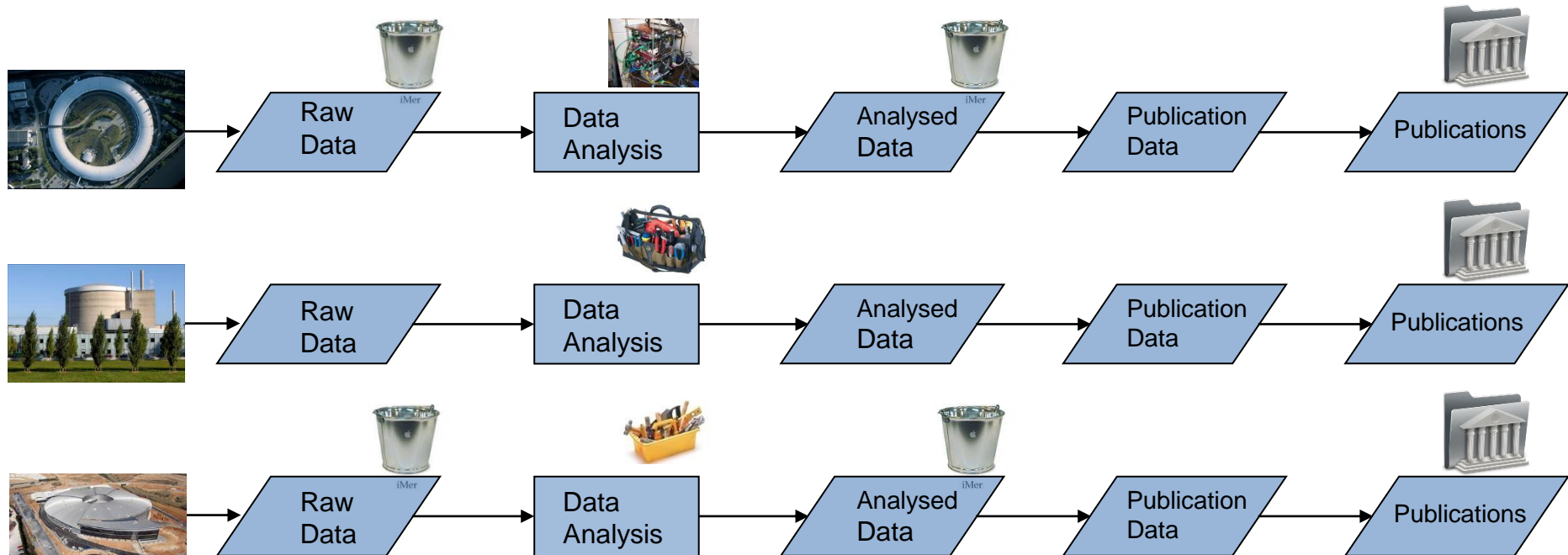
- Come from everywhere with many small, scattered & volatile collaborations
 - Research is often highly competitive
 - Many users come without experiment or compute experience
 - Large fraction of one-time or first-time users
 - Experiments are often very short, but essential in the scientific lifecycle
 - Disciplines like Arts and Humanities need full support
- **Data Infrastructure needs to comply with very inexperienced users**
- **DM/AAI across all facilities, many disciplines, communities and the same users**

