PaNOSC + ExPaNDS outcomes boosting Interoperability Data and Services for Photon and Neutron (PaN) Science

Andy Götz (ESRF, PaNOSC coordinator)

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.

Talk outline

1. **Data interoperability** for **PaN cluster**
2. **Data interoperability** with **science clusters**
3. **Service interoperability** for all clusters

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.

Data interoperability for PaN cluster

1. **Standardized metadata (Nexus + PaNET)**
2. **FAIR assessment** of data catalogues
3. **Visualisation** tools for standard data files
4. **Federated data portal** for PaN data catalogues
5. **Validating** interoperability of scientific data

Walk before you run …

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

Data Interoperability for PaN Cluster

1. **PaN Cluster members are all producers of huge (petabytes) quantities of raw data from experiments with photons and/or neutrons**

2. **Experiments can use one of hundreds of different techniques for very diverse samples e.g. batteries, crocodiles, human organs, proteins, …**

3. **Data Interoperability operates at multiple levels:**

    1. **Raw data** produced during the experiment
    2. **Processed data** produced from the raw data
    3. **Results** from the data analysis

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.

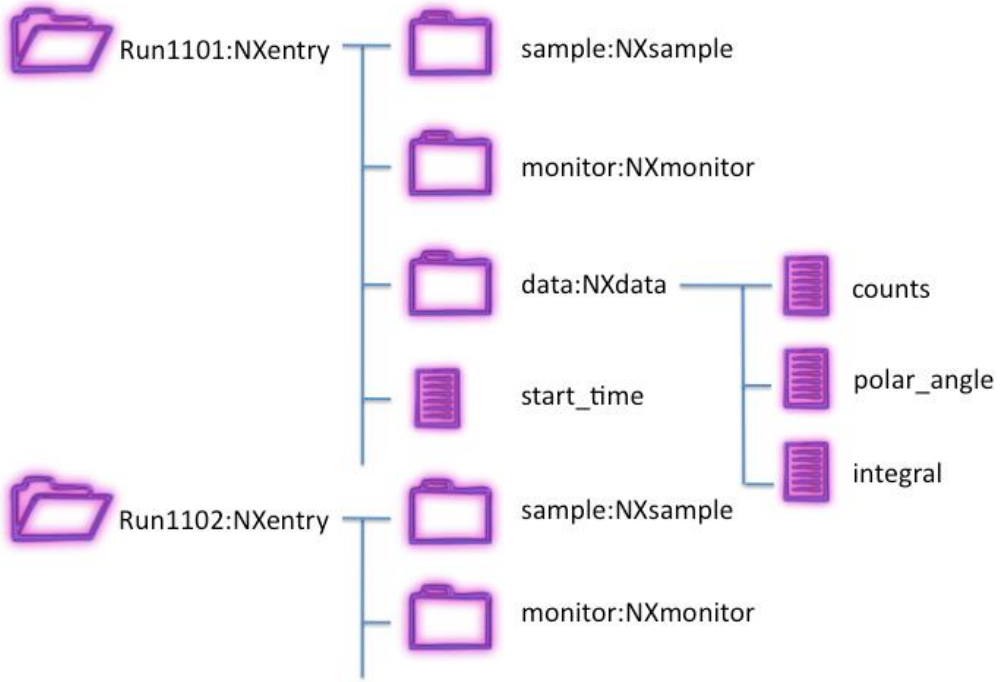
PaNET defining a set of standard techniques

- A set of **opaque PIDs**, each representing a scientific technique; e.g. http://purl.org/pan-science/PaNET/PaNET01168 → serial femtosecond crystallography
- Techniques have **a label** (currently just English).
- Techniques are organised into **a hierarchy**:
- – "**x-ray tomography**" has less specific terms: "tomography", "x-ray probe".
- – "tomography" has more specific terms: "x-ray tomography", "fluorescence
- tomography", "absorption tomography", …

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.

Nexus is the standard PaN ontology

1. NeXus format has been under developed since 1990s
2. Originally for raw data from detectors
3. Now applied to processed data too
4. Although most PaN facilities have adopted Nexus in some form it is still not universally adopted

