



**NOBUGS 2022**

# **Building a PaN Data Commons on the PaNOSC and ExPaNDS outcomes**

**Andy Götz** (ESRF + PaNOSC coordinator) + **Patrick Fuhrman** (DESY + ExPaNDS coordinator)  
on behalf of and with contributions by  
**PaNOSC + ExPaNDS WP2, WP3 + WP4, LEAPS WG3**



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

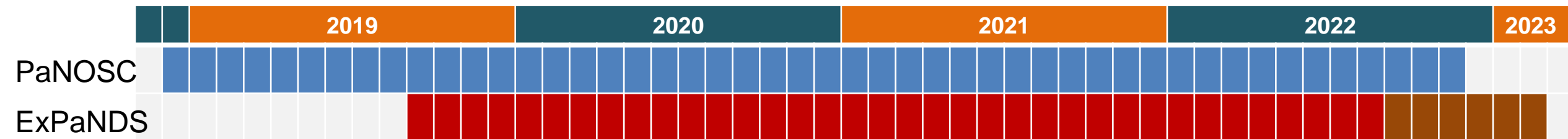
1. What are PaNOSC + ExPaNDS?
2. What are their outcomes?
3. What is a Data Commons?
4. Why build a Data Commons?
5. How to build the Data Commons?
6. Sustaining the Data Commons?
7. What is your role in the Data Commons?
8. Conclusion



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 PaNOSC and ExPaNDS H2020 EOSC projects in numbers



EU Call	HORIZON 2020 INFRA-EOSC-04	HORIZON 2020 INFRA-EOSC-5B
Description	<b>Cluster of ESFRI PaN Sources</b>	<b>EOSC PaN Data Services</b>
Partners	<b>ESRF</b> , ILL, ESS, EU-XFEL, CERIC-ERIC, ELI-DC, EGI	<b>DESY</b> , ALBA, DLS, ELETTRA, EGI, HZB, HZDDR, Max IV, PSI, Soleil, UKRI
Observers	GEANT EU-DAT National RI's	
Linked 3 <sup>rd</sup> Party	DESY STFC CESNET	
Start – End (Duration)	2018-12-01 – 2022-11-30 [4 Years]	2019-09-01 – 2023-02-28 [ 3 ½ Years]
Coordinators	A. Götz, G. Boder	P. Fuhrmann, S. Servan, J. Maraуска
Budget	12 M Euros	6 M Euros
Home Page	<a href="https://panosc.eu">https://panosc.eu</a>	<a href="https://expands.eu">https://expands.eu</a>
Twitter	@PaNOSC_eu #PaNOSC	@ExPaNDS_eu #ExPaNDS
GitHUB	<a href="https://github.com/panosc-eu">https://github.com/panosc-eu</a>	<a href="https://github.com/expands-eu">https://github.com/expands-eu</a>



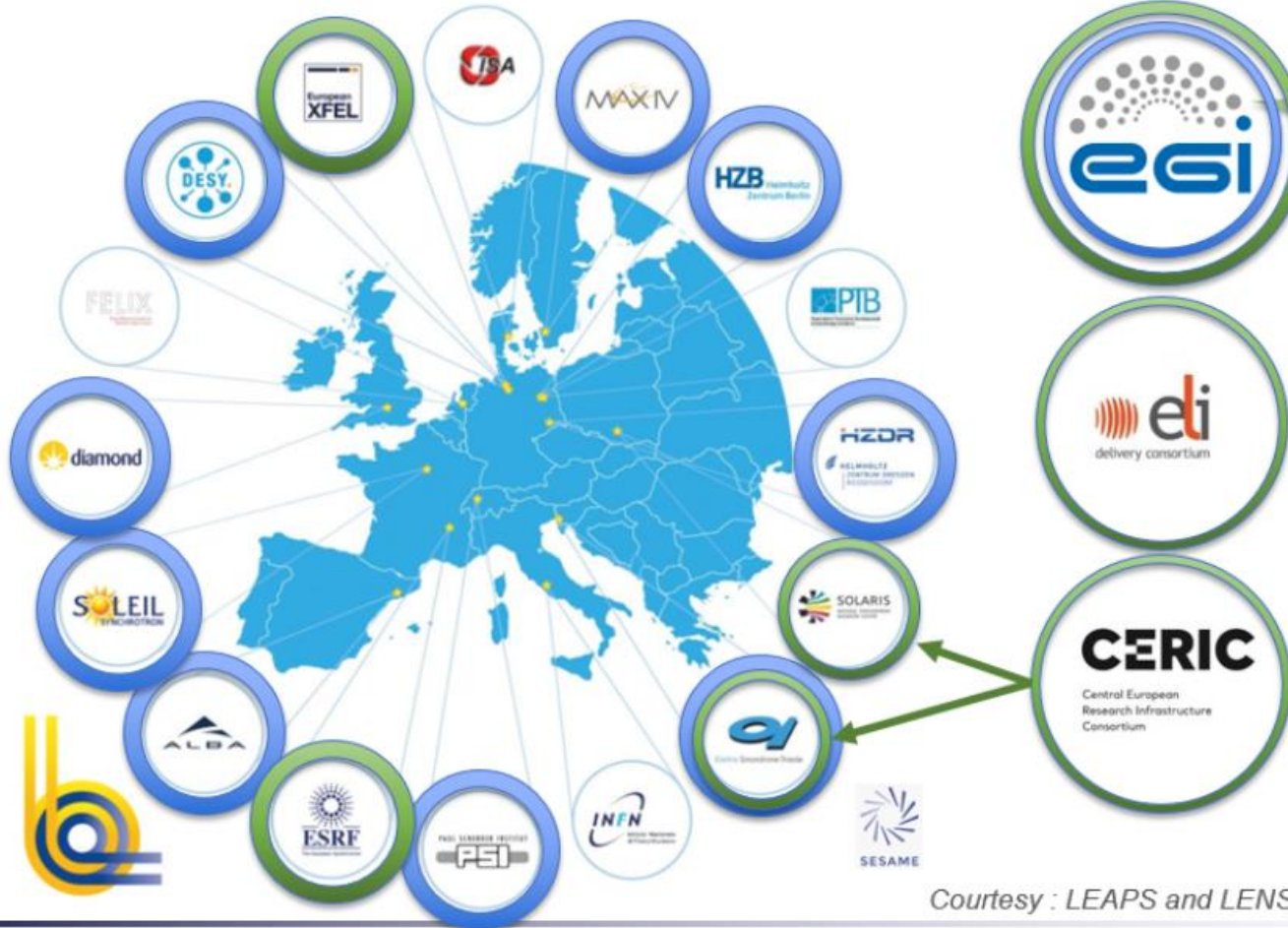
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.



# ExPaNDS and PaNOSC in the PaN EU project landscape



## Photon (LEAPS)



## Neutron (LENS)



Courtesy : LEAPS and LENS Web Pages



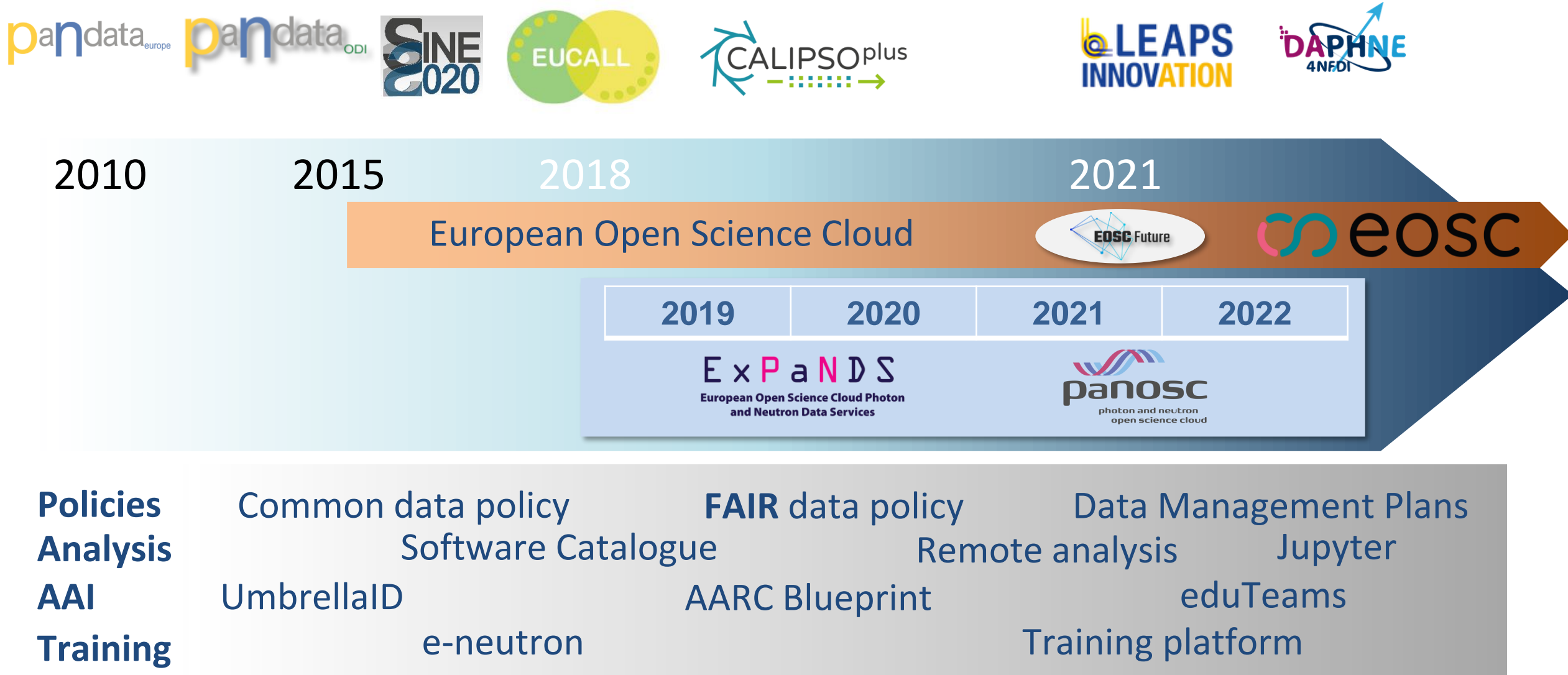
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

