Data Strategy - the future of the ExPaNDS and PaNOSC projects

LEAPS General Assembly - 17/09/2021

Presenters: **Andy Götz** (ESRF) + **Sophie Servan** (DESY)

Contributors: Patrick Fuhrmann (DESY), Jordi Bodera (ESRF)
on behalf of PaNOSC + ExPaNDS partners

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Project Factsheets

PaNOSC

Call: Horizon 2020 INFRAEOSC-04

Partners: ESRF, ILL, XFEL.EU, ESS, CERIC-ERIC, ELI-DC, EGI

Description: Cluster of ESFRI Photon and Neutron sources

Observers/non-funded: GÉANT, EUDAT, national RIs

Linked 3rd parties via EGI: DESY, STFC, CESNET

Status: Started 1/12/2018

Github: https://github.com/panosc-eu

Home page: https://panosc.eu

Twitter: @PaNOSC_eu #PaNOSC

Budget: 12 M€

Coordinator: ESRF (A.Götz + J.Bodera)

Started: 1/12/2018

Duration: 4 years

End: 1/12/2022

ExPaNDS

Call: Horizon 2020 INFRAEOSC-5b

Partners: ALBA, DESY, DLS, Elettra, EGI, HZB, HZDR, MAXIV, PSI, SOLEIL, UKRI

Description: European Open Science Cloud Photon and Neutron Data Services

Status: Started 1/9/2019

Github: https://github.com/expands-eu

Home page: https://expands.eu

Twitter: @ExPaNDS_EU #ExPaNDS

Budget: 6 M€

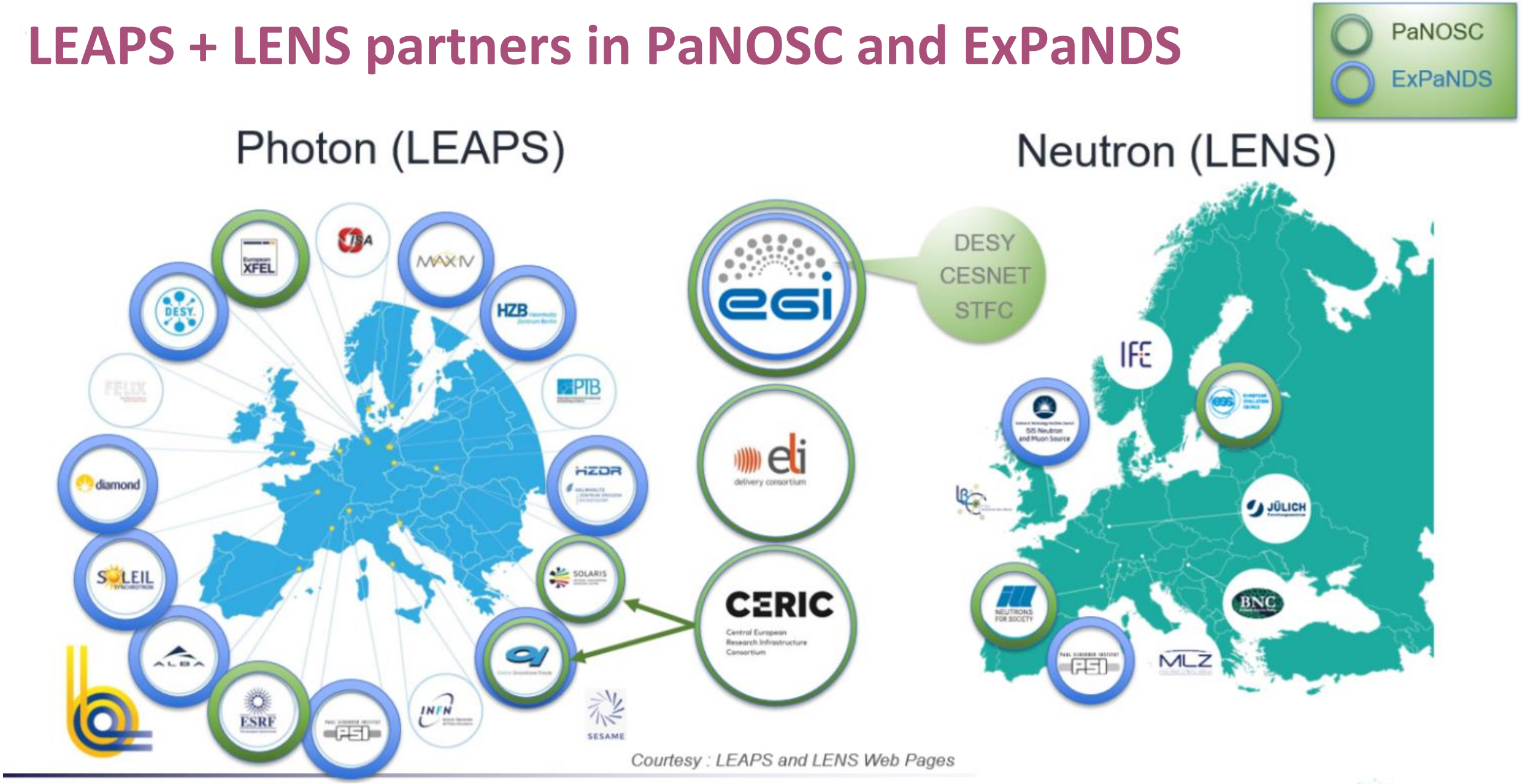
Coordinator: DESY (P.Fuhrman + S.Servan)

