PaNdata ODI

Photon and Neutron Open Data Infrastructure Project

European federated data infrastructure for research with photons and neutrons

http://www.PaN-data.eu

The PaN-data ODI project is partly funded by the European Commission under the 7th Framework Programme, Grant Agreement Number RI-283556
Outline

- Introduction to PaN-data
- PaN-users and implications for IdM
- Standards
- Policies
- Example for an Open Access Database
Consortium of 13 Photon and Neutron RIs

- **Focus** entirely on topics around data management
- **Represented** mostly by IT, but also user offices & beamline scientists
# PaNdata Facts

## Consortium of 13 Photon and Neutron RIs

- **Focus** entirely on topics around data management
- **Number of instruments** > 200
- **Number of users** > 30,000 users/yr
- **Investment** > €4,000,000,000*
- **Running costs** > €500,000,000/yr*
- **Publications** > 10,000/yr*
- **RCosts/Publication** ~ €50,000*%
- **Data volume** >> 10PB/yr*

* wild guess

---

% ESRF 2011

---

* wild guess
Basic Idea

Any user could with a single credential/identity (SSO)

- submit or manage proposals and beamtime
- manage, access, control his/her data
- combine x-ray and neutron data
- share data within collaborations
- archive, index & publish data
- analyze data
- access resources
- capture the process in data catalogues
### Basic Idea

Any user could with a single credential/identity (SSO)

- submit or manage proposals and beamtime
- manage, access, control his/her data
- combine x-ray and neutron data
- share data within collaborations
- archive, index & publish data
- analyze data
- access resources
- capture the process in data catalogues

regardless where experiments were conducted
• Users are at the core of the proposed data infrastructure
  • Would they actually benefit from a data infrastructure?
    • We need one anyway for in-house users
  • Would they actually benefit from a federation?
    • Means: are they using more than just a single facility/instrument?
  • Where do users come from?
  • And so on ...

• To answer some of the questions
  • Anonymous survey in 2011 and 2012
  • See: http://www.pan-data.eu/Users2012
  • See: http://www.pan-data.eu/Users2011
Users

• Counting only active users:

<table>
<thead>
<tr>
<th></th>
<th>Total number of unique users:</th>
<th>30958</th>
<th>Photons:</th>
<th>24121</th>
<th>Neutrons:</th>
<th>9040</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using only Neutrons:</td>
<td>6837</td>
<td></td>
<td></td>
<td></td>
<td>or 22.0% of all unique users</td>
<td></td>
</tr>
<tr>
<td>Using only Photons:</td>
<td>21918</td>
<td></td>
<td></td>
<td></td>
<td>or 70.7% of all unique users</td>
<td></td>
</tr>
<tr>
<td>Using Neutrons and Photons:</td>
<td>2203</td>
<td></td>
<td></td>
<td></td>
<td>or 7.1% of all unique users</td>
<td></td>
</tr>
<tr>
<td>Using more than one facility:</td>
<td>6973</td>
<td></td>
<td></td>
<td></td>
<td>or 22.5% of all users</td>
<td></td>
</tr>
<tr>
<td>Using more than one Photon source:</td>
<td>4527</td>
<td></td>
<td></td>
<td></td>
<td>or 18.7% of all photon users</td>
<td></td>
</tr>
<tr>
<td>Using more than one Neutron source:</td>
<td>1604</td>
<td></td>
<td></td>
<td></td>
<td>or 17.7% of all neutron users</td>
<td></td>
</tr>
<tr>
<td>Total number of returning users:</td>
<td>21222</td>
<td></td>
<td>Photons:</td>
<td>16457 (68.2%)</td>
<td>Neutrons:</td>
<td>6026 (66.7%)</td>
</tr>
<tr>
<td>Total number of new users:</td>
<td>9736</td>
<td></td>
<td>Photons:</td>
<td>7664 (31.8%)</td>
<td>Neutrons:</td>
<td>3014 (33.3%)</td>
</tr>
</tbody>
</table>

* Incomplete figures. Some facilities missing & can’t capture collaborators not registered as users