**panosc**  
photon and neutron  
open science cloud

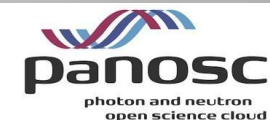




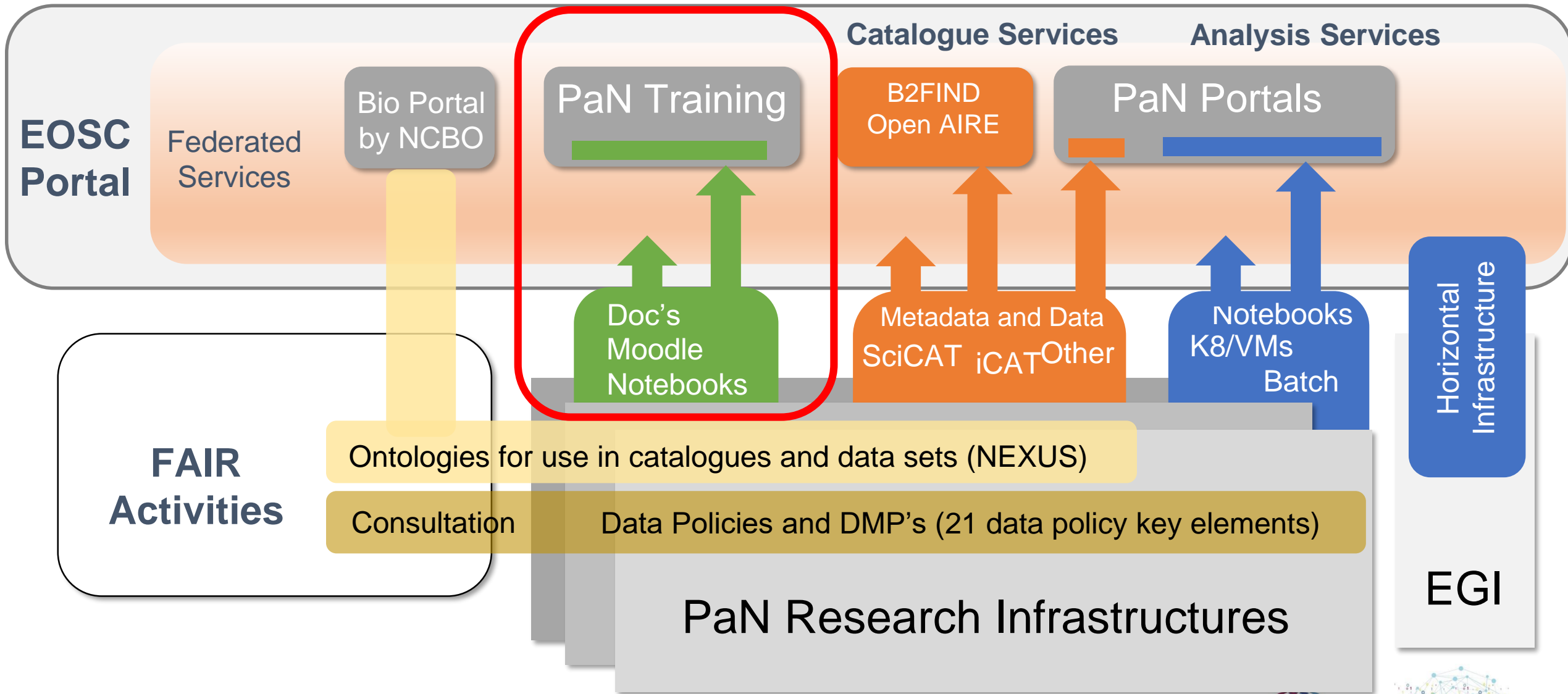
# ExPaNDS and PaNOSC in the PaN EU project landscape



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 Big Picture of EOSC in PaNOSC and ExPaNDS in EOSC



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.

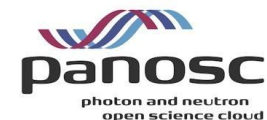


# 10 Primary Outcomes of PaNOSC and ExPaNDS

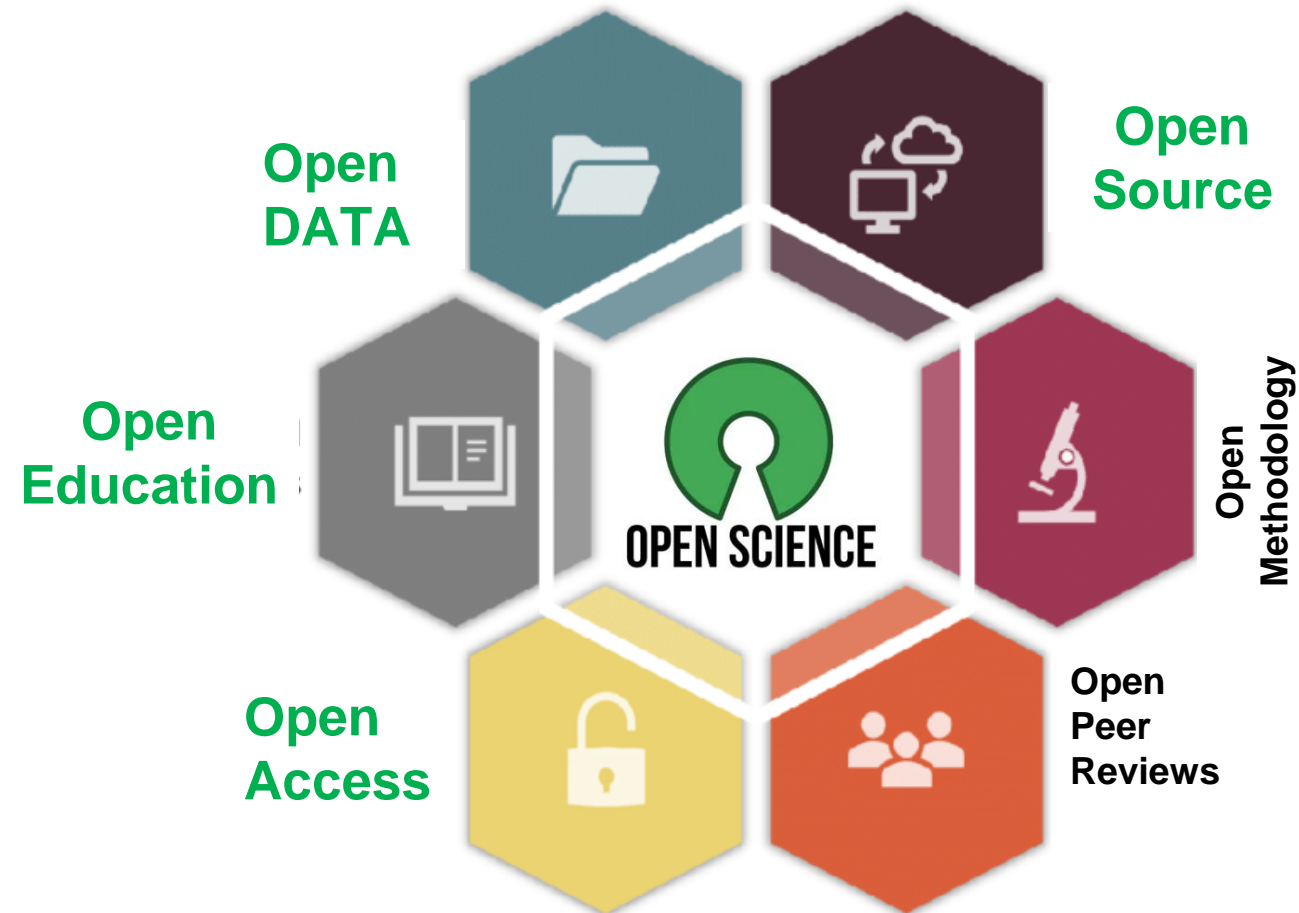
1. **FAIR data policy** and **DMPs**
2. **FAIR assessment** and common **PID** framework
3. Standardised metadata (**Nexus/HDF5**, PaN ontologies)
4. **Federated search API** for PaN data catalogues
5. **Open Data portal** for searching + downloading data
6. Community **AAI UmbrellaId**
7. **JupyterLab notebooks** and **Nexus/HDF5** 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)



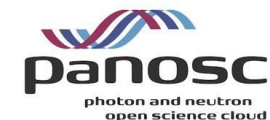
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.