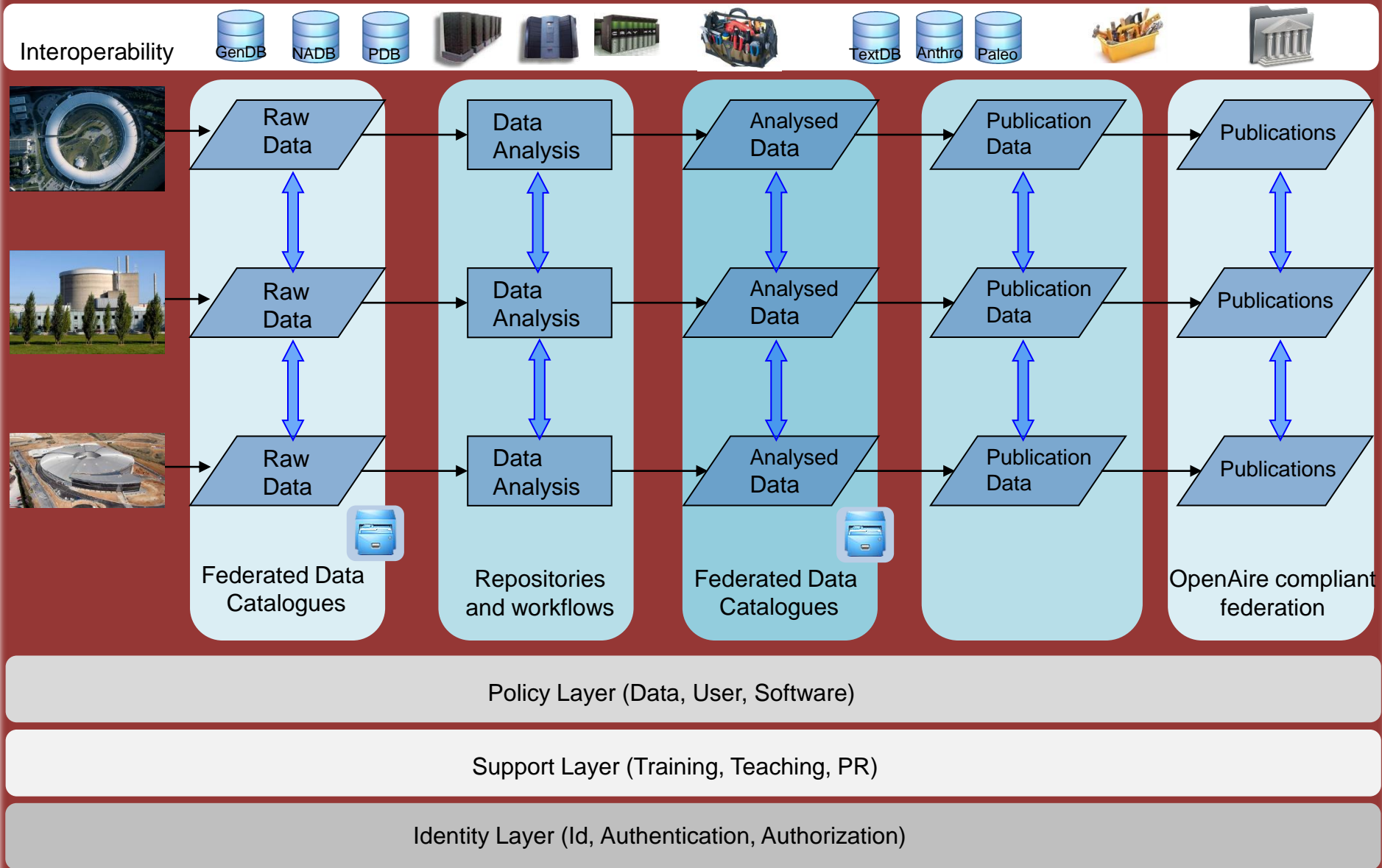
Users of a national Photon Source



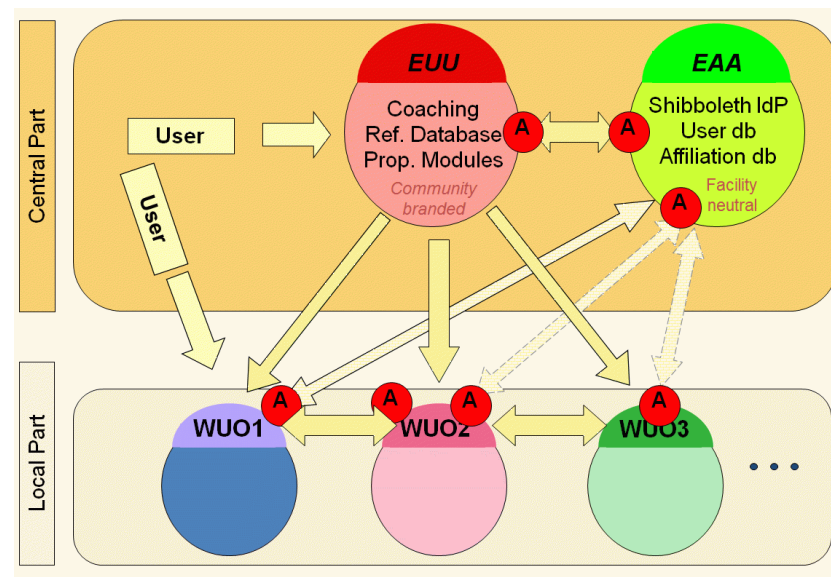
Collaborations of Swiss ESRF users







- Unique, persistent ID for all (European) Photon and Neutron Source users
 - Uniqueness is the real problem
 - Records dating back to the 60th
- Shibboleth based authentication system
 - So called *Umbrella*
 - Provides automatic Id mapping across facilities
- Prototype developed by EuroFEL/PSI
 - Pilot phase starting soon
 - Federation and CAS integration works
 - iCAT (1) integration works
 - Moonshot (2) integration ??
- Proposed system for related projects
- Proposal module in preparation
- Teaching and Community modules missing