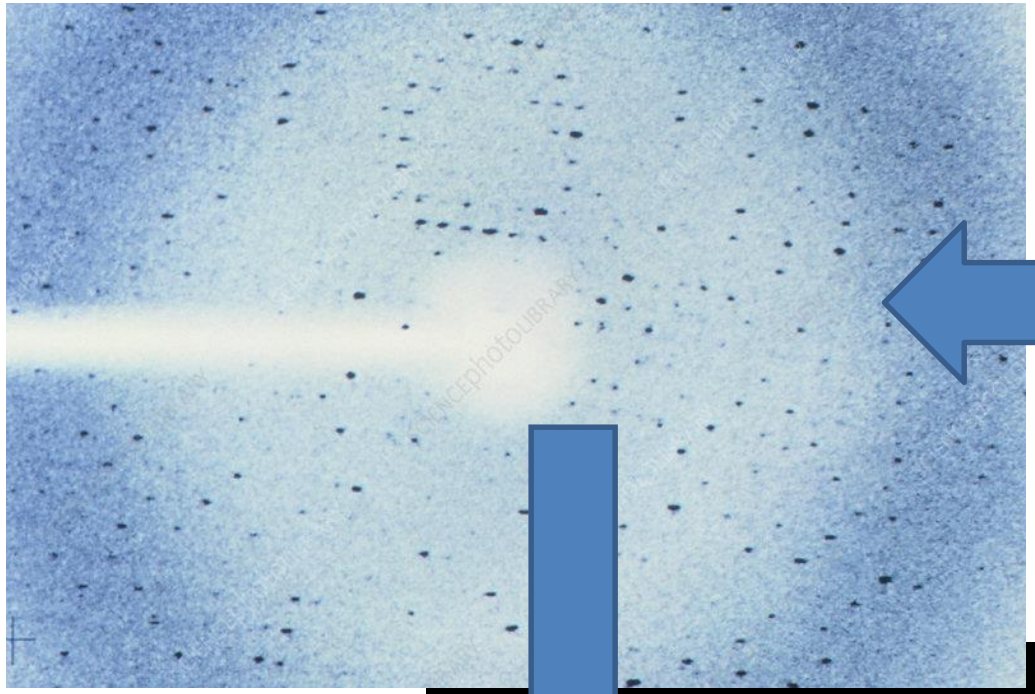


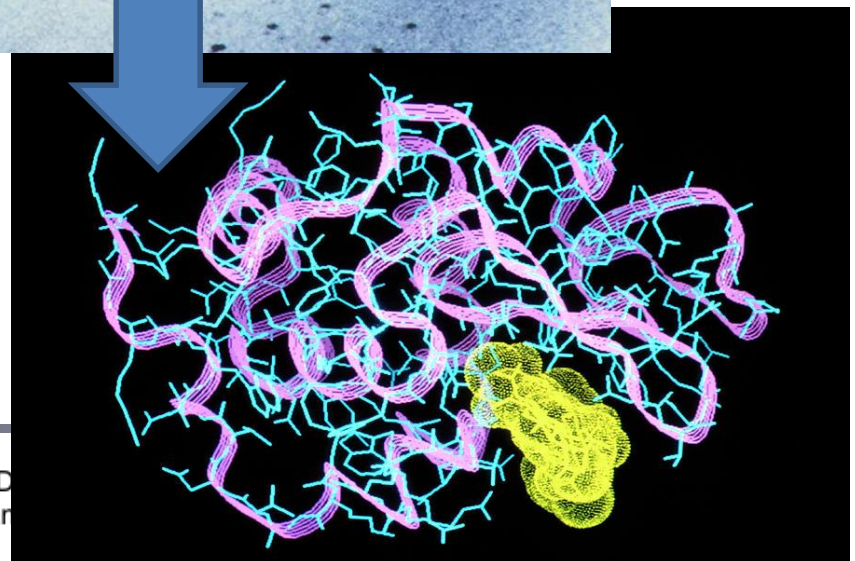
PaNOSC+ExPaNDS reinforced the adoption of Nexus/HDF5 in the **PaN facilities**

