FAIR DATA IN PHOTON SCIENCE.

DESY-SESAME Scientific Computing Collaboration Meeting

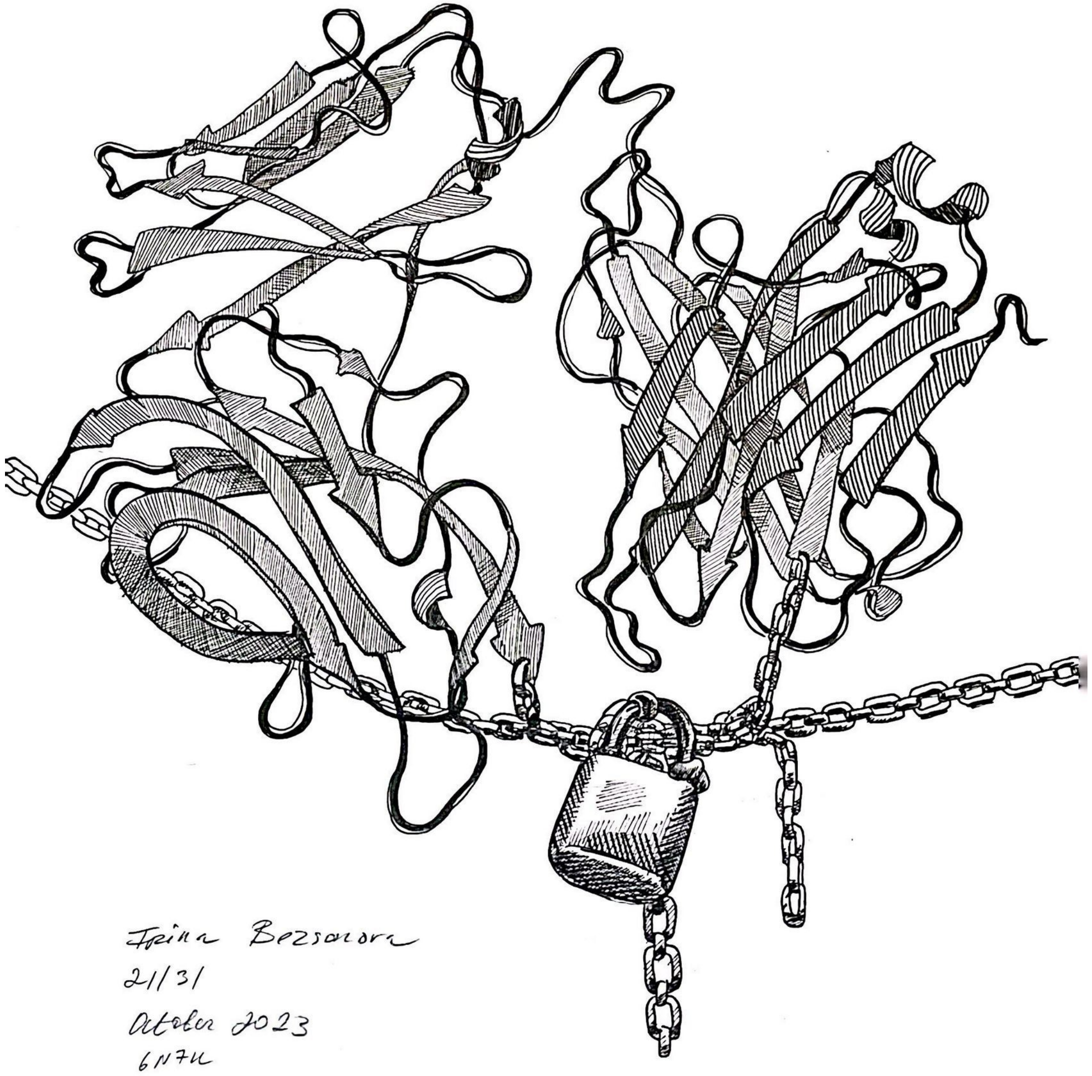
Sophie Servan
DESY Research and Innovation in Scientific Computing

13.08.2025

The Goal for FAIR Data in Photon Science

Making AlphaFold-level breakthroughs the rule

Image by Irina Bezsonova, free for use under a CC-BY-4.0 licence.



In 2020, a team at DeepMind released **AlphaFold**—a system that could predict the 3D structure of proteins with astonishing accuracy. Overnight, it solved a 50-year grand challenge in biology: how a chain of amino acids folds into a functional protein.

It was trained on **hundreds of thousands of protein structures**—data painstakingly collected over decades by scientists around the world.

And crucially, that data was **public, structured, and accessible**—thanks to the **Protein Data Bank (PDB)**, a public repository where researchers deposit their findings.

Image by Irina Bezsonova, free for use under a CC-BY-4.0 licence.

The Goal for FAIR Data in Photon Science

Making AlphaFold-level breakthroughs the rule

Structural biologists populated the PDB thanks to Photon Science instruments and techniques.

They are one in hundreds of communities using light sources for their science.

The next Alphafold could be feeding on the perovskite database, the human organ atlas, etc.

See a curated list of open data resources related to Photon (and Neutron) Science here:

https://leaps-wg3.desy.de/open-data-resources.html

FAIR[i] Data in Photon Science

Topics touched upon in this talk

FA

Policies

Journals editorial policies, funders, data policies of PaN RIs[i]

Implementation

DMPs, Metadata catalogues, PaN data portal, OAI-PMH

I

PaN standards

Formats, Metadata framework, PaNET, PaN-training

Communities standards

OSCARS projects

R

Big data

Visualisation, slicing, VISA

[i] „FAIR" as in Findable, Accessible, Interoperable and Reusable – *https://www.go-fair.org/fair-principles/*
[ii] „PaN RIs" are Photon and Neutron Research Infrastructures, i.e. synchrotrons, FELs, neutron sources.

FA. the Findable and Accessible in FAIR

Researchers around the globe make 3D structures of proteins freely available from the PDB archive. Why?

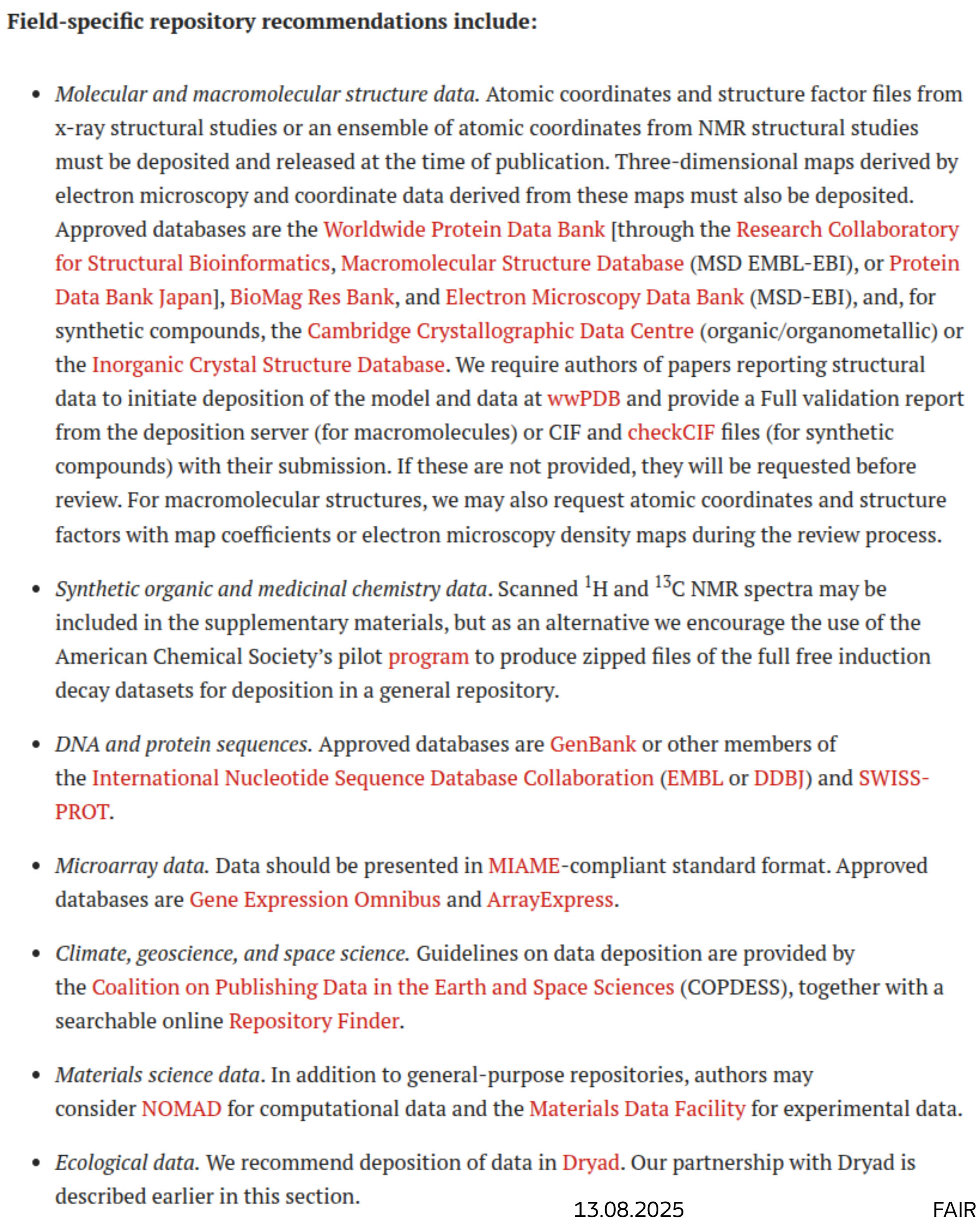
The structural biology community has long embraced the principle of open data sharing. Structural data are considered **foundational scientific knowledge**.