Started: 1/9/2019

Duration: 3 years + 6 months extension

End: 1/2/2023

LEAPS + LENS partners in PaNOSC and ExPaNDS



Question: What is EOSC ?

EUROPEAN OPEN SCIENCE CLOUD

panosc photon and neutron open science cloud

ExPaNDS European Open Science Cloud Photon and Neutron Data Services

Question: What is EOSC ?

EUROPEAN OPEN SCIENCE CLOUD

1. FAIR and Open Data
2. Common User Identity (AAI)
3. Connecting Infrastructures
4. Adopting Open Science
   - making science reproducible by sharing notebooks, workflows, software

credits: http://eurodoc.net/news/2018/handbook-on-how-to-be-an-open-scientist-for-early-career-researchers

Answer: WE are EOSC !

panosc photon and neutron open science cloud

ExPaNDS European Open Science Cloud Photon and Neutron Data Services

Common

GOAL(s)

Our most important goal is to provide FAIR Data

- One of the **main objectives** of PaNOSC and ExPaNDS is to avoid this:

Data availability

The data supporting this study can be made available from the corresponding author upon request.

Article | Open Access | Published: 07 February 2020

4D imaging of lithium-batteries using correlative neutron and X-ray tomography with a virtual unrolling technique

Ralf F. Ziesche, Tobias Arlt, Donal P. Finegan, Thomas M. M. Heenan, Alessandro Tengattini, Daniel Baum, Nikolay Kardjilov, Henning Markötter, Ingo Manke, Winfried Kockelmann, Dan J. L. Brett & Paul R. Shearing

Nature Communications **11**, Article number: 777 (2020) | Cite this article

9747 Accesses | **34** Citations | **223** Altmetric | Metrics

What Users want

- Good (meta)data + logbooks
- Performant Download services
- Digital Object Identifiers for Data
- Remote data analysis
- Access to Open Data
- Credit for Data re-use

What Funders want

- FAIR Data
- Open Science
- Digital Object Identifiers for Data
- Reproducible Publications
- Participate in the EOSC
- Metrics about Data Re-Use

10 Primary Outcomes of PaNOSC and ExPaNDS

1. **FAIR data policy** and **DMPs**
2. **FAIR assessment** and common **PID** framework
3. Standardised metadata (**Nexus/HDF5**, PaN ontologies)
4. **Federated search API** for PaN data catalogues
5. **Open Data portal** for searching + downloading data
6. Community **AAI UmbrellaId**
7. **JupyterLab notebooks** and HDF5/NeXus files visualisation
8. **Remote data analysis** with VISA + data analysis pipelines
9. **Simulation** software for simulating experimental data (SIMEX)
10. **PaN-learning** platform (pan-learning.org + pan-training.org)

42 outcomes in total

PaN Data getting more and more attention

- Why it's important
  - attribution of published data to our facilities: **impact**, visibility
  - globalisation of research
- What we do
  - **FAIR data policy** framework for PaN
  - support for implementation: active **DMPs**, **PID** infrastructures...
  - open data harvested and searchable in **EOSC**