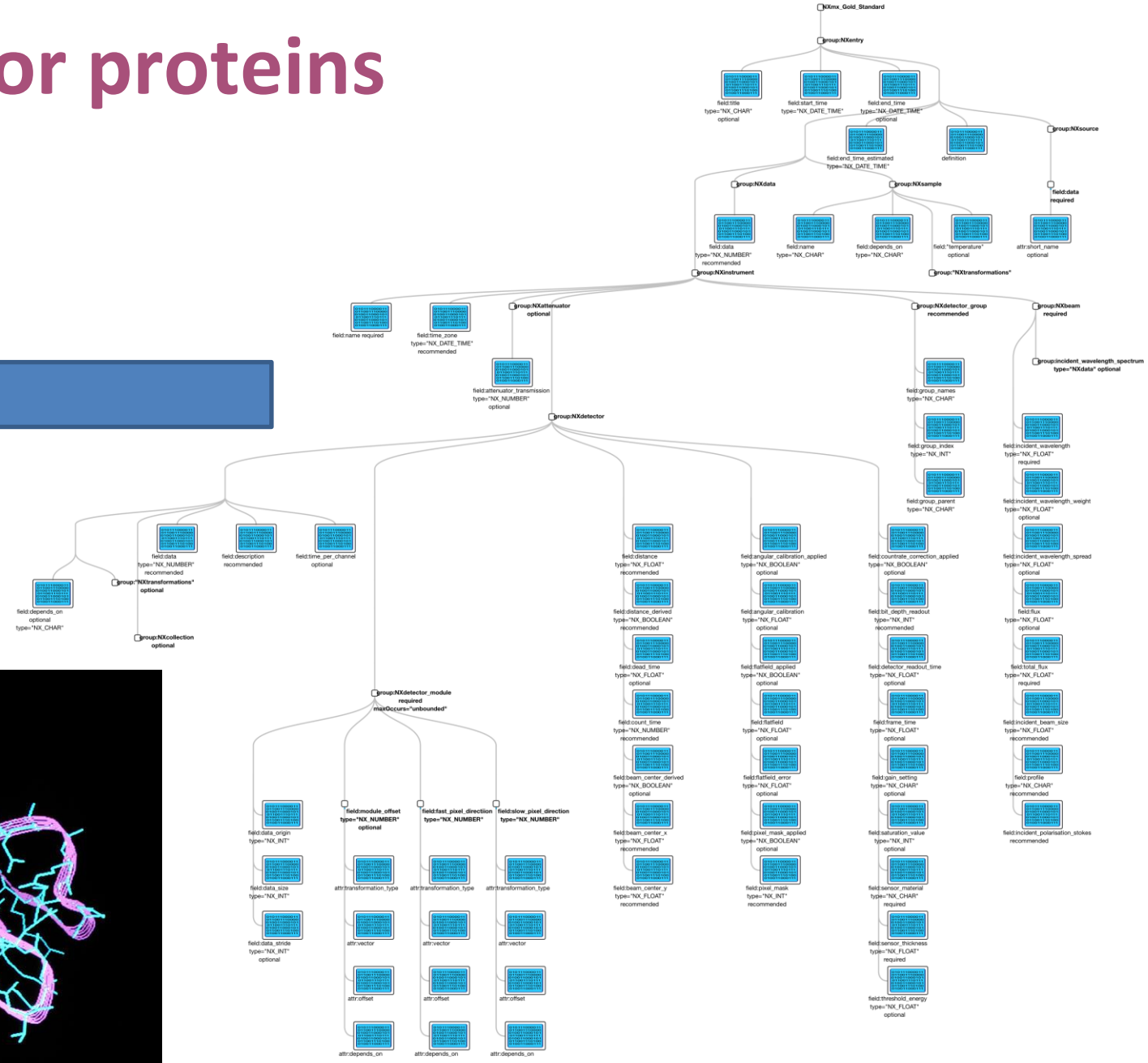
J. Appl. Cryst. (2015). **48**, 301-305 doi:10.1107/S1600576714027575

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.

NxMx – Gold Standard for proteins







IUCrData's Raw Data Letters

editorial

IUCrDATA

Volume 7 | Part 9 | September 2022 | x220821
https://doi.org/10.1107/S2414314622008215
OPEN ACCESS

ISSN: 2414-3146

IUCrData launches Raw Data Letters

L. M. J. Kroon-Batenburg,[a]* J. R. Helliwell[b] and J. R. Hester[c]

[a]Department of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands, [b]Department of Chemistry, The University of Manchester, Manchester M13 9PL, United Kingdom, and [c]Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC, NSW 2232, Australia
*Correspondence e-mail: l.m.j.kroon-batenburg@uu.nl

Keywords: Raw Data Letters; imgCIF.

raw data letters

IUCrData

ISSN 2414-3146

Received 20 April 2021
Accepted 1 May 2021

Keywords: twinning; diffuse scattering; tetraspanin $CD9_{EC2}$.

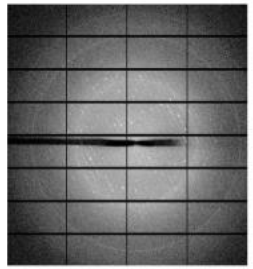
Second extracellular domain of human tetraspanin CD9: twinning and diffuse scattering

Viviana Neviani, Martin Lutz, Wout Oosterheert, Piet Gros and Loes Kroon-Batenburg*

Department of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands. *Correspondence e-mail: l.m.j.kroon-batenburg@uu.nl

Remarkable features are reported in the diffraction pattern produced by a crystal of tetraspanin $CDCD9_{EC2}$, the structure of which was described previously [Oosterheert *et al.* (2020). *Life Sci. Alliance*, **3**, e202000883]. $CD9_{EC2}$ crystallized in space group *P*1 and was twinned. Concurrent with the twinning, diffuse streaks were seen in the direction perpendicular to the twinning interface. Preliminary conclusions are made on packing disorder and potential implications for the observed molecular structure. It is envisaged that the raw diffraction images could be very useful for methods developers in trying to remove the diffuse scattering to extract accurate Bragg intensities or by using it to model the effect of packing disorder on the molecular structure.

Raw data



Structure

Raw diffraction data

HDF5 data file, DOI: https://doi.org/10.5281/zenodo.1234567

he European Union's Horizon 2020 research
nd 857641, respectively.

