

ESRF

10 years of Open Data management effort.

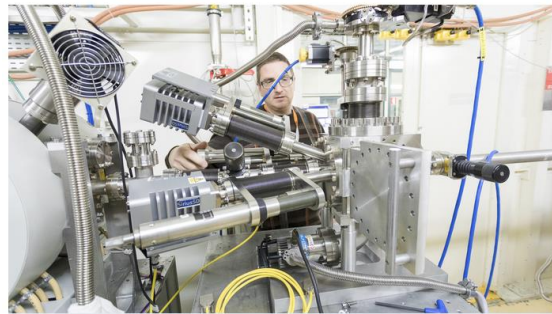
An international collaboration of 22 partners

Low-emittance, high-energy synchrotron light source (4th Generation)

51 Instruments

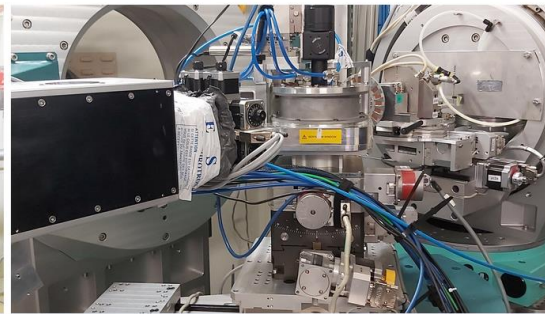
150+ different techniques

Diverse scientific community: energy, materials, biology, health, cultural heritage, ...



03-10-2022
New versatile instrument for magnetism research at ID12

→ [Read more](#)



03-10-2022
New chemical vapour deposition reactor at ID10

→ [Read more](#)



03-10-2022
ID18 offers new possibilities for Synchrotron Mössbauer Source spectroscopy

→ [Read more](#)

2010 – Community (Photon and Neutron User Facilities - PaN) policy framework on scientific data

2011 – First adoption in the community by Neutron facilities

2015 – Open Data Policy validated by ESRF council

- Open data after a 3 embargo period
- Archiving **Raw Data** and Metadata
- Implementation completion expected by 2020

The ESRF Data Policy

The ESRF aims to implement a Data Policy starting as soon as possible in 2016. The main elements of this policy comprise:

- Data ownership
- Data curation
- Data archiving
- Open access to data

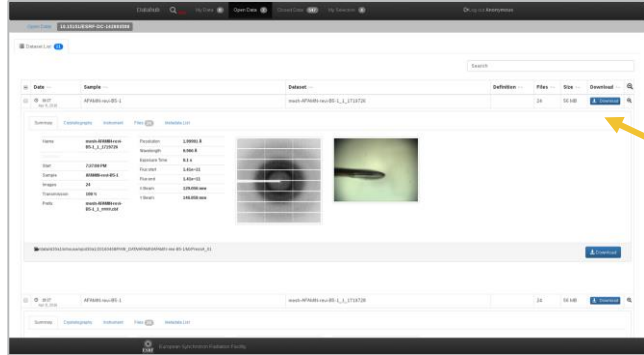
This policy follows largely the recommendations of the PaN-data Europe Strategic Working Group laying out a common framework for scientific data management at photon and neutron facilities (Deliverable D2.1, PaN-data Europe, co-funded by the European Commission under the 7th Framework Programme)

1. General Principles

- 1.1. The present data management policy pertains to the ownership of, the curation of and access to experimental data and metadata collected and/or stored at the ESRF.
- 1.2. Acceptance of this policy is a condition for the award of beam time.
- 1.3. Users must not attempt to access, exploit or distribute raw data or metadata unless they are entitled to do so under the terms of this policy.
- 1.4. Deliberate infringements of the policy may lead to denial of access to raw data or metadata and/or denial of future beam time requests at the ESRF, as well as actions of the ESRF in the court of law.
- 1.5. All data and metadata will be subject to the data protection legislation of France.

2. Definitions

For the purposes of this policy:



Data Catalogue
<https://data.esrf.fr>

**HDF5
Nexus**



Data Service
Explore and Download data

Data Catalogue

Download Data Service

Assessment Information

[CoreTrustSeal Requirements 2020–2024](#)

Repository: European Synchrotron Radiation Facility
Website: <https://data.esrf.fr>
Certification period: June 1, 2023 - 31 May 2026
Requirements version: CoreTrustSeal Requirements 2020-2022