There’s a strong **ethical and cultural expectation** to share them publicly.

Researchers get **credit and recognition** for their published structures.

And as a result…

https://www.science.org/content/page/science-journals-editorial-policies#data-and-code-deposition checked Aug, 6th 2025



Field-specific repository recommendations include:

- *Molecular and macromolecular structure data.* Atomic coordinates and structure factor files from x-ray structural studies or an ensemble of atomic coordinates from NMR structural studies must be deposited and released at the time of publication. Three-dimensional maps derived by electron microscopy and coordinate data derived from these maps must also be deposited. Approved databases are the Worldwide Protein Data Bank [through the Research Collaboratory for Structural Bioinformatics, Macromolecular Structure Database (MSD EMBL-EBI), or Protein Data Bank Japan], BioMag Res Bank, and Electron Microscopy Data Bank (MSD-EBI), and, for synthetic compounds, the Cambridge Crystallographic Data Centre (organic/organometallic) or the Inorganic Crystal Structure Database. We require authors of papers reporting structural data to initiate deposition of the model and data at wwPDB and provide a Full validation report from the deposition server (for macromolecules) or CIF and checkCIF files (for synthetic compounds) with their submission. If these are not provided, they will be requested before review. For macromolecular structures, we may also request atomic coordinates and structure factors with map coefficients or electron microscopy density maps during the review process.

- *Synthetic organic and medicinal chemistry data*. Scanned ^{1}H and ^{13}C NMR spectra may be included in the supplementary materials, but as an alternative we encourage the use of the American Chemical Society's pilot program to produce zipped files of the full free induction decay datasets for deposition in a general repository.

- *DNA and protein sequences.* Approved databases are GenBank or other members of the International Nucleotide Sequence Database Collaboration (EMBL or DDBJ) and SWISS-PROT.

- *Microarray data.* Data should be presented in MIAME-compliant standard format. Approved databases are Gene Expression Omnibus and ArrayExpress.

- *Climate, geoscience, and space science.* Guidelines on data deposition are provided by the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS), together with a searchable online Repository Finder.

- *Materials science data*. In addition to general-purpose repositories, authors may consider NOMAD for computational data and the Materials Data Facility for experimental data.

- *Ecological data*. We recommend deposition of data in Dryad. Our partnership with Dryad is described earlier in this section.

AAAS
Science
31 JULY 2025
How Ukraine's scientists are mobilizing for war p. 446
Deep brain stimulation target to restore walking wins Science & PINS Prize p. 468
Tropical tree growth resilience to past droughts p. 532
MACHINE LINKING
Molecular motor makes catenated rings pp. 454 & 526

FA. And in Photon Science in general?

➔ A lot of progress, mostly driven by funders and convinced communities.

Many communities, many expectation levels

Still at very different points on the road towards FAIR and open data. Great benefits to coordination efforts and **transfers**.

Recognition slowly taking off

In 2014, the **European Commission** began requiring open data sharing through a Horizon 2020 pilot program.
It expanded over time until open data became the default for all projects in Horizon Europe. „As open as possible, as closed as necessary."
Helmholtz is recognising data (and SW) publication as a research outcome from 2028 onwards. Calls for benchmarking data, data hubs, …

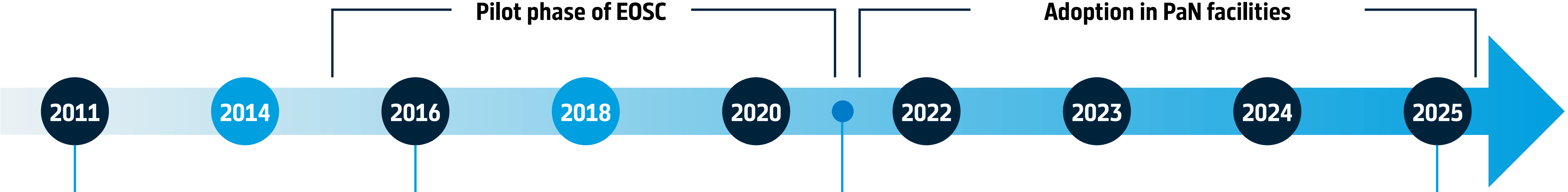
FA. Data Policies at PaN RIs and at DESY

An early recognition that PaN RIs are becoming **data producers** with high data flow and need to alter their practices.

panosc
photon and neutron
open science cloud

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

LEAPS
WORKING GROUP 3



Pilot phase of EOSC

Adoption in PaN facilities

2011 2014 2016 2018 2020 2022 2023 2024 2025

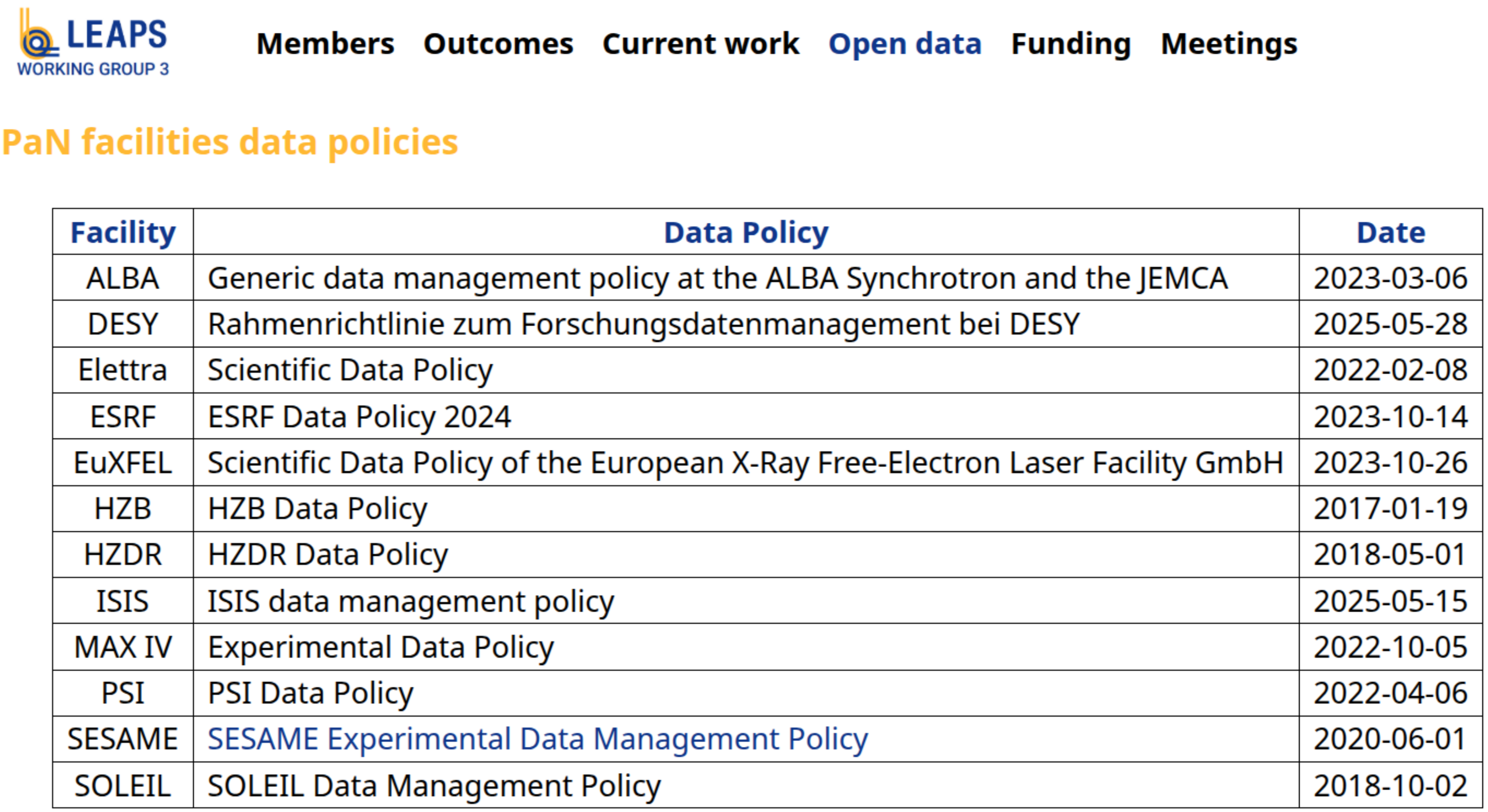
PaN-data Europe.
Common policy framework on scientific data

Wilkinson, M et.al.
The **FAIR guiding principles** for scientific data management and stewardship