# PaNOSC and ExPaNDS and EOSC are contributing to:



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.





# We had One Common GOAL



# We now have one new Common GOAL

**Making a PaN Data Commons of  
FAIR Data from the PaN Community**

# What is a Data Commons

## Data Commons



**BERKMAN  
KLEIN CENTER**  
FOR INTERNET & SOCIETY  
AT HARVARD UNIVERSITY



# What is a Data Commons?

## *Building a public data commons*

The “data sharing for public good” narratives can be traced at least back to 2011, when the United Nations popularized the concept of “data commons”: using privately-owned big data for sustainable development and humanitarian action.<sup>[16]</sup> The concept of the data commons is crucial, as it defines both values and institutional setups necessary for valuing access and freedom to operate, over the power to appropriate.<sup>[17]</sup>

<https://openfuture.eu/publication/public-data-commons/>

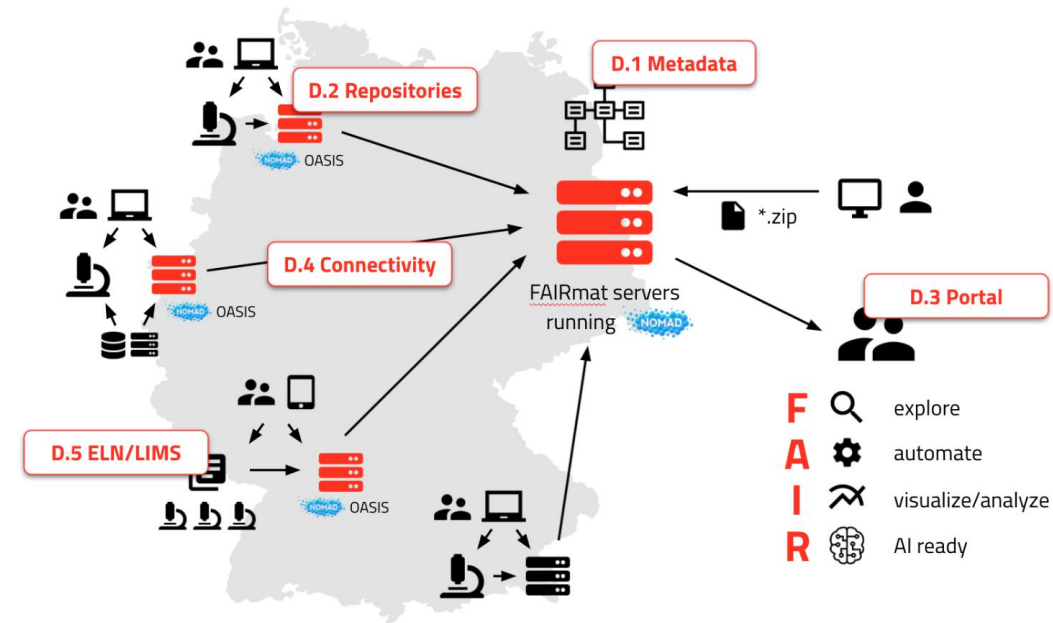
*The overarching goal of the NIH Data Commons was to accelerate new biomedical discoveries by developing and testing a cloud-based platform where investigators could store, share, access, and interact with digital objects (data, software, etc.) generated from biomedical and behavioral research.*

<https://commonfund.nih.gov/commons>

Data Commons aggregates data from a **wide range of sources** into a unified database to make it more accessible and useful. More on **why we are building Data Commons**.

<https://datacommons.org/>

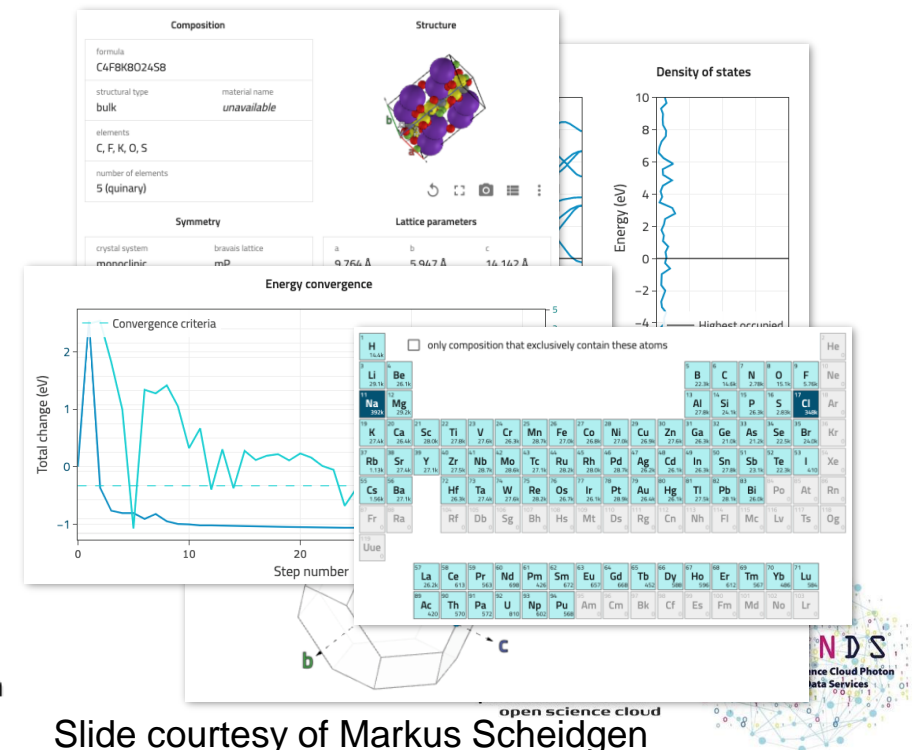
# Example of a Data Commons



## NOMAD: Publishing research data

*More than 12 million of simulations (22 billion quantities) from over 500 authors world-wide*

- Free publication and sharing data of data
- Extracts **rich metadata** for more than **50 codes**
- All data in a **raw** and a common **machine readable** form
- Use integrated tools to **explore, visualize, and analyze**



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.

Slide courtesy of Markus Scheidgen



# Google Data Commons

## <https://datacommons.org>

Data Commons

Explore ▾

Documentation ▾

About ▾



Data Commons aggregates data from a **wide range of sources** into a unified database to make it more accessible and useful. More on **why we are building Data Commons**.

### Explore the data



Scatter Plots



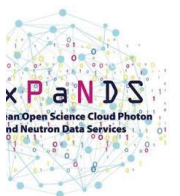
Map Explorer



Timelines



Place Explorer



# Google Dataset Search:

The image shows a screenshot of the Google Dataset Search interface. At the top, the Google logo is on the left, and a search bar contains the text 'esrf'. To the right of the search bar are icons for a grid and a lightbulb. Below the search bar, there are filter buttons: 'Last updated', 'Download format', 'Usage rights', 'Topic', and 'Free'. On the far right, there is a 'Saved datasets' button. The search results section shows '100+ datasets found'. The first result is a 'FAIRsharing record for: European Synchrotron Radiation Facility Data Portal' from 'fairsharing.org', updated on May 10, 2021. The second result is 'XRF mapping of samples, ESRF' from 'doi.pangaea.de', updated on Apr 11, 2017, with file formats 'html, tsv'. A third result is partially visible: 'X-ray data collection (ESRF) and refinement statistics'. A large, dark purple diagonal banner with the text 'A poor man's Data Commons' is overlaid across the center of the page. At the bottom right, there is a URL: <https://datasetsearch.research.google.com/>.