This repository is owned by: **ESRF**



Search Service

for Big Data



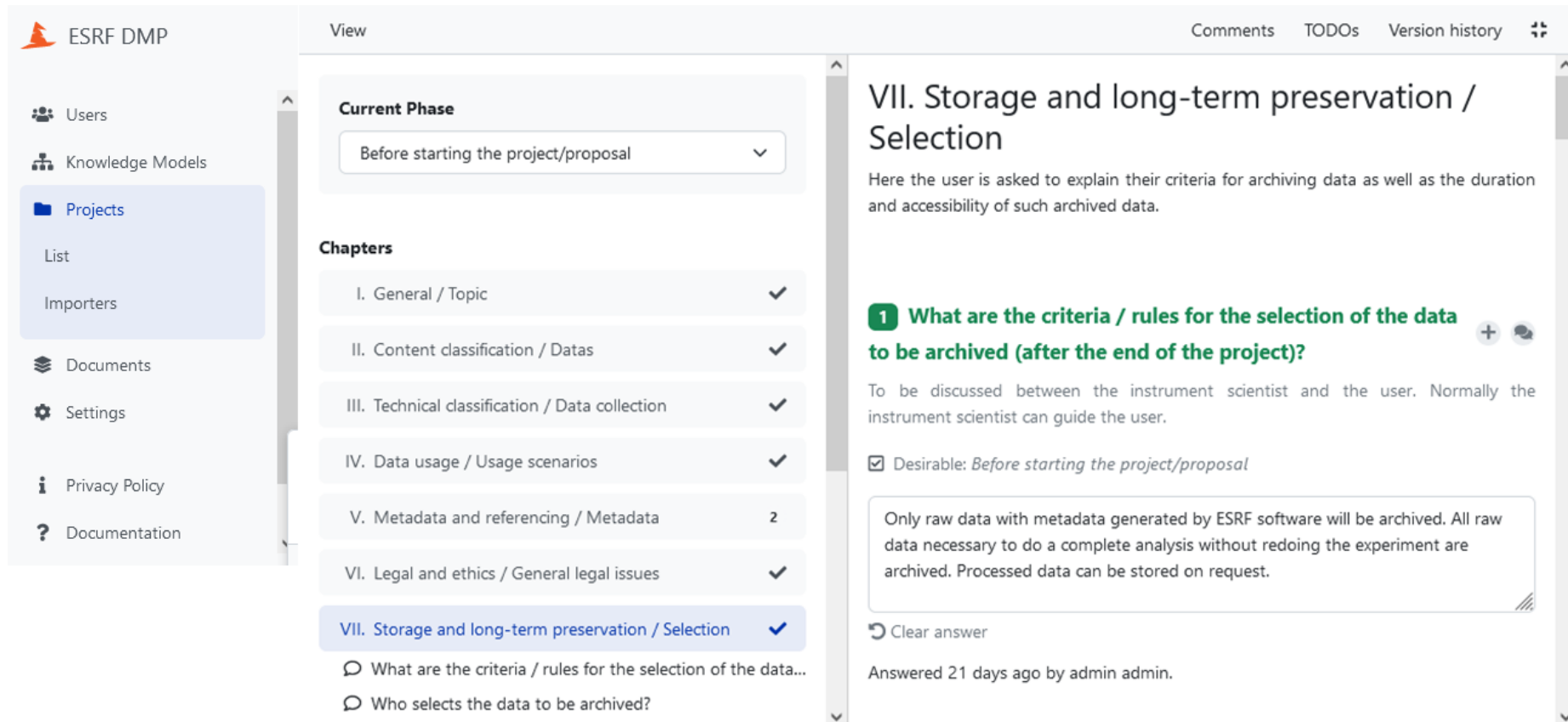
Identifiers
Data findable

Annotate your experiments

and searchable

DATA MANAGEMENT PLANS – WHAT FOR?

- ESRF has implemented **Data Management Plans** – <https://dmp.esrf.fr> + linked from data portal
- **DMPs are pre-filled for all proposals** for users following the data policy + statistics of data produced
- **DMP service** is available for projects requiring a **DMP for the funders** e.g. **Horizon Europe, ANR, ...**
- **Next step** : (1) provide a summary of the DMP, (2) request users to fill in DMPs for certain experiments



The screenshot displays the ESRF DMP (Data Management Plan) web interface. On the left is a navigation sidebar with options: Users, Knowledge Models, Projects (selected), List, Importers, Documents, Settings, Privacy Policy, and Documentation. The main content area is titled 'View' and shows the 'Current Phase' as 'Before starting the project/proposal'. Below this is a list of 'Chapters' with checkboxes and counts:

- I. General / Topic ✓
- II. Content classification / Datas ✓
- III. Technical classification / Data collection ✓
- IV. Data usage / Usage scenarios ✓
- V. Metadata and referencing / Metadata 2
- VI. Legal and ethics / General legal issues ✓
- VII. Storage and long-term preservation / Selection ✓

Chapter VII is expanded to show a question: '1 What are the criteria / rules for the selection of the data to be archived (after the end of the project)?'. The text below the question explains that users should discuss criteria with instrument scientists. A checkbox is checked for 'Desirable: Before starting the project/proposal'. A text box contains the answer: 'Only raw data with metadata generated by ESRF software will be archived. All raw data necessary to do a complete analysis without redoing the experiment are archived. Processed data can be stored on request.' The interface also shows a 'Clear answer' button and a comment: 'Answered 21 days ago by admin admin.' At the top right of the main area are links for 'Comments', 'TODOs', and 'Version history'.

The screenshot displays a log of commands in the OPTICS environment. The date is November 4th 2018. The log includes the following entries:

- 00:16:17 OPTICS> New dataset: fe2streptor2_main_root
- 23:59:17 OPTICS> zapxiaimage samy 8.064 7.124 94 samz 25.426 26.056 63 60 0 (zap: #1, spec: #1)
- 23:58:41 OPTICS> New dataset: fe2streptor2_coarse
- 23:58:38 OPTICS> New sample: fe2streptor2 ("Fe2 strepto replica 2")
- 23:58:38 OPTICS> New spec file: /data/visitor/ev280/id21/fe2streptor2/fe2streptor2_spec01/fe2streptor2_spec01.dat
- 23:58:38 OPTICS> New dataset: fe2streptor2_spec01
- 23:53:58 no new data collected
- 23:50:43 OPTICS> moveE 7.2
- 23:50:29 OPTICS> Fexanes_ev280
- 23:03:53 OPTICS> New dataset: fe3prior3_sec_root
- 23:03:26 OPTICS> moveE 7.2
- 22:55:29 OPTICS> New dataset: fe3prior3_sr_coarse
- 22:48:31