PaNOSC is working with the International Union of Crystallographers to validate metadata of raw data before publishing

[CheckCif for Raw Data]

checkImgCIF report

```
ImgCIF checker version 2022-07-16

Checking block 5886687 in he4557img.cif

Running checks (no image download)
=====================================

Testing: Required items: PASS

Testing: Data source: PASS

Testing: Axes defined: PASS

Testing: Our limitations: PASS

Testing: Detector translation: PASS

Testing: Scan range: PASS

Testing: All frames present: PASS
   All frames present and correct for SCAN1

Testing: Detector surface axes used properly: PASS

Testing: Pixel size and origin described correctly: PASS

Testing: Check calculated beam centre: PASS

Testing: Check principal axis is aligned with X: PASS
Testing presence of archive:

Testing: All archives are accessible: PASS

Running checks with downloaded images
=====================================

Testing image 4: Image type and dimensions: PASS

Testing image 4: Overloaded values present: PASS

====End of Checks====
```

Raw data table generated from the CIF

Raw data	
DOI	https://doi.org/10.5281/zenodo.5886687
Data archive	Zenodo
Data format	HDF5
Data collection	
Beamline	Diamond I04
Detector	
Temperature (K)	
Radiation type	Synchrotron X-ray source
Wavelength (Å)	0.979491
Beam centre (mm)	−166.874, 172.497
Detector axis	−Z
Detector distance (mm)	-287.22
Swing angle (°)	
Pixel size (mm)	0.075 × 0.075
No. of pixels	4148 × 4362
No. of scans	1
Exposure time per frame (s)	
Scan axis	ω, X
Start angle, increment per frame (°)	0.0, 0.1
Scan range (°)	360.0
No. of frames	3600

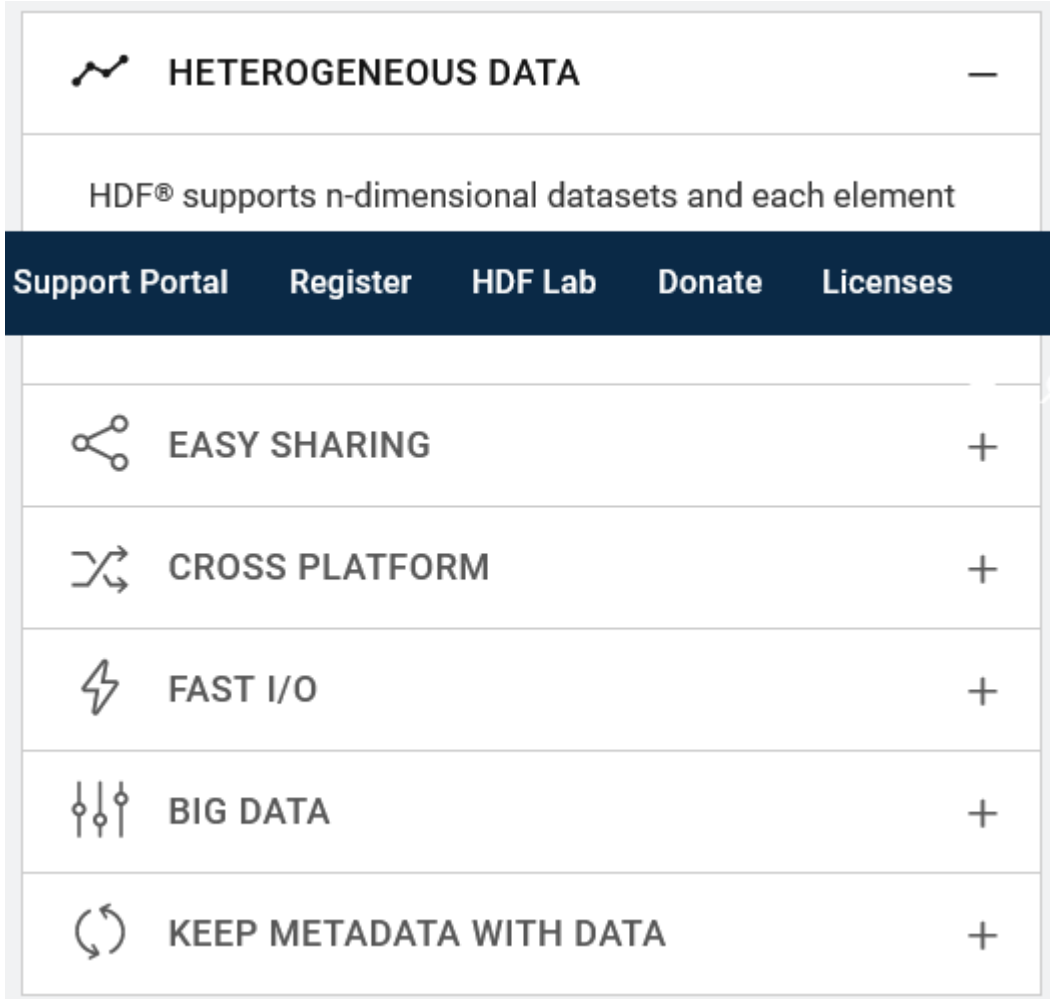
Slide by **John Helliwell, Chair of CommDat, IUCr, 2022**