August 2021
Data Policy framework for PaN RIs

May 2025
DESY Data Policy

FA. Data Policies at PaN RIs and at DESY



PaN facilities data policies

Facility	Data Policy	Date
ALBA	Generic data management policy at the ALBA Synchrotron and the JEMCA	2023-03-06
DESY	Rahmenrichtlinie zum Forschungsdatenmanagement bei DESY	2025-05-28
Elettra	Scientific Data Policy	2022-02-08
ESRF	ESRF Data Policy 2024	2023-10-14
EuXFEL	Scientific Data Policy of the European X-Ray Free-Electron Laser Facility GmbH	2023-10-26
HZB	HZB Data Policy	2017-01-19
HZDR	HZDR Data Policy	2018-05-01
ISIS	ISIS data management policy	2025-05-15
MAX IV	Experimental Data Policy	2022-10-05
PSI	PSI Data Policy	2022-04-06
SESAME	SESAME Experimental Data Management Policy	2020-06-01
SOLEIL	SOLEIL Data Management Policy	2018-10-02

https://leaps-wg3.desy.de/open-data-resources.html#gotodatapolicy

FA. Data Policies at PaN RIs and at DESY

„The follow-up costs for research data management and long-term archiving of data […] will be borne by DESY as an institution […]“

„A data management plan (DMP) should be created for all activities that generate research data. […] DESY provides suitable tools for planning, creating, implementing and managing DMPs.“

Translated from German by DeepL

„Published research data that has been assigned a persistent identifier must be stored and made accessible indefinitely […].“

„Research data should be made publicly available in a timely manner, and associated metadata must be made publicly available in a timely manner. […] the data should be made available under a CC-BY or CC-0 licence. Published metadata must be licensed under a CC-0 licence or an even less restrictive licence.“

Rahmenrichtlinie zum Forschungsdatenmanagement bei DESY

Rahmenrichtlinie zum Forschungsdatenmanagement bei DESY

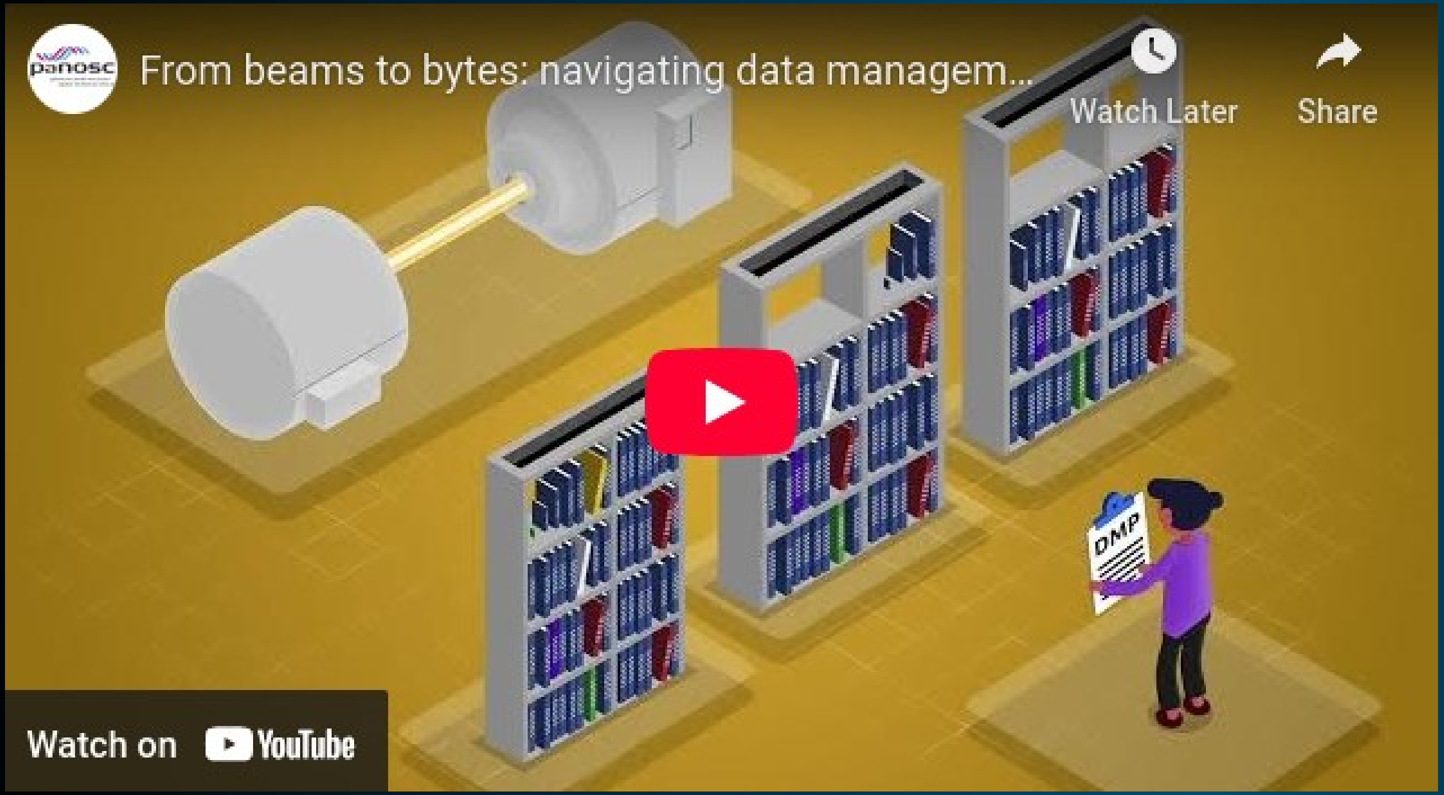
DESY-Forschungsdatenmanagement Version 1.5

„DESY supports its researchers […] in managing their research data […] by appointing one or more **research data managers**.“

Stand: 28.05.2025

1

FA. Talking about Data Management Plans…



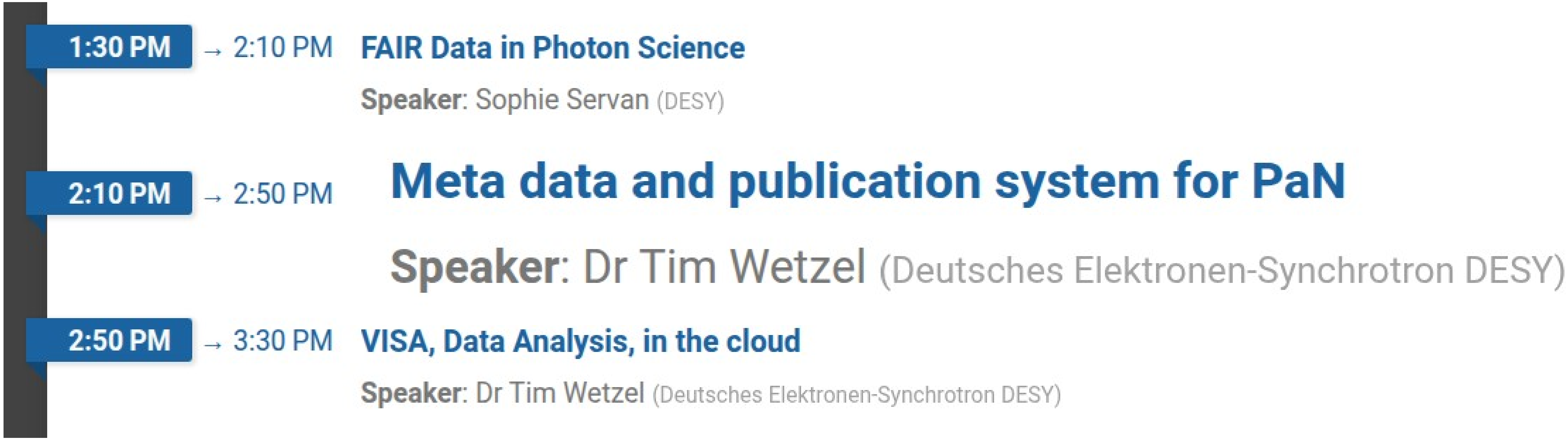
https://www.youtube.com/watch?v=9wCh2z8e7DI

FA. the Findable and Accessible in FAIR

Implementation for (meta)data access in Tim’s talk right after me

SciCat

advance your data to advance your science



1:30 PM → 2:10 PM **FAIR Data in Photon Science**

Speaker: Sophie Servan (DESY)

2:10 PM → 2:50 PM **Meta data and publication system for PaN**

Speaker: Dr Tim Wetzel (Deutsches Elektronen-Synchrotron DESY)

2:50 PM → 3:30 PM **VISA, Data Analysis, in the cloud**

Speaker: Dr Tim Wetzel (Deutsches Elektronen-Synchrotron DESY)