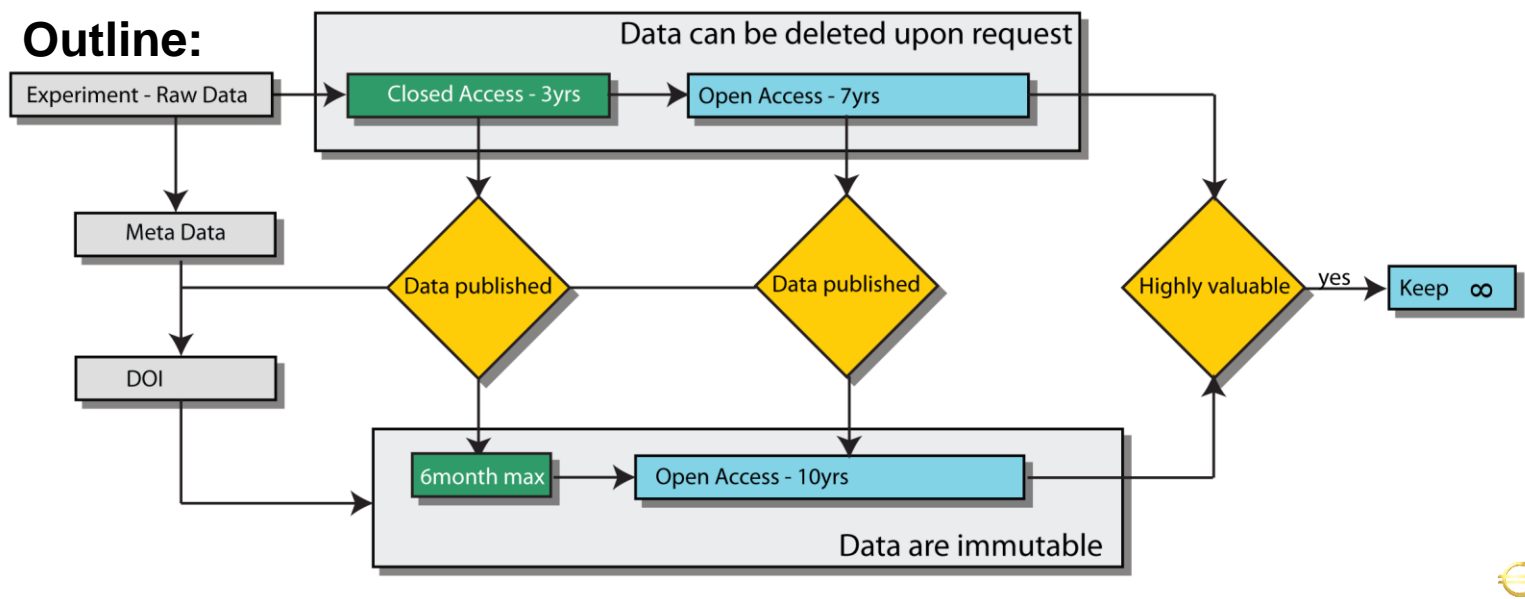


EUU: European Unified Umbrella (3)

Support Layer (Training, Teaching, PR)