EOSC support for standard data formats

FAIR Data will be stored for decades (maybe longer) i.e. Interoperability must be ensured over same period

Some widely used data formats depend on a single company e.g. HDF5

EOSC should provide assistance (funding+working group) to sustain data formats over the many decades



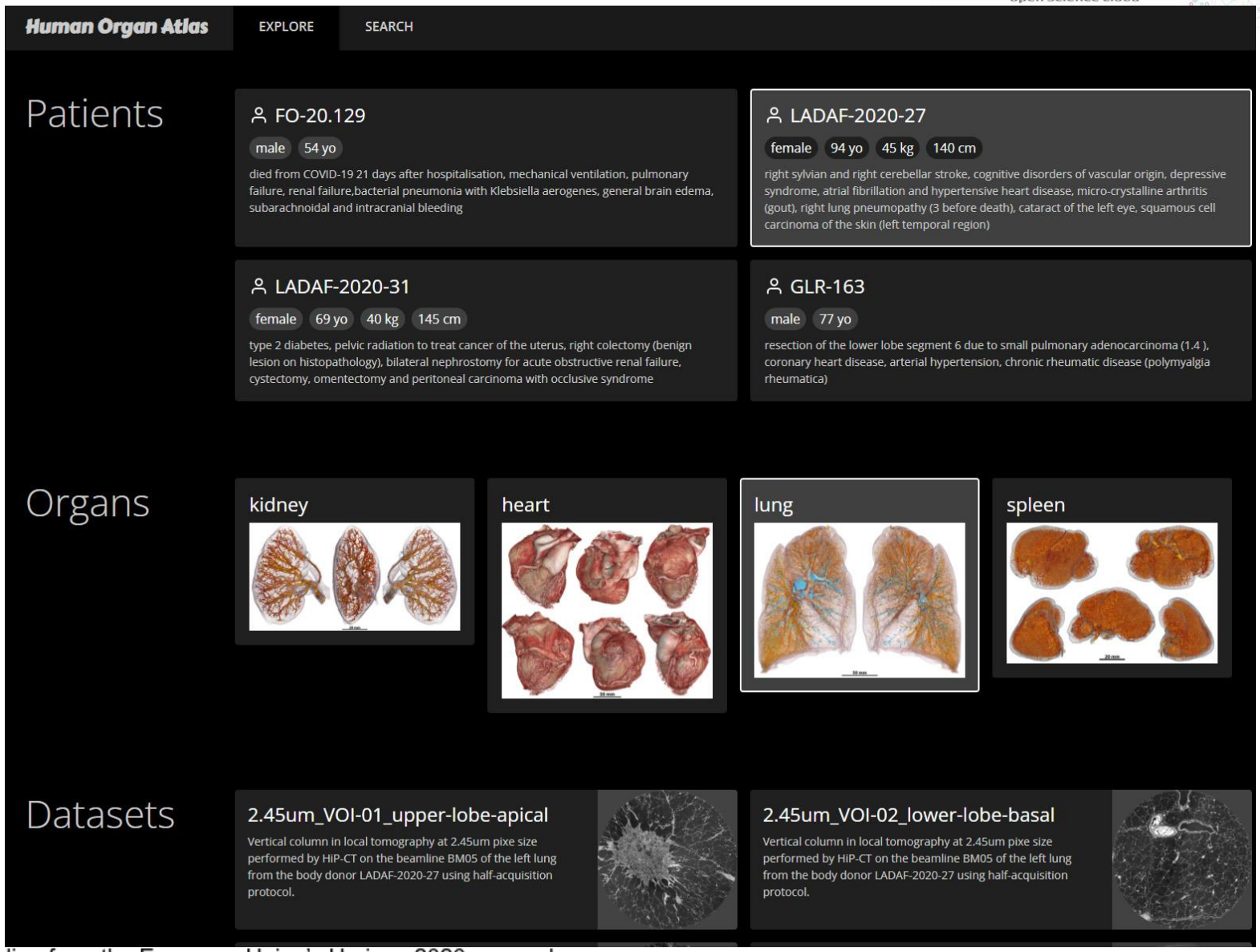
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

The Human Organ Atlas

An open access database, developed as part of the EU PaNOSC project.

Published online on 4/11/2021
https://human-organ-atlas.esrf.eu/

The Human Organ Atlas uses Hierarchical Phase-Contrast Tomography to span a previously poorly explored scale in the understanding of human anatomy, the micron to whole intact organ scale.



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.