☞ cf. demo at ExPaNDS mid-term review

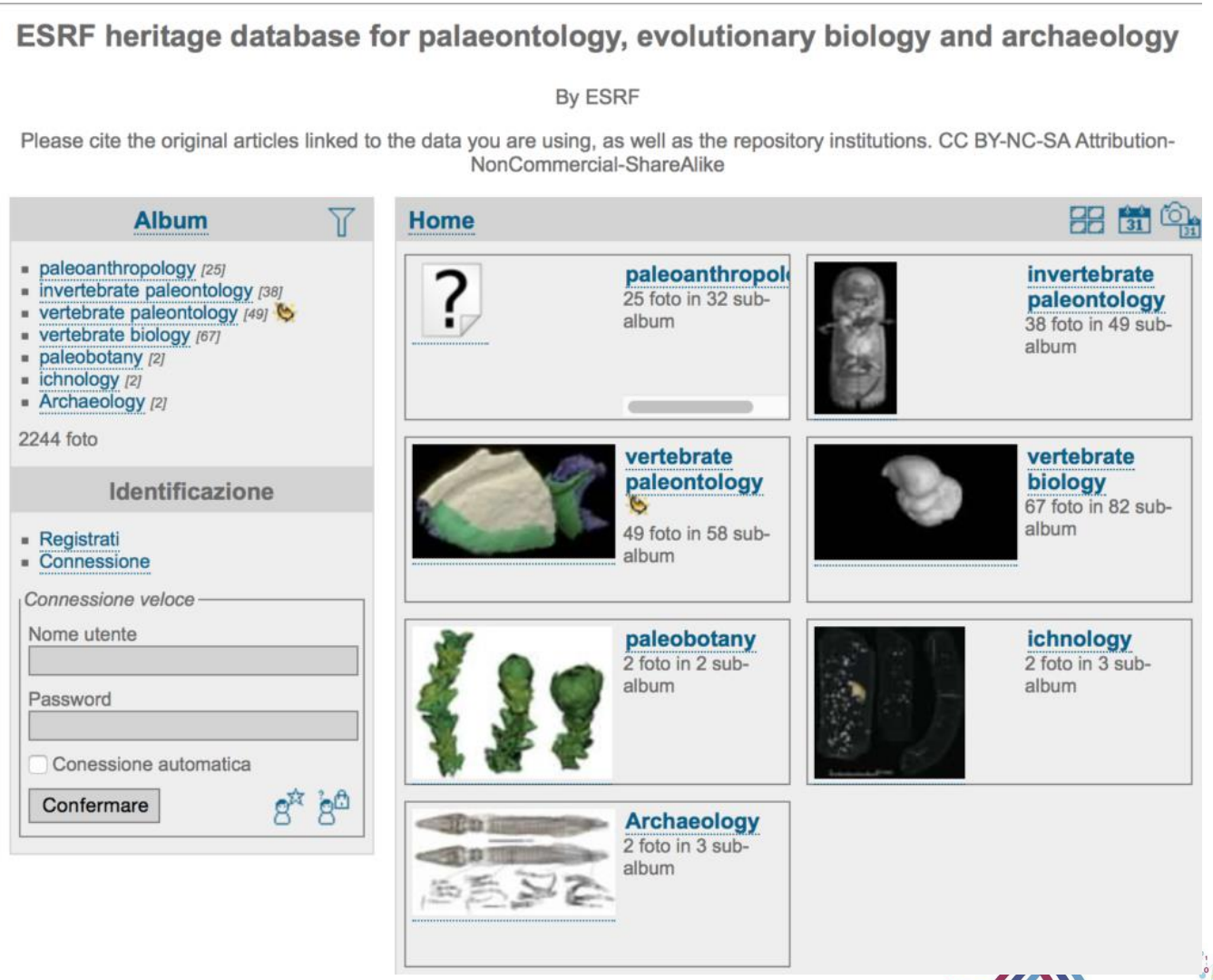
- What we need from LEAPS
  - commitment to FAIR data management
  - means to implement the policies

→ recommend updating of policies and hiring data managers

Example of data publishing

Paleontology @ ESRF
(http://paleo.esrf.eu)