---

**Basic facts**

**Federation:** ~ $\frac{1}{4}$ of users use more than one facility

**Scalable:** ~ $\frac{1}{3}$ new users every year

**Simplicity:** Majority of users come at most twice a year

**Compatibility:** >2000 use Neutrons and Photons
Basic facts

**Federation:** $\sim \frac{1}{4}$ of users use more than one facility

**Scalable:** $\sim \frac{1}{3}$ new users every year

**Simplicity:** Majority of users come at most twice a year

**Compatibility:** $>2000$ use Neutrons and Photons

**Global:** Users come from everywhere (almost)

The majority from European countries operating PaN sources (not surprisingly)
• Requirements
  • Unique and persistent Identity
  • Global coverage
  • Highly scalable & simple to manage
  • Simple to use
  • Single sign on capabilities across domains and facilities
Requirements

- Unique and persistent Identity
- Global coverage
- Highly scalable & simple to manage
- Simple to use
- Single sign on capabilities

Solution: Umbrella

- Initiated & developed in the EuroFEL project
- PaN-data/CRISP establish Identity System under UmbreillaID.org
- Standard technology (shibboleth/saml2)
- Capable to permit login at any of the Digital User Offices (DOOR)
- Capable to permit searches across data catalogues
- Easy to integrate into existing systems
- Like any such jointly operated system quite a task, but on a good way
Standards

- Data mining and discovery across facility
  - Requires a standard meta-data scheme & persistent identifier (doi)

- Implementation of workflows & analysis tools
  - Standard data format highly favorable
Standards

• Data mining and discovery across facility
  • Requires a standard meta-data scheme & persistent identifier (doi)

• Implementation of workflows & analysis tools
  • Standard data format highly favorable

• HDF5 is an EC proposed ISO standard for all binary data
  • NeXus adds a standard meta-data layer to HDF5, but is plain HDF5
  • netcdf became effectively HDF5
  • cif is becoming NeXus compliant (or vice versa)
  • Many commercial tools support HDF5
  • Huge advantage storing/archiving an entire dataset in a single file
Standards

• Data mining and discovery across facility
  • Requires a standard meta-data scheme & persistent identifier (doi)

• Implementation of workflows & analysis tools
  • Standard data format highly favorable

• HDF5 is an EC proposed ISO standard for all binary data
  • NeXus adds a standard meta-data layer to HDF5, but is plain HDF5
  • netcdf became effectively HDF5
  • cif is becoming NeXus compliant (or vice versa)
  • Many commercial tools support HDF5
  • Huge advantage storing/archiving an entire dataset in a single file

• Substantial investment by different parties to accelerate HDF5 processing

• Developments at DESY (in co-operation with PNI-HDRI)
  • Open Source NeXus C++ API: http://sourceforge.net/p/libpninx/wiki/Home/
  • Open source NeXus Data writer: http://code.google.com/p/nexdatas/
NeXus metadata scheme

- NeXus permits full description of an experiment
  - So called application definition
  - Automatic collection of metadata
  - Automatic ingestion into data catalogue

- A tedious process to define a common scheme
  - Traditionally a Neutron data format
  - Neutrons and X-rays have different approaches
  - Different labs have different vocabularies

- A tedious process to adopt detectors to HDF5
  - But it’s progressing, for example Dectris

- A tedious process to adopt analysis software
  - Many experiments use software grown over decades, won’t easily change
  - But: Analysis frameworks like DawnScience or Mantid are NeXus aware

- ILL/ISIS: everything in NeXus
  - 100,000’s of datasets catalogued
  - Everything ever collected
• We’ve got
  • Identity management systems (Umbrella)
  • Standards (HDF5/NeXus)
  • Data catalogues to be federated (ICAT/TopCat/Portals)
  • Storage infrastructure (dCache at DESY)
  • Analysis frameworks supporting standards & catalogues (Mantid/DawnScience/DPDAK etc)

• No point if data are not being archived

• So what are the rules (policies) for data archival?
• PaN-data has developed a simple policy framework.
  • One component is a policy for scientific data:
    • http://wiki.pan-data.eu/imagesGHD/0/08/PaN-data-D2-1.pdf

3.1.1. All raw data and the associated metadata obtained as a result of publically funded access to the research facilities are open access, with the research facility acting as the custodian.

