

“Grand challenges in AI and data science”
EIROForum Workshop
Heidelberg (Germany)
27 April 2022

Data at ESRF

Andy Götz¹

¹ ESRF, 71 avenue des Martyrs, 38000 Grenoble (France)



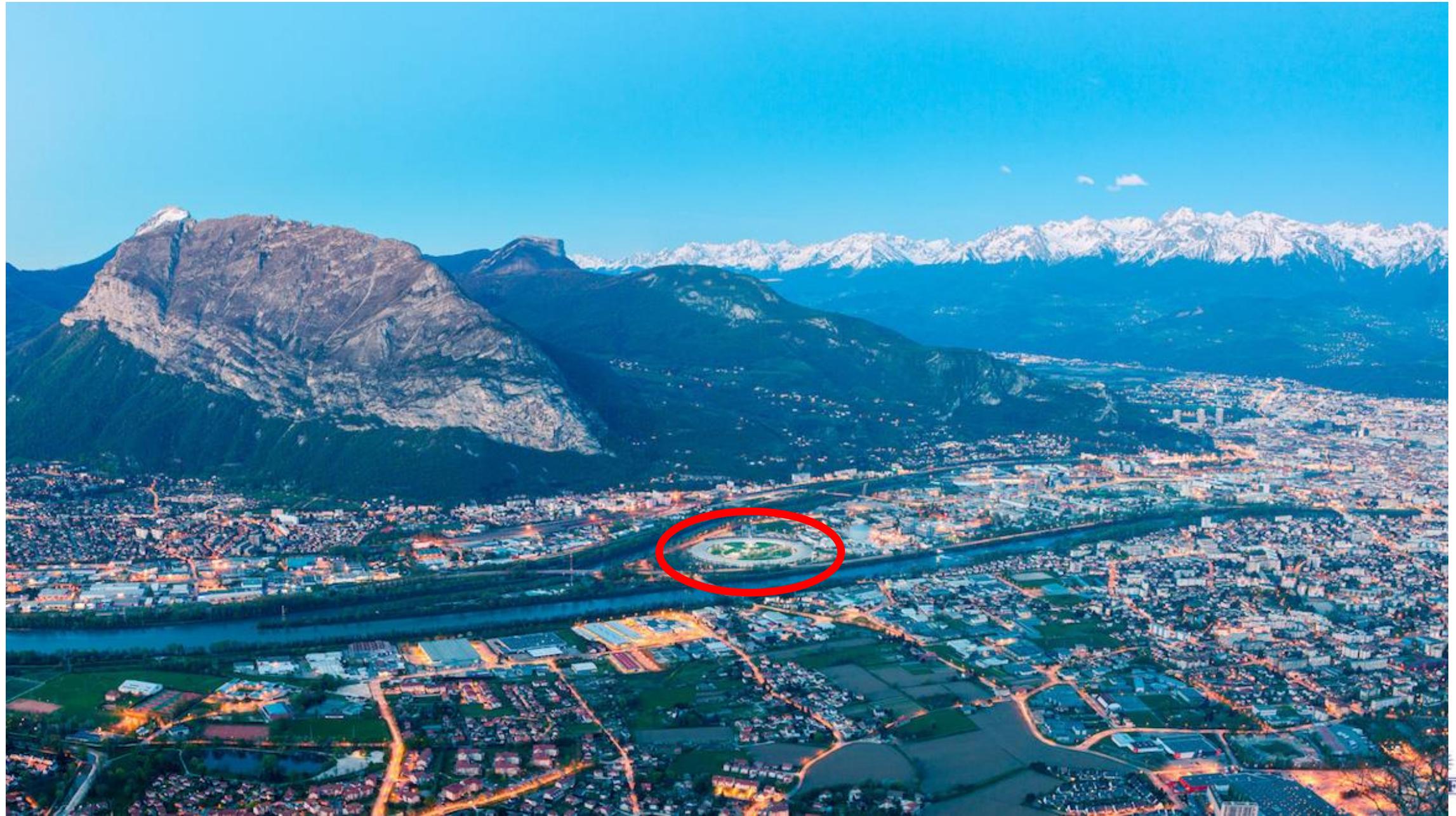
PIONEERING SYNCHROTRON SCIENCE



TALK LAYOUT – ANSWER THESE QUESTIONS

1. *Background on organisation structure and mission*
2. *History of open data*
3. *What do these data look like, types? Scale, formats, heterogeneity - pre-survey?*
4. *Who are the users (e.g. data generators, reusers, unanticipated users)?*
5. *Showcase a best example of data use*
6. *Tools that are needed to use the data -*
7. *At what level of use (raw data → accessible visualisation)*
8. *What do you need to do to make the data available to a variety of different audiences*
9. *How do they find the data? F of fair.*
10. *How to share [really] big data for analysis, and pitfalls*
11. *Impact of policies (open science - funders, journals, our own ..) DMPs?*
12. *Community-organised metadata data standards - how to do it, successes, challenges*
13. *Link with publications? Other research outputs? Demonstration of value of open data? Reuse examples (as well as primary use case) COVID as an example?*
14. *Showing impact e.g. tracing use, measuring use and outcomes*
15. *What do we do (and mean by) for citizen science (definition: scientists asking questions and providing infrastructure and engaging the public to contribute to prescribed tasks, the data from which is then analysed again by the scientists), public engagement and broader comms*
16. *How are data being used by AI / planned to be used*
17. *Strengths and areas for learning*

ESRF – THE EUROPEAN SYNCHROTRON



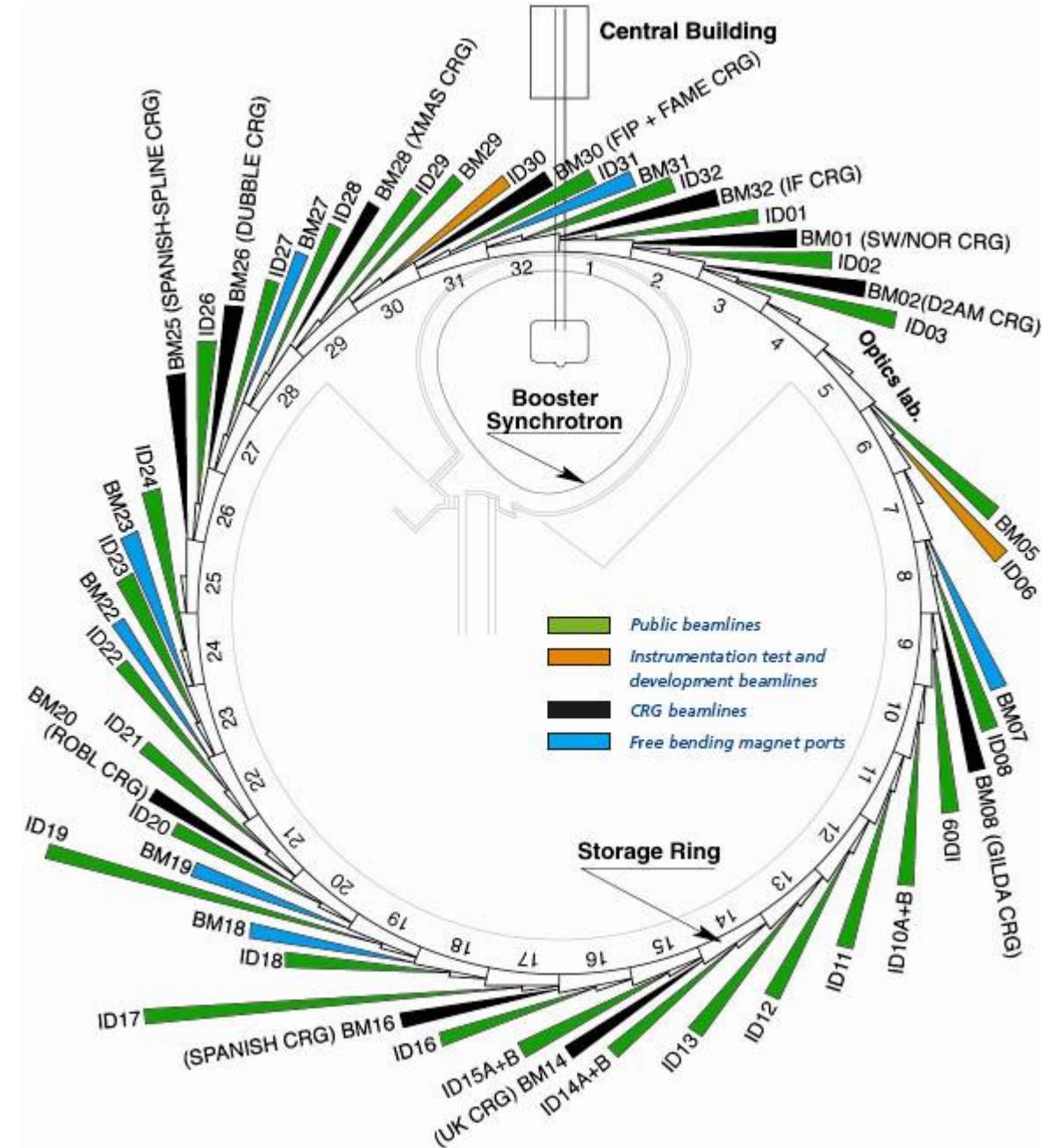
BACKGROUND ON ORGANISATION STRUCTURE AND MISSION



BACKGROUND ON ORGANISATION STRUCTURE AND MISSION

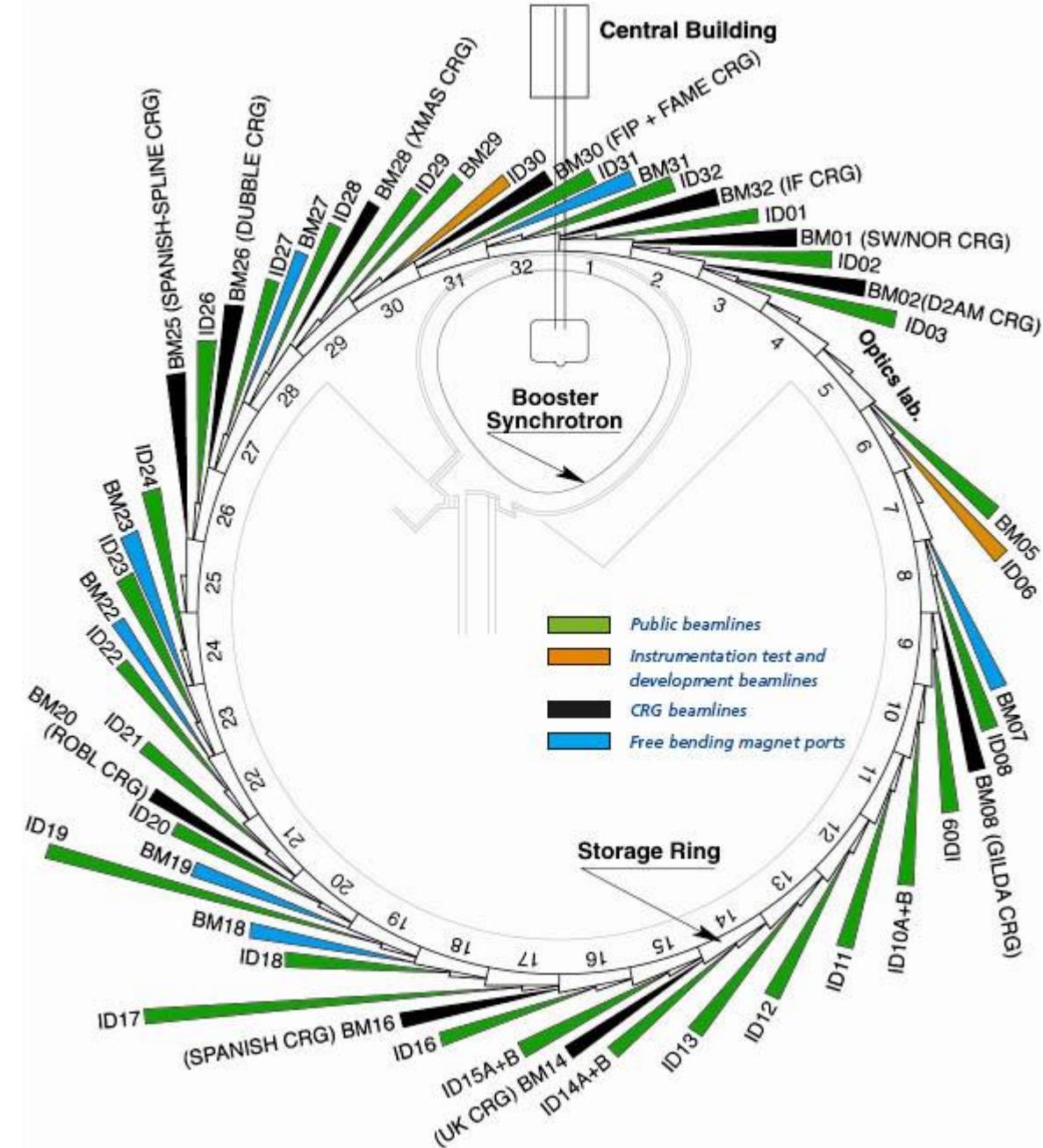
Mission

1. Produce synchrotron radiation in the hard x-ray region (10 keV – 250 keV) for doing experiments on applied science.
2. Provide visiting scientists with a hardware and software support for running experiments (free of charge for users for peer-reviewed experiments).
3. Provide users with the data from their experiments and support on how to process them.
4. Make data open and FAIR for at least 10 years