← Notify commands

← Notify errors

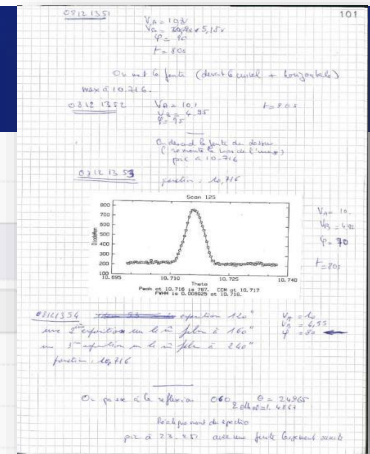
← Comment on a command

Annotations

- Created by humans
- Editable

Notifications

- Created by software:
 - Bliss
 - SPEC
 - Flint
- It can only be commented



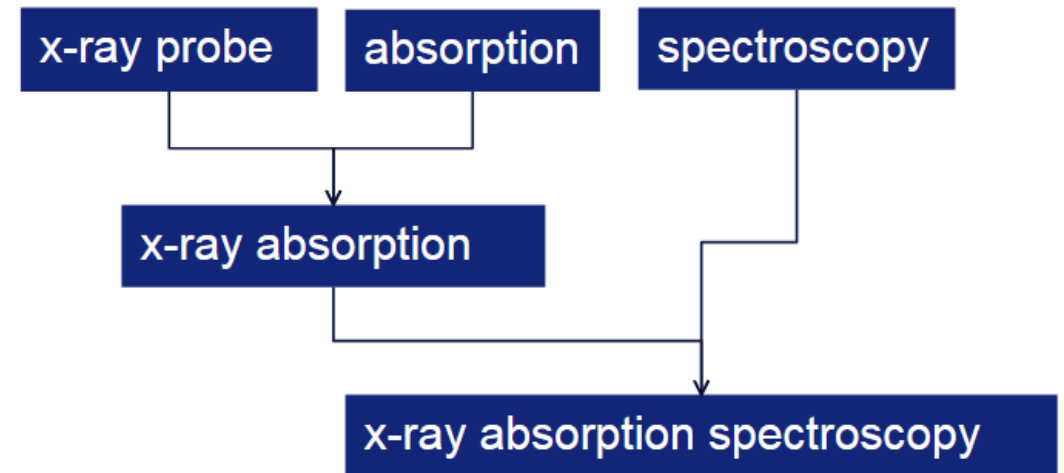
The PaNET ontology provides a taxonomy (classification) and thesaurus (synonyms and acronyms) of Photon and Neutron Experimental Techniques.







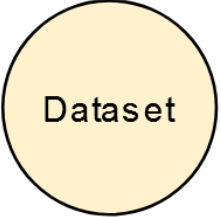











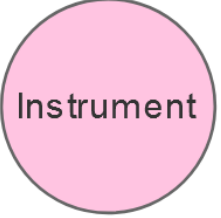







EXAFS	EXAFS	extended x-ray absorption fine structure	PaNET01198
FT-SSX	FT-SSX	fixed target serial synchrotron crystallography	PaNET01310
FTIR	FTIR	fourier transform infrared spectroscopy	PaNET01320
GISANS	GISANS	grazing incidence small angle neutron scattering	PaNET01276
GISAS	GISAS	grazing incidence small angle scattering	PaNET01099
GISAXS	GISAXS	grazing incidence small angle x-ray scattering	PaNET01162
GIWAXS	GIWAXS	grazing incidence wide angle scattering	PaNET01316
GIXD	GIXD	grazing incidence x-ray diffraction	PaNET01161
HAXPES	HAXPES	hard x-ray photoelectron spectroscopy	PaNET01103
WAXS	WAXS	wide angle x-ray scattering	PaNET01191
XAFS	XAFS	x-ray absorption fine structure	PaNET01197
XANES	XANES	x-ray absorption near edge structure	PaNET01199
XAS	XAS	x-ray absorption spectroscopy	PaNET01196

PaNET Ontology (not just taxonomy but relations as well)

- x-ray probe (PaNET01012)
- spectroscopy (PaNET01125)
- absorption technique (PaNET00202)
- x-ray absorption (PaNET01227)
- x-ray absorption spectroscopy (PaNET01196) - XAS



PIDS, VOCABULARIES, ONTOLOGIES...

	 Institution	 ESRF : https://ror.org/02550n020		 Dataset	 ESRF data doi : https://doi.org/10.15151/ESRF-DC-572196058
	 Person	 G. Bouchard : 0000-0003-0705-6026		 Grant/ Proposal	
	 Publication	 https://doi.org/10.1038/s41592-021-01317-x		 Instrument	
	 Journal	 Nature Communications : 2041-1723		 PaN Technique	 PaNET vocabulary/ontology Photon and Neutrons Experiment Techniques SAXS - Small Angle X-ray Scattering : http://purl.org/pan-science/PaNET/PaNET01188



WHERE IS OUR PROCESSED DATA ?

Human Organ Atlas EXPLORE SEARCH TUTORIALS 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](#)

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**


Team

- UCL, London, England: **Peter D Lee, Claire Walsh, David Stansby**
- ESRF, Grenoble, France: **Paul Tafforeau, Guillaume Gaisne, Joseph Brunet, Theresa Urban, Axel Bocciarelli**

Related projects

HOAHub is the wider project that organises and manages data collection for the Human Organ Atlas.