FA. the Findable and Accessible in FAIR

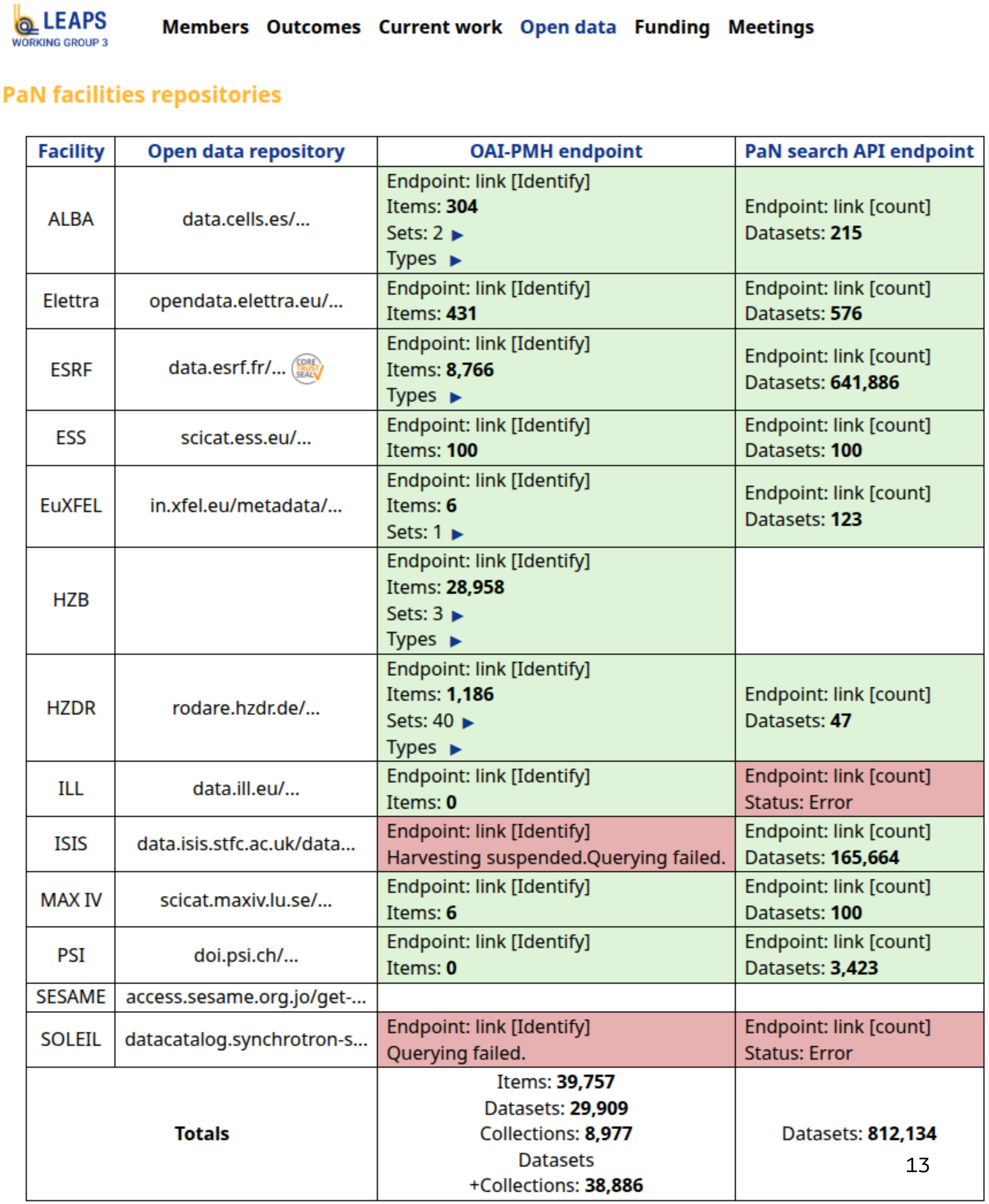
OAI-PMH endpoint: standard for metadata harvesting

Another outcome of ExPaNDS and PaNOSC is the addition of a module in SciCat and ICAT for OAI-PMH.

PaN search API endpoint

And for a PaN-specific search API.

https://leaps-wg3.desy.de/open-data-resources.html#gotofacility

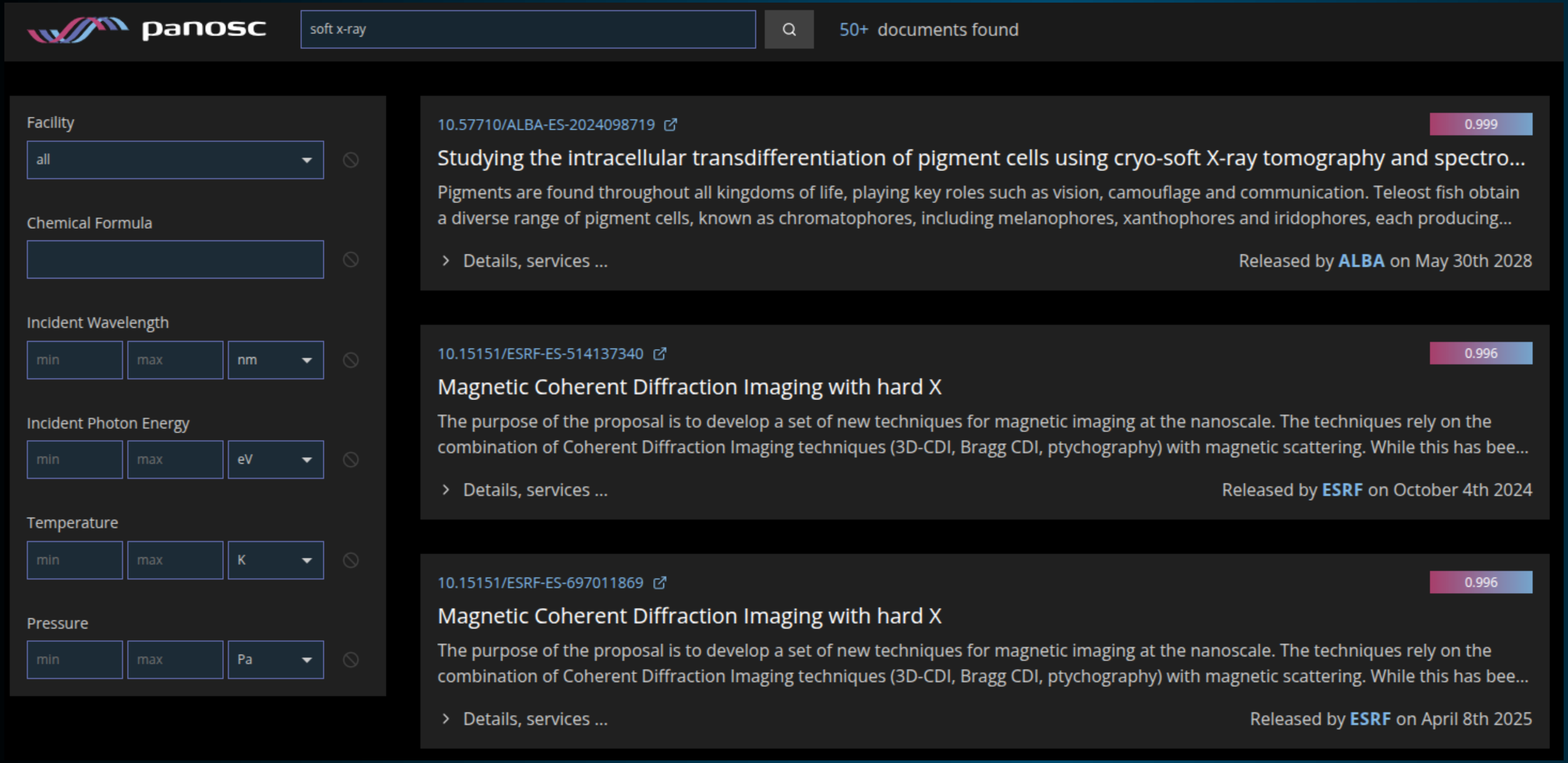


PaN facilities repositories

Facility	Open data repository	OAI-PMH endpoint	PaN search API endpoint
ALBA	data.cells.es/...	Endpoint: link [Identify] Items: **304** Sets: 2 ▶ Types ▶	Endpoint: link [count] Datasets: **215**
Elettra	opendata.elettra.eu/...	Endpoint: link [Identify] Items: **431**	Endpoint: link [count] Datasets: **576**
ESRF	data.esrf.fr/...	Endpoint: link [Identify] Items: **8,766** Types ▶	Endpoint: link [count] Datasets: **641,886**
ESS	scicat.ess.eu/...	Endpoint: link [Identify] Items: **100**	Endpoint: link [count] Datasets: **100**
EuXFEL	in.xfel.eu/metadata/...	Endpoint: link [Identify] Items: **6** Sets: 1 ▶	Endpoint: link [count] Datasets: **123**
HZB		Endpoint: link [Identify] Items: **28,958** Sets: 3 ▶ Types ▶	
HZDR	rodare.hzdr.de/...	Endpoint: link [Identify] Items: **1,186** Sets: 40 ▶ Types ▶	Endpoint: link [count] Datasets: **47**
ILL	data.ill.eu/...	Endpoint: link [Identify] Items: **0**	Endpoint: link [count] Status: Error
ISIS	data.isis.stfc.ac.uk/data...	Endpoint: link [Identify] Harvesting suspended.Querying failed.	Endpoint: link [count] Datasets: **165,664**
MAX IV	scicat.maxiv.lu.se/...	Endpoint: link [Identify] Items: **6**	Endpoint: link [count] Datasets: **100**
PSI	doi.psi.ch/...	Endpoint: link [Identify] Items: **0**	Endpoint: link [count] Datasets: **3,423**
SESAME	access.sesame.org.jo/get-...		
SOLEIL	datacatalog.synchrotron-s...	Endpoint: link [Identify] Querying failed.	Endpoint: link [count] Status: Error
Totals		Items: **39,757** Datasets: **29,909** Collections: **8,977** Datasets +Collections: **38,886**	Datasets: **812,134**

FA. the Findable and Accessible in FAIR

A few words on the PaN data portal <data.panosc.eu>



soft x-ray

50+ documents found

Facility: all

Chemical Formula

Incident Wavelength: min max nm

Incident Photon Energy: min max eV

Temperature: min max K

Pressure: min max Pa

10.57710/ALBA-ES-2024098719 — 0.999

Studying the intracellular transdifferentiation of pigment cells using cryo-soft X-ray tomography and spectro...