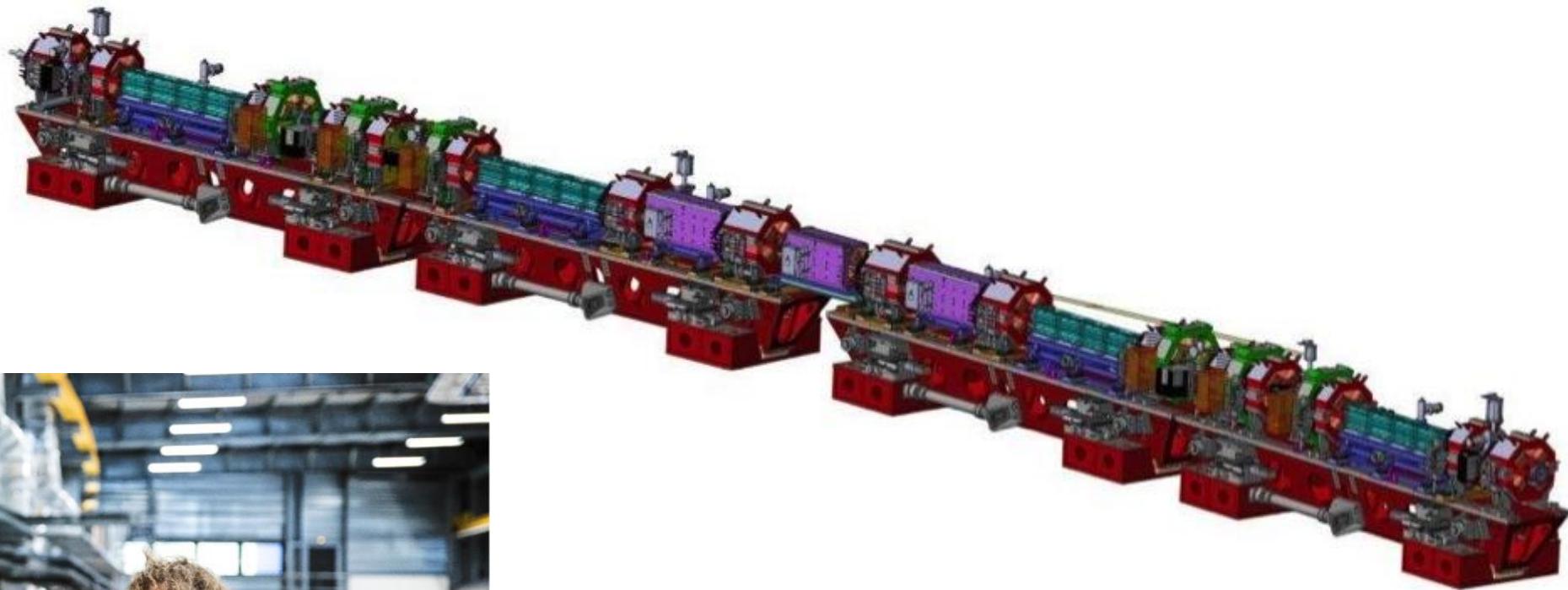


Experiment Categories

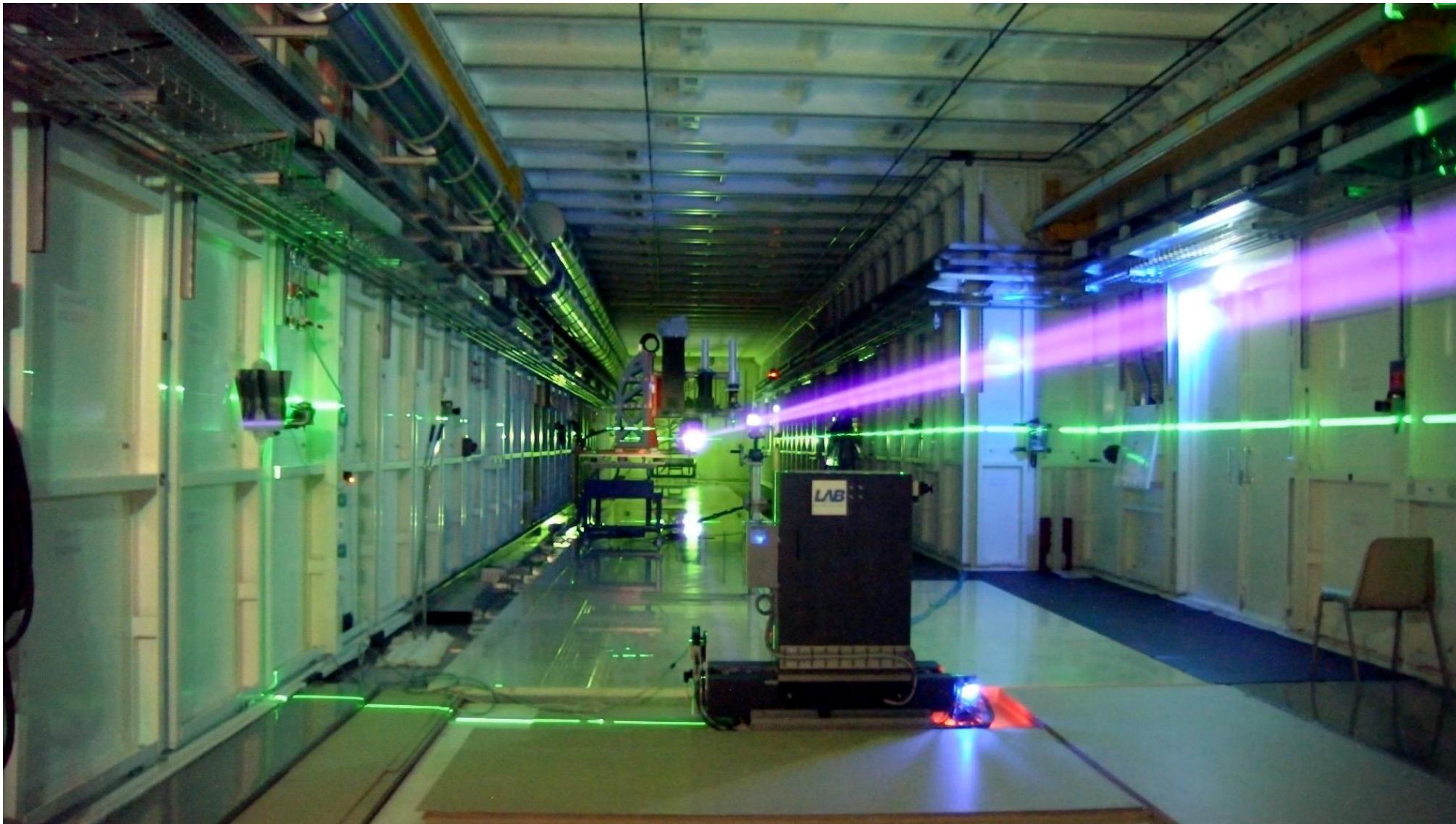
1. CH (Chemistry)
2. ES (Earth Science)
3. EV (Environment)
4. HC (Hard Condensed Matter Science)
5. HG (Cultural Heritage)
6. LS (Life Sciences)
7. MA (Applied Material Science)
8. MD (Medicine)
9. ME (Engineering)
10. MI (Methods and Instrumentation)
11. MX (Structural Biology) –
12. SC (Soft Condensed Matter Science)



NEW SOURCE EBS = 100 X MORE COHERENT BEAMS



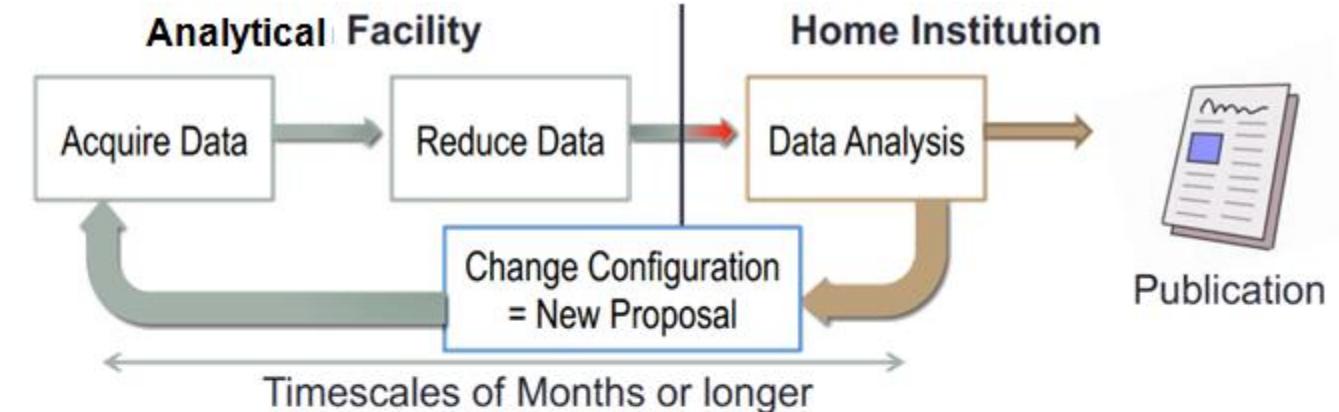
NEW SOURCE BM18 = 100 X MORE COHERENT BEAMS



HISTORY OF OPEN DATA

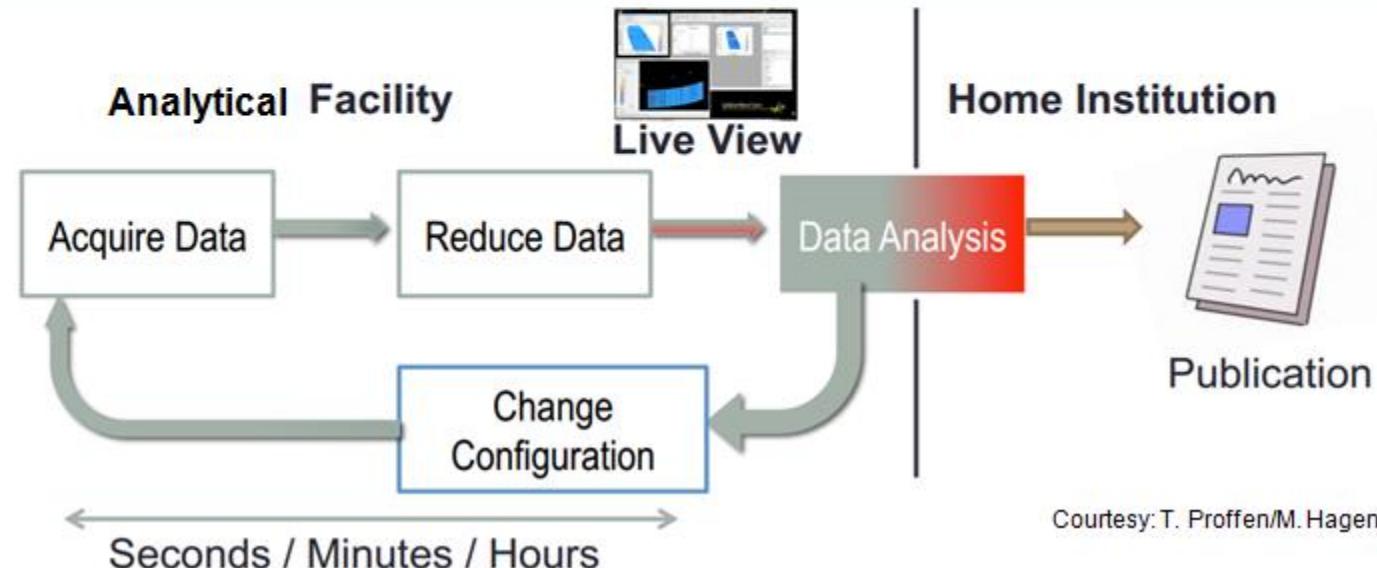
Key Dates

1. **1992 – 2015** : data were acquired, processed, exported and then deleted.
2. **2015** : ESRF Data Policy was adopted by Council.
3. **2015 – now** : data are acquired, curated, processed, exported and **archived and open**



