



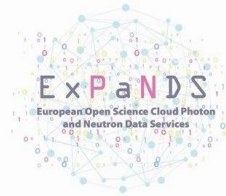
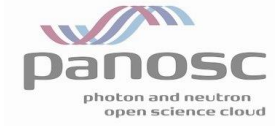
PaNOSC + ExPaNDS outcomes boosting Interoperability Data and Services for Photon and Neutron (PaN) Science

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**



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 ...



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



PaNET defining a set of standard techniques

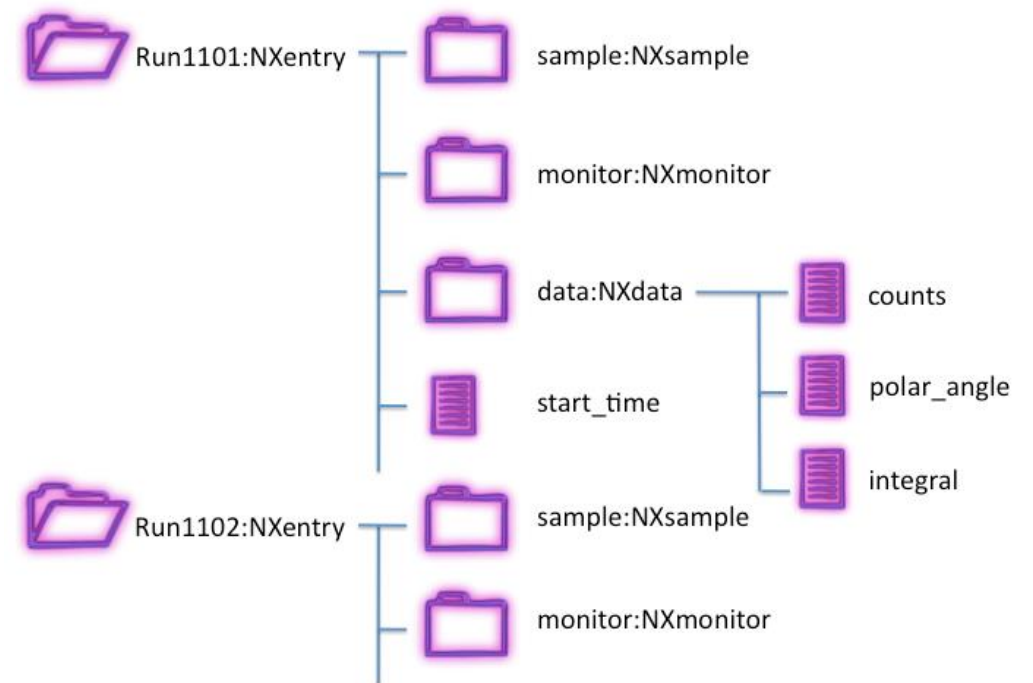
- A set of **opaque PIDs**, each representing a scientific technique; e.g. <http://purl.org/pan-science/PaNET/PaNET01168> → serial femtosecond crystallography
- Techniques have a **label** (currently just English).
- Techniques are organised into a **hierarchy**:
- – “**x-ray tomography**” has less specific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