Google

esrf

▼ Last updated ▼ Download format ▼ Usage rights ▼ Topic Free

Saved datasets

100+ datasets found

FAIRsharing record for:  
European Synchrotron  
Radiation Facility Data Portal  
fairsharing.org  
Updated May 10, 2021

XRF mapping of  
samples, ESRF  
doi.pangaea.de  
html, tsv  
Updated Apr 11, 2017

X-ray data collection (ESRF)  
and refinement statistics

A poor man's Data Commons

<https://datasetsearch.research.google.com/>

# Genomics Data Commons -

## <https://portal.gdc.cancer.gov/>

Harmonized Cancer Datasets

## Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

Q e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

### Data Portal Summary

[Data Release 34.0 - July 27, 2022](#)

PROJECTS

72

PRIMARY SITES

67

CASES

86,046

FILES

863,977

GENES

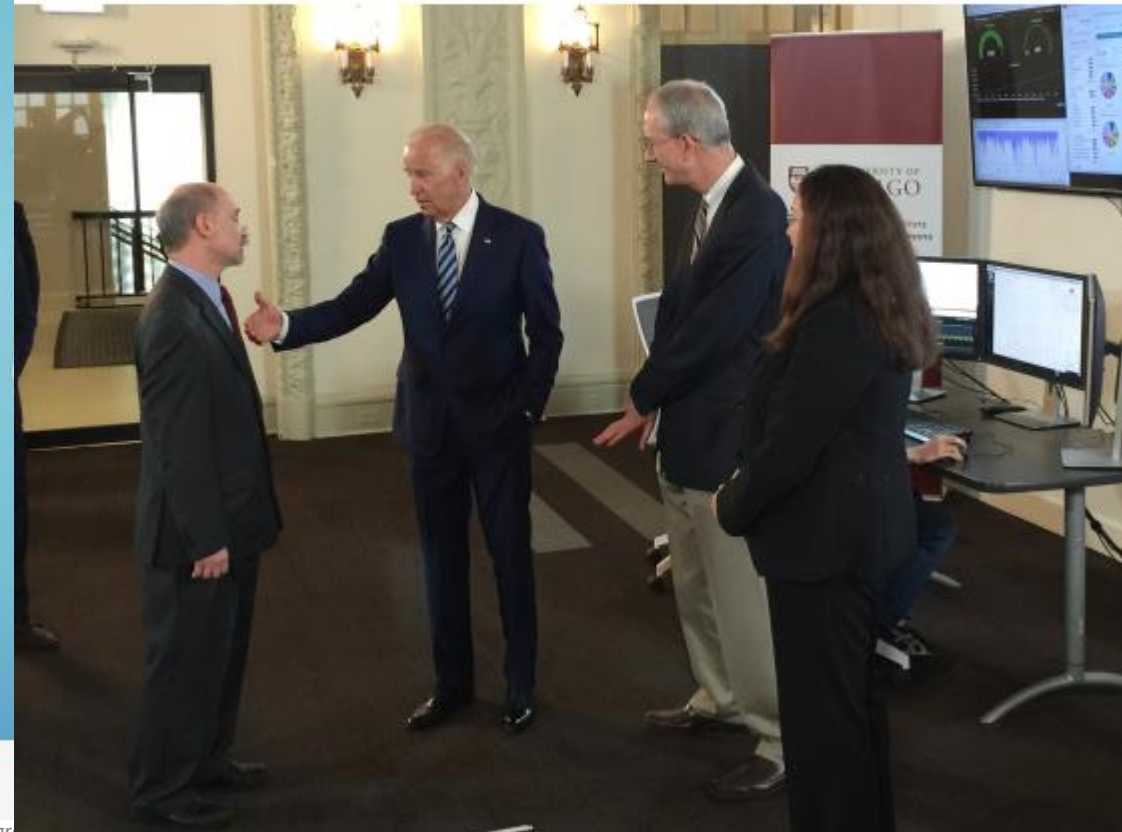
21,773

MUTATIONS

2,730,388

I can't tell you how excited I am about this

— Current President of the United States, Joseph R. Biden



and innovation programme under grant agreements 823852 and 857641, respectively under grant ag



# Example of domain specific open data publishing:

## Human Organ Atlas

\*Walsh, C.L., \* Tafforeau, P., \* Wagner, W.L., Jafree, D.J., Bellier, A., Werlein, C., Kühnel, M.P., Boller, E., Walker-Samuel, S., Robertus, J-L., Long, D.A., Jacob, J., Marussi, S., Brown, E., Holroyd, N., Jonigk#, D.D., Ackermann#, M., Lee#, P.D. **Imaging intact human organs locally resolving cellular structures using hierarchical phase- contrast tomography.** Nat Methods (2021) Accepted

*refer to PaNOSC Use Case 23 for more info*

### Welcome to the Human Organ Atlas

The Human Organ Atlas uses **Hierarchical Phase-Contrast Tomography** (HiP-CT) to span a previously poorly explored scale in our understanding of human anatomy, the micron to whole intact organ scale. Histology using optical and electron microscopy images cells and other structures with sub-micron accuracy but only on small biopsies of tissue from an organ, while clinical CT and MRI scans can image whole organs, but with a resolution only down to just below a millimetre. HiP-CT bridges these scales in 3D, imaging intact organs with ca. 20 micron voxels, and locally down to microns. We hope this open access Atlas, enabled by the ESRF-EBS, will act as a reference to provide new insights into our biological makeup in health and disease.

This project has been made possible by funding from:

- The [European Synchrotron Radiation Facility \(ESRF\)](#) — funding proposal MD-1252
- The [Chan Zuckerberg Initiative](#), a donor-advised fund of the Silicon Valley Community Foundation
- The [German Registry of COVID-19 Autopsies](#) (DeRegCOVID), supported by the German Federal Ministry of Health
- The Royal Academy of Engineering, UK
- The UK Medical Research Council
- The Wellcome Trust

#### Collaborators

- [UCL](#), London, England: **Peter D Lee, Claire Walsh, Simon Walker-Samuel, Rebecca Shipley, Sebastian Marussi, Joseph Jacob, David Long, Daniyal Jafree, Ryo Torii, Charlotte Hagen**
- [ESRF](#), Grenoble, France: **Paul Tafforeau, Elodie Boller**
- Medizinische Hochschule Hannover, Germany: **Danny D Jonigk, Christopher Werlein, Mark Kuehnel**
- Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Germany: **M Ackermann**
- University Hospital of Heidelberg, Germany: **Willi Wagner**
- Grenoble Alpes University, Department of Anatomy, French National Center for Scientific Research: **A Bellier**
- [Diamond Light Source](#), Harwell, UK: **Andy Bodey, Robert C Atwood**
- Imperial College London, UK: **JL Robertus**



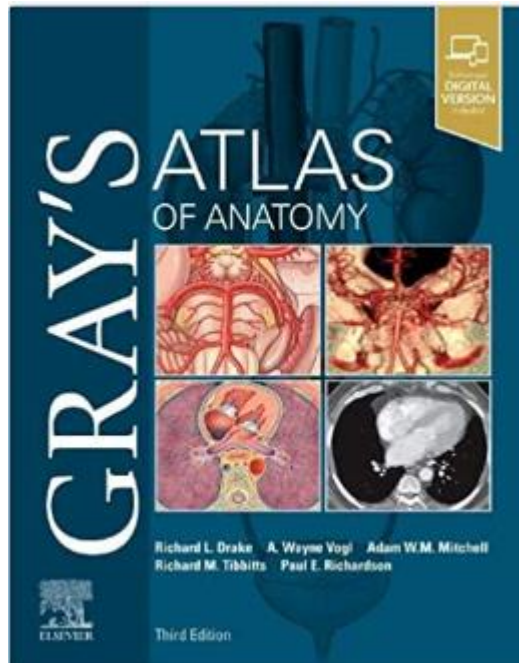
#### Aknowledgements

The development of this portal has been done as part of the [PaNOSC project](#). PaNOSC has received funding from the European Union's [Horizon 2020](#) research and innovation programme under grant agreement No. 823852. The following people were involved in the development: Paul Tafforeau, Alejandro De Maria Antolinos, Axel Bocciarelli, Marjolaine Bodin and Andrew Götz from the ESRF, Jiří Majer from ELI, as well as the broader PaNOSC and ICAT communities.