Pigments are found throughout all kingdoms of life, playing key roles such as vision, camouflage and communication. Teleost fish obtain a diverse range of pigment cells, known as chromatophores, including melanophores, xanthophores and iridophores, each producing...

Details, services ... Released by **ALBA** on May 30th 2028

10.15151/ESRF-ES-514137340 — 0.996

Magnetic Coherent Diffraction Imaging with hard X

The purpose of the proposal is to develop a set of new techniques for magnetic imaging at the nanoscale. The techniques rely on the combination of Coherent Diffraction Imaging techniques (3D-CDI, Bragg CDI, ptychography) with magnetic scattering. While this has bee...

Details, services ... Released by **ESRF** on October 4th 2024

10.15151/ESRF-ES-697011869 — 0.996

Magnetic Coherent Diffraction Imaging with hard X

The purpose of the proposal is to develop a set of new techniques for magnetic imaging at the nanoscale. The techniques rely on the combination of Coherent Diffraction Imaging techniques (3D-CDI, Bragg CDI, ptychography) with magnetic scattering. While this has bee...

Details, services ... Released by **ESRF** on April 8th 2025

Further developed in the frame of OSCARS

PaN-Finder project, introducing an AI-powered search tool to improve findability.

I. The road towards Interoperability

The primary information stored in the PDB consists of coordinate files for biological molecules: a list of the atoms in each protein and their 3D location in space. The PDB provides **uniform formats** (PDBx/mmCIF files) with **standard metadata**, automatic validation tools, quality checks and **training material**. Everyone can understand the data in the PDB.

I. Formats and standards for Photon Science

ascii

CSV

XDI

EDFF

jpeg

TIFF

mrc

raw

CBF

NeXus

dat

txt

CXI

openPMD

json

Zarr

pxt

HDF5

Current strategy

1. Keep advertising NeXus as the standard in PaN
2. Investigate LLMs for metadata extraction from e.g. ELN
3. Investigate transcoding modules

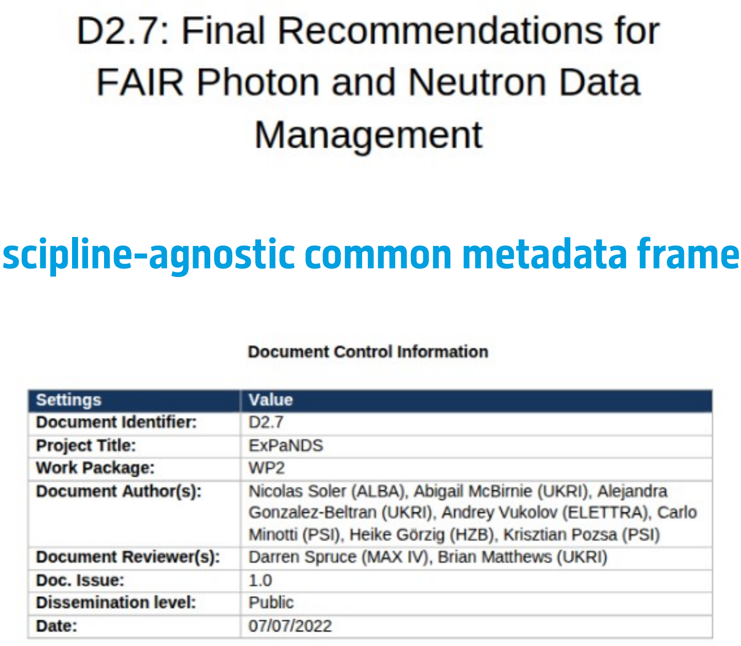
bluesky

Verification and validation of Nexus files

NXvalidate

punx

See *https://manual.nexusformat.org/validation.html*



D2.7: Final Recommendations for FAIR Photon and Neutron Data Management

→ **A discipline-agnostic common metadata framework**

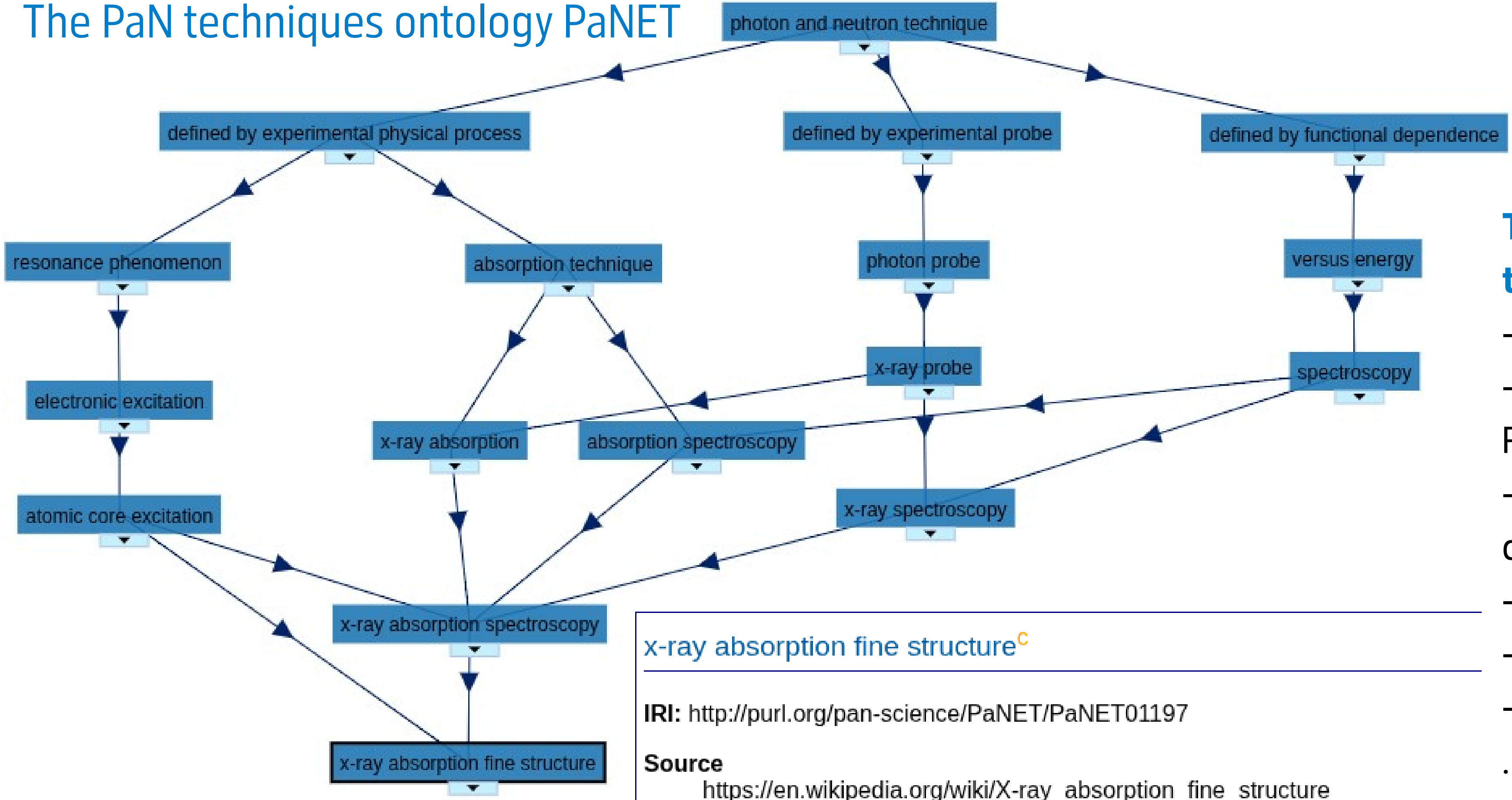
Document Control Information

Settings	Value
Document Identifier:	D2.7
Project Title:	ExPaNDS
Work Package:	WP2
Document Author(s):	Nicolas Soler (ALBA), Abigail McBirnie (UKRI), Alejandra Gonzalez-Beltran (UKRI), Andrey Vukolov (ELETTRA), Carlo Minotti (PSI), Heike Görzig (HZB), Krisztian Pozsa (PSI)
Document Reviewer(s):	Darren Spruce (MAX IV), Brian Matthews (UKRI)
Doc. Issue:	1.0
Dissemination level:	Public
Date:	07/07/2022