~300 TB in 10 yrs

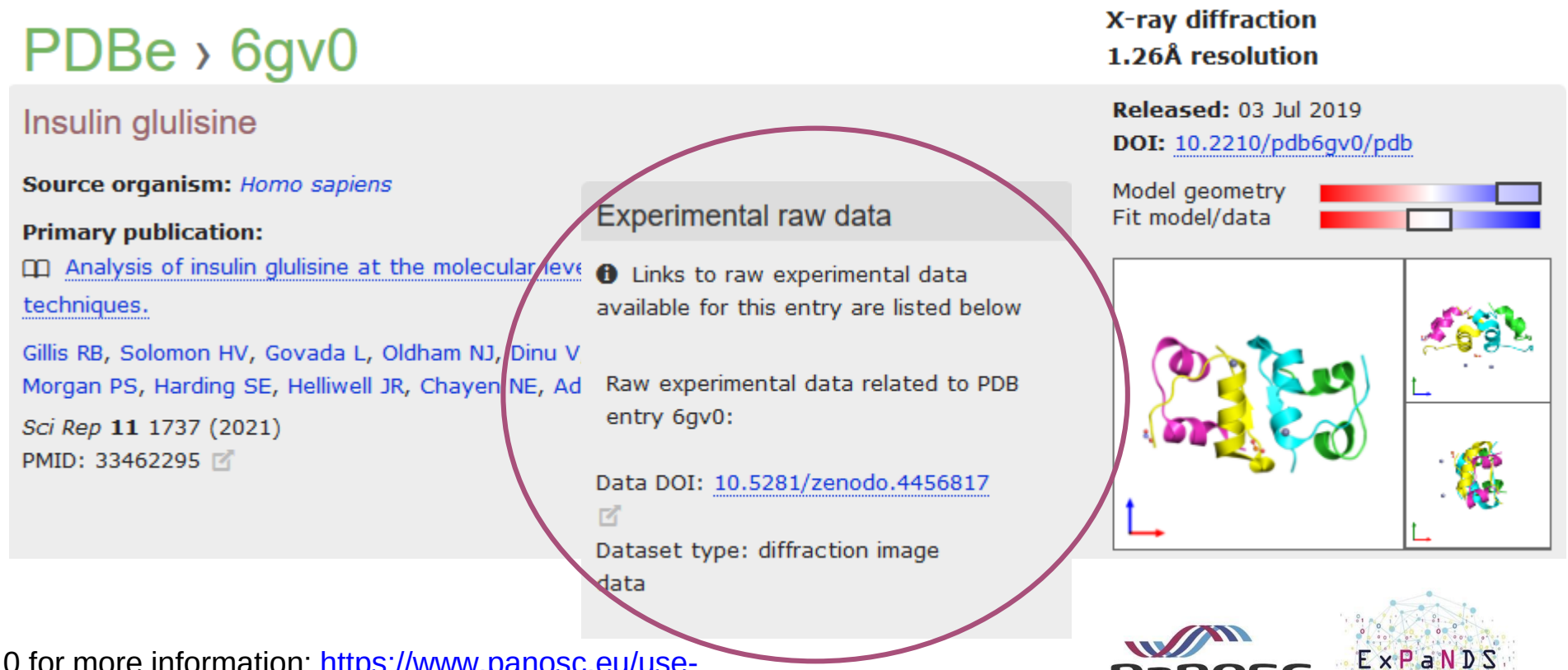


refer to PaNOSC Use Case 11 for more information: https://www.panosc.eu/use-cases/panosc-use-case-11-create-dois-for-datasets-in-paleontology-database/

Linking raw data to the Protein Data Bank

Up to 50% of experiments at synchrotrons are MX
84.4% of PDB entries (148k!) have resulted from these