<https://human-organ-atlas.esrf.eu>

# Human Organ Atlas FAIR data reuse

Examples of domain specific open data reuse by commercial companies:



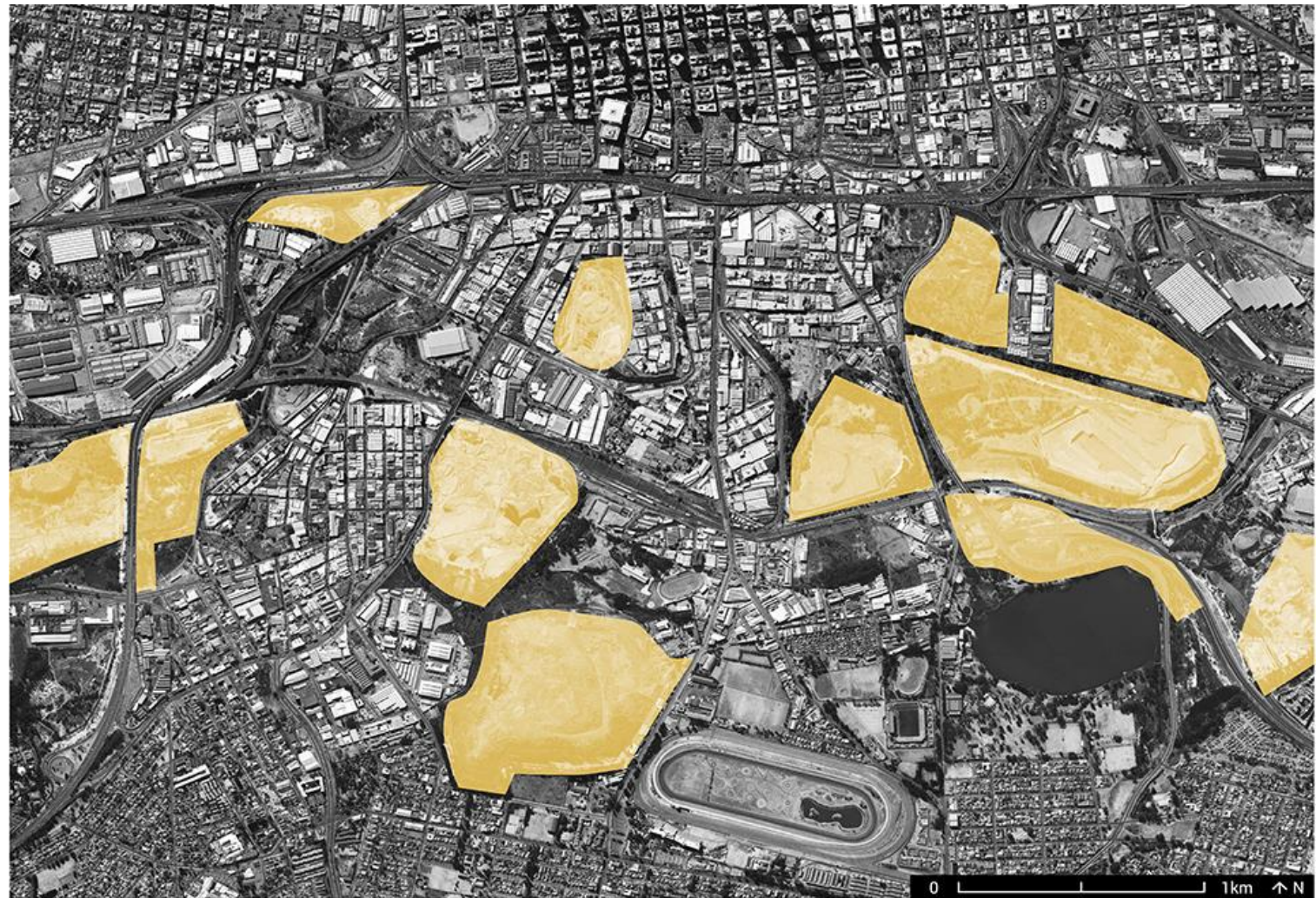
[https://www.linkedin.com/posts/isabellewegmannhachette\\_sciences-3d-visualisation-activity-6975790784402837504-ym8z](https://www.linkedin.com/posts/isabellewegmannhachette_sciences-3d-visualisation-activity-6975790784402837504-ym8z)

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.



# Why make a Data Commons?

Data is sitting around  
like gold mine dumps ...



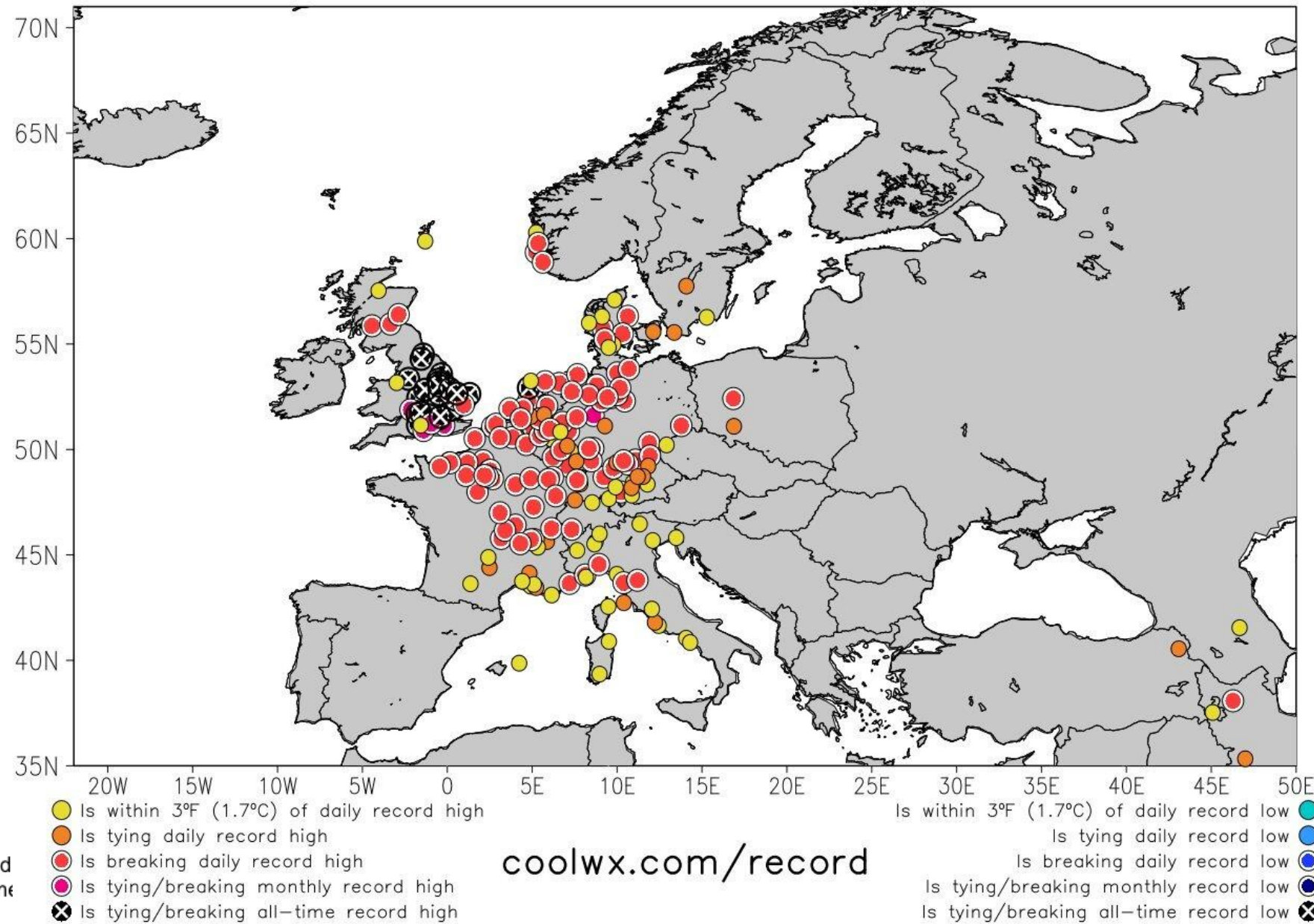
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.





Locations approaching or surpassing unofficial  
daily (19 Jul) temperature records based on temperature  
at 1300 UTC 19 Jul 2022

Climate  
change is  
more and  
more visible ...



PaNOSC and ExPaNDS projects have received  
and innovation programme under grant agreem

# Energy crisis is hitting our labs ...

SCIENCE|BUSINESS®

## Energy crisis is starting to hit Europe's big science labs

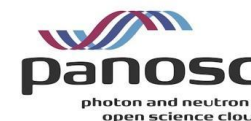
20 Sep 2022 | News

*Research infrastructures are worried about the rising cost of running large scientific experiments and are looking for help with paying sky-high electricity bills. One lab has seen a 60% increase in its tariff this year*

By Florin Zubaşcu



ALBA synchrotron. Photo: [albasynchrotron / Facebook](#)



PaNOSC and ExF <https://sciencebusiness.net/news/energy-crisis-starting-hit-europes-big-science-labs>  
and innovation programme under grant agreements 823852 and 857641, respectively.