Nexus is the standard PaN ontology

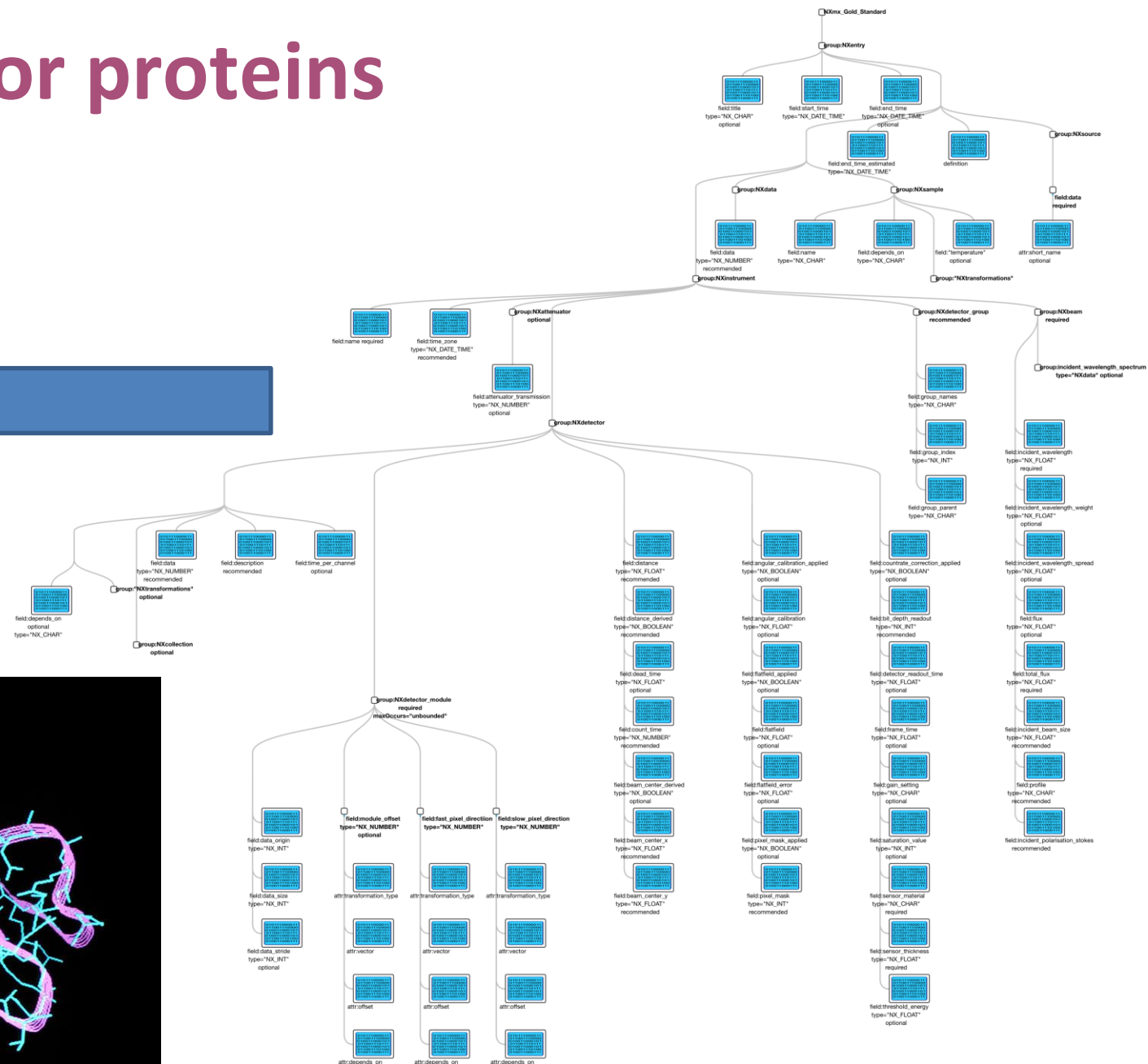
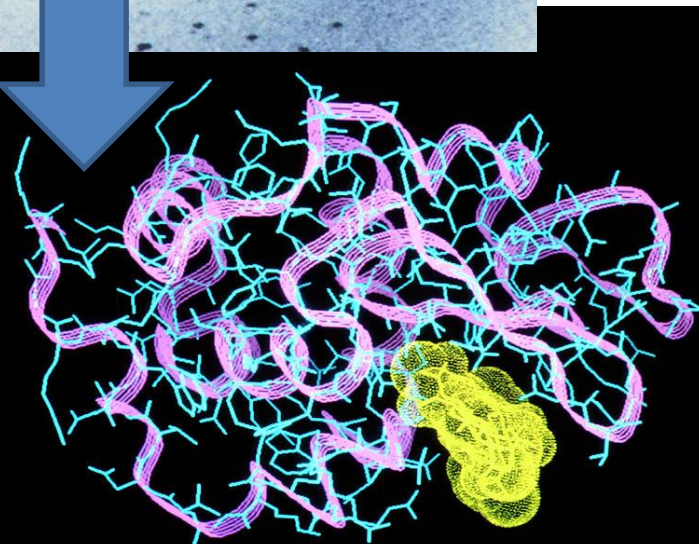
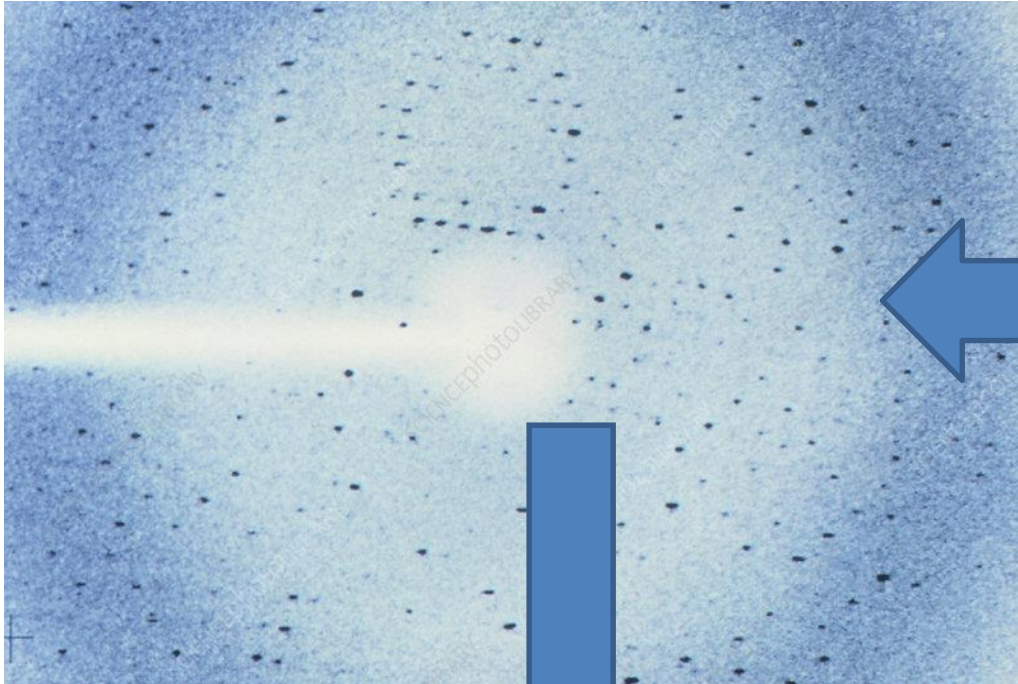
1. NeXus format has been under developed since 1990s
2. Originally for raw data from detectors
3. Now applied to processed data too
4. Although most PaN facilities have adopted Nexus in some form it is still not universally adopted