ESRF Open Data Policy

1. Data are archived for at least 10 years
2. Metadata are archived forever
3. Data are under embargo for 3 years after which they are made open
4. All data are referenced by a DOI
5. Users can create bespoke DOIs



Courtesy: T. Proffen/M. Hagen

HISTORY OF OPEN DATA – CONTRIBUTION OF EU PROJECTS

Common data policy

FAIR data policy

Data Management Plans

Software Catalogue

Remote analysis

Jupyter

UmbrellaID

AARC Blueprint

eduTeams

e-neutron

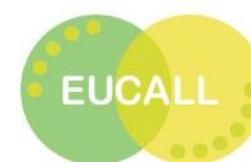
Training platform

2010

2015

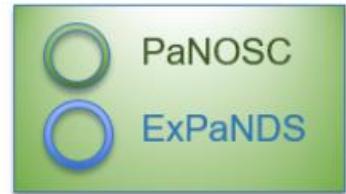
2018

2021

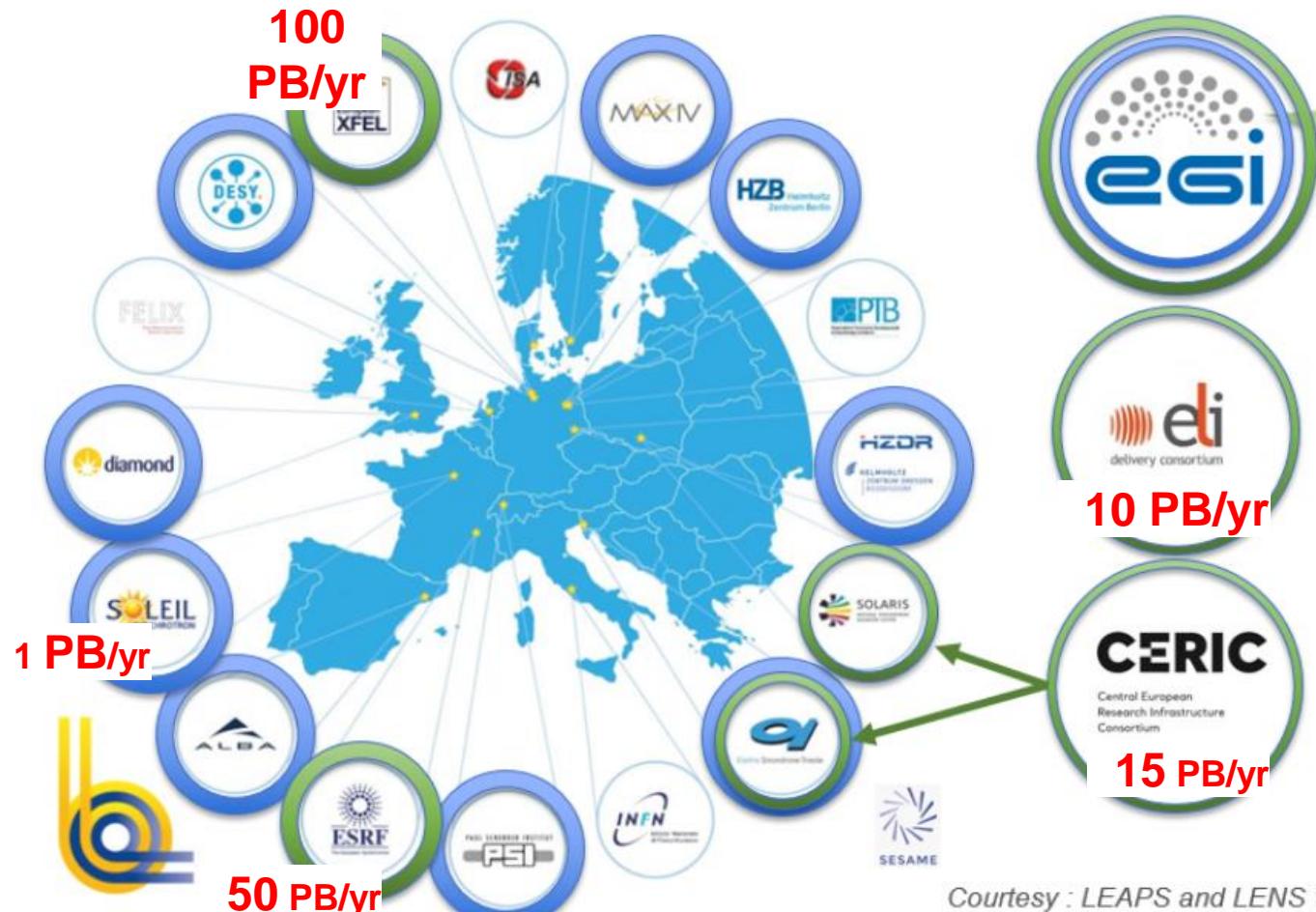


Slide courtesy of Patrick Fuhrman (DESY)

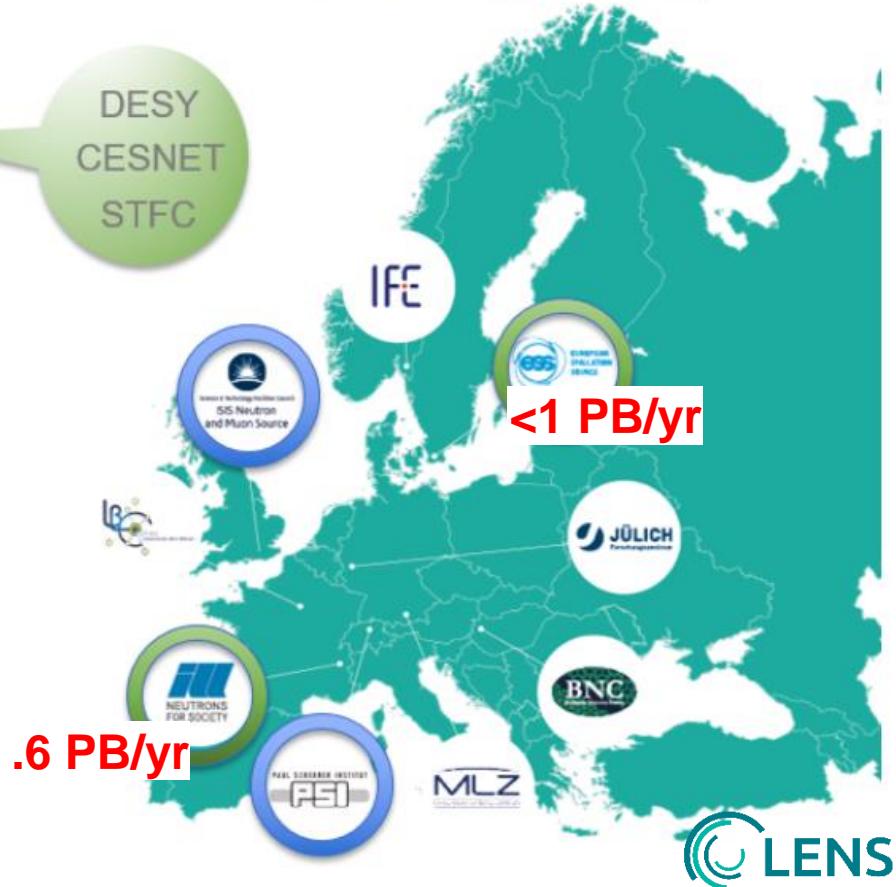
League of Photon Sources (LEAPS) and Neutrons (LENS) partners in PaNOSC and ExPaNDS



Photon (LEAPS)



Neutron (LENS)



Courtesy : LEAPS and LENS Web Pages

Slide courtesy of Patrick Fuhrman (DESY)



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.

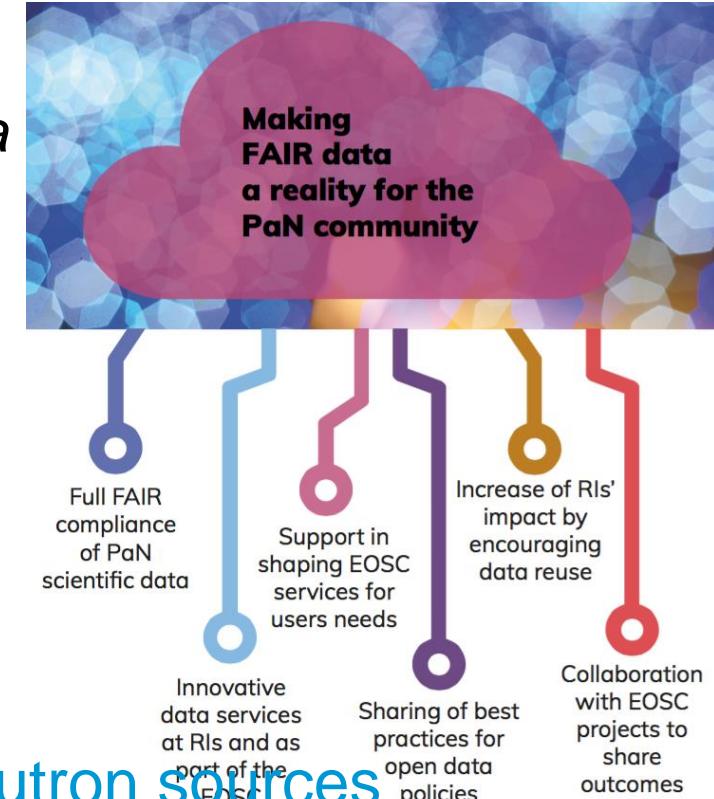
PaNOSC
photon and neutron
open science cloud



EOSC IS BOOSTING FAIR DATA AT ESRF

PaNOSC - ESRF coordinates the Photon and Neutron Open Science Cloud (PaNOSC), Key Outcomes are:

- *Open data policy framework for raw + processed data*
- *Open data + standardised metadata (Nexus)*
- *Federated search API + data portal*
- *Remote desktop cloud platform (VISA)*
- *Jupyter notebook service*
- *Simulation framework (ViNYL)*
- *Community AAI (*Umbrellald*) on eduTEAMS*
- *Training + e-learning platform*



ExPaNDS is a similar project for national Photon and Neutron sources

- *ExPaNDS is collaborating with PaNOSC to adopt the PaNOSC outcomes*

European Open Science Cloud impact

- *EOSC has boosted the adoption and implementation of the FAIR principles*



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

The European Synchrotron



WHAT DO ESRF DATA LOOK LIKE ?

Data types

- Preferred format is HDF5 a hierarchical binary format for storing all data and metadata. HDF5 is used for archiving raw and processed data. We have developed tools for browsing, viewing and accessing HDF5 files.
- Additional formats are used for analysis programs e.g. tif, cif, ascii, ...