# Calculating the carbon footprint of data

- **User Travel** - 3 users fly from Copenhagen to ESRF (380+10 kg CO<sub>2</sub>e) = 3 x 390 kg
- **Beamtime energy consumption** – 1 week of beamtime (8MW/42) = 190 kWh
- **Data stored on disk** – 100 GB stored on disk (10W x 100 days)
- **Data processing on site** – 1 week of processing on 64 cores (1kW x 1 week)
- **Data transfer** – transfer 100 GB of data back to user (31 kWh)

**CO<sub>2</sub>e per kWh in France** (2022) = 75 g/kWh





# Estimated carbon footprint of experiment

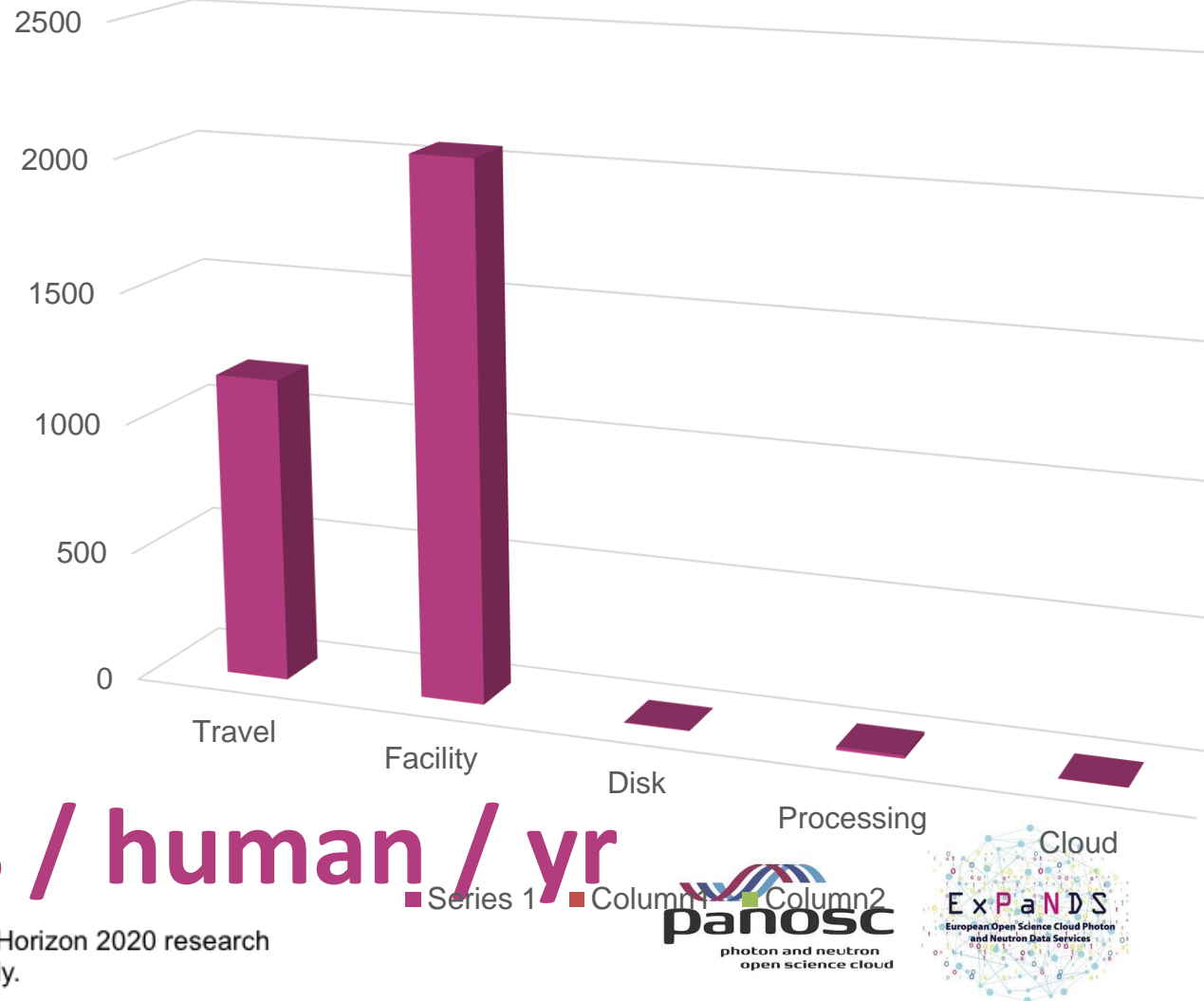
- User Travel = **1170 kg**
- Beamtime energy consumption = **2056 kg**
- Data stored on disk = **1.8 kg**
- Data processing on site = **12.6 kg**
- Cloud transfer = **2.3 kg**

CO<sub>2</sub>e per kWh in France = **75 g/kWh**

**TOTAL = 3.253 tons !**

**Sustainable Goal = 5 tons / human / yr**

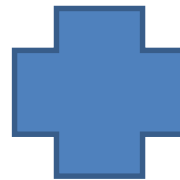
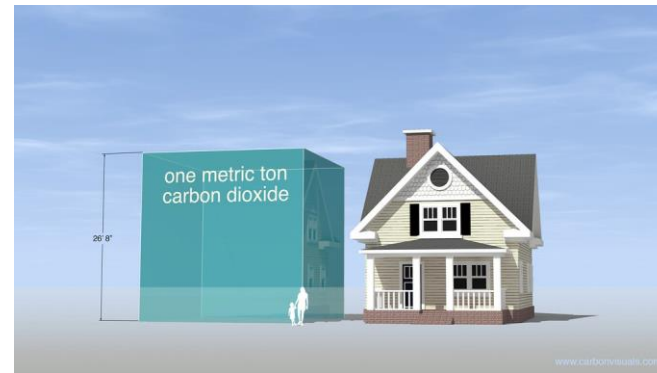
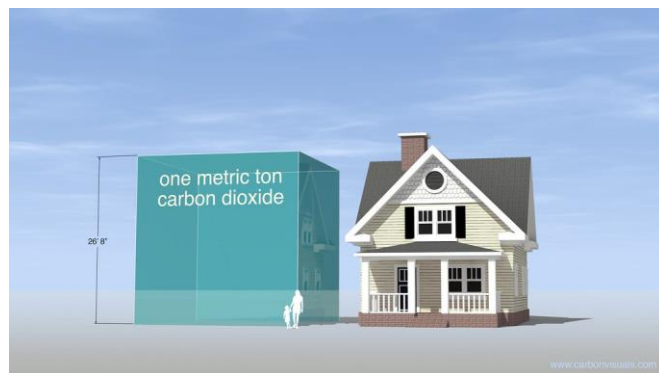
Carbon footprint for 1 week experiment @ ESRF



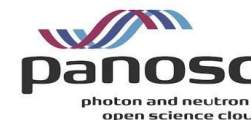
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.



# 1 week of experiment is equivalent to a cube 30x30x30 metres of CO<sub>2</sub>



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.



# Carbon footprint of archiving data

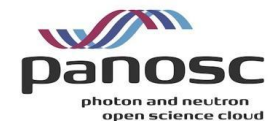
- Data stored on tape for 10 years  $\sim 200 \text{ g} * 35 = 7 \text{ kg}$

CO<sub>2</sub>e per kWh in France = 75 g/kWh

**ARCHIVING for 10 years  $\sim 7 \text{ kgs}$   
i.e. 0.2% of the raw data!**



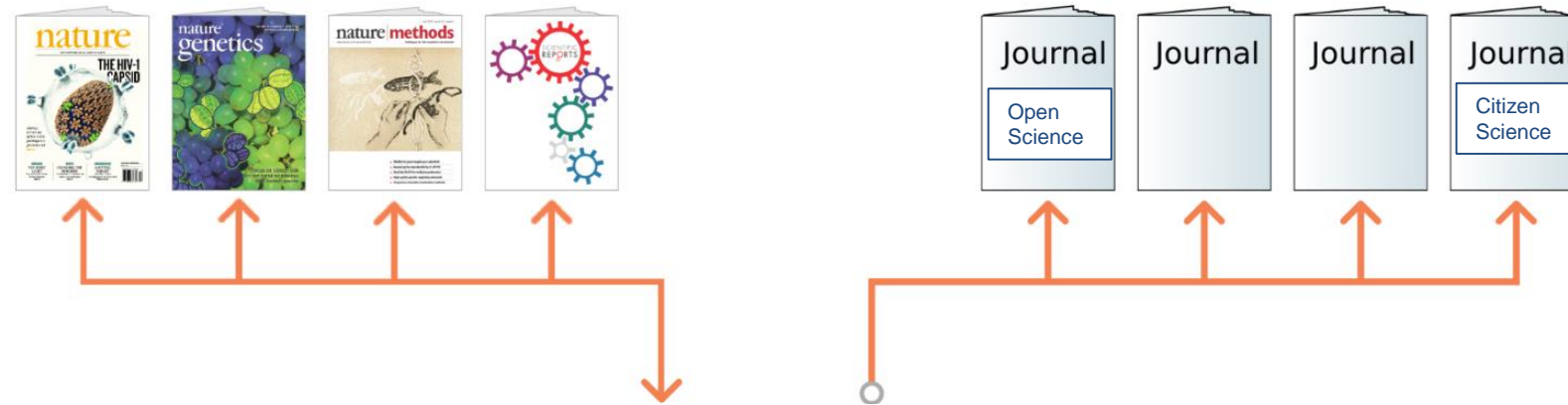
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.