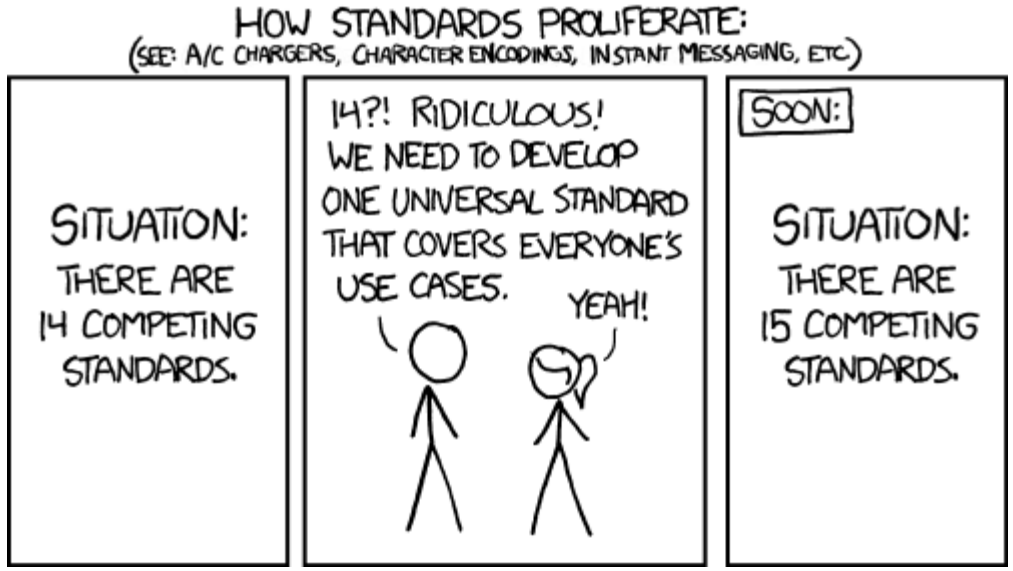
PaNOSC will link raw data to PDB entries



refer to PaNOSC Use Case 10 for more information: https://www.panosc.eu/use-cases/panosc-use-case-10-linking-raw-data-to-the-protein-data-bank-in-europe-pdbe/

Standards on metadata and NeXus format

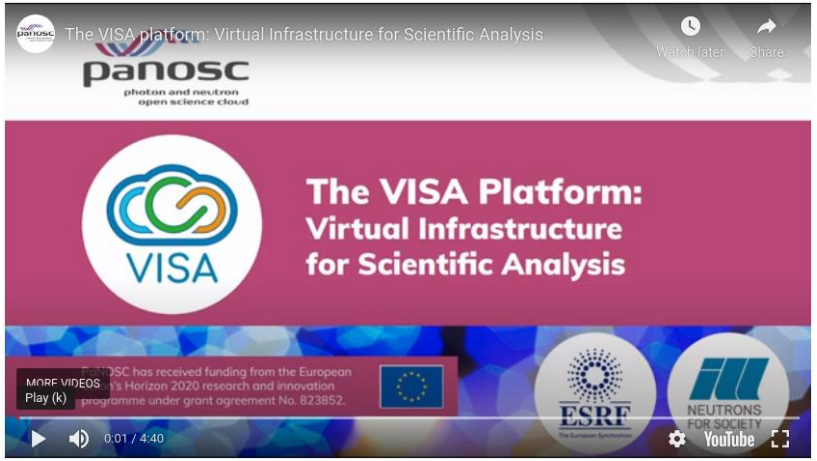
- Why it's important
  - interoperability
  - feeding **machine learning** algorithms
- What we do
  - recommend **metadata lifecycle**
  - develop and publish **PaN ontologies**
  - enrich the **NeXus** format
- What we need from LEAPS
  - recommend to **reuse** our work
  - encourage **implementation** in data catalogues and beamlines