Raw Data → Covered by Data Policy

- 2D images from detectors (cameras) from 1 megapixel to 64 megapixels
- 2D movies of particles (cryo-electron microscopy)
- 1D and 0D arrays (spectroscopy)

Processed Data → Covered by Data Policy on ‘best effort’

- 3D volumes representing models of the sample
- 3D models of electron distribution of proteins
- 2D movies of samples reactions to changes
- 2D maps of elemental distributions in samples
- 1D plots of diffraction images / spectroscopy

TOTAL CURATED DATA PRODUCTION SINCE 2015

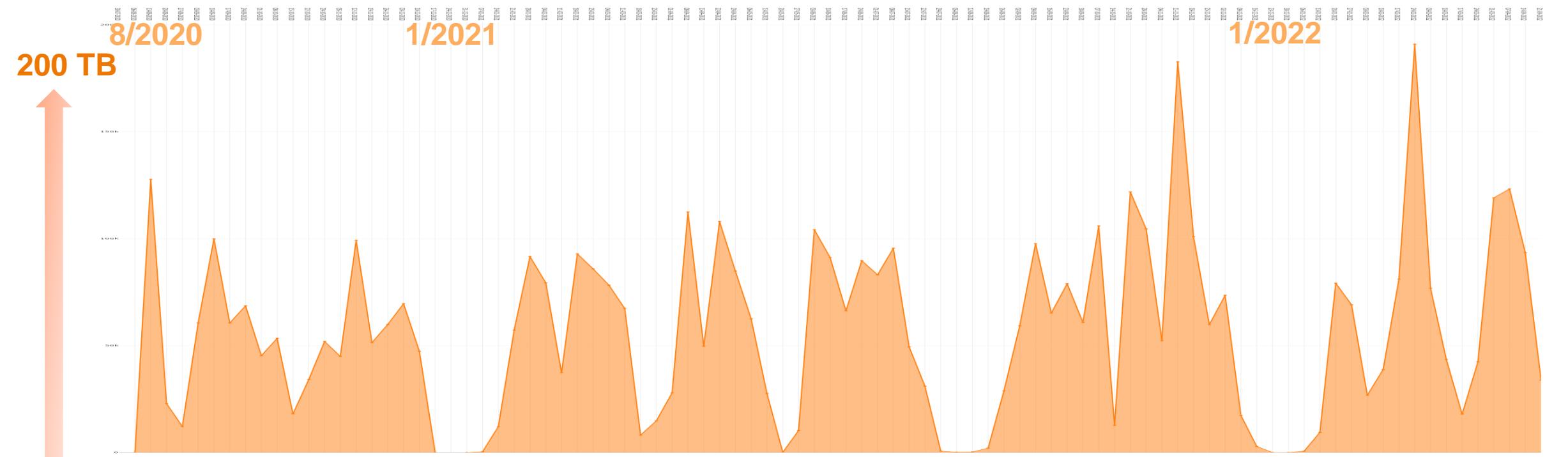
Summary

Datasets	1250310
Beamlines	47
Total Volume	7.0 PB
Total Number of files	484124421

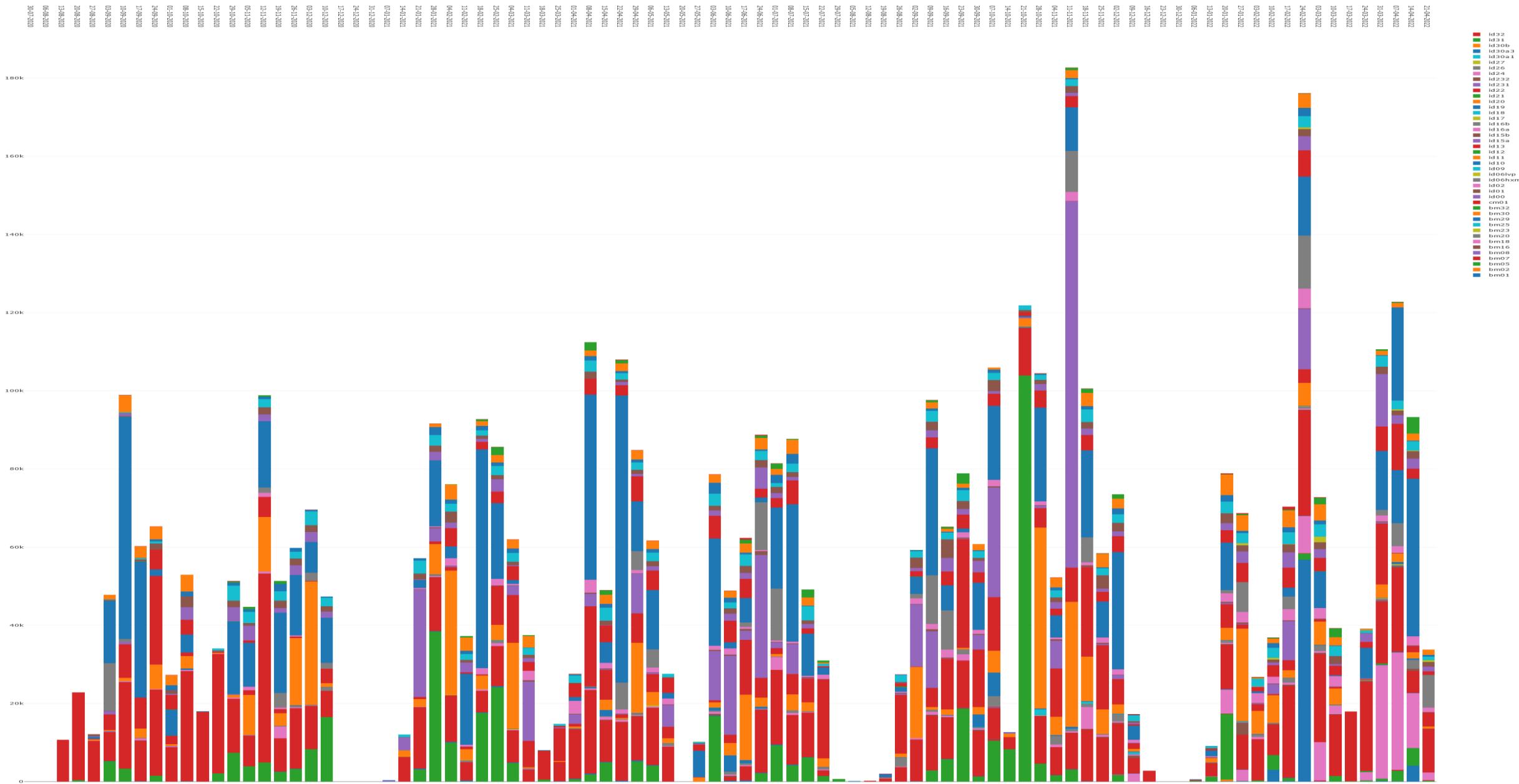
Dataset

Average file count	387
Max files	200002
Average volume	5.9 GB
Max volume	8.1 TB
Average metadata	25.2

DATA CURATED WEEKLY SINCE 1/8/2020 – PEAK = 200 TB



DATA PER BEAMLINE SINCE 1/8/2020



WHO ARE THE USERS OF ESRF DATA AND SERVICES?

ESRF USERS

- Experimental team who generated the data profit most from data services:
 1. *Rich metadata collected automatically + curated*
 2. *Raw data curated for (at least) 10 years*
 3. *Exclusive access for (at least) 3 years*
 4. *Efficient download of large volumes*
 5. *DOI for raw and processed data*
 6. *Searchable electronic logbook*
 7. *Data searching + viewing*