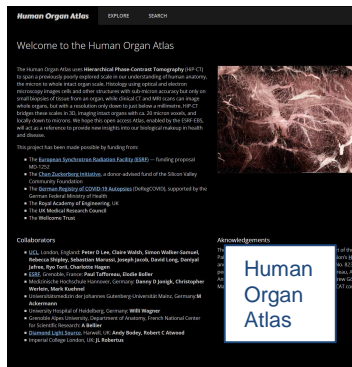
# PaN Data Commons - Concept

- **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 + Improved 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** and **climate change**)

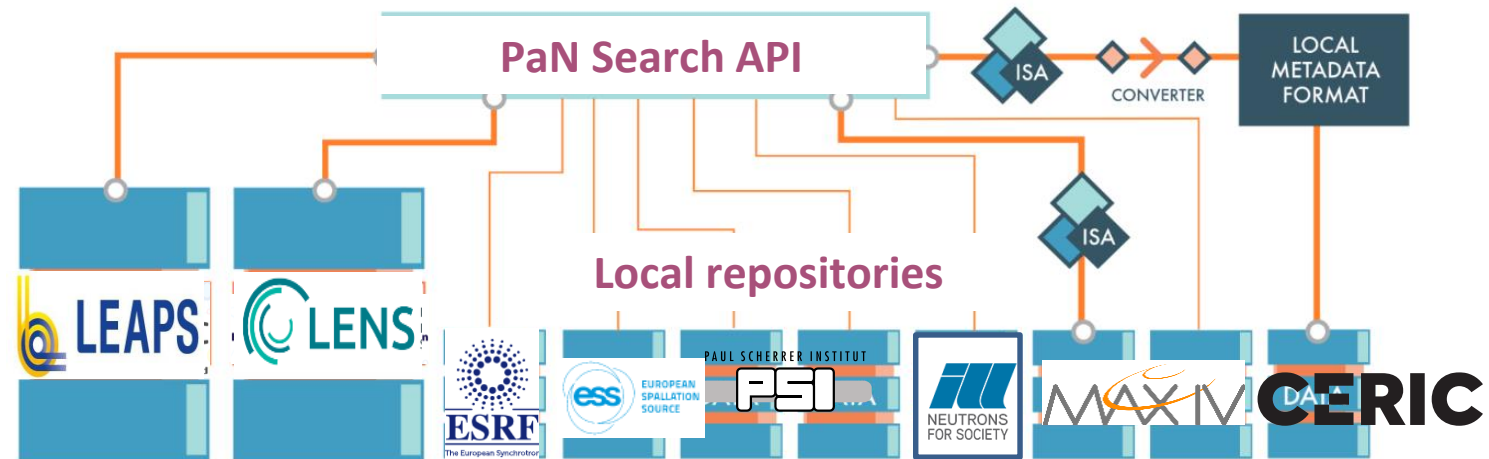
# Sustain Published Data ➡ Through PaN Repositories



## PaN Open Data Repository



PaN Open Data Portal



# Building the PaN Data Commons on the PaNOSC search portal



## European Photon and Neutron Open Data Search Portal

Type a query to search for open data from photon and neutron sources:

... or try one of these queries: *diffraction, lung*

The European Photon and Neutron sources are working together in the [PaNOSC](#) and [ExPaNDS](#) projects financed by the European Commission to build the **European Open Science Cloud**. One of the main objectives of the EOSC is to make **Open Data** from these facilities FAIR. This portal implements the F(indable) part of FAIR via a **federated search engine** from the following facilities:

- European Synchrotron Radiation Facility
- European Spallation Source
- MAX IV
- Paul Scherrer Institut
- Central European Research Infrastructure Consortium

Additional facilities will be included in the federated search as their search engines come online locally. The goal is to include all photon and neutron facilities who provide open data by the end of the two projects PaNOSC and ExPaNDS.

The mission of the PaN data search portal is to contribute to the realization of a data commons for Neutron and Photon science. The search results provide a link to the landing page of the data DOIs through which the other data services provided by PaNOSC and ExPaNDS for data downloading, analysis, notebooks and simulation can be accessed. The aim of the portal is to facilitate using data from photon and neutron sources for the many



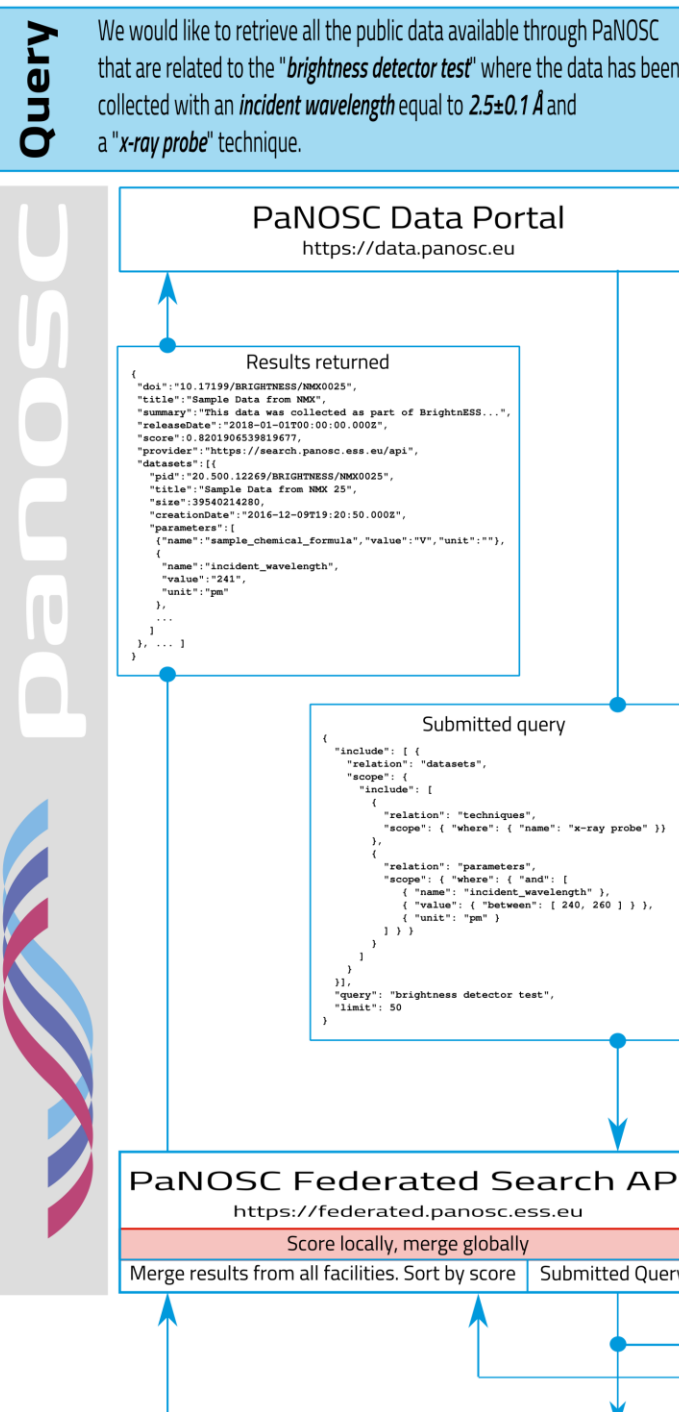
PaNOSC a  
and innova





# PaNOSC Search API portal @ <https://data.panosc.eu> DEMO!

Source: “**SciCat implementation of the PaNOSC Search**” by **Max Novelli** (ESS) et al,  
<https://indico.psi.ch/event/12738/contributions/38937/>



**panosc**

European Photon and Neutron Open Data Search Portal

Type a query to search for open data from photon and neutron sources:

brightness detector test

1 Submit the query alone: *brightness detector test*

... or try one of these queries: diffraction, lung

**panosc** brightness detector test 50+ documents found

Type

proposal  
publication

Technique

Select a technique...

Incident Wavelength

min max nm

Incident Photon Energy

min max eV

Chemical Formula

Temperature

min max K

Pressure

min max Pa

10.17199/BRIGHTNESS/MB0039  
Sample Data from multiblade  
This data was collected as part of BrightnESS, funded by the European Union Framework Programme for Research and Innovation Horizon 2020, under grant agreement 676548. It consists of test data f...  
Released on January 1st 2018 by the ESS

10.17199/BRIGHTNESS/MB0034  
Sample Data from multiblade  
This data was collected as part of BrightnESS, funded by the European Union Framework Programme for Research and Innovation Horizon 2020, under grant agreement 676548. It consists of test data f...  
Released on January 1st 2018 by the ESS

10.17199/BRIGHTNESS/MB00  
Sample Data from multibl  
This data was collected as part of BrightnESS, funded by the European Union Framework Programme for Research and Innovation Horizon 2020, under grant agreement 676548. It consists of test data f...