3.2.1. All raw data will be curated in well-defined formats [...].

3.2.3. Data will be read-only for the duration of its life-time.

3.2.5. [...] each data set will have a unique identifier (doi).

3.3.3. Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of 3 years after the end of the experiment.

Thereafter, it will become openly accessible.

Any PI that wishes their data to remain restricted access for a longer period [needs a good reason].

If data can only be stored at the facility for less than three years, then access is exclusive to the PI up to the end of the storage period.

Data can always be made openly accessible earlier on request of the PI.
Policies in human readable form

• Store everything worth to be kept for at least 10 years
  • Don’t keep crap (in view of the costs of SR-beamtime: that better be a small fraction)

• Store data which are maintainable
  • Don’t keep data too big to be kept
  • Don’t keep data no one bothered to annotate properly

• Make data freely accessible after 3 years
  • Extend the embargo period if necessary
  • Don’t release sensitive data (e.g. medical data)
  • Only applies to academic research
  • Enable rapid release of data after publication
  • Encourage re-using parties to give credit and seek collaborations with original authors

• Pretty much in-line for example with recent “Executive Orders” of US government

---

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM:  John P. Holdren
        Director

SUBJECT:  Increasing Access to the Results of Federally Funded Scientific Research

February 22, 2013

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

Obama Administration backs open access to all federal research

Papers available one year after publication, data behind them preserved.

by John Timmer - Feb 22 2013, 10:03pm +0100

SCIENTIFIC METHOD / SCIENCE & EXPLORATION

GODMENT SCIENCE POLICY AND EDUCATION
Data ownership

• The policy identifies a data responsible (the *principle investigator*)
• The policy avoids any implications on ownership, obligations or IPRs
  • Ownership is entirely unknown (even to the owner) and might change over time
  • PI=Owner? Probably rarely the case (certainly conflicts best practice recommendations)
Policy adoption

• Facilities with a such a policy (un)officially in place (afaik):
  • ILL, ISIS and DLS

• Why so few?
  • Many obstacles and objections
  • Took ILL ~2 yrs to implement (despite governmental support)
Obstacles

• From users point of view (*):
  • Effort to curate data is too high; Quality of meta-data too low
  • Data could be misused
  • No credit for publishing data
  • 3 year embargo is too short
  • Data are not worth to be kept; data won’t be re-used anyway.

(*) DCC/JISC 2008: INFRASTRUCTURE PLANNING AND DATA CURATION: A COMPARATIVE STUDY OF INTERNATIONAL APPROACHES TO ENABLING THE SHARING OF RESEARCH DATA.
Obstacles

• From users point of view (*)
  • Effort to curate data is too high; Quality of meta-data too low
  • Data could be misused
  • No credit for publishing data
  • 3 year embargo is too short
  • Data are not worth to be kept; data won’t be re-used anyway.

• From facilities point of view
  • Effort to archive and curate data is high
  • Some sites are simply lacking the infrastructure and/or appropriate services & tools
  • Costs to archive and curate data on a long term not negligible, but there is no budget
    • Either requires funding through users (cumbersome) or other sources
  • Facilities don’t really have a mandate to archive data
  • ‘Official’ recommendations and policies are too vague to derive mandates or obligations

(*) DCC/JISC 2008: INFRASTRUCTURE PLANNING AND DATA CURATION: A COMPARATIVE STUDY OF INTERNATIONAL APPROACHES TO ENABLING THE SHARING OF RESEARCH DATA.
Obstacles

• From users point of view (*)
  • Effort to curate data is too high; Quality of meta-data too low
  • Data could be misused
  • No credit for publishing data
  • 3 year embargo is too short
  • Data are not worth to be kept; data won’t be re-used anyway.

• From facilities point of view
  • Effort to archive and curate data is high
  • Some sites are simply lacking the infrastructure and/or appropriate services & tools
  • Costs to archive and curate data on a long term not negligible, but there is no budget
    • Either requires funding through users (cumbersome) or other sources
  • Facilities don’t really have a mandate to archive data
  • ‘Official’ recommendations and policies are too vague to derive mandates or obligations

• From the “agencies” point of view
  • Publicly funded research results are public property
  • 3 years embargo is way too long (EC suggests a maximum of 6 month)
  • Proper annotation is a basic “best practice” requirement
  • Data will be explored if properly annotated
  • What costs?