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



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

NxMx – Gold Standard for proteins



IUCrData's Raw Data Letters



editorial



ISSN: 2414-3146

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

IUCrData launches Raw Data Letters

L. M. J. Kroon-Batenburg,^{a*} J. R. Helliwell^b and J. R. Hester^c

^aDepartment of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands, ^bDepartment of Chemistry, The University of Manchester, Manchester M13 9PL, United Kingdom, and ^cAustralian 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.



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

raw data letters



ISSN 2414-3146

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

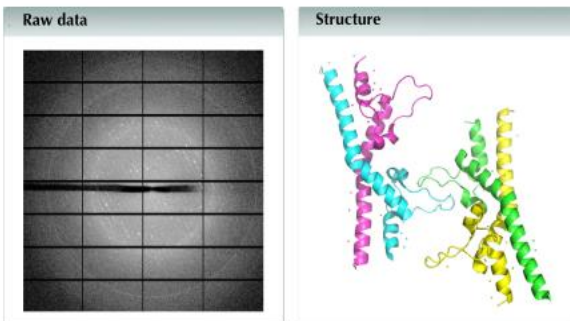
Viviana Neviani, Martin Lutz, Wout Oosterheert, Piet Gros and Loes Kroon-Batenburg*

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

Received 20 April 2021
Accepted 1 May 2021

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

Remarkable features are reported in the diffraction pattern produced by a crystal of tetraspanin CD9_{EC2}, the structure of which was described previously [Oosterheert *et al.* (2020). *Life Sci. Alliance*, **3**, e202000883]. CD9_{EC2} crystallized in space group *P1* 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 diffraction data
HDF5 data file, DOI: <https://doi.org/10.5281/zenodo.1234567>

the European Union's Horizon 2020 research
and 857641, respectively.

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

 HETEROGENEOUS DATA	—
HDF® supports n-dimensional datasets and each element	
Support Portal	Register
HDF Lab	Donate
Licenses	
 EASY SHARING	+
 CROSS PLATFORM	+
 FAST I/O	+
 BIG DATA	+
 KEEP METADATA WITH DATA	+



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.

Human Organ Atlas
EXPLORE
SEARCH

Patients

FO-20.129

male 54 yo

died from COVID-19 21 days after hospitalisation, mechanical ventilation, pulmonary failure, renal failure, bacterial pneumonia with Klebsiella aerogenes, general brain edema, subarachnoidal and intracranial bleeding

LADAF-2020-27

female 94 yo 45 kg 140 cm

right sylvian and right cerebellar stroke, cognitive disorders of vascular origin, depressive syndrome, atrial fibrillation and hypertensive heart disease, micro-crystalline arthritis (gout), right lung pneumopathy (3 before death), cataract of the left eye, squamous cell carcinoma of the skin (left temporal region)

LADAF-2020-31

female 69 yo 40 kg 145 cm

type 2 diabetes, pelvic radiation to treat cancer of the uterus, right colectomy (benign lesion on histopathology), bilateral nephrostomy for acute obstructive renal failure, cystectomy, omentectomy and peritoneal carcinoma with occlusive syndrome

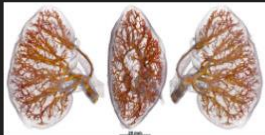
GLR-163

male 77 yo

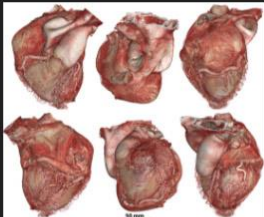
resection of the lower lobe segment 6 due to small pulmonary adenocarcinoma (1.4), coronary heart disease, arterial hypertension, chronic rheumatic disease (polymyalgia rheumatica)

Organs


kidney



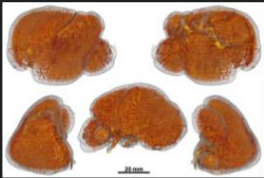
heart



lung



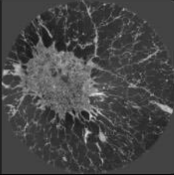
spleen



Datasets

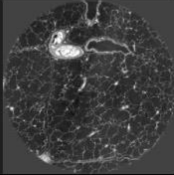
2.45um_VOI-01_upper-lobe-apical

Vertical column in local tomography at 2.45um pixe size performed by HIP-CT on the beamline BM05 of the left lung from the body donor LADAF-2020-27 using half-acquisition protocol.