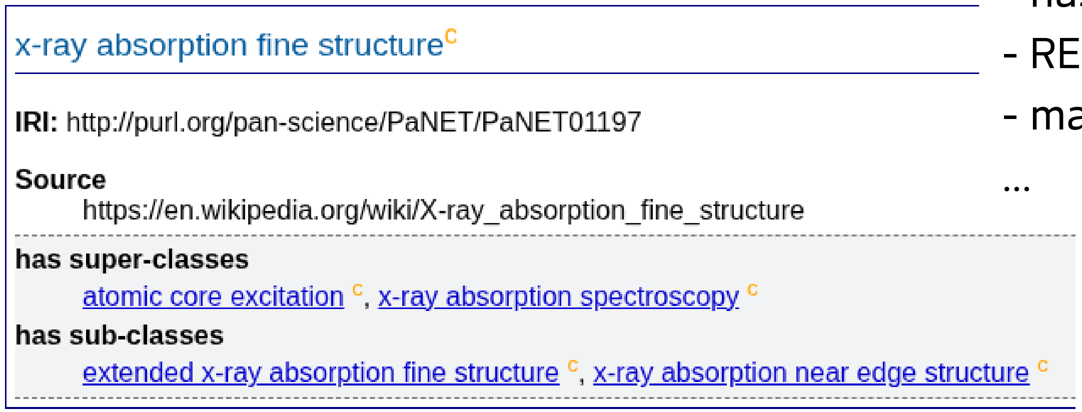
DOI 10.5281/zenodo.6821676

I. Formats and standards for Photon Science

The PaN techniques ontology PaNET



photon and neutron technique
defined by experimental physical process
defined by experimental probe
defined by functional dependence
resonance phenomenon
absorption technique
photon probe
versus energy
electronic excitation
x-ray probe
spectroscopy
x-ray absorption
absorption spectroscopy
atomic core excitation
x-ray spectroscopy
x-ray absorption spectroscopy
x-ray absorption fine structure



x-ray absorption fine structure^{c}

IRI: http://purl.org/pan-science/PaNET/PaNET01197

Source
https://en.wikipedia.org/wiki/X-ray_absorption_fine_structure

has super-classes
atomic core excitation ^{c}, x-ray absorption spectroscopy ^{c}

has sub-classes
extended x-ray absorption fine structure ^{c}, x-ray absorption near edge structure ^{c}

The ontology for experimental techniques

- facilitate consistent semantics
- provides synonyms, references and PIDs
- enhances search results from PaN data portal
- has a defined maintenance process
- REST API
- mappings

…

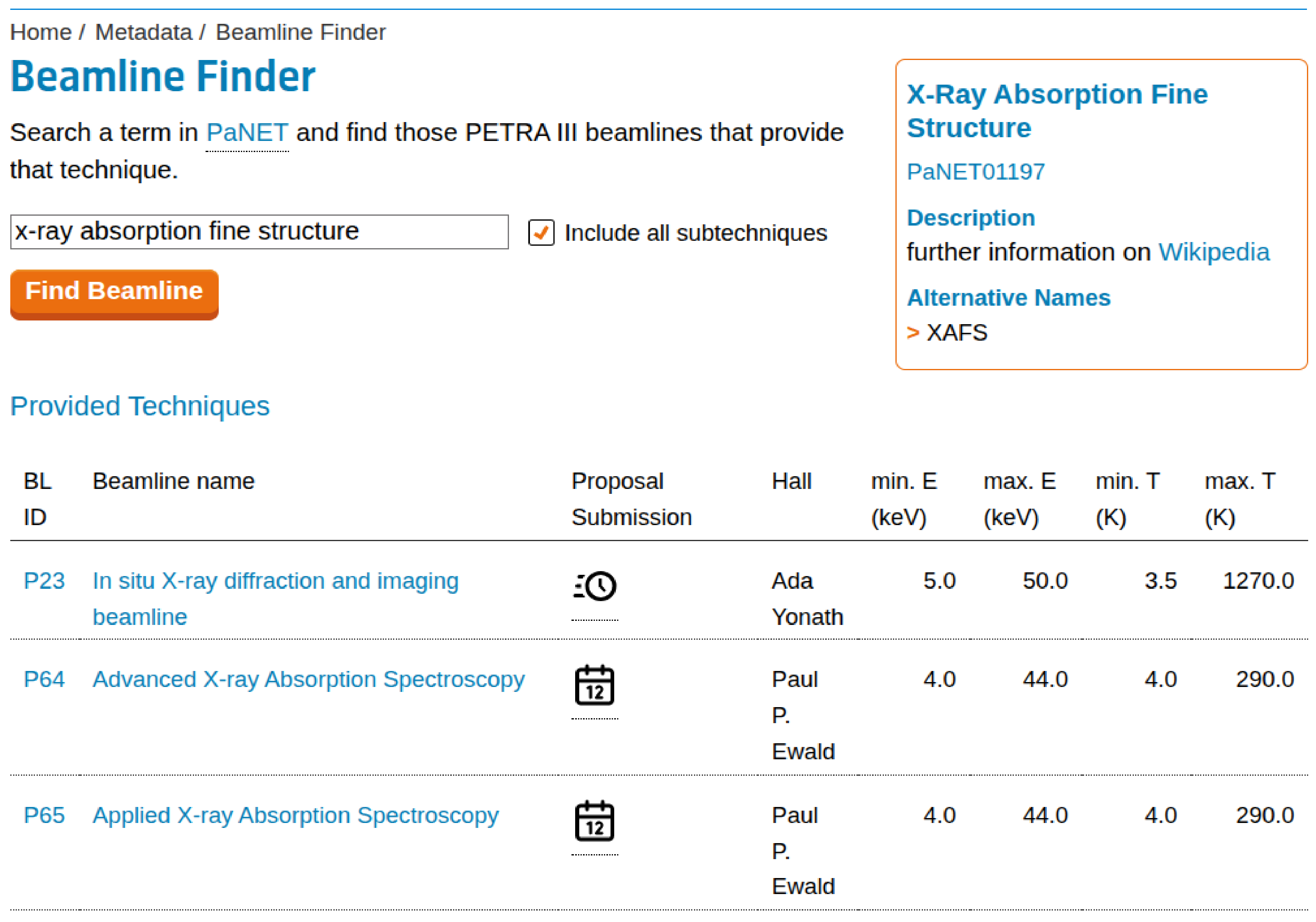
http://purl.org/pan-science/PaNET/

I. Formats and standards for Photon Science

An example application for PaNET

Developed by Melanie in the frame of

OSCARS



Home / Metadata / Beamline Finder

Beamline Finder

Search a term in PaNET and find those PETRA III beamlines that provide that technique.