From XKDC: https://xkcd.com/927/

Users asking for remote data analysis

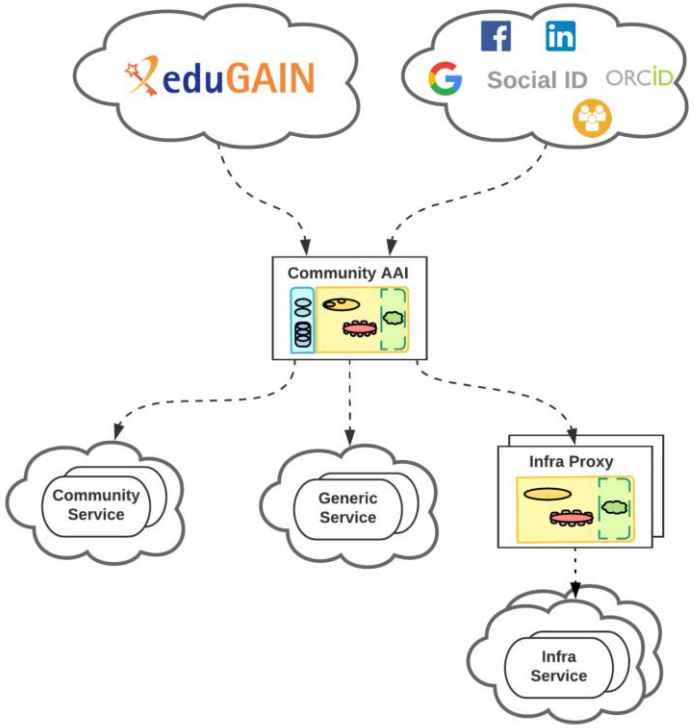
- Why it's important
  - next generation of data analysis in globalised research
  - **remote access** to facilities
- What we do
  - develop and deploy **VISA** platform
  - make **Jupyter** notebooks available at all sites
  - make data analysis pipelines **interoperable**
- What we need from LEAPS
  - commitment to VISA
  - endorse the **shift to cloud computing**



https://bit.ly/VISA-video

One Umbrella ID vs. many accounts

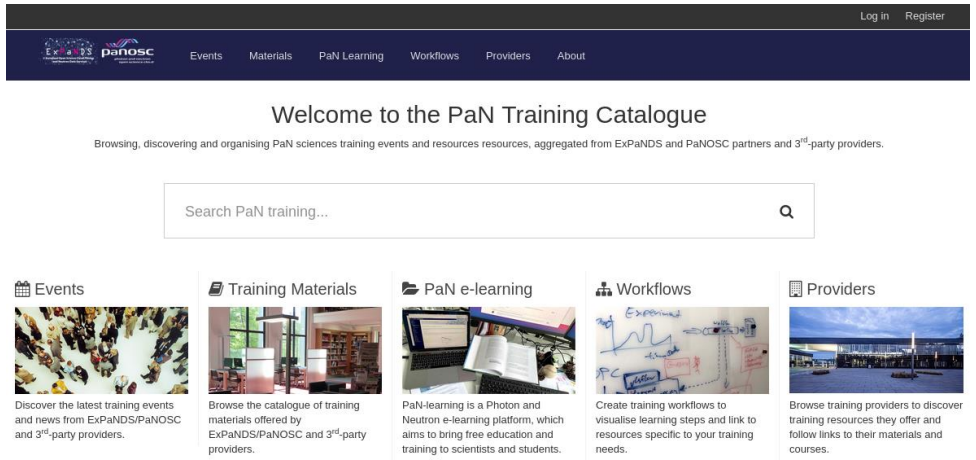
- Why it's important
  - easier use for everyone
  - easier **group access** management
- What we do
  - solve technical challenges for the sustainability of the community AAI
  - and its compatibility with **EOSC AAI**
- What we need from LEAPS
  - services should use the new UmbrellaID proxy



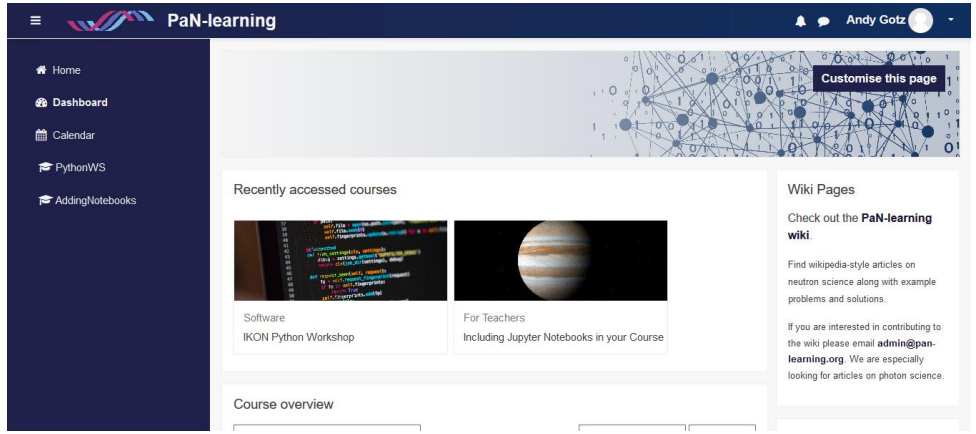
https://aarc-project.eu/

Training material

- Why it’s important
  - a **gap** identified at the facilities
  - a **golden thread** in the current disparity of content available
- What we do
  - a PaN training platform to **create**/store courses and to **collect** existing material
  - reusing successful projects developed by **Elixir** and **SINE2020** e-neutrons
- What we need from LEAPS
  - facilities to **use the PaN training platform**
  - new projects to build on it