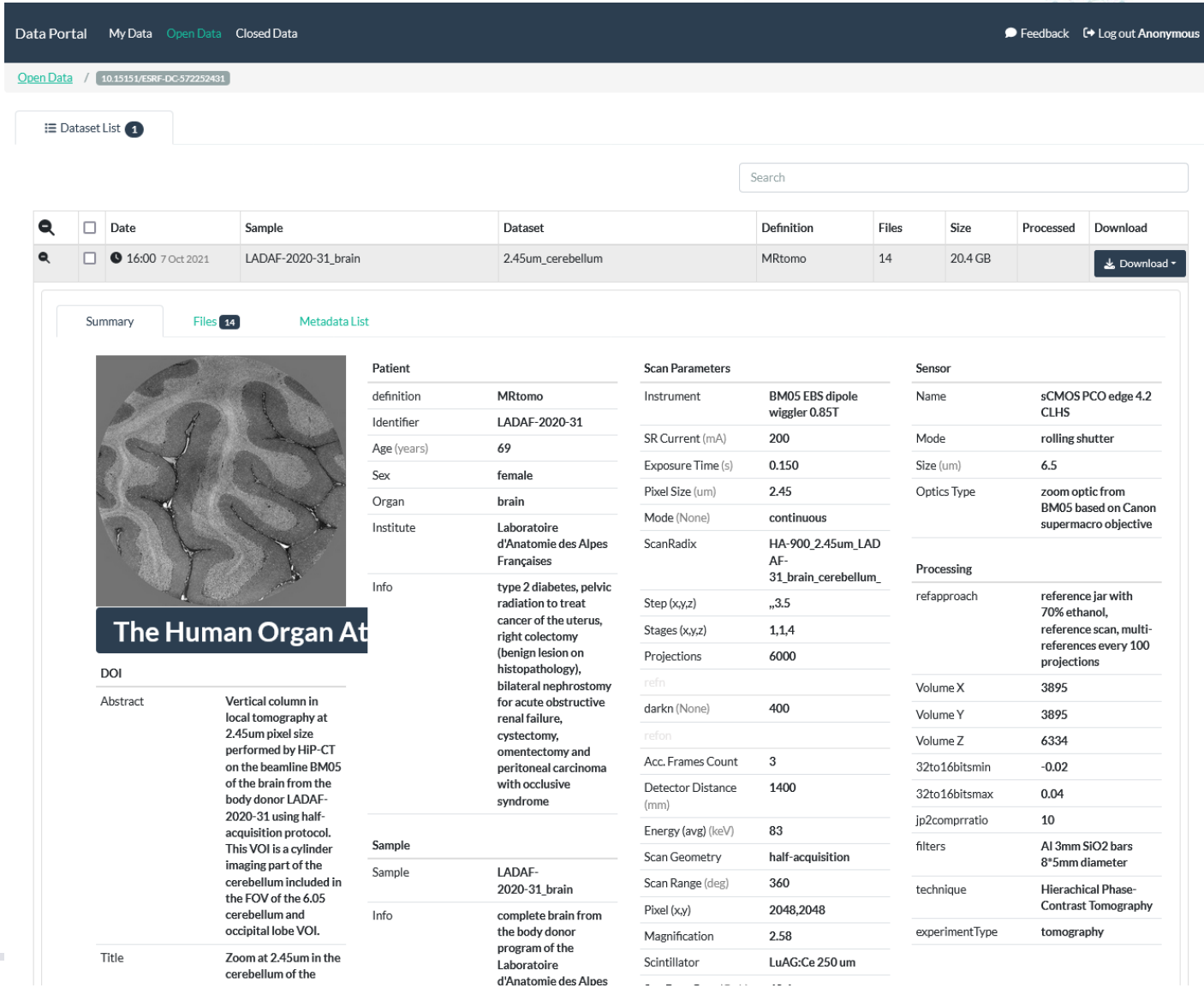
Next step is domain specific vocabularies

1. **https://human-organ-atlas** is an example of **processed data** which is **inter-domain**

2. **Rich sample and experiment metadata** are provided

3. **Data** are proving very **useful** for **medical researchers but**

Doctors require help to use the data with current tools



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.

https://data.esrf.fr/public/10.15151/ESRF-DC-572252431

LEAPS facilities are fighting COVID-19

Photon sources in Europe, represented by LEAPS, joined actively in the fight against the COVID-19 to determine the protein structure, develop vaccines, develop better masks etc.