Research Institutions
species geographic distributions

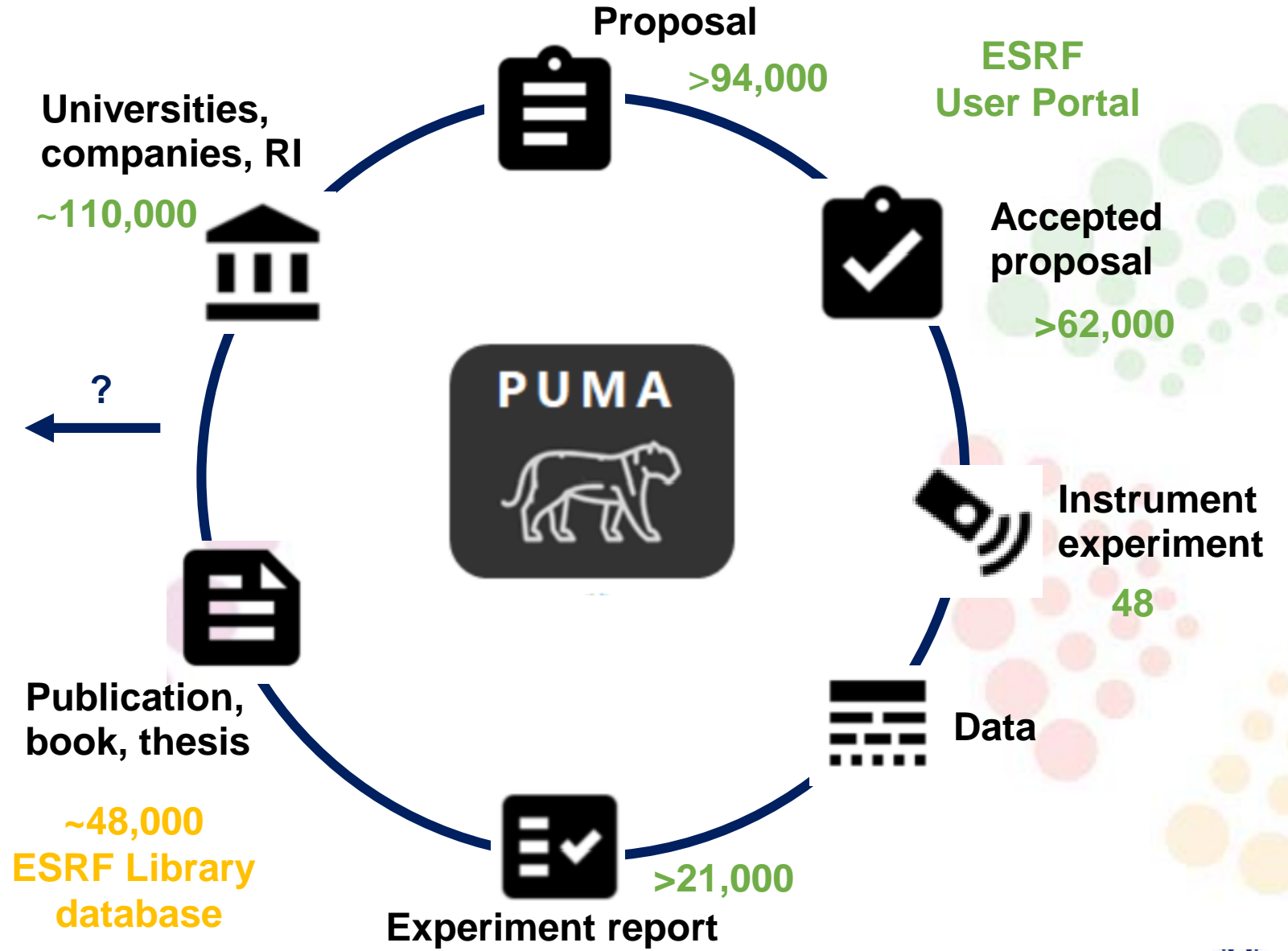
 **Paleobiology Database (PBDB)**

Museum collections



ESRF SCIENTIFIC ACTIVITY IN DETAILS: PUMA



DATAcite API METADATA – RELATED IDENTIFIERS

DataCite Metadata Properties

Overview	10. ResourceType
1. Identifier	11. AlternateIdentifier
2. Creator	12. RelatedIdentifier
3. Title	13. Size
4. Publisher	14. Format
5. PublicationYear	15. Version
6. Subject	16. Rights
7. Contributor	17. Description
8. Date	18. GeoLocation
9. Language	19. FundingReference
	20. RelatedItem

Controlled List Values:

- **IsCitedBy**
- Cites
- IsSupplementTo
- IsSupplementedBy
- IsContinuedBy
- Continues
- IsDescribedBy
- Describes
- HasMetadata
- IsMetadataFor
- HasVersion
- IsVersionOf
- IsNewVersionOf
- IsPreviousVersionOf
- IsPartOf
- HasPart
- IsPublishedIn
- IsReferencedBy
- References
- IsDocumentedBy
- Documents
- IsCompiledBy
- Compiles
- IsVariantFormOf
- IsOriginalFormOf
- IsIdenticalTo
- IsReviewedBy
- Reviews
- **IsDerivedFrom**
- IsSourceOf
- IsRequiredBy
- Requires
- IsObsoletedBy
- Obsoletes
- **IsCollectedBy**
- Collects

relatedIdentifierType

- [ARK](#)
- [arXiv](#)
- [bibcode](#)
- [DOI](#)
- [EAN13](#)
- [EISSN](#)
- [Handle](#)
- [IGSN](#)
- [ISBN](#)
- [ISSN](#)
- [ISTC](#)
- [LISSN](#)
- [LSID](#)
- [PMID](#)
- [PURL](#)
- [UPC](#)
- [URL](#)
- [URN](#)
- [w3id](#)

```
"relatedIdentifiers": [  
  {  
    "relatedIdentifier": "https://doi.org/10.1038/yyyy",  
    "relatedIdentifierType": "DOI",  
    "relationType": "IsCitedBy",  
    "resourceTypeGeneral": "JournalArticle"  
  }  
]
```











```
"relatedIdentifiers": [  
  {  
    "relatedIdentifier": "https://doi.org/10.15151/ESRF-INSTR-123456",  
    "relatedIdentifierType": "DOI",  
    "relationType": "IsCollectedBy",  
    "resourceTypeGeneral": "Instrument"  
  }  
]
```



```
"relatedIdentifiers": [  
  {  
    "relatedIdentifier": "https://www.gbif.org/fr/occurrence/1699073315",  
    "relatedIdentifierType": "URL",  
    "relationType": "IsDerivedFrom",  
    "resourceTypeGeneral": "PhysicalObject"  
  }  
]
```



ESRF-EBS DETECTORS

Detector	Type	Max. data rate	# beamlines
	PCO EDGE	1 GB/s	10
	PCO DIMAX	8 GB/s	2
	Eiger2	4 GB/s	8
	PSI Eiger	2 GB/s	2
	Pilatus	2 GB/s	7
	Frelon	1 GB/s	8
	Jungfrau	9 GB/s	1
	Medipix	1 GB/s	6

2024 - 2025 :

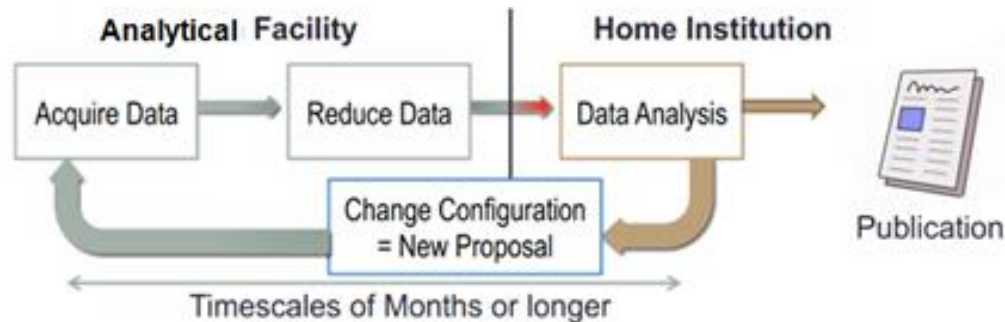
- Pilatus4 (x3)
- Jungfrau 1M
- Jungfrau upgrade
- Sensicam
- Citius
- Rigaku

...