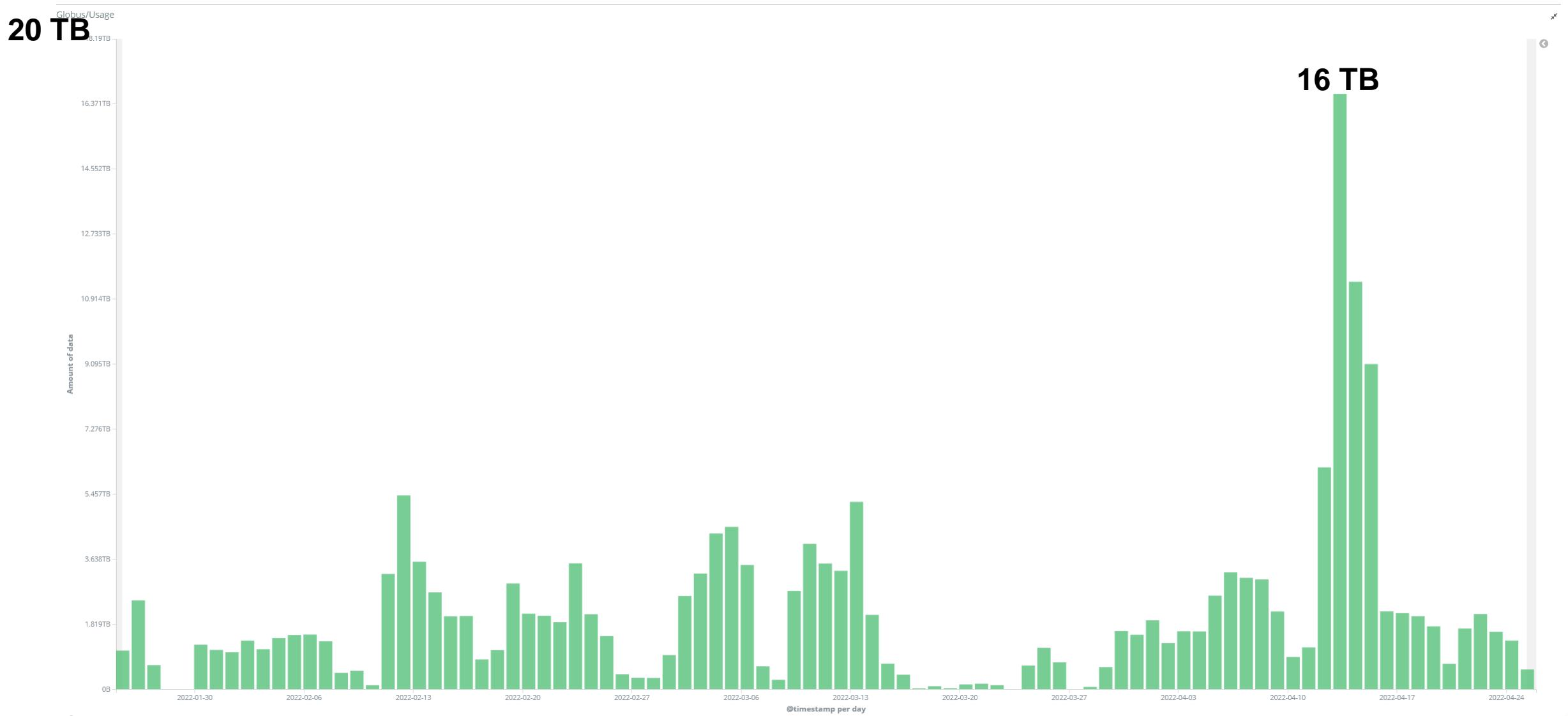
OPEN DATA USERS

- *All the above services as soon as data are made open (after 3 years)*

OPEN DATA ACCESS for MACHINES

- *Currently the above services are optimized for machines*

BIG DATA TRANSFERS – GLOBUS ONLINE DAILY USE



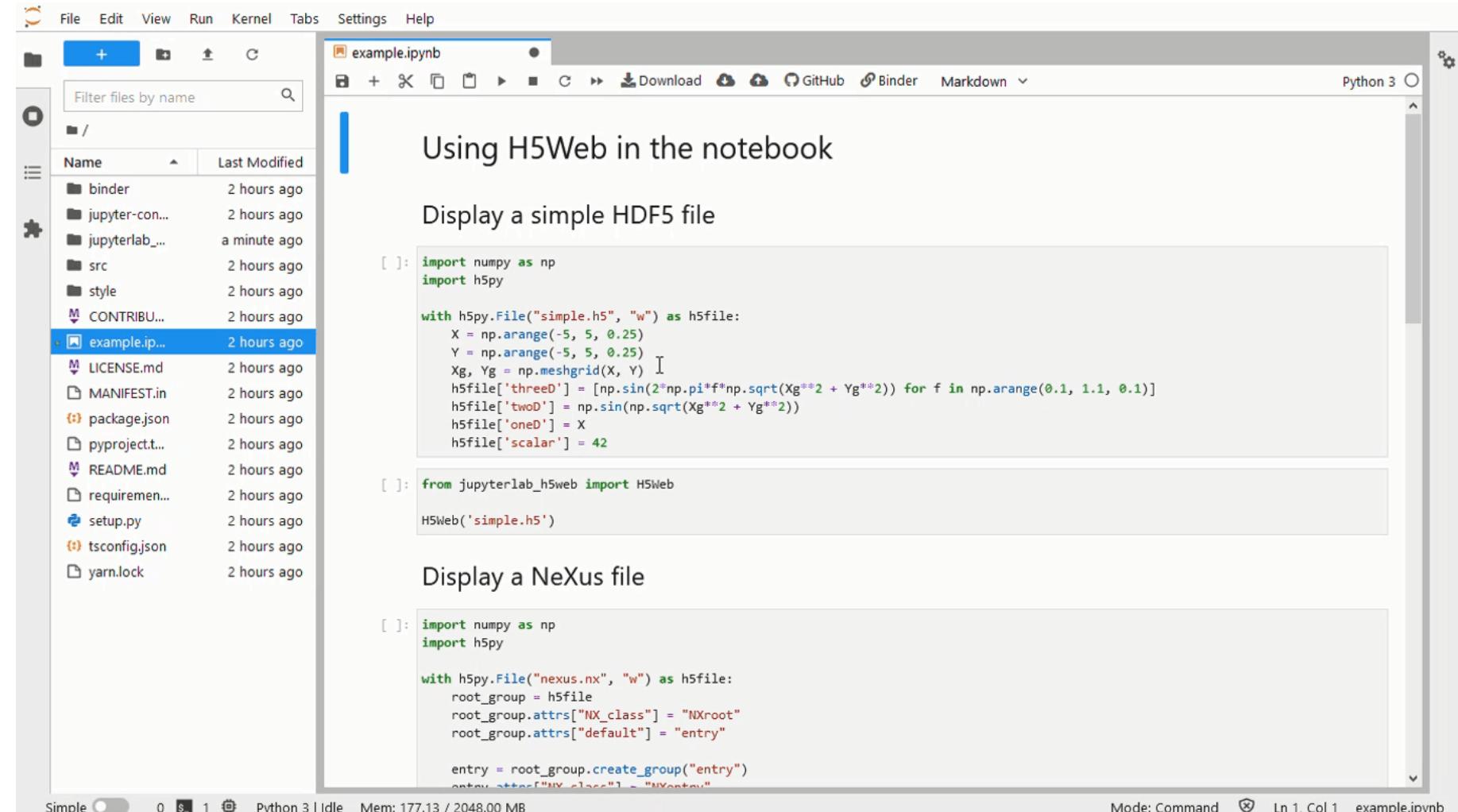
TOOLS TO VISUALIZE DATA – WEB BASED

PaNOSC developed
H5web web-based viewer
of HDF5 files and
integrated it in Jupyterlab,
data portals, + web
applications:

<https://github.com/silx-kit/h5web>

<https://h5web.panosc.eu/>

Next step : 3D viewer



Using H5Web in the notebook

Display a simple HDF5 file

```
[ ]: import numpy as np
import h5py

with h5py.File("simple.h5", "w") as h5file:
    X = np.arange(-5, 5, 0.25)
    Y = np.arange(-5, 5, 0.25)
    Xg, Yg = np.meshgrid(X, Y)
    h5file['threeD'] = [np.sin(2*np.pi*f*np.sqrt(Xg**2 + Yg**2)) for f in np.arange(0.1, 1.1, 0.1)]
    h5file['twoD'] = np.sin(np.sqrt(Xg**2 + Yg**2))
    h5file['oneD'] = X
    h5file['scalar'] = 42
```

[]: from jupyterlab_h5web import H5Web
H5Web('simple.h5')

Display a Nexus file

```
[ ]: import numpy as np
import h5py

with h5py.File("nexus.nx", "w") as h5file:
    root_group = h5file
    root_group.attrs["NX_class"] = "NXroot"
    root_group.attrs["default"] = "entry"

    entry = root_group.create_group("entry")
    entry.attrs["NX_class"] = "NXentry"
```

Simple 0 Python 3 | Idle Mem: 177.13 / 2048.00 MB Mode: Command 0 Ln 1, Col 1 example.ipynb

<https://h5web.panosc.eu/>

EXAMPLE OF ESRF DATA – HUMAN ORGAN ATLAS

A recent example of data from the **ESRF** is the **Human Organ Atlas**
<https://human-organ-atlas.esrf.eu/>

The data represent the highest resolution scanning of individual human organs made possible by the new **4th generation source - EBS**

The data are being made **open** as soon as they are processed. **Over 30 groups world-wide are using the data.**

The goal is to make a complete atlas of the human body.

Human Organ Atlas EXPLORE SEARCH HELP

Welcome to the Human Organ Atlas

The Human Organ Atlas uses **Hierarchical Phase-Contrast Tomography** 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. To stay up to date, follow [@HiP-CT](#)



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

HiP-CT imaging and 3D reconstruction of a [complete brain](#) from the body donor LADAF-2020-31.
More videos can be viewed on the [HiP-CT YouTube channel](#).

Funding

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



Reference

Walsh, C.L., Tafforeau, P., Wagner, W.L. *et al.* Imaging intact human organs with local resolution of cellular structures using hierarchical phase-contrast tomography. *Nat Methods* (2021).
<https://doi.org/10.1038/s41592-021-01317-x>

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



Acknowledgements

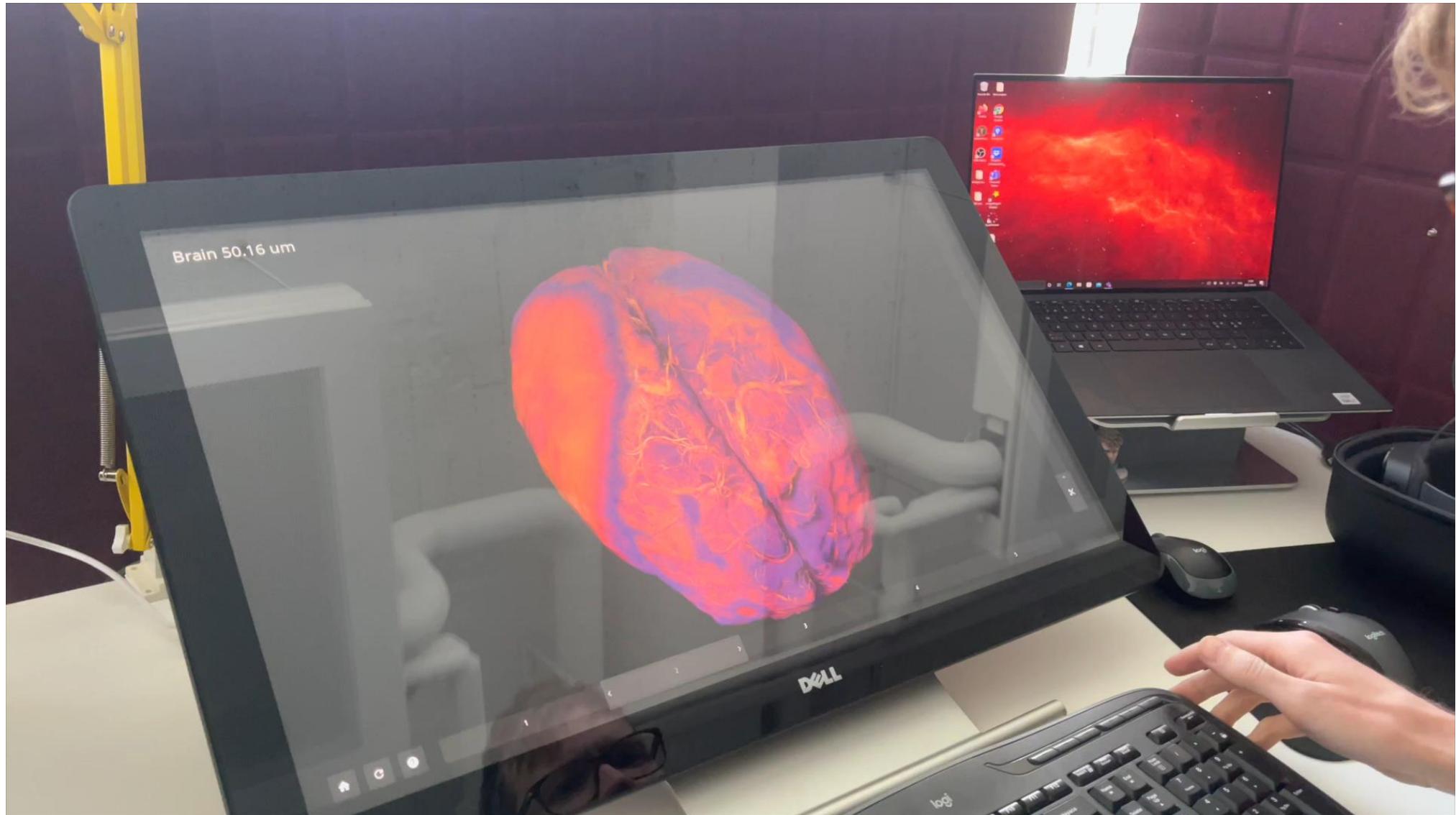
The development of this portal has been done as part of the [PaNOSC project](#). PaNOSC has received funding from the European Union's [Horizon 2020](#) research and innovation programme under grant agreement No. 823852. The following people were involved in the development: Paul Tafforeau, Alejandro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

TOOLS TO VISUALIZE DATA – 3D VIEWER CLOUD BASED

PaNOSC has collaborated with **InterSpectral** to use their cloud service to give users access to a 3D viewer in the browser without downloading the data

Next step :
integrate 3D viewing in data portal

DEMO



BRIDGE THE GAP BETWEEN SCIENCE AND OUTREACH



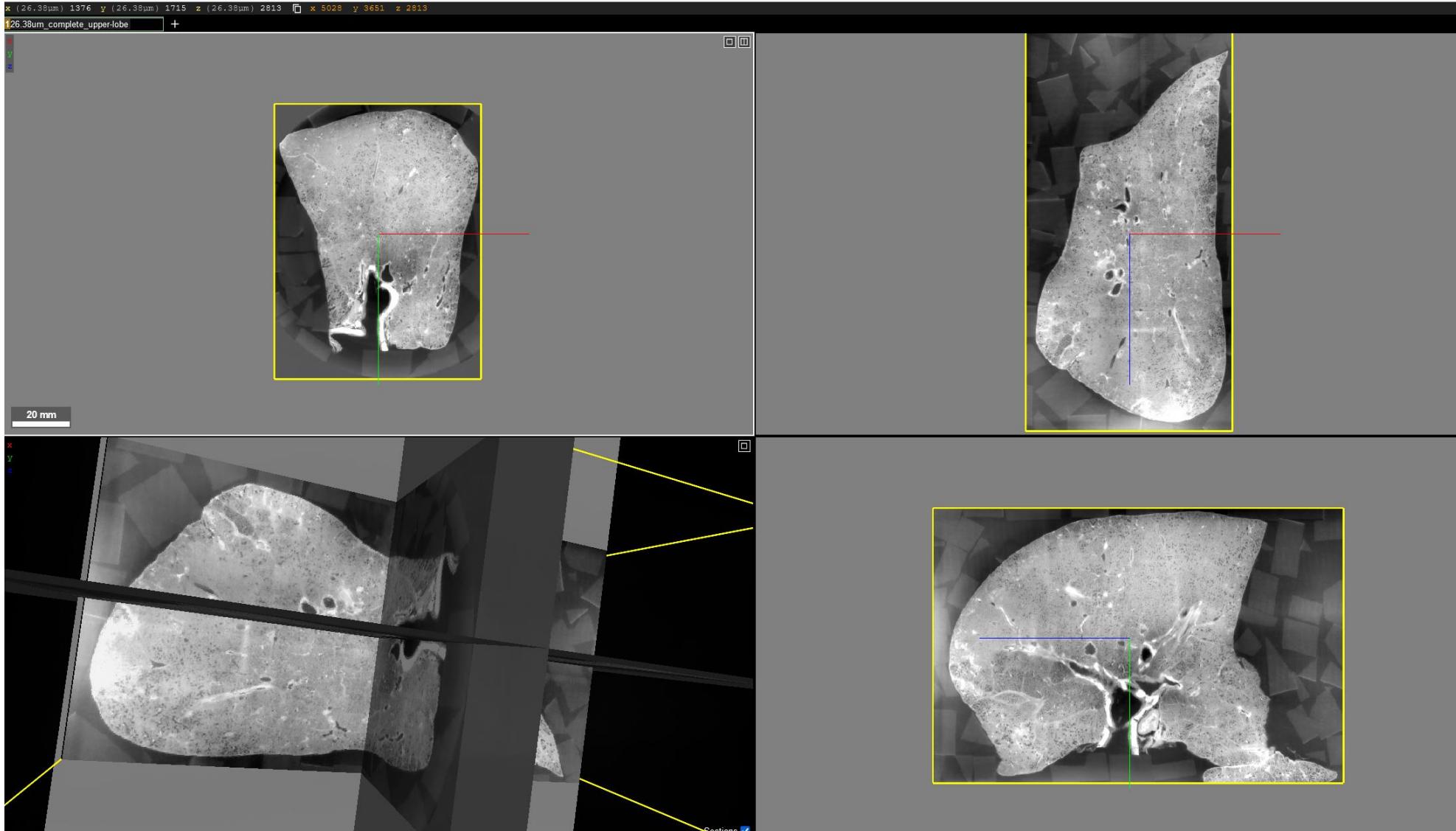
Interspectral – provides a cloud service for 3d volume rendering e.g.:

<https://interspectral.com/dataset/complete-brain-from-the-body-donor-ladaf-2020-31/>

TOOLS TO ANALYZE DATA – NEUROGLANCER CLOUD-BASED

Google has collaborated with Human Organ Atlas to provide online collaborative analysis tools in the browser without downloading the data

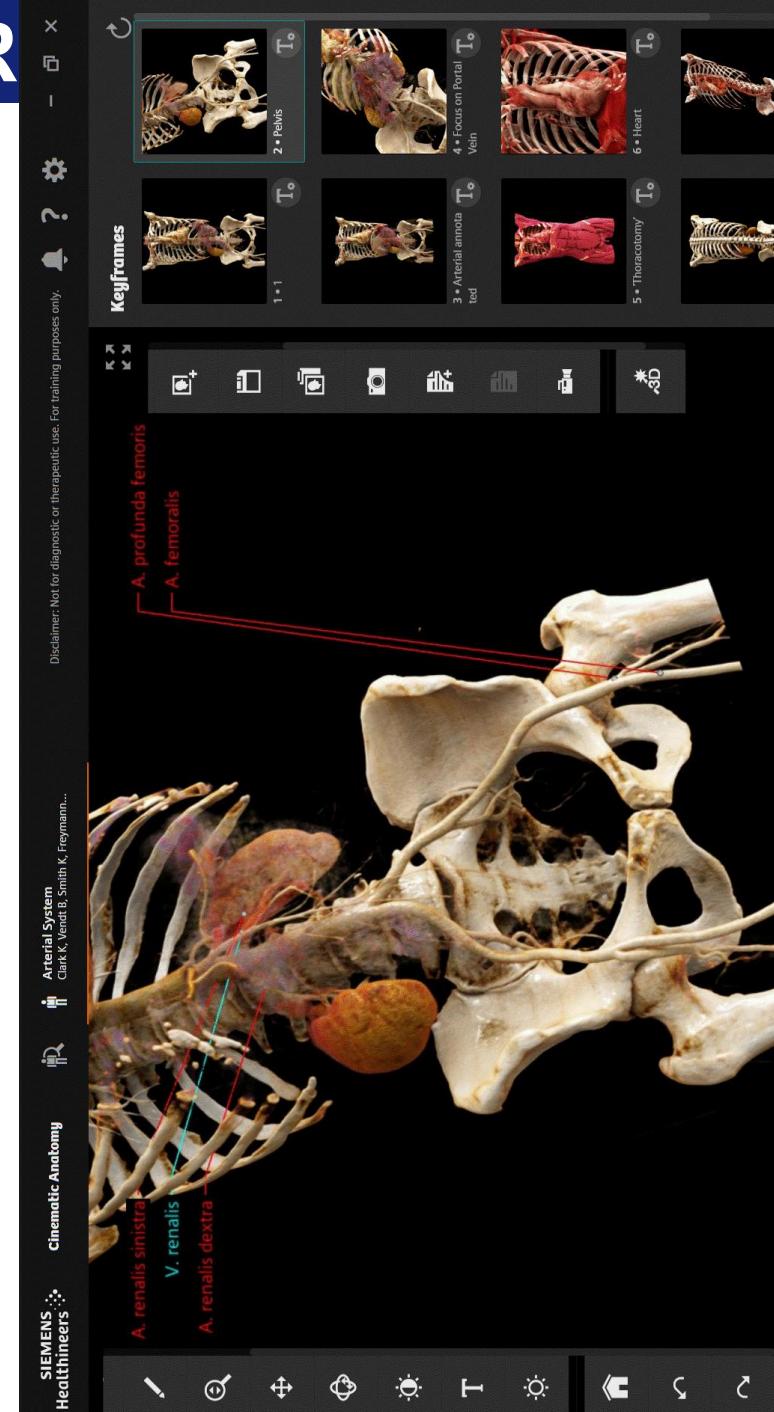
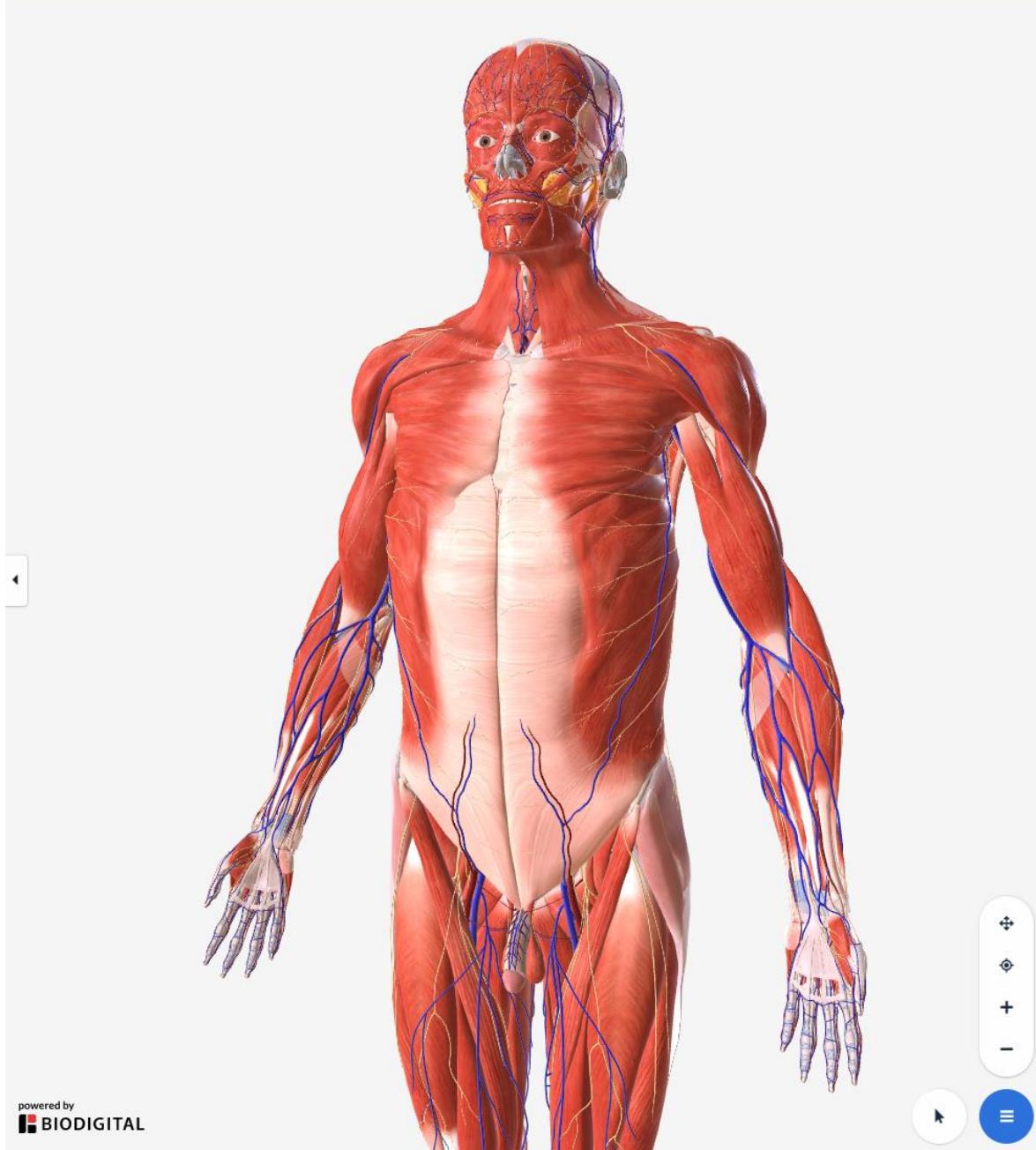
Next step : integrate link to 3D analysis in data portal



CITIZEN SCIENCE – 3D HUMAN VIEWER

Companies have expressed interest to incorporate the **Human Organ Atlas** into their database to provide high resolution volumes of human organs for training medical staff on virtual human bodies

Next step : integrate HOA data into virtual human tools



FINDING DATA LOCALLY - ELASTICSEARCH

Finding data is essential for ESRF users and re-use of Open Data.

Two approaches today:

(1) Local Elasticsearch engine

(2) Federated search of Open Data across facilities

Next step : introduce tracking of data re-use

Data Portal My Data Open Data Closed Data Shipping ▾ My Beamlines ▾ Manager ▾

Search

90 results found in 15ms

Filter by date -

Beamlines

SXMDATASET

#2: FeC_00_Kc_00_Cac_00

Name: cFe2_sec_root4
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 395
Volume: 163.8 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_sec_root3
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 241
Volume: 156.9 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_sec_root2
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 321
Volume: 87.6 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_sec_root1
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 405
Volume: 190.7 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_sec_root0
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 365
Volume: 241.6 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_mr_cortexmed
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 341
Volume: 199.5 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_mr_phloem
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 655
Volume: 251.8 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_mainroot
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 429
Volume: 372.1 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_coarse
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 17
Volume: 26.8 MB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: cFe2_spec01
Sample: cFe2
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 2
Volume: 37.6 KB
Mode: 7/8 multibunch
Type: SI

SXMDATASET

#1: FeC_00_Kc_00_Cac_00

Name: align_0101
Sample: align
Proposal: EV-280
Date: 31/10/2018
Beamline: id21
Files: 33.6 KB
Volume: 7/8 multibunch
Mode: SI

Prev 1 2 3 4 Next

ESRF

FINDING DATA – FEDERATED SEARCH API + PORTAL

Finding data is essential
for ESRF users and re-
use of Open Data.

Two approaches today:

(1) Local Elasticsearch
engine

(2) Federated search of
Open Data across
facilities

Next step : deploy
federated search across
all photon and neutron
facilities

The screenshot shows the panosc federated search portal interface. At the top, there is a logo and a search bar with the word "data". To the right of the search bar, it says "50+ results". On the left, there is a "FILTER" sidebar with the following options:

- Type:
 - proposal (radio button selected)
 - publication
- Technique:
 - Select a technique...
- Incident Wavelength:
 - min (input field)
 - max (input field)
 - nm (unit dropdown)
- Incident Photon Energy:
 - min (input field)
 - max (input field)
 - eV (unit dropdown)
- Chemical Formula:
 - Input field
- Temperature:
 - min (input field)
 - max (input field)
 - K (unit dropdown)
- Pressure:
 - min (input field)
 - max (input field)
 - Pa (unit dropdown)

The main area displays five search results, each with a title, a snippet, and a "Relevancy = 0.000" indicator. The results are:

- 10.17199/BRIGHTNESS/MG0006
Sample Data from multigrid
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 for the detector.
Released: 04/24/2022
- 10.17199/BRIGHTNESS/MG0032
Sample Data from multigrid
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 for the detector.
Released: 04/24/2022
- 10.17199/BRIGHTNESS/MG0029
Sample Data from multigrid
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 for the detector.
Released: 04/24/2022
- 10.17199/BRIGHTNESS/MG0028
Sample Data from multigrid
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 for the detector.
Released: 04/24/2022
- 10.17199/BRIGHTNESS/MG0027
Sample Data from multigrid
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 for the detector.
Released: 04/24/2022

<https://data.panosc.eu>

DOI - EXAMPLE



DOI > 10.15151/ESRF-DC-572252655

Data collection

Dataset Open access

COMPLETE BRAIN FROM THE BODY DONOR LADAF-2020-31

Paul Tafforeau ; Claire Walsh ; Willi L. Wagner ; Daniyal J. Jafree ; Alexandre Bellier ; Christopher Werlein ; Mark P. Kühnel ; Elodie Boller ; Simon Walker-Samuel ; Jan Lukas Robertus ; David A. Long ; Joseph Jacob ; Sebastian Marussi ; Eeline Brown ; Natalie Holroyd ; Danny D. Jonigk ; Maximilian Ackermann ; Peter D. Lee.