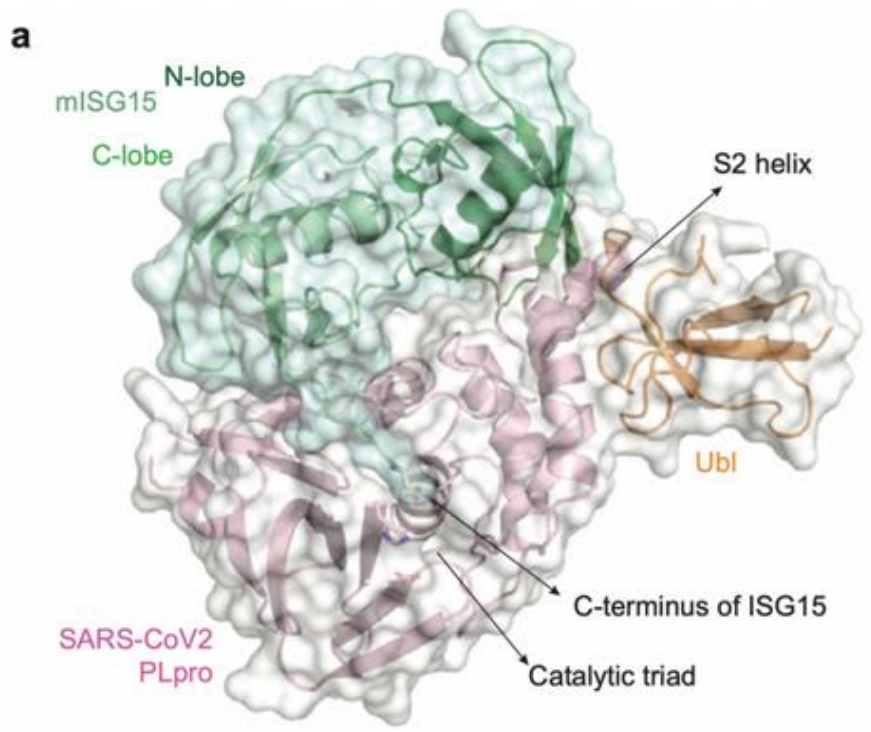
WHERE IS THE DATA?