https://pan-training.hzdr.de/



https://pan-learning.org

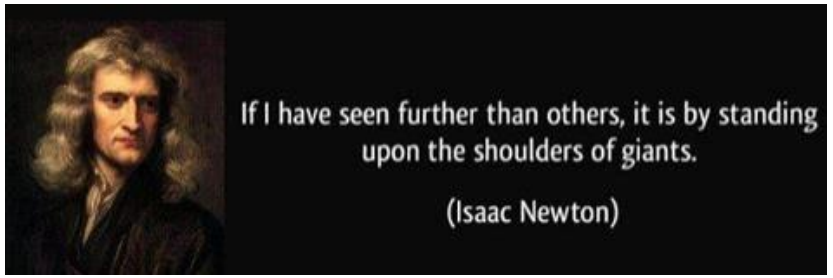
The road to Sustainability is via …

1. **Reuse** the outcomes by integrating them in our operational frameworks

2. **Reduce** the number of single-site solutions, using community ones

3. **Recycle** the outcomes, so we do not start from zero with each new project

A rich community we keep building on



UmbrellaID

Common data policy

Software Catalogue

Jupyter Notebooks

Training platform

FAIR data policy

VISA Remote analysis

DMPs

PaN Data portal

2010 **2015** **2018** **2021** **2022**

pandata europe

pandata ODI

EUCALL

SINE 2020

CALIPSOplus

panosc photon and neutron open science cloud

ExPaNDS European Open Science Cloud Photon and Neutron Data Services

EOSC Future

LEAPS INNOVATION

DiTARI

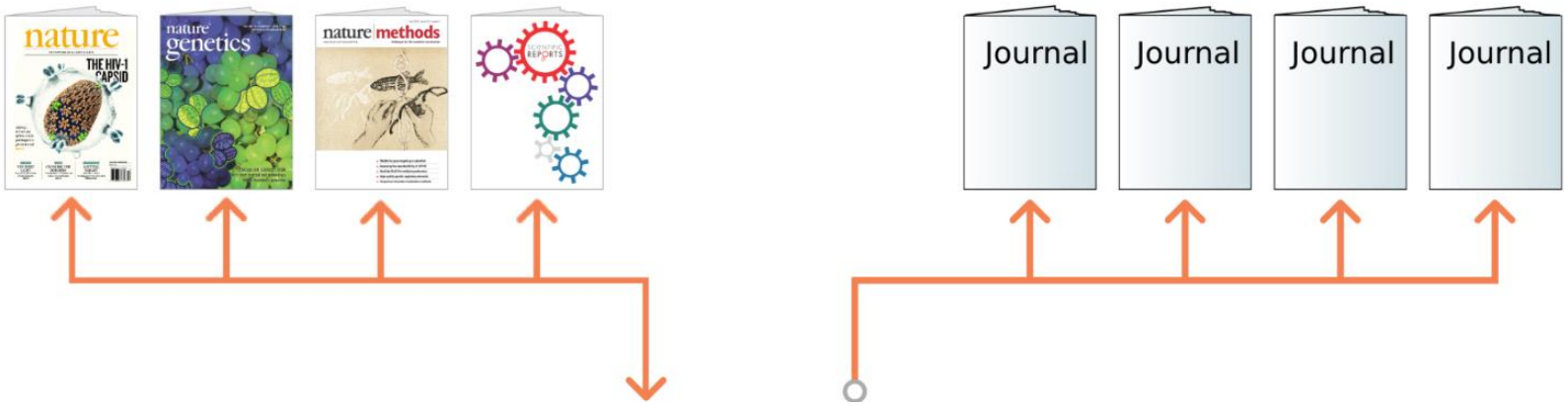
DAPHNE 4NFDI

eRImote

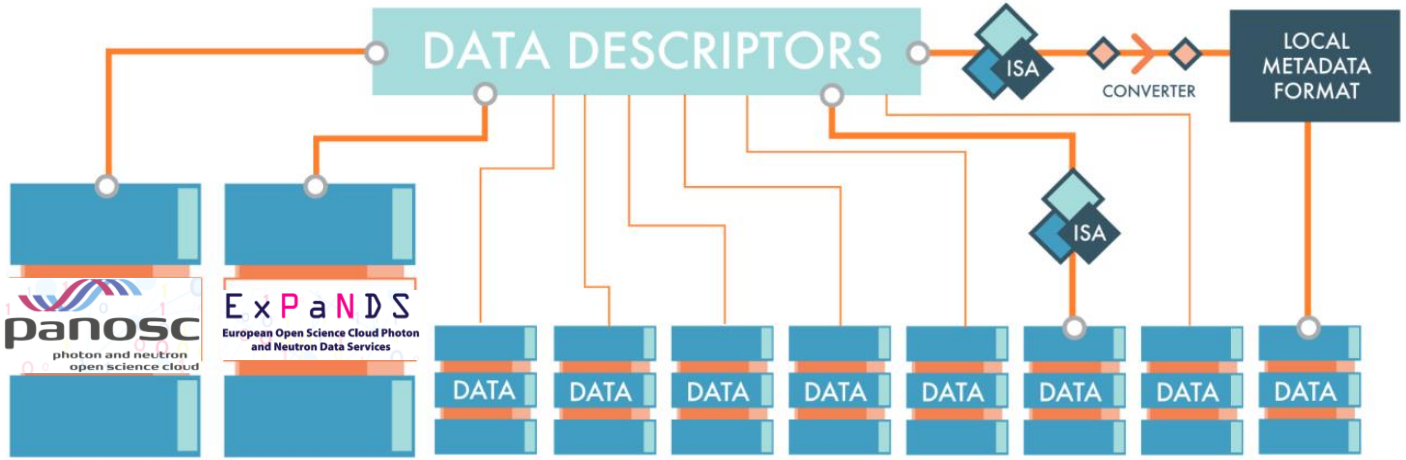
panosc photon and neutron open science cloud

ExPaNDS European Open Science Cloud Photon and Neutron Data Services

Sustain Data → By Publishing



PaN Open Data Commons



credits: http://blogs.nature.com/scientificdata/2013/07/23/scientific-data-to-complement-and-promote-public-data-repositories/

PaN Open Data Commons - Concept

- **Vision** – create a common space for PaNOSC and ExPaNDS facilities where petabytes of PaN FAIR data, analysis software, notebooks, workflows, and training material can be **F**ound, **A**ccessed (downloaded and/or executed), **R**e-Used + **I**mproved i.e. **FAIR**

- **Remote access** – the PaN commons will be accessible remotely while being executed locally (close to the data) or via the EOSC (data needs to be moved)