DOI

DOI 10.15151/ESRF-DC-572252655

Licence (for files)

Creative Commons Attribution 4.0

Abstract

Complete scan at 25.08um performed by HiP-CT on the beamline BM05 of the brain from the body donor LADAF-2020-31 using quarter-acquisition protocol.

Proposals Beamlines

MD-1252 BM05

Publication year

2021

Experimental report

One report has been found for this proposal.

94736_C.pdf

Experimental data

The data can be accessed by clicking on the link below

Access data

Reference

Below is the recommended format for citing this work in a research publication.

Tafforeau P., Walsh C., Wagner W. L., Daniyal J. Jafree, 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. (2021). Complete brain from the body donor LADAF-2020-31. European Synchrotron Radiation Facility (ESRF). doi:10.15151/ESRF-DC-572252655

How do your DOI landing pages look like?

What is considered a good landing page?

Metadata catalogue

- **ICAT Catalogue is developed by STFC**
- **ICAT provides:**
 - Generic data model
 - Robust fine-grained user authorization
- **ESRF added:**
 - New User Interface
 - SSO login via openid
 - DOI landing page support
 - Sample shipping + tracking
 - Search based on Elasticsearch
 - E-logbook for experiments+beamlines
- **For more info: <https://github.com/icatproject>**

Data Portal My Data Open Data Closed Data Shipping ▾ My Beamlines ▾ Manager ▾

Logout Andy GOTZ

Open Data / 10.15151/ESRF-DC-572252655

Dataset List 1

Search

Date	Sample	Dataset	Definition	Files	Size	Download
16:03 7 Oct 2021	LADAF-2020-31_brain	25.08um_complete-organ[2021-10-07 16:03:18]	MRtomo	23	52.5 GB	

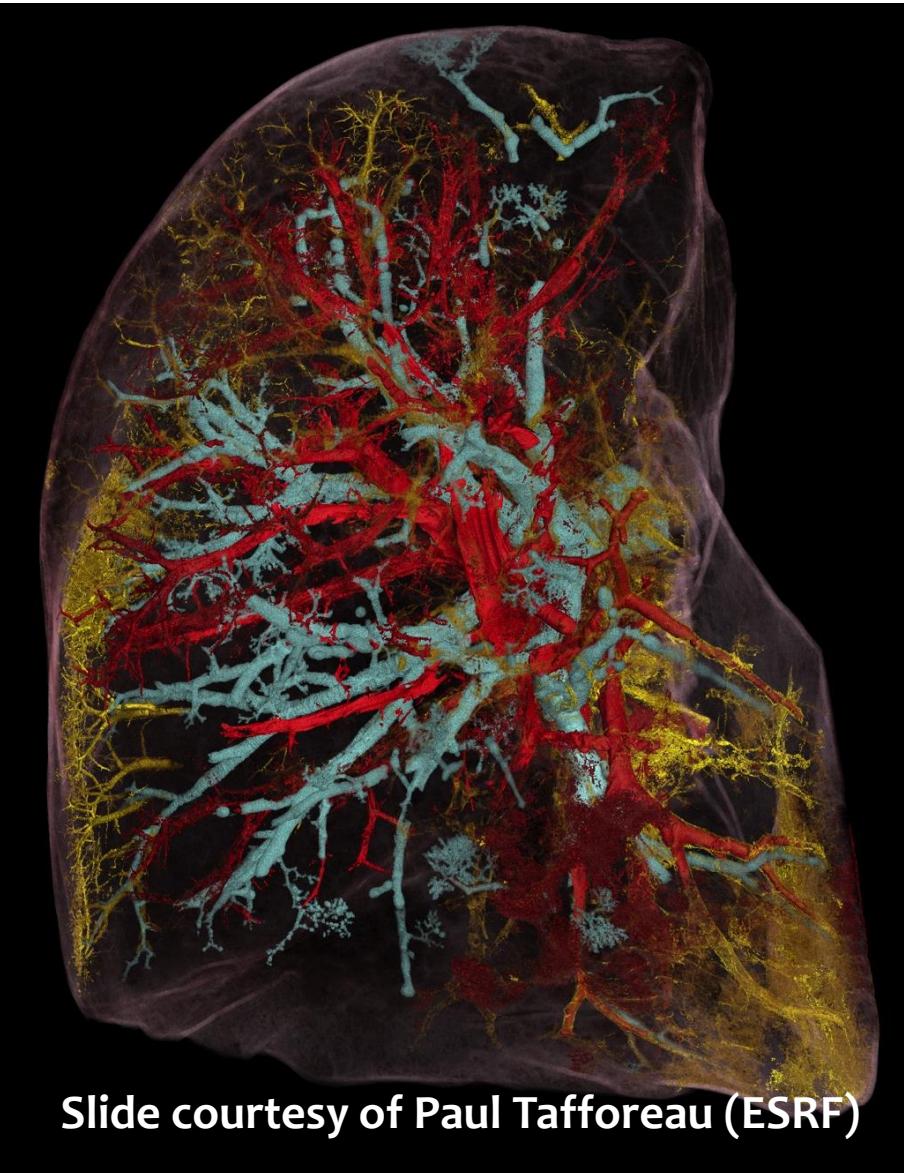
Summary Files 23 Metadata List

The Human Organ A

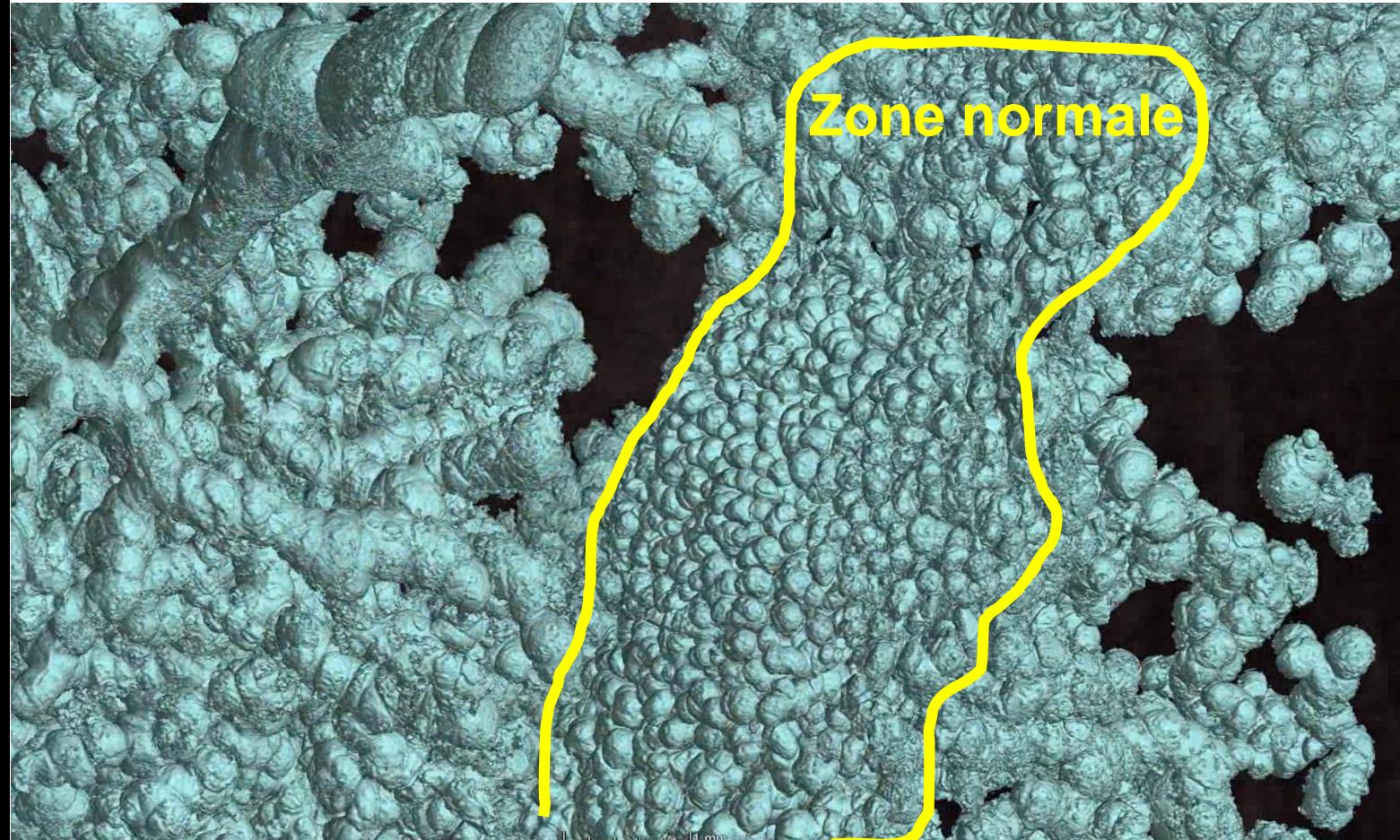
Patient	Scan Parameters	Sensor
definition	Instrument	Name
Identifier	BM05 EBS dipole	sCMOS PCO edge 4.2
Age (years)	wiggler 0.85T	CLHS
Sex	200	rolling shutter
Organ	Exposure Time (s)	6.5
Institute	Pixel Size (um)	Optics Type
Laboratoire d'Anatomie des Alpes Françaises	25.08	dzoom optic from BM05 based on Hasselblad 120mm F/4 macro objective
Info	Mode (None)	
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	continuous	
	ScanRadix	
	25.08um_LADAF-2020-31_brain_	
	Step (x,y,z)	Processing
	,,2	
	Stages (x,y,z)	refapproach
	1,1,2x79	reference jar with 70% ethanol, single reference
	Projections	Volume X
	9990	5965
	refin	Volume Y
	Click to edit	5965
	darkn (None)	Volume Z
	400	6990
	refon	32to16bitsmin
	Click to edit	-0.04
	Acc. Frames Count	32to16bitsmax
	6	0.1
	Detector Distance (mm)	jp2compratio
	3475	10
	Energy (avg) (keV)	filters
	93	Mo 0.1mm Al 2mm SiO2 bars 12°5mm diameter
	Scan Geometry	technique
	quarter-acquisition, one scan in half-acquisition plus one annular scan	Hierarchical Phase-Contrast Tomography
		experimentType
		tomography
Abstract	Complete scan at 25.08um performed by HIP-CT on the beamline BM05 of the brain from the body donor LADAF-2020-31 using quarter-acquisition protocol.	
Title	Complete brain from the body donor LADAF-2020-31	
Users	Paul Tafforeau, Claire Walsh, Willi L. Wagner, Daniyal J. Jafree, Alexandre Bellier, Christopher Werlein, Mark P. Kühnel, Elodie Boller, Simon Walker-Samuel, Jan Lukas Robertus, David A. Long, Joseph Jacob, Sebastian Marussi, Emmeline Brown, Natalie Holroyd, Danny D. Jonigk, Maximilian Ackermann, Peter D. Lutz	
Sample	LADAF-2020-31_brain	
Info	complete brain from the body donor program of the Laboratoire d'Anatomie des Alpes Française (LADAF)	
Preparation	formalin fixed, progressive transfer to ethanol 70% with gentle vacuum degassing at each step, mounted with mixed agar gel at 70% ethanol, n.b. some damages due to the	

CORE TRUST SEAL

OPEN DATA – COVID EXAMPLE



Q: should we try to link with other COVID databases?



<https://human-organ-atlas.esrf.eu/datasets/571998122>

NeXus

NeXus is developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data.

[Home](#)

[GitHub Organisation](#)

© 2022 NIAC

representing major scientific facilities in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data.