2.45um_VOI-02_lower-lobe-basal

Vertical column in local tomography at 2.45um pixe size performed by HIP-CT on the beamline BM05 of the left lung from the body donor LADAF-2020-27 using half-acquisition protocol.



photon and neutron open science cloud

and Neutron Data Services

 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

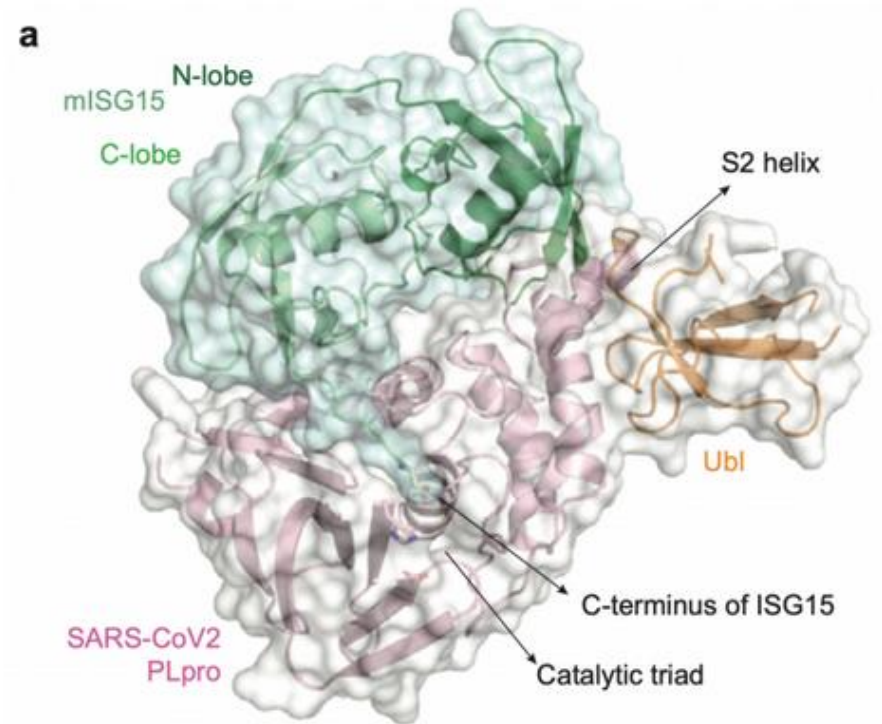
The screenshot shows the Data Portal interface. At the top, there are navigation links: Data Portal, My Data, Open Data, and Closed Data. A search bar is present. Below the navigation, there is a 'Dataset List' section with a search bar and a table of datasets. The table has columns for Date, Sample, Dataset, Definition, Files, Size, Processed, and Download. One dataset is highlighted: LADAF-2020-31_brain, 2.45um_cerebellum, MRtomo, 14 files, 20.4 GB. Below the table, there is a 'Summary' section with a thumbnail image of a brain scan and a title 'The Human Organ Atlas'. To the right of the image is a table of metadata for the patient and scan parameters. The patient information includes: definition (MRtomo), identifier (LADAF-2020-31), age (69), sex (female), organ (brain), and institute (Laboratoire d'Anatomie des Alpes Françaises). The scan parameters include: instrument (BM05 EBS dipole wiggler 0.85T), SR current (200 mA), exposure time (0.150 s), pixel size (2.45 um), mode (continuous), scan radix (HA-900_2.45um_LAD AF-31_brain_cerebellum), step (3.5), stages (1,1,4), projections (6000), refn, darkn (400), acc. frames count (3), detector distance (1400 mm), energy (83 keV), scan geometry (half-acquisition), scan range (360 deg), pixel size (2048,2048), magnification (2.58), and scintillator (LuAG:Ce 250 um). To the right of the patient and scan parameters is a table of sensor information: name (sCMOS PCO edge 4.2 CLHS), mode (rolling shutter), size (6.5 um), optics type (zoom optic from BM05 based on Canon supermacro objective), processing (refapproach: reference jar with 70% ethanol, reference scan, multi-references every 100 projections), volume X (3895), volume Y (3895), volume Z (6334), 32to16bitsmin (-0.02), 32to16bitsmax (0.04), jp2compratio (10), filters (Al 3mm SiO2 bars 8*5mm diameter), technique (Hierarchical Phase-Contrast Tomography), and experimentType (tomography).