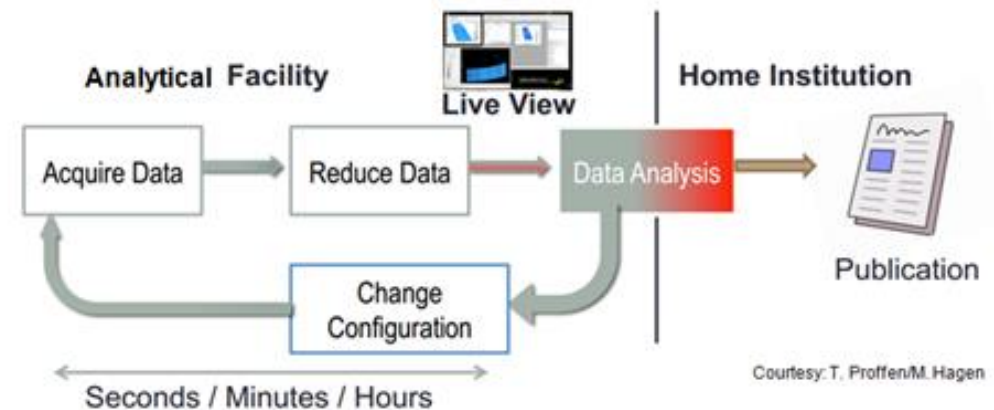
But we need processed data and we need it fast since we acquire data at alarmous rate

- To make informed decisions during acquisition
- To provide to users to help/encourage further data analysis

To go from this...



...to this



Courtesy: T. Proffen/M. Hagen

VISA - REMOTE DATA PROCESSING/ANALYSES PLATFORM



New compute instance

Please fill in the details below to create a new compute instance


Experiments


Select the experiments you wish to associate with your instance


Instance not associated to any specific experiments

Computing Environment

Choose an environment


Desktop staging


Desktop


Bliss

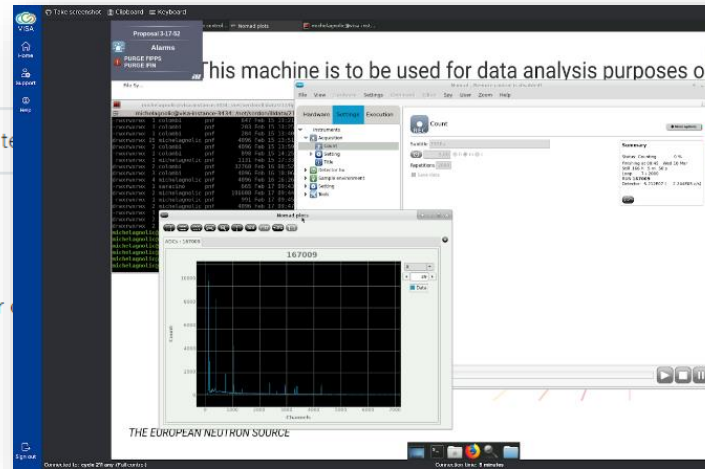
Choose hardware requirements

4 Cores
4GB memory
esrf.medium

8 Cores
16GB memory
esrf.large

16 Cores
32GB memory
esrf.gpu.a40

32 Cores
128GB memory
esrf.gpu.a40.xlarge



Infrastructure for remote data processing / analysis

Users dedicated VM

Access to data

Access to Provisioning of scientific SW using CVMFS and Containers

Access to the GPUs, HPC cluster

Infrastructure based on OpenStack

Development led by ILL in the scope of the PaNOSC project



PROCESSED DATA VISUALIZATION

Data Portal Data Logistics Instruments My selection 1 datasets 1 samples My Jobs 41 Search investigation... Feedback Alejandro DE MARIA

Home / OH-OH ID23-1 04/10/2023-05/10/2023 / Datasets

Investigation

- Experiment
- Statistics
- Datasets 296
- Jobs 18
- Logbook
- Prepare

Datasets

View as List Summary

Sample Select sample

MX

Groups collection workflow Show actions

Ranking shell Overall

Ranking parameter I/s(l)

Sample A

Best result for sample A

Quality indicator

Best auto processing [Show all](#)

Friedel pairs unmerged
Monoclinic system (P2)

	a	b	c			
	71.4 Å	120.9 Å	80.5 Å			
β	112.3°					
Compl.	Res. low	Res. high	Rmerge	I/s(l)	cc1/2	ccAno
inner	99.7%	120.9	8.9	4.7	26.3	1.0
outer	97.1%	3.1	3.0	238.2	0.6	0.2
overall	99.3%	120.9	3.0	13.6	5.4	1.0

Datacollection 21:51:36

Sample snapshot **Diffraction** **Quality indicator** **Best auto processing** [Show all](#)

Friedel pairs unmerged
Monoclinic system (P2)

	a	b	c			
	71.4 Å	120.9 Å	80.5 Å			
β	112.3°					
Compl.	Res. low	Res. high	Rmerge	I/s(l)	cc1/2	ccAno
inner	99.7%	120.9	8.9	4.7	26.3	1.0
outer	97.1%	3.1	3.0	238.2	0.6	0.2
overall	99.3%	120.9	3.0	13.6	5.4	1.0

X-ray centering on id23eh1 21:48:13

Diffraction **Mesh** **Line**

Characterisation 21:46:12

Sample snapshot **Diffraction** **Characterisation** **Quality indicator**

Data Portal Data Logistics Instruments My selection 1 datasets 1 samples My Jobs 41 Search investigation... Feedback Alejandro DE MARIA

Investigation

- Experiment
- Statistics
- Datasets 296
- Jobs 18
- Logbook
- Prepare

Datasets

View as List Summary

Sample Select sample

MX

Groups collection workflow Show actions

Ranking shell Overall

Ranking parameter I/s(l)

datacollection 01/10/2023 21:51:36

Sample snapshot **Diffraction** **Quality indicator** **Best auto processing** [Show all](#)