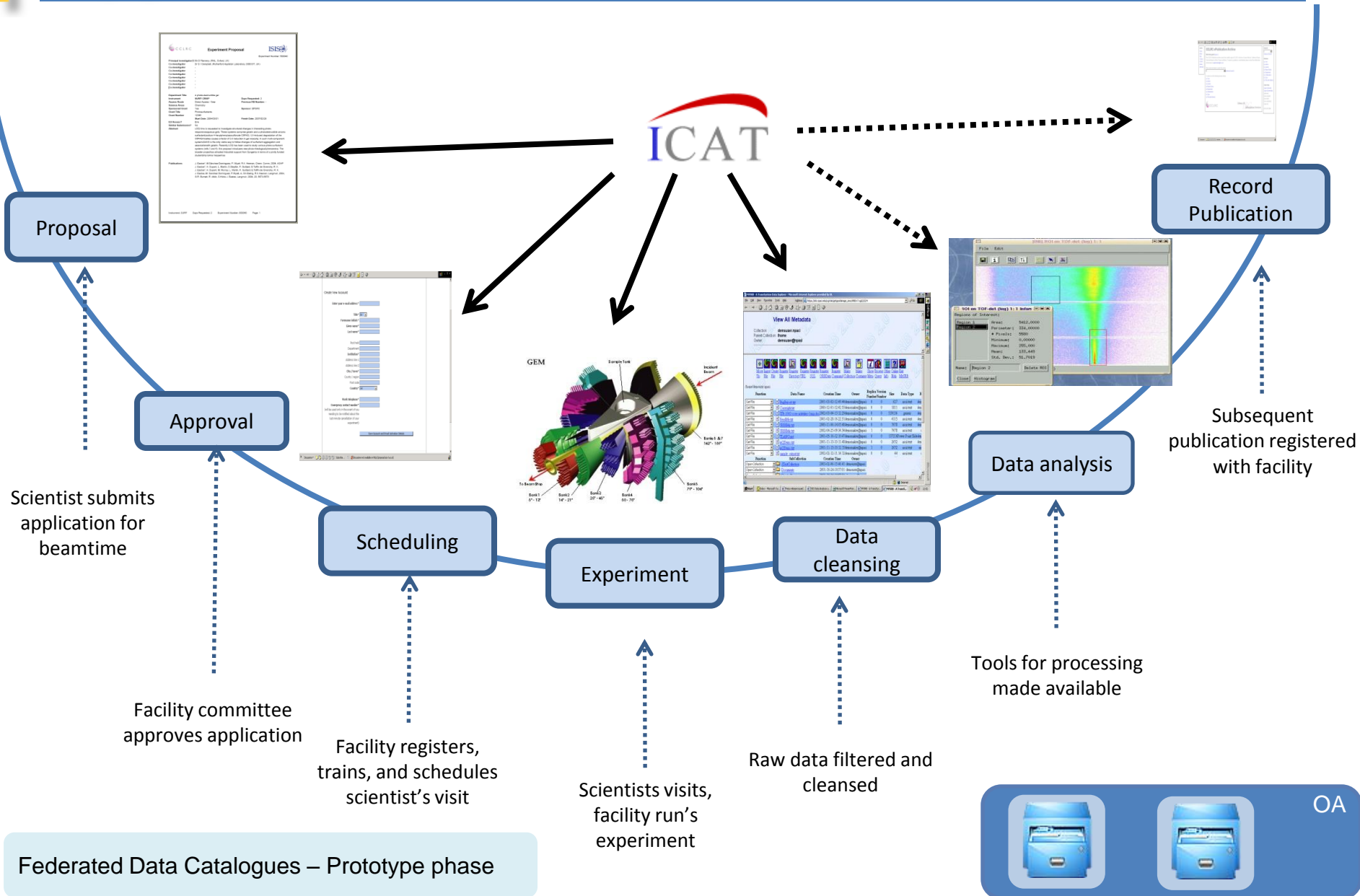
Identity Layer (Id, Authentication, Authorization)