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.

Research at LEAPS facilities fighting COVID-19

14 May, 2020 by [Cristina Pereira](#)



WHERE IS THE DATA?

<https://leaps-initiative.eu/>



The answer is the PaN Data Commons

Federated Data Search Service

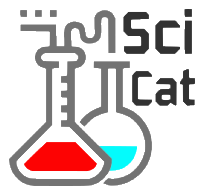
PaN Search API

Adapter

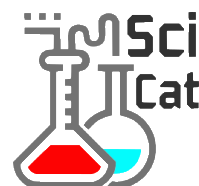
Proprietary
Meta Data
System



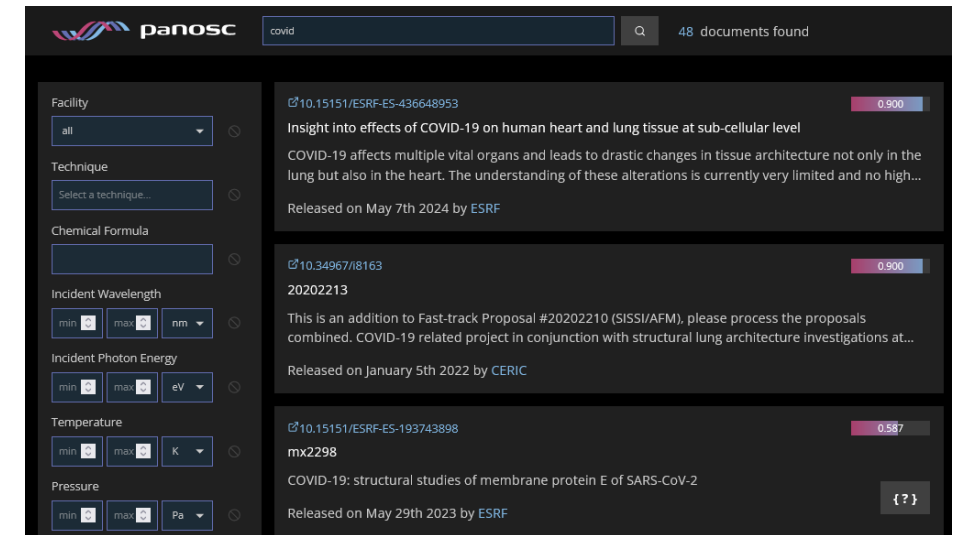
Adapter



Adapter



Adapter



panosc covid 48 documents found

Facility: all

Technique: Select a technique...

Chemical Formula:

Incident Wavelength: min max nm

Incident Photon Energy: min max eV

Temperature: min max K

Pressure: min max Pa

10.15151/ESRF-ES-436648953 0.900
Insight into effects of COVID-19 on human heart and lung tissue at sub-cellular level
COVID-19 affects multiple vital organs and leads to drastic changes in tissue architecture not only in the lung but also in the heart. The understanding of these alterations is currently very limited and no high...
Released on May 7th 2024 by ESRF

10.34967//8163 0.900
20202213
This is an addition to Fast-track Proposal #20202210 (SISSI/AFM), please process the proposals combined. COVID-19 related project in conjunction with structural lung architecture investigations at...
Released on January 5th 2022 by CERIC

10.15151/ESRF-ES-193743898 0.587
mx2298
COVID-19: structural studies of membrane protein E of SARS-CoV-2
Released on May 29th 2023 by ESRF