x-ray absorption fine structure ☑ Include all subtechniques

Find Beamline

X-Ray Absorption Fine Structure

PaNET01197

Description

further information on Wikipedia

Alternative Names

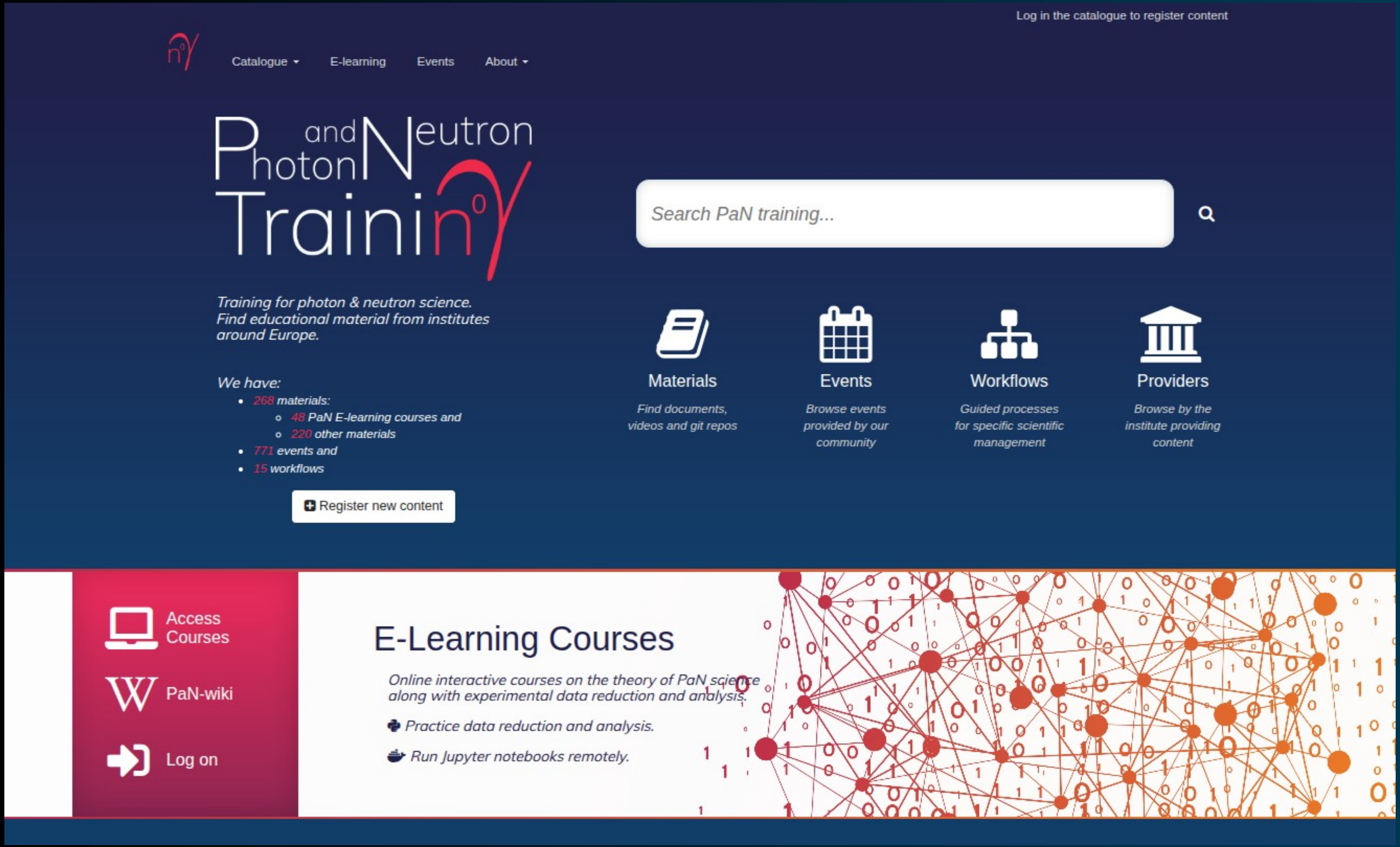
> XAFS

Provided Techniques

BL ID	Beamline name	Proposal Submission	Hall	min. E (keV)	max. E (keV)	min. T (K)	max. T (K)
P23	In situ X-ray diffraction and imaging beamline	🕓	Ada Yonath	5.0	50.0	3.5	1270.0
P64	Advanced X-ray Absorption Spectroscopy	12	Paul P. Ewald	4.0	44.0	4.0	290.0
P65	Applied X-ray Absorption Spectroscopy	12	Paul P. Ewald	4.0	44.0	4.0	290.0

I. Training material for PaN

A few words on the PaN training portal <pan-training.eu>



Training for photon & neutron science. Find educational material from institutes around Europe.

We have:

- *268 materials:*
  - *48 PaN E-learning courses and*
  - *220 other materials*
- *771 events and*
- *15 workflows*

Materials: *Find documents, videos and git repos*

Events: *Browse events provided by our community*

Workflows: *Guided processes for specific scientific management*

Providers: *Browse by the institute providing content*

E-Learning Courses

Online interactive courses on the theory of PaN science along with experimental data reduction and analysis.

Practice data reduction and analysis.

Run Jupyter notebooks remotely.

Further developed in the frame of

OSCARS

mTeSS-X project, enhancing its interoperability with other content aggregators.

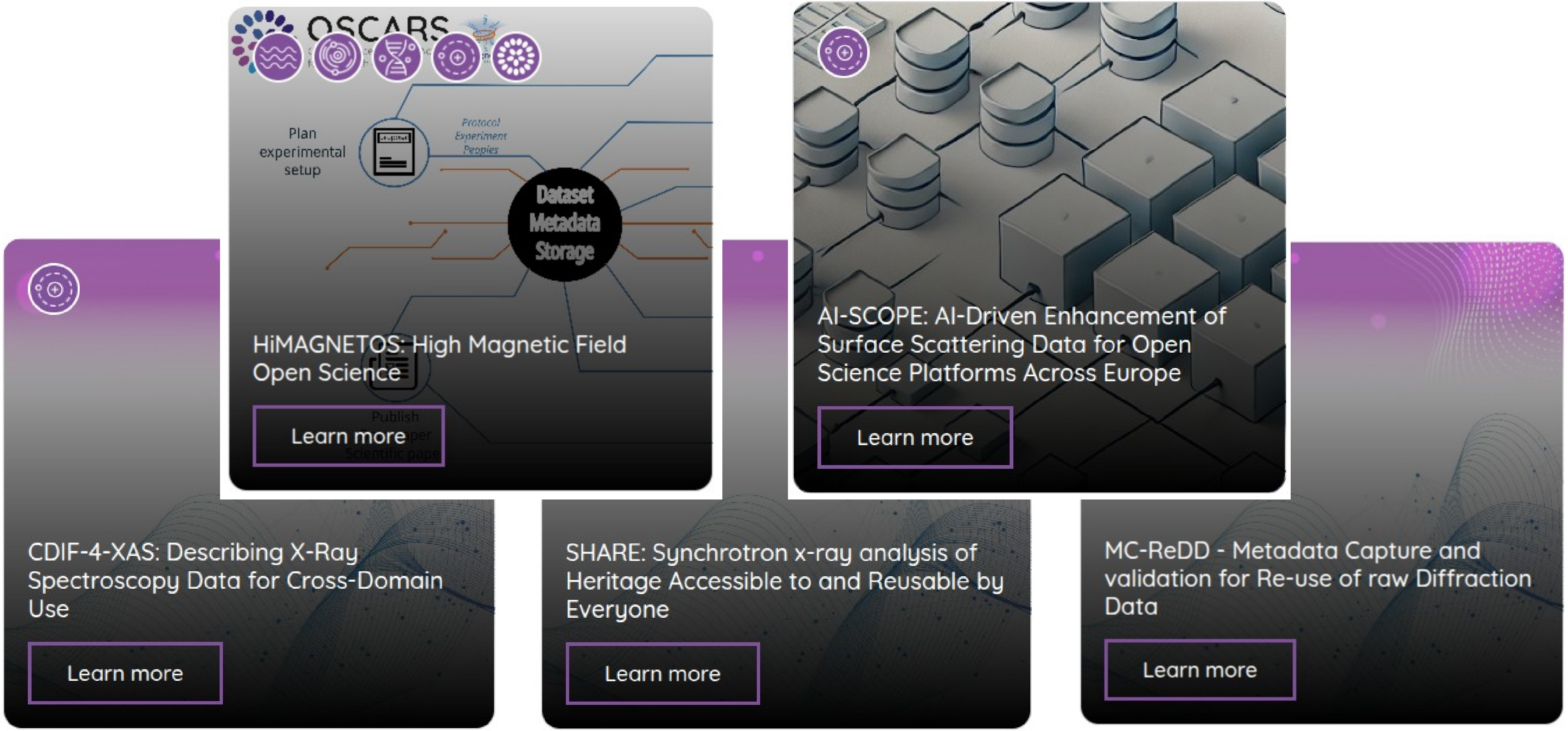
I. Formats and standards for Photon Science

Community-specific endeavours towards interoperability

OSCARS *https://oscars-project.eu/projects/*

DAPHNE 4NFDI

ORSO



HiMAGNETOS: High Magnetic Field Open Science

Learn more

AI-SCOPE: AI-Driven Enhancement of Surface Scattering Data for Open Science Platforms Across Europe

Learn more

CDIF-4-XAS: Describing X-Ray Spectroscopy Data for Cross-Domain Use

Learn more

SHARE: Synchrotron x-ray analysis of Heritage Accessible to and Reusable by Everyone