Outline:



Status:

- **Common Data Policy proposed for all participating facilities**
 - Implementation ongoing (ISIS, Diamond, ILL)
 - EU, APA, NIH, OECD, HGF (name it) recommendation compliant
- **Common Standard Data format**
 - HDF5/NeXus as the common standard
 - Self-describing, standardized Metadata (dublin core based)
 - Implementation and development ongoing
 - Agreement for all future application development to support the standard
- **Common Software Repository in prototype phase**



Work in progress

AUTHENTICATION PROCESS

1. CLIENT REQUEST
2. AUTHENTICATION REDIRECT
3. TICKET FORWARDING
4. TICKET VALIDATION

TRANSMITTED DATA

- T TICKET
- S SERVICE ID
- C COOKIE (CAS SSO)
- A AUTHENTICATED ID

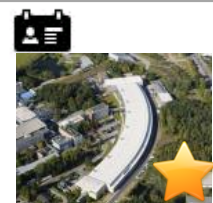
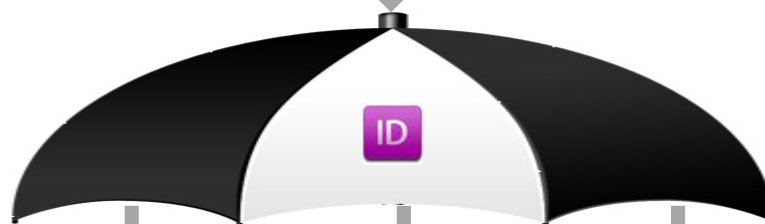
DOI



Federation

dDOI

Nothing done yet

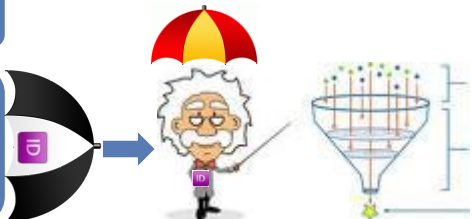
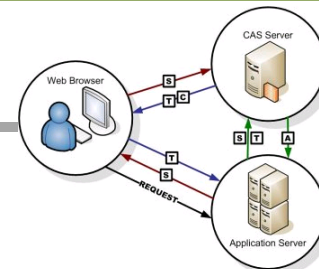