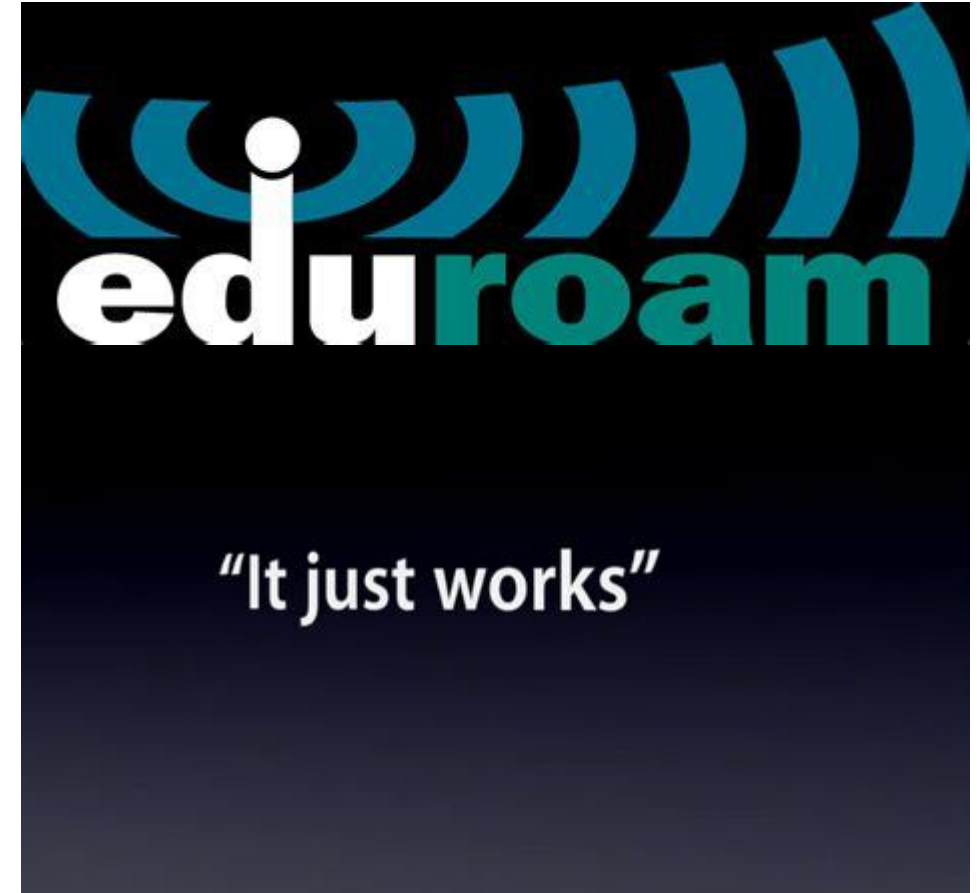
<https://data.panosc.eu>

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



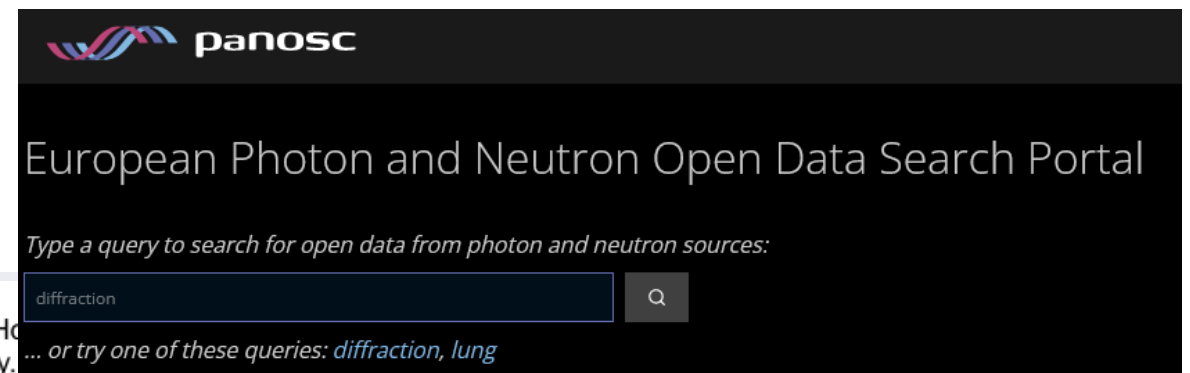
Interoperability of services

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**



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



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**