LEAPS

About Consortium News Events Careers

Research at LEAPS facilities fighting COVID-19

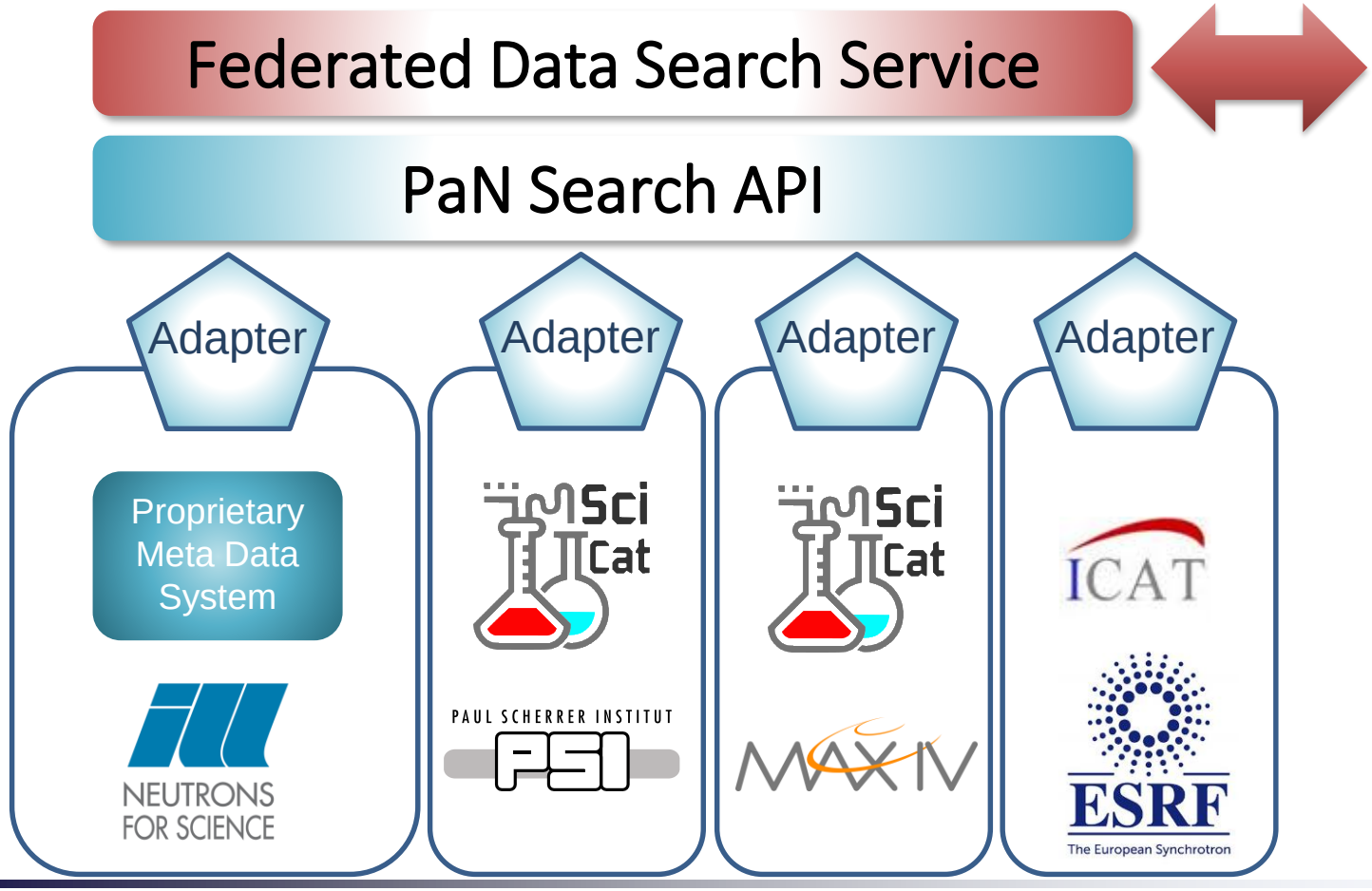
14 May, 2020 by Cristina Pereira

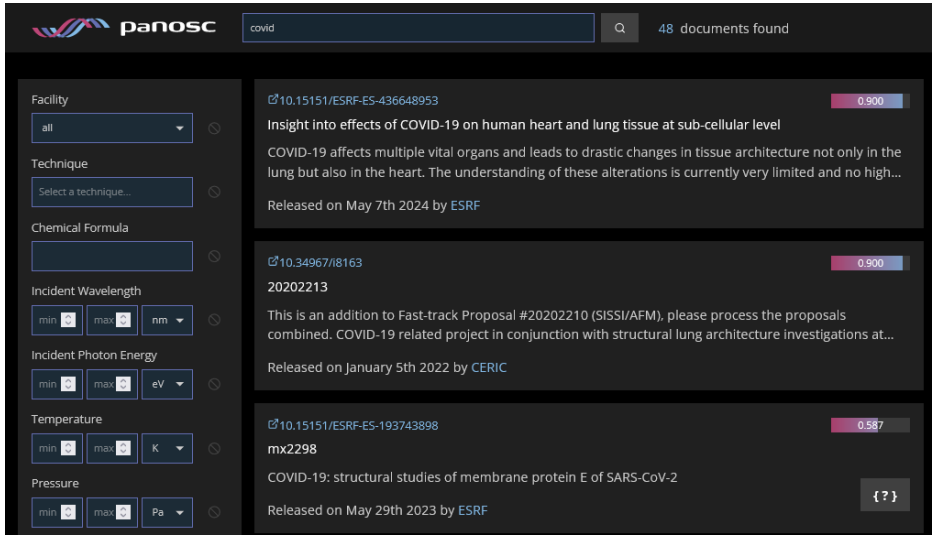


https://leaps-initiative.eu/

PaNOSC and ExPaNDS projects have received funding from the European U
and innovation programme under grant agreements 823852 and 857641, respectively.

The answer is the PaN Data Commons





https://data.panosc.eu

Find data on: *covid, alzheimers, protein xyz, dinosaurs, batteries, …*

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.

Interoperability of services

1. **AAI – GÉANT has integrated the PaN AAI (UmbrellaID) in eduTEAMS**

2. **Data transfer –** PaN facilities have chosen **Globus Online** for data transfer

3. **Data processing – deployed and developed generic tools for Jupyter, developed VISA for remote data analysis**

4. **Training – deployed a PaN training + e-learning platform**

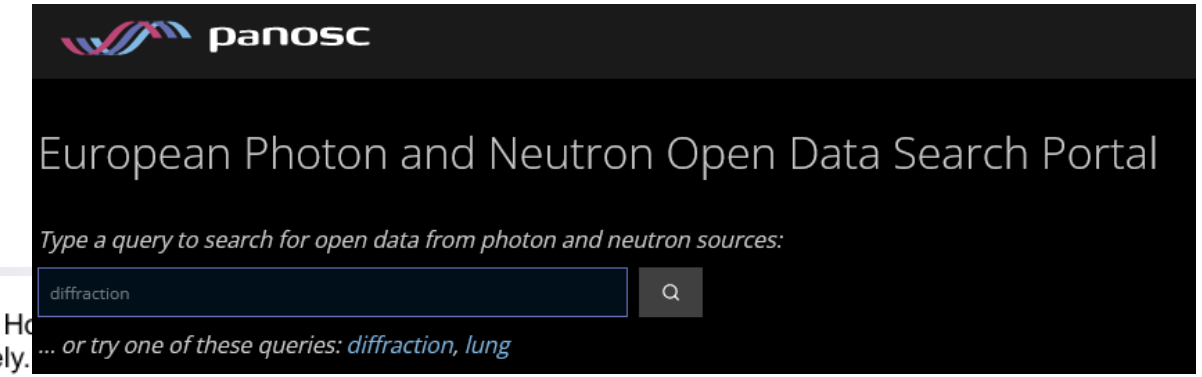
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Remaining interoperability challenges

1. Standardising and gathering sample metadata is challenging due to the large diversity of samples e.g. from crocodiles to quantum dots

2. Standardising metadata for new experimental techniques and new application domains

3. Making data tools easier to use and available in the cloud



4. Improving data search engines

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

Conclusion

1. Interoperability has progressed in the PaN community but still needs work to reuse intra-domain + inter-domain data.
2. Following actions boost Interoperability
   - Automating data processing
   - Working hand-in-hand with publishers
   - Making data findable and accessible via a PaN data commons boosts interoperability (“put your data out there”)
3. EOSC should have a working group on data formats

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.