Learn more

MC-ReDD - Metadata Capture and validation for Re-use of raw Diffraction Data

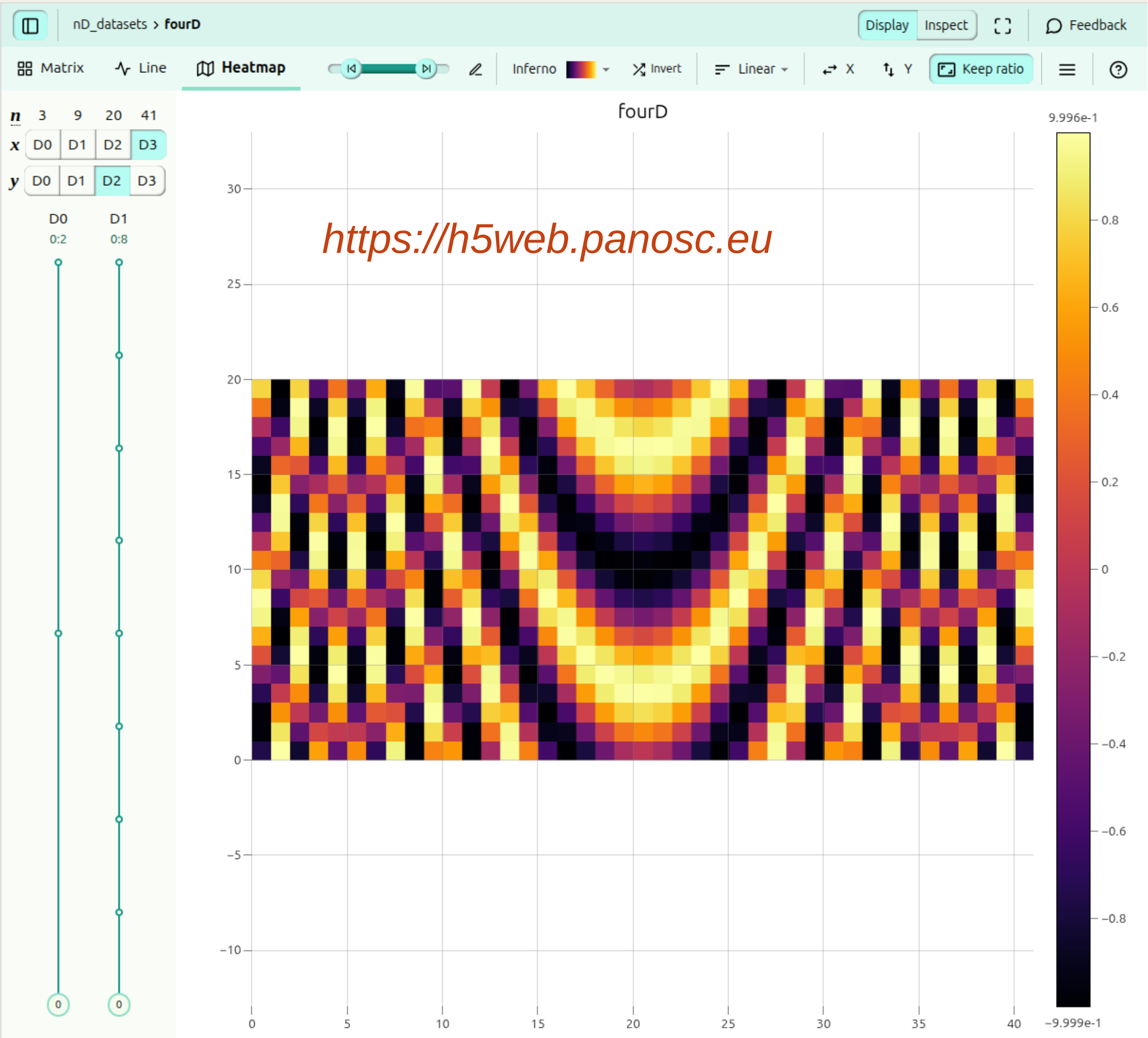
Learn more

R. Handling big data to make it reusable

PDB files are a few KB for small proteins, several MB for large multi-chain assemblies. The wwPDB policy states that data files contained in the PDB archive are available under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.

R. Reusable Data: the R in FAIR

Data volumes in Photon Science and solutions



https://h5web.panosc.eu

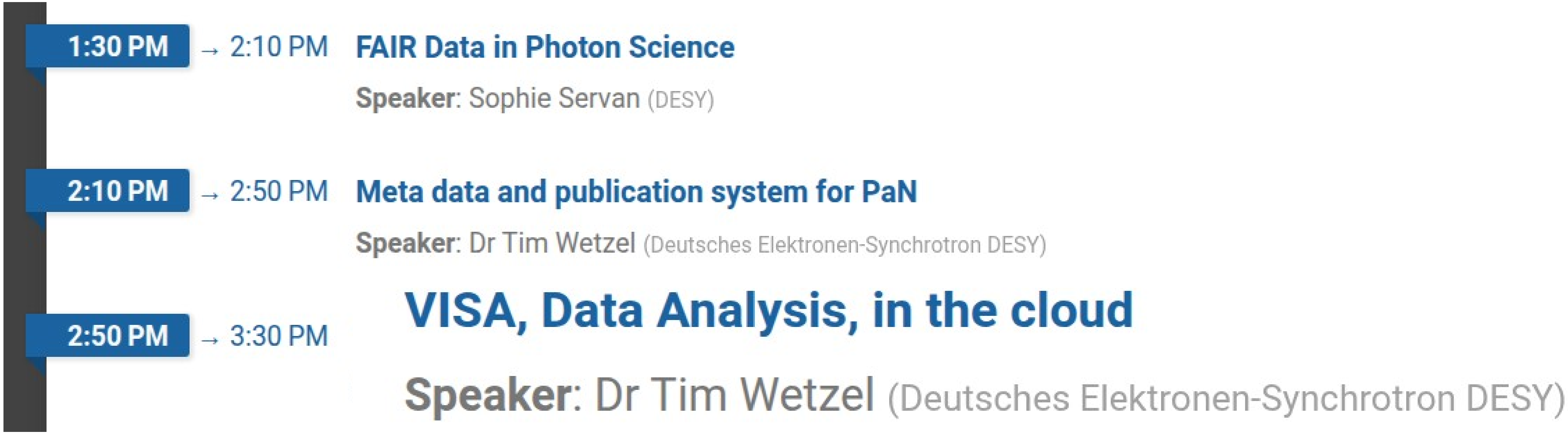
FEL experiments generate the largest per-run data volumes in photon science: **PB-scale** vs. TB-scale for synchrotrons. Until now, users mostly need to be on site.

Working on off-premises access and reuse

Tools such as online visualisation and data slicing open up new perspectives for data reuse in PaN, transferring only the data of interest.

R. Reusable Data: the R in FAIR

Data Analysis Services in second Tim’s talk right after me



1:30 PM → 2:10 PM **FAIR Data in Photon Science**

Speaker: Sophie Servan (DESY)

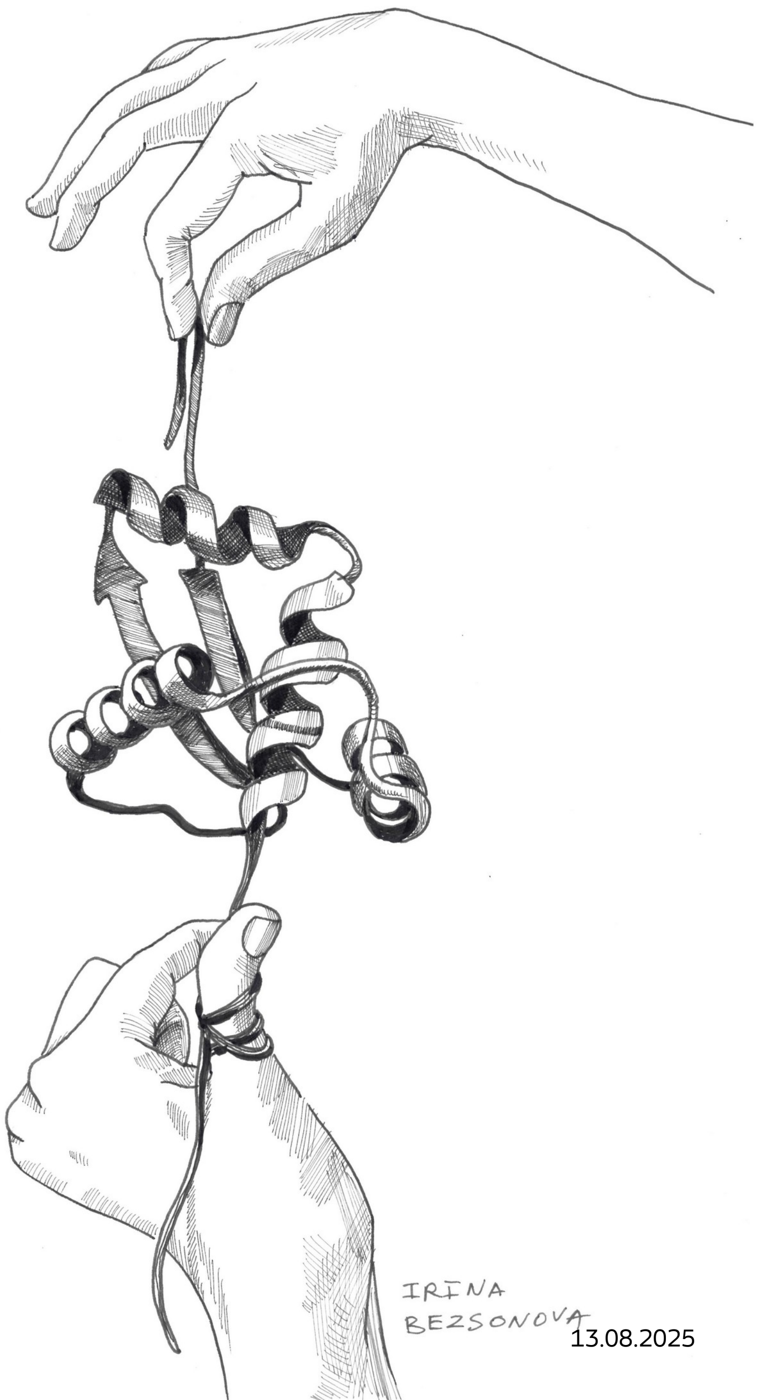
2:10 PM → 2:50 PM **Meta data and publication system for PaN**

Speaker: Dr Tim Wetzel (Deutsches Elektronen-Synchrotron DESY)

2:50 PM → 3:30 PM **VISA, Data Analysis, in the cloud**

Speaker: Dr Tim Wetzel (Deutsches Elektronen-Synchrotron DESY)

Image by Irina Bezsonova, free for use under a CC-BY-4.0 licence.



FAIR Data in Photon Science

Making AlphaFold-level breakthroughs the rule

Importance of **Data Managers** to make sure data usage complies with our policies and standards are kept up-to-date. NIAC, RDA, NFDI, LEAPS.

Several outcomes of previous EU projects have become **long-term resources** for PaN: data.panosc.eu, PaNET, PaN-training.eu, SciCat, VISA.

LEAPS WG3 is a good vehicle for the necessary coordination effort.

Direct data access is our next big task.

Thank you.

Acknowledgements

The images from Irina Bezsonova were taken from *PDB-101, the training and outreach portal of the Protein Data Bank*.

p1, p7 and p22: pictures from DESY.

European Commission for funding *ExPaNDS*, *PaNOSC* and *OSCARS*.