Friedel pairs unmerged
Monoclinic system (P2)

	a	b	c			
	71.4 Å	120.9 Å	80.5 Å			
β	112.3°					
Compl.	Res. low	Res. high	Rmerge	I/s(l)	cc1/2	ccAno
inner	99.7%	120.9	8.9	4.7	26.3	1.0
outer	97.1%	3.1	3.0	238.2	0.6	0.2
overall	99.3%	120.9	3.0	13.6	5.4	1.0

datacollection 05/10/2023 21:51:52

datacollection 05/10/2023 21:50:46

datacollection 05/10/2023 21:50:46

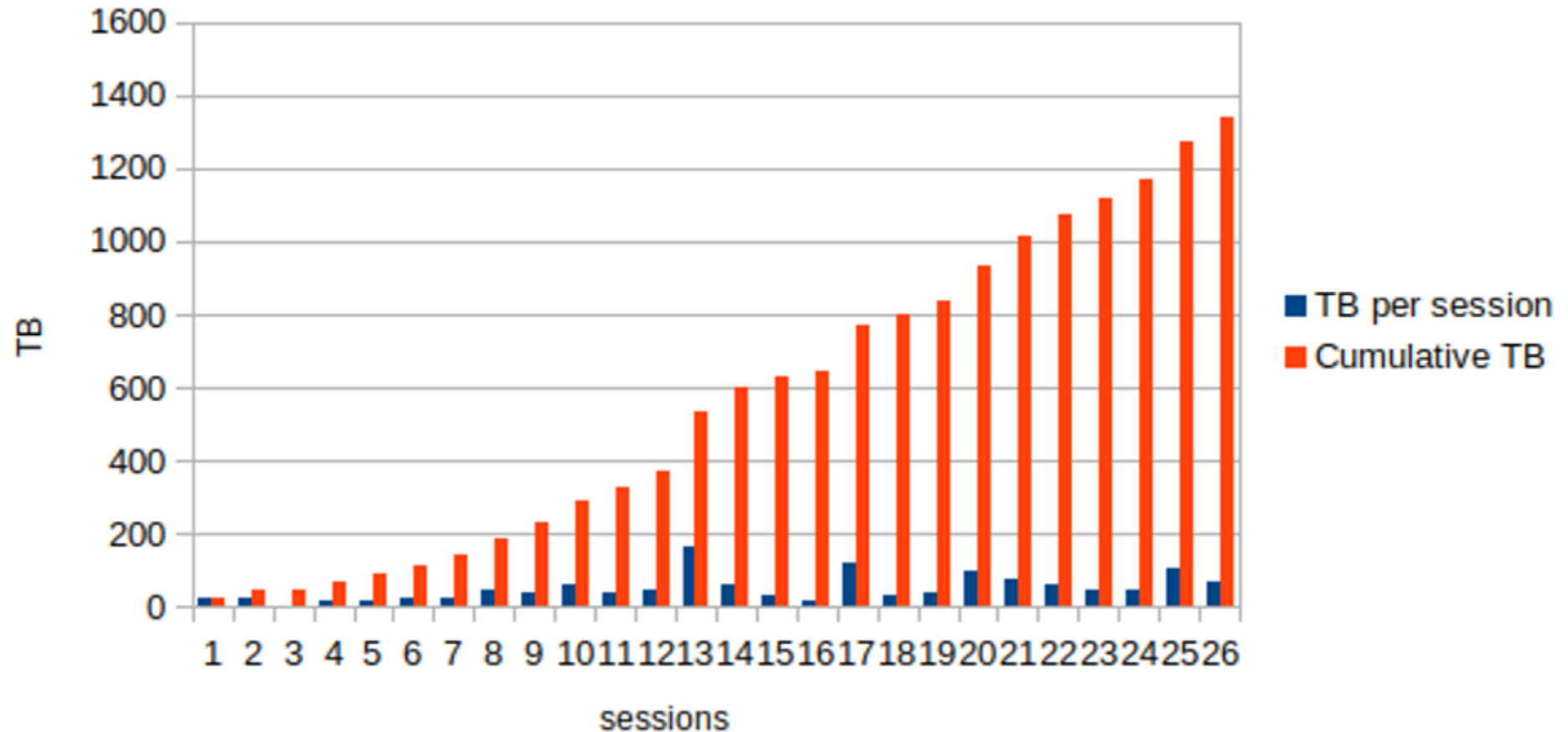
FAIR data

Configurable

EBS DATA PRODUCTION – ID29 SCALING UP ...

ESRF-EBS serial crystallography beamline (ID29) produced > 1.2 PB in first 6 months
(high priority to switch on data triage and lossy compression)

TB per session vs cumulative TB



EBS DATA PRODUCTION MEASURED

	2020	2021	2022	2023
Raw Data volume	1.3 PB	2.8 PB	3.1 PB	3.6 PB
Datasets	203k	352k	471k	554k
No. files	87m	166m	138m	118m
Ave / Max volume	6.8 GB	8.4 GB	6.9 GB	6.8 GB
	8.1 TB	7.8 TB	8.5 TB	11.1 TB
Ave / Max files	428	474	293	213
	39k	43k	51k	195k
Main Beamline	CM01	ID19	BM18	BM18
	40%	20%	CM01 21%	19%
Archived (tape)	--	3.86 PB	6.52 PB	7.75 PB

ARCHIVING

Mediaflux software

```
235 lines (200 slots) | 7.99 KB
1  #! /usr/bin/perl -w
2  # This file is part of the LibArchive project.
3  # It is subject to the terms of the Mozilla Public
4  # License, v. 2.0. If a copy of the MPL was not distributed with this
5  # file, you can obtain one at http://mozilla.org/MPL/2.0/.
6
7  use Archive::LibArchive;
8  use Archive::LibArchive::IO;
9  use Archive::LibArchive::IO::File;
10 use Archive::LibArchive::IO::File::File;
11 use Archive::LibArchive::IO::File::File;
12 use Archive::LibArchive::IO::File::File;
13 use Archive::LibArchive::IO::File::File;
14 use Archive::LibArchive::IO::File::File;
15 use Archive::LibArchive::IO::File::File;
16 use Archive::LibArchive::IO::File::File;
17 use Archive::LibArchive::IO::File::File;
18 use Archive::LibArchive::IO::File::File;
19 use Archive::LibArchive::IO::File::File;
20 use Archive::LibArchive::IO::File::File;
21 use Archive::LibArchive::IO::File::File;
22 use Archive::LibArchive::IO::File::File;
23 use Archive::LibArchive::IO::File::File;
24 use Archive::LibArchive::IO::File::File;
25 use Archive::LibArchive::IO::File::File;
```

ARCITECTA®



Disk storage



SpectraLogic BlackPearl object storage & tape library gateways



SpectraLogic T-Finity tape libraries & drives



2023: UPDATED ESRF DATA POLICY

After 8 years of the initial Data Policy the **GOAL** of the new data policy was to simplify the text, make data **FAIR**, include processed data and enable data triage and compression.