API

Policies

OA

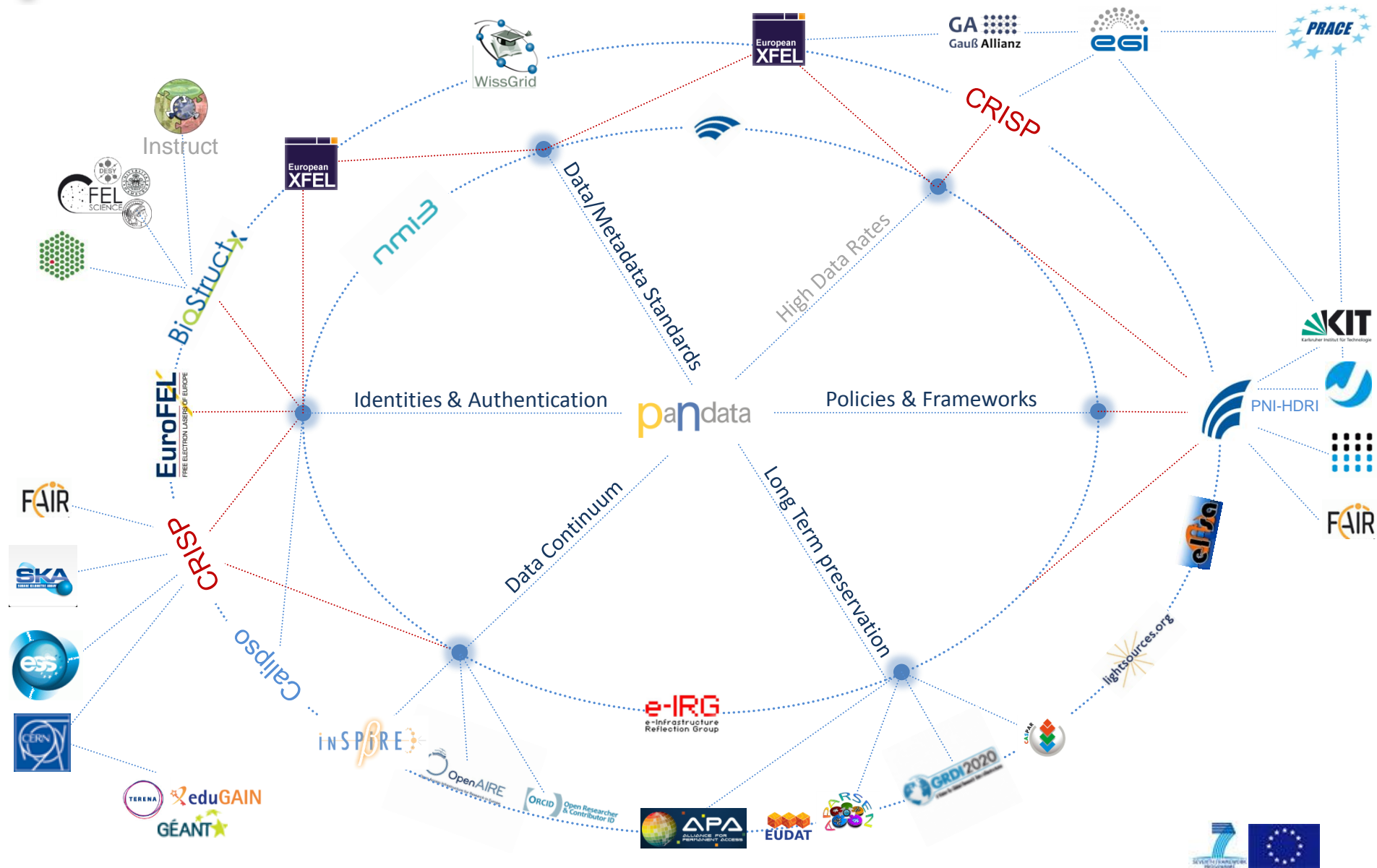
Federation



- We are on a good way establishing a cross-facility interdisciplinary ODI
 - Making it real easy for our user to manage and share data
 - Making it more difficult not to share
 - Scientists don't like to though (1) , so ...
- *Encouragement* (rather than mere recommendations) could be helpful
 - Funding agencies
 - Policy makers
 - Publishers
- Interoperability
 - We are heavily depending on interoperability of compute and data infrastructures and services across all scientific disciplines
 - Means heavily dependent on quite a number of seemingly unrelated projects
 - Therefore very interested on more joint activities
 - PaNdata kickoff meeting with (hopefully) CRISP, PNI-HDRI, EUDAT, ORCID, OpenAire, Eur.XFEL on the 3rd of November at RAL, right after the 2nd Workshop on Federated Identity Systems (2).

(1) <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0007078>

(2) <http://indico.cern.ch/conferenceDisplay.py?confId=157486>



Thanks for your attention!