- **Remote users** – the PaN commons will enable and encourage remote users and experiments (urgently required in the post-COVID-19 phase)

Sustaining the **PaN Open Data Commons**

- Option 1 **- Local implementation ($$$)** – **all sites** implement a local data repository and the PaNOSC API which supports federated searching of Open Data

- Option 2 **- Centralised implementation ($) -** all sites contribute data (and money) to **one site** which implements a PaN data repository for Open Data + Open Science

- Option 3 **- Hybrid implementation ($$)** - **some sites** implement a local data repository and make Open Data available via the PaNOSC search API, sites without a data repository contribute Open Data (and money) to **a centralised site**

Business Models for **PaN Open Data Commons**

1. Project funding → R&D not Operation
2. Collaboration Contracts between RIs
3. Agreements including other funding agencies
4. New legal entity → e.g. an ERIC

Conclusion / What we need from LEAPS

1. **Primary outcomes to be adopted at a maximum number of sites** → Reuse, Recycle, Reduce

2. **LEAPS facilities to make FAIR data reality** → Implement FAIR data policies

3. **A centralized PaN Open Data Commons** → Financing

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Points for discussion

1. How many LEAPS sites will adopt the PaNOSC + ExPaNDS outcomes?

2. Can LEAPS commit to a PaN Open Data Commons?

3. Should we publish a LEAPS Open Science and Data Strategy paper?

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Thank you!

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.