2 Review results and refine search

**panosc** brightness detector test 4 documents found

Type

proposal  
publication

Technique

x-ray probe

Incident Wavelength

240 260 pm

Incident Photon Energy

min max eV

Chemical Formula

Temperature

min max K

Pressure

min max Pa

10.17199/BRIGHTNESS/NMX0025  
Sample Data from NMX  
This data was collected as part of BrightnESS, funded by the European Union Framework Programme for Research and Innovation Horizon 2020, under grant agreement 676548. It consists of test data f...  
Released on January 1st 2018 by the ESS

10.17199/BRIGHTNESS/NMX0024  
Sample Data from NMX  
This data was collected as part of BrightnESS, funded by the European Union Framework Programme for Research and Innovation Horizon 2020, under grant agreement 676548. It consists of test data f...  
Released on January 1st 2018 by the ESS

10.17199/BRIGHTNESS/NMX0  
Sample Data from NMX  
This data was collected as part of BrightnESS, funded by the European Union Framework Programme for Research and Innovation Horizon 2020, under grant agreement 676548. It consists of test data f...

3 Apply additional filters:

- incident wavelength between 2.40Å (240pm) and 2.60Å (260pm)
- technique is x-ray probe or any of its siblings

Other facilities with other data catalogues

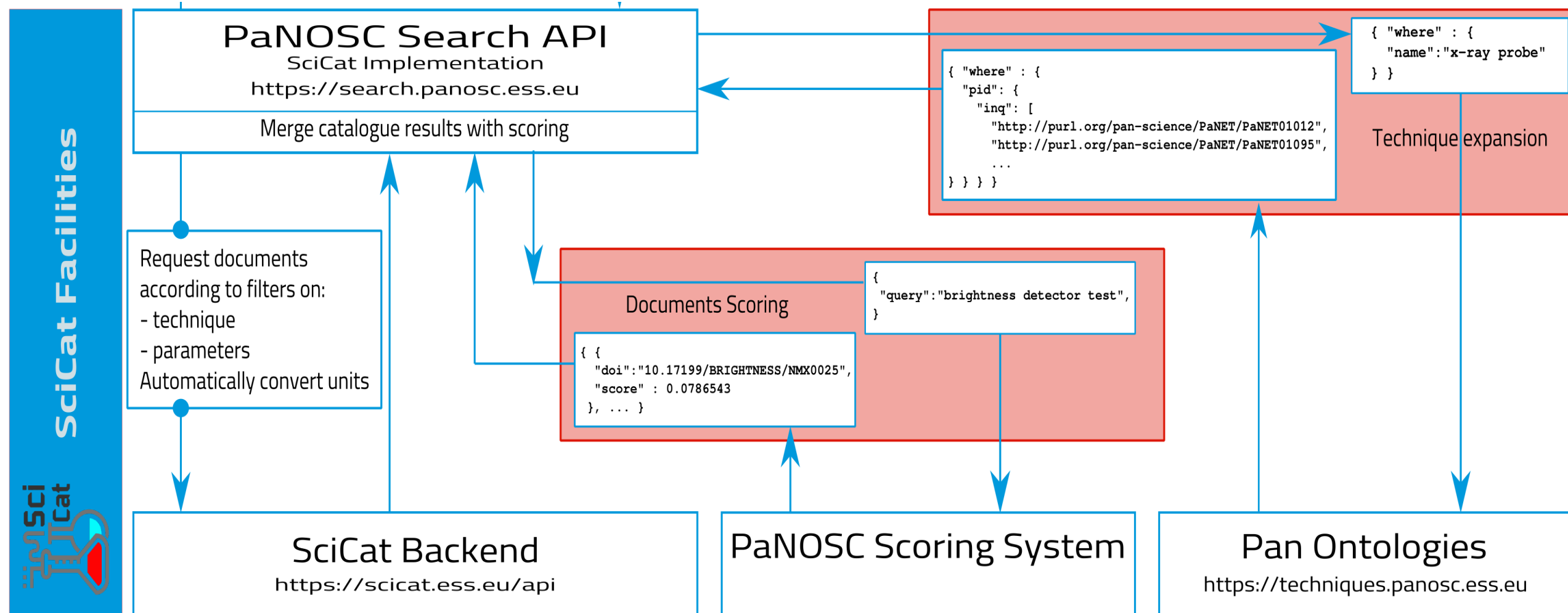
ICAT

open science cloud



PaNOSC and ExPaNDS projects have received funding from the European Union Horizon 2020 research and innovation programme under grant agreements 82381

# PaNOSC Search API – implemented on top of SciCat



Source: “**SciCat implementation of the PaNOSC Search**” by **Max Novelli** (ESS) et al, <https://indico.psi.ch/event/12738/contributions/38937/>



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.

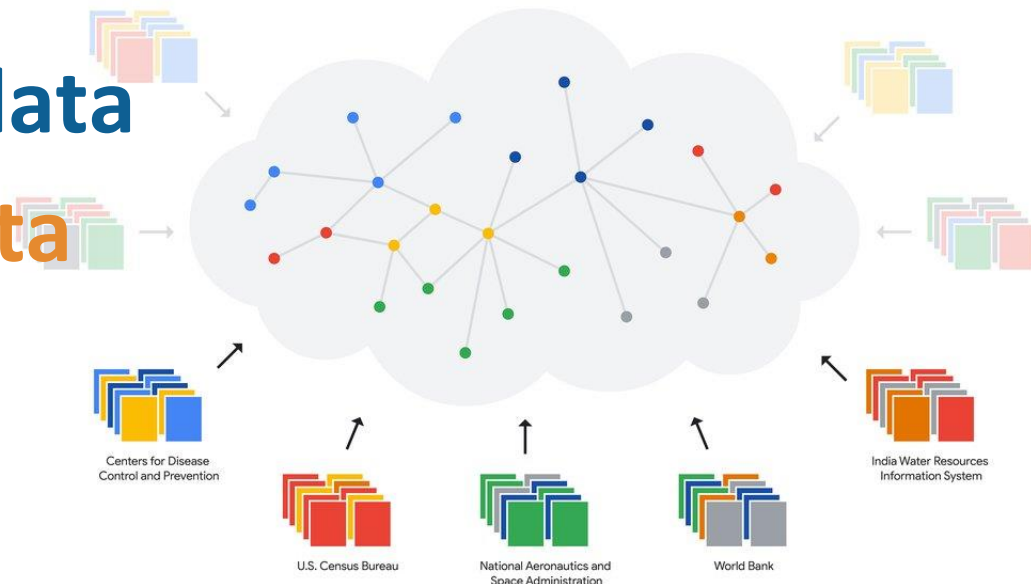


# What we need from YOU:

1. **Implement a open data repository**
2. **Deploy the PaNOSC Search API + scoring**
3. **Connect your search endpoint to the PaN Search Portal**
4. **Data stewards to curate metadata**
5. **Train your scientists in FAIR data**
6. **Help build a knowledge graph**



Data Commons Knowledge Graph



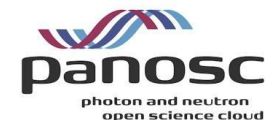
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

# Conclusion

1. ExPaNDS and PaNOSC have laid the foundations for a PaN Data Commons
2. The outcomes of the two projects will enable a PaN Data Commons of FAIR data
3. A PaN Data Commons will preserve and increase data reuse
4. Finance to sustain a Data Commons will come from facilities + EOSC
5. The PaN community is on the road to becoming part of the FAIR data landscape
6. Saving our data helps fight climate change and supports open science



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.

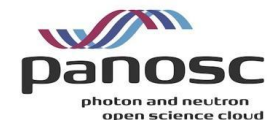


# Useful links

- ExPaNDS - <https://expands.eu/>
- PaNOSC - <https://www.panosc.eu/>
- EOSC Association - <https://eosc.eu/>
- PaNOSC data portal – <https://data.panosc.eu>



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.





# Sources used for carbon footprint estimates

- **User Travel** - <https://calculator.carbonfootprint.com/calculator.aspx?tab=3>
- **Beamtime energy consumption** – ESRF electrical monitor + control system
- **Data stored on disk** – <https://www.buildcomputers.net/power-consumption-of-pc-components.html>
- **Data processing on site** – <https://www.buildcomputers.net/power-consumption-of-pc-components.html>
- **Data transfer+storage in cloud** – <https://medium.com/stanford-magazine/carbon-and-the-cloud-d6f481b79dfe>
- **Tape storage** - <https://datastorage-na.fujifilm.com/reducing-carbon-emissions-through-the-data-tape-ecosystem/>
- **CO2 by kWh in France** - <https://www.rte-france.com/eco2mix/les-emissions-de-co2-par-kwh-produit-en-france#>