New Data Policy endorsed by Council in Dec 2023.

Summary of changes:

1. A separate policy for **proprietary research**;
2. Inclusion of the possibility of **more extensive data curation**, including **triage** and the use of **lossy compression** (when acceptable), or **limiting the amount of data stored** for very high throughput experiments;
3. Inclusion of **processed data**, which can be published on the data portal and **possibly replace raw data**;
4. Addition of a **reference to FAIR** (Findable, Accessible, Interoperable and Reusable) principles;
5. Mention of the ESRF **data portal** (<https://data.esrf.fr>) as the **main access** to the data;
6. **3-year embargo** rule also to the **experiment report**, which will be included as part of the metadata;
7. **Access to all experimental data by ESRF personnel**, e.g. for algorithm development and performance;
8. Use of a **Digital Object Identifier (DOI)** to reference datasets - the **DOI must be quoted** in any publication exploiting ESRF data;
9. CC-BY-4



<https://www.esrf.fr/datapolicy>

CITING DATA IN PUBLICATIONS

Users **MUST** mint DOIs for data cited – do not use the automatically minted DOIs!
Watch <https://www.youtube.com/watch?v=dPeN855-Mu4>

PDB can link to raw data but Users **MUST** mint bespoke DOIs

Scientists are encouraged to publish raw data in the IUCr Raw Data Journal :

Crystal structure of the second extracellular domain of human tetraspanin D9: twinning and diffuse scattering

The screenshot shows the IUCrData article page for the crystal structure of the second extracellular domain of human tetraspanin D9. The page includes the journal title 'IUCrData', ISSN 2414-3146, and the authors: Viviana Neviani, Martin Lutz, Wout Oosterheert, Piet Gros and Loes Kroon-Batenburg*. The article is dated Received 20 April 2021 and Accepted 1 May 2021. The abstract describes the structure of CD9_{EC2}} crystallized in space group P1 and twinned, with diffuse streaks observed. The article includes two images: 'Raw data' showing a diffraction pattern with diffuse streaks, and 'Structure' showing a ribbon diagram of the protein structure. At the bottom, there is a section for 'Raw diffraction data' with HDF5 data file and Metadata ImgCIF file DOIs.

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:

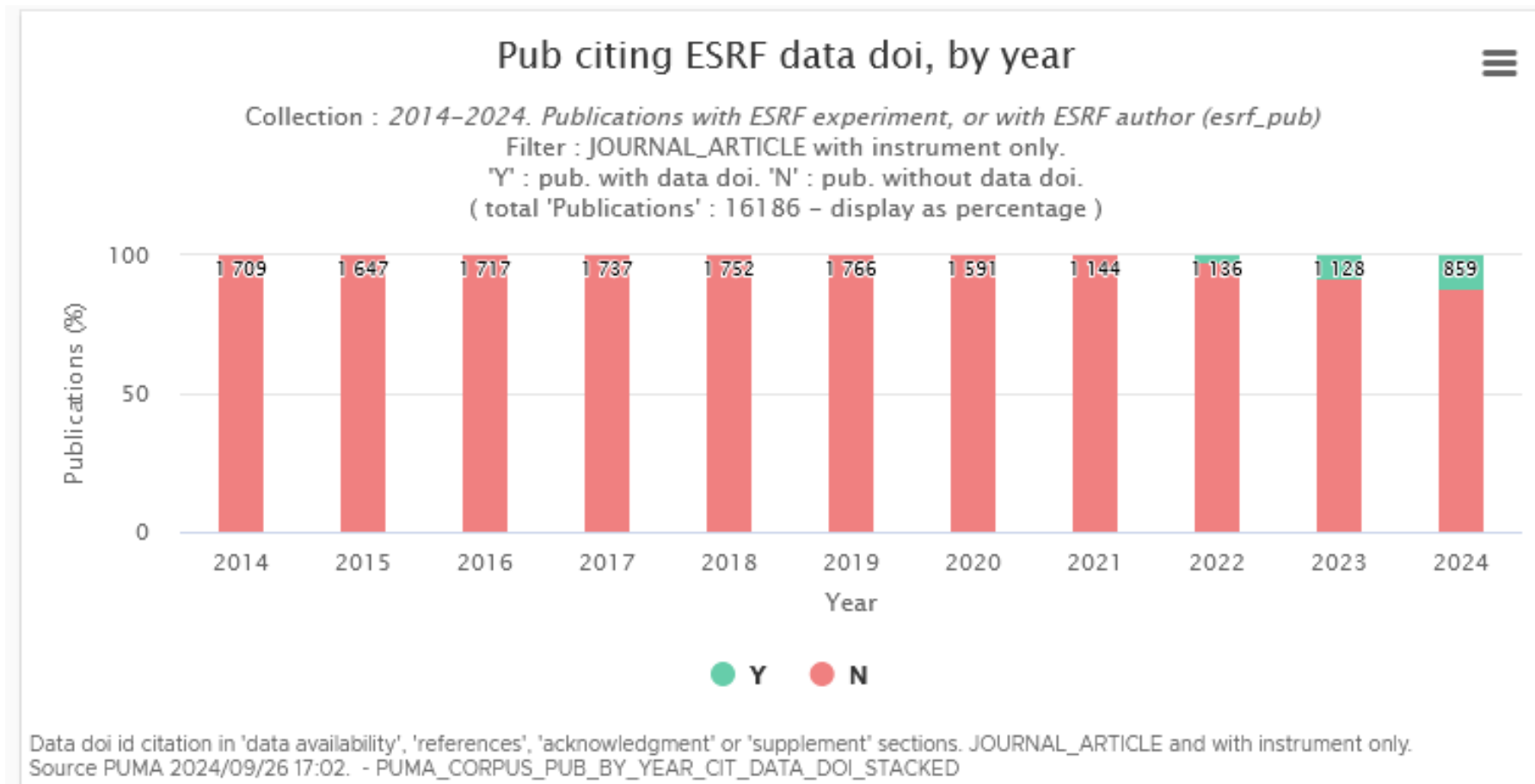
Testing: All archives are accessible: PASS
Running checks with downloaded images
=====