(*) DCC/JISC 2008: INFRASTRUCTURE PLANNING AND DATA CURATION: A COMPARATIVE STUDY OF INTERNATIONAL APPROACHES TO ENABLING THE SHARING OF RESEARCH DATA.
• Is there a place for open scientific data repositories in the PaN domain?
• There are some examples …

  • Coherent Xray Imaging Database [http://cxidb.org/](http://cxidb.org/)
    • CXIDB is dedicated to further the goal of making data from Coherent X-ray Imaging (CXI) experiments available to all, as well as archiving it.

  • Joint Center for Structural Genomics [http://www.jcsg.org/](http://www.jcsg.org/)
    • >1400 protein structures including all meta-data and raw images

  • IXAS XAFS database [http://ixs.iit.edu/database/](http://ixs.iit.edu/database/)
    • “The EXAFS database, which consists of just about every scan of data ever taken by Farrel Lytle, is a service of the International X-ray Absorption Society”

  • Protein DataBank [http://www.pdb.org](http://www.pdb.org)
    • Raw data not stored, but under debate …

  • Science3D.org … not yet …
• **SRµCTomography**
  • Extremely powerful and versatile tool to study 3-D composition of a variety of materials and biological specimen
  • Excellent tool for comparative studies
  • Investigation of paleontological samples
  • Investigation of extinct insects captured in opaque amber and so on ...
  • Effectively it’s a time machine

Top: Australopithecus sediba from South Africa
Bottom: Australopithecus afarensis from Ethiopia
Imaged on: ID17/ESRF
SRµCT at BW2

42 million-year-old fossil †Mengea tertiara (Insecta, Strepsiptera)

In cooperation with H. Pohl: Entomology Group, FSU Jena
Institut für Spezielle Zoologie und Evolutionsbiologie mit Phyletischem Museum
[Slides from Felix Beckmann, HZG]
42 million-year-old fossil †Mengea tertiara (Insecta, Strepsiptera)

Twisted-wing parasites. The early-stage larvae and the short-lived adult males are free-living, but most of their lives are spent as endoparasites in other insects, such as bees, wasps, leafhoppers, silverfish, and cockroaches.

Caecilians (Gymnophiona)

Limbless fossorial amphibians

Materials from: Thomas Kleinteich & Alexander Haas
Universität Hamburg, Biozentrum Grindel
Slides from: Felix Beckmann, HZG
Pantropical distribution

173 species; total length 10 – 150 cm

based on Himstedt, 1996
Ichthyophis cf. kohlaensis
Comparison between species

*Ichthyophis cf. kohtaoensis*  
*Boulengerula taitana*
Use case: science3d.org

- Lots of efforts goes into proper 3-D modelling
- Even more in 3-D segmentation
- Perfect material for education
  - 3-D printing got quite affordable
  - Take skulls, amphibia, extinct insects into classroom
  - But mostly end up as 2-D illustrations
  - Lots of additional material
  - But hard to discover
Use case: science3d.org

- **VL3: Create an Open Access Tomography database**
  - Well annotated meta-data (needs users support)
  - Discoverable through datacite (for example)
  - Segmented models as well as raw-data
  - Suitable for 3-D visualization and rapid prototyping
  - Free to use for educational purposes

- **Feedback so far**
  - User groups of HZG very positive (F.Beckmann)
  - Teacher education very interested (T.Kracht)
  - Hope to launch science3d.org with a kickoff later this year
Caecilians (Lissamphibia: Gymnophiona) are unique among vertebrates in having two sets of jaw-closing muscles, one on either side of the jaw joint. Using data from high-resolution X-ray radiation computed tomography scans, we modelled the effect of these two muscle groups (mm. levatores mandibulae and m. interhyoideus posterior) on bite force over a range of gape angles, employing a simplified lever arm mechanism that takes into account muscle cross-sectional area and fibre angle.
Thank you for our attention!