Documentation:

- Most recent publication to cite:
J. Appl. Cryst. (2015). **48**, 301-305 [doi:10.1107/S1600576714027575](https://doi.org/10.1107/S1600576714027575)
- [User Manual](#):
 - [Introduction](#) to the concepts behind the NeXus data format
 - [Design](#): The hierarchical design of NeXus files
 - [NeXus Class Definitions](#): description of each NXDL specification
 - [base classes](#): components that might be used in any NeXus data file
 - [application definitions](#): layout specifications for a specific purpose
 - [contributed definitions](#): propositions from the community
 - [Utilities](#): Software applications that browse, plot, and analyze NeXus data
 - [FAQ](#): Commonly asked questions about NeXus
- [Facilities](#) using NeXus

<https://www.nexusformat.org/>

Discussion and Development:

- Next Meetings: [Code Camp 2022](#) and [Autumn NIAC2022](#)

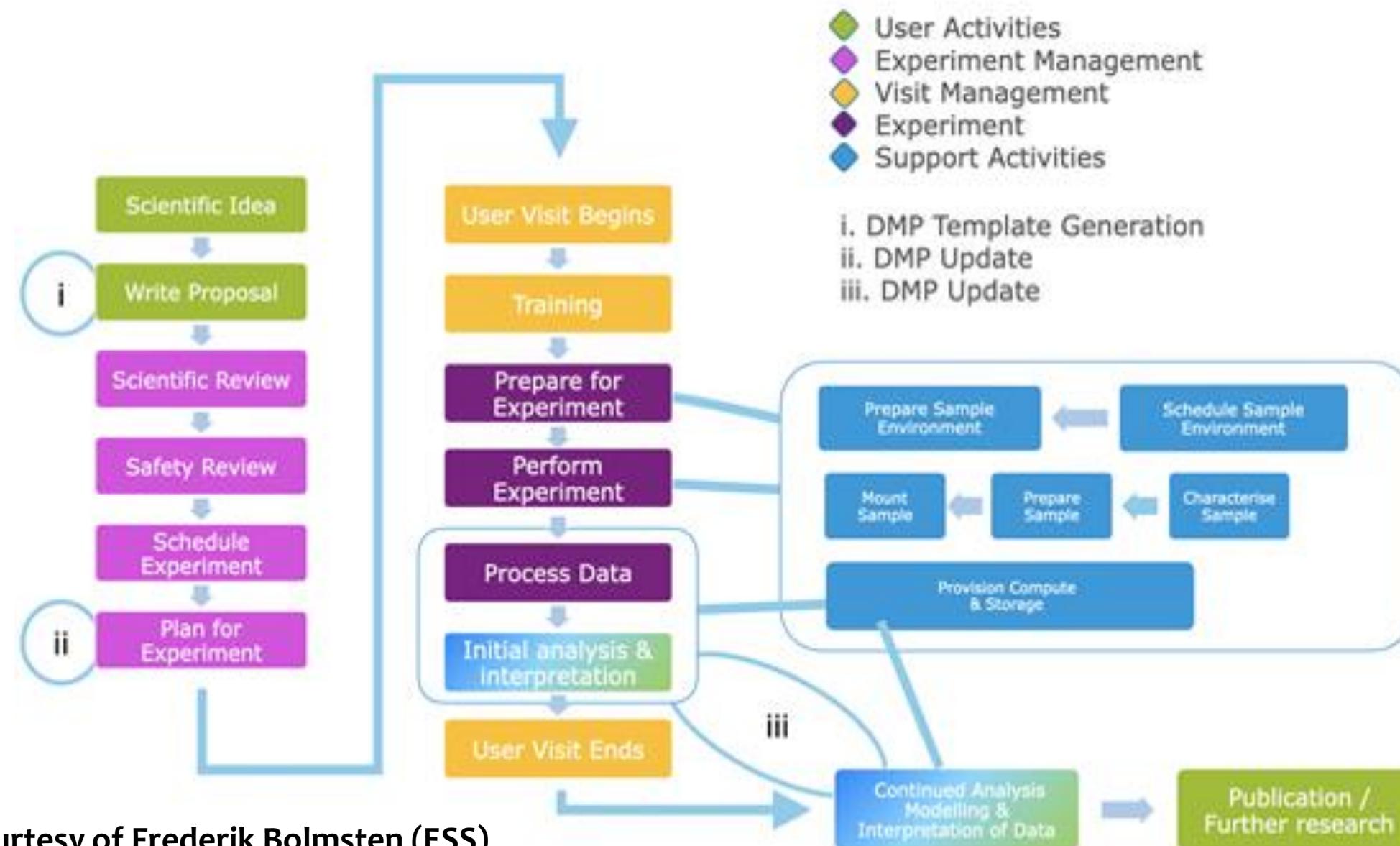
DMP as a Service

- PaNOSC developed a DMP service based on DS Wizard (developed by Elixir)
- Provided as a service for Users since 01/2022
- DMP templates for ESRF, H2020 and ANR already available
- DMP service pre-fills in 80% of the DMP
- DMPs are updated during the life-cycle of the project i.e. active DMPs
- DMP is stored as part of the output of the experiment

Feedback from friendly users

- DMPs are new for most users
- DMPs are time-consuming to fill in
- Applying the facility data policy to fill in the DMP is not obvious
- The DMP service is a time server and therefore appreciated
- Next step is to provide templates for Horizon Europe proposals
- Rollout DMP service for all experiments at ESRF as an information

ACTIVE DATA MANAGEMENT PLANS



Examples for AI being used for Data Analysis

- Connectomics - segmentation
- Human Organs - segmentation
- Spectroscopy – element identification
- AlphaFold – protein structures

Questions

- How to make access to data easier for machines
- Who can profit from ML trained algorithms
- What is the cost of training AI/ML
- How to make data for machines findable

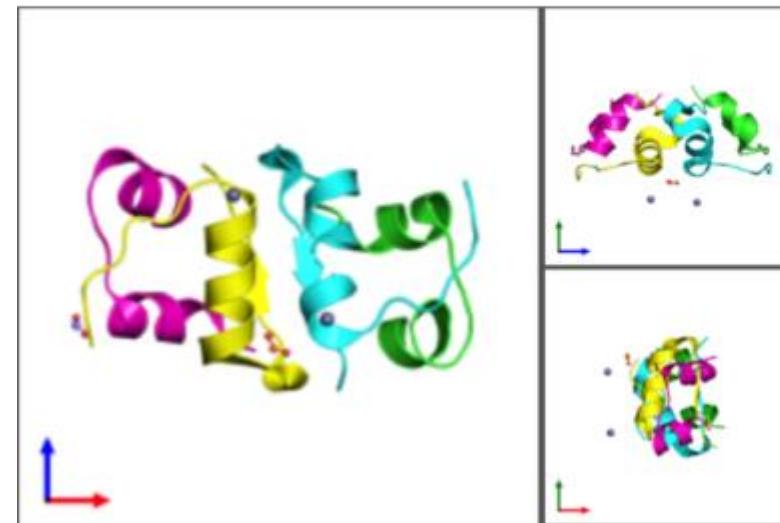
DOMAINS OF OPEN DATA

- Paleontology
- Additive manufacturing
- Neuronal connectomes
- Batteries research
- Cultural heritage

Macromolecular crystallography – Protein Data Bank

154,291 (84.7%) of structures in the PDB are from synchrotrons

PDB 6gv0 coloured by chain and viewed from the front



Experimental raw data

ⓘ Links to raw experimental data available for this entry are listed below

Raw experimental data related to PDB entry 6gv0:

Data DOI: [10.5281/zenodo.4456817](https://doi.org/10.5281/zenodo.4456817)



Dataset type: diffraction image data

<https://www.panosc.eu/use-cases/panosc-use-case-10-linking-raw-data-to-the-protein-data-bank-in-europe-pdbe/>

PALEONTOLOGY OPEN DATA



SOME ARTEFACTS ANALYZED AT THE ESRF



Egyptian cosmetics
Louvre museum,
(18th dynasty, 1570-1292 BC)
Ph. Walter,
Paris

Inks in Egyptian papyri
Carlsberg Collection
(1st-3rd C.),
T. Christiansen,
Copenhagen

Ming porcelains
Beijing Archaeological
Institute (15th-16th C.),
Ph. Sciau,
Toulouse

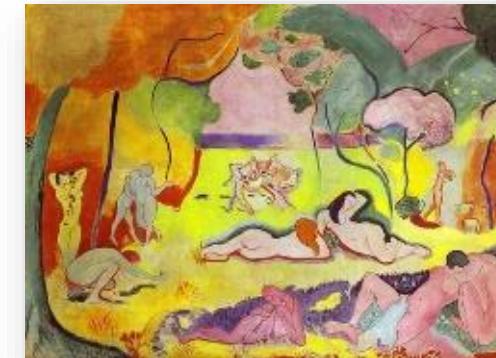
Rembrandt's impasto
Rijksmuseum
(1634),
V. Gonzalez,
Delft

Van Gogh's Sunflowers
van Gogh Museum,
(1888-1889),
L. Monico,
Perugia, Antwerp

Munch's Scream,
Munch Museum
(1910),
L. Monico,
Perugia

Rodin's sculptures
Rodin museum
(1908-1913),
J. Langlois,
Paris

Antiquity → Today



Opaque Egyptian glasses
Louvre museum and
British museum,
(18th dynasty, 1570-1292 BC)
I. Biron,
Paris

Bamiyan Buddhist paintings
(6th-9th C.)
Y. Taniguchi,
Tokyo

Leonardo da Vinci's whites
Louvre Museum
(1452–1519)
V. Gonzalez,
Paris

Ultramarine pigment
in *Girl with a Pearl Earring*,
Mauritshuis
(1665),
A. Gambardella,
Amsterdam

Matisse's Joy of Life
Barnes Foundation,
(1905-1906),
E. Pouyet,
Grenoble

Italian design
objects in plastics
private collection,
1960's, L. Toniolo,
Milano

Slide courtesy of Marine Cote (ESRF)

QUESTIONS TO DISCUSS DURING WORKSHOP

- 1. What solution do you provide for searching – can it be used for other data sources?**
- 2. What guidelines do you have for DOI landing pages?**
- 3. How do you support access for machines i.e. non-humans?**
- 4. Do you have open source solutions for viewing high-resolutions 3D volumes?**
- 5. Do you have experts who could help us on managing sensitive data?**
- 6. How do you estimate the cost of managing and archiving FAIR data?**
- 7. What metrics do you have in place for tracking data use + re-use?**
- 8. Should we try to link data between the Eiroforums?**

WHAT WE CAN SHARE

- 1. Experience with ICAT metadata catalogue + e-logbooks**
- 2. Tools and experience with Jupyter notebooks**
- 3. Experience with Nexus/HDF5 for online and offline use**
- 4. Web-based viewer for HDF5 + plotting**
- 5. Federated search API + scoring**
- 6. OPEN DATA!**

CONCLUSION

- 1. Data has become a first class citizen @ ESRF**
- 2. Data services help mainly the producers of the data**
- 3. Open data allows data to be verified and re-used and is becoming the norm but still encounters resistance**
- 4. Young scientists are showing the way ...**

YOUNG SCIENTIST SHOWING THE WAY

***“If you don't
want to share
data why
become a
scientist?”***



PaNOSC
@Panosc_eu

...

← Thread

Interview w @Claire_J_Walsh, Senior Research Fellow at @uci & @esrfsynchrotron, and part of the team who developed (w PaNOSC contribution) the #HumanOrganAtlas, an online data portal of human organs scanned w @hip_ct technology at 20µm per voxel

🎥 bit.ly/3Mk3yc5

The thumbnail image features the PaNOSC logo at the top left. In the center, there is a circular portrait of Claire Walsh, a young woman with glasses and a smile. To the right of the portrait, the text reads: "Interview with Claire Walsh (UCL - ESRF) on the Human Organ Atlas". At the bottom, there is a small paragraph about funding from the European Union's Horizon 2020 research and innovation programme, followed by the EU flag logo, the ESRF logo, the UCL logo, and the ESRF logo.

PaNOSC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

ESRF

UCL

ESRF

PIONEERING SYNCHROTRON SCIENCE



WE
DATA



THANK YOU FOR YOUR ATTENTION

Thanks to all users and colleagues involved in making research at ESRF possible!



This work is supported by PaNOSC, an European project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 870313, in the H2020-INFRADEV-2019-2 call.