Testing image 4: Image type and dimensions: PASS
Testing image 4: Overloaded values present: PASS
```



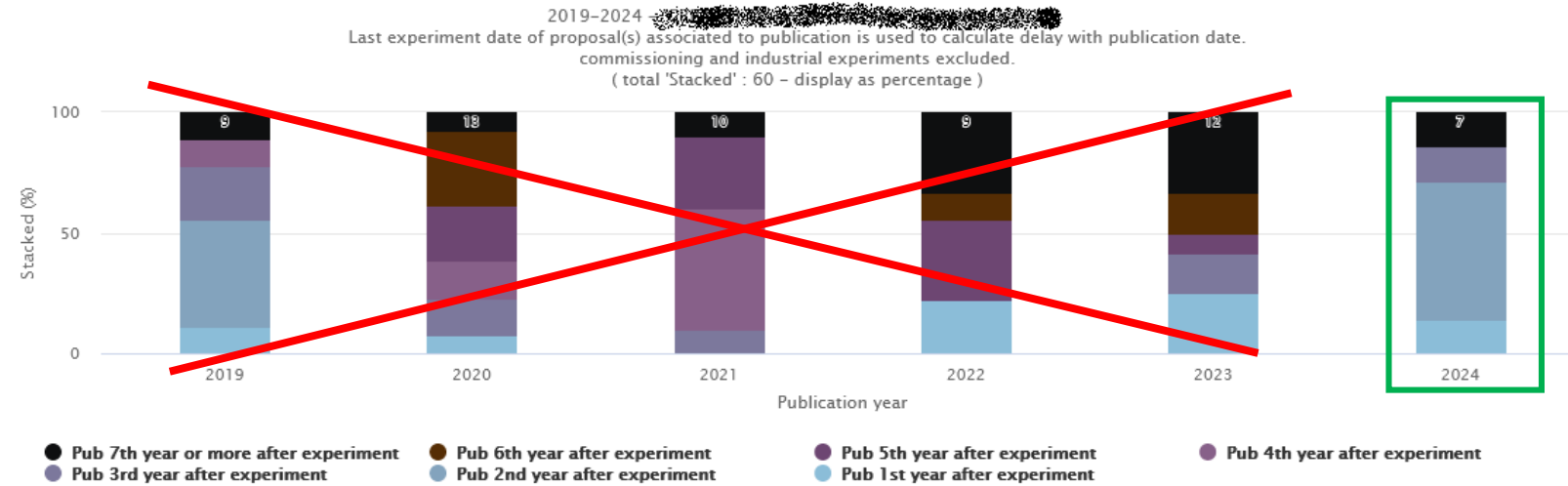
<https://iucrdata.iucr.org/x/>

WHERE IS THE ESRF PROCESSED DATA ??

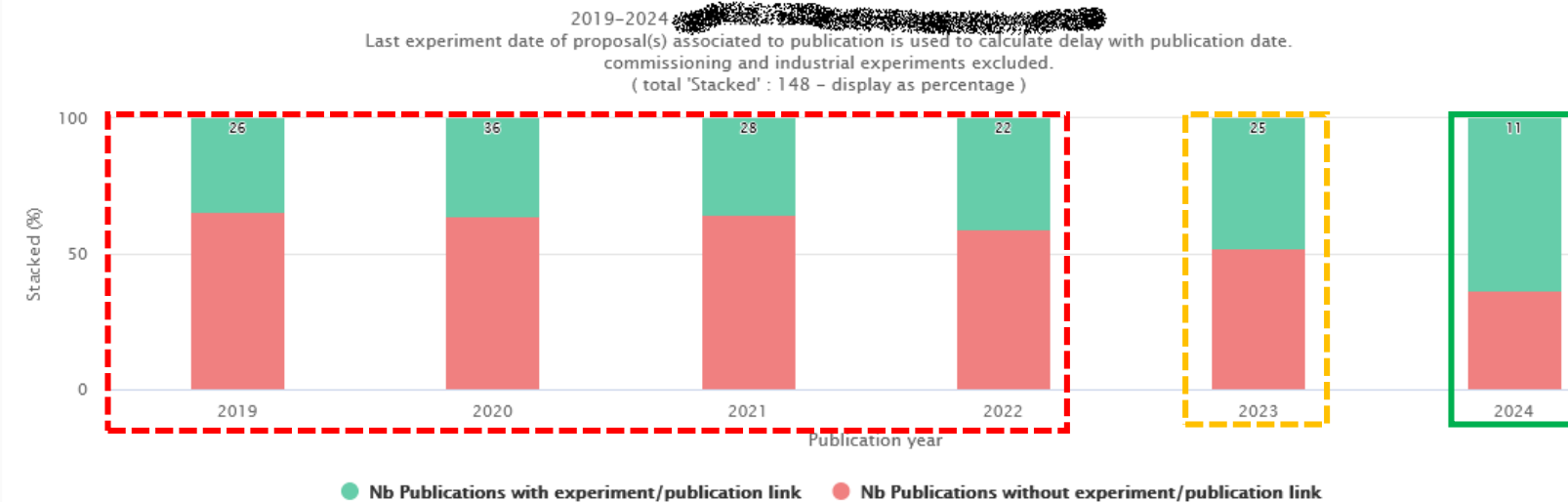


TIME BETWEEN EXPERIMENT AND PUBLICATION

Time between publication and experiment



Percentage of publications with link to experiment



- Library email campaign : we now ask the users for missing proposal reference in publication ex: MX-12345. Flora **library** database is updated.
- Email sent 08/2024 to publication corresponding authors (for 2023 and 2024 publications).
- Good user responses : we see impact on data quality for these years.
- First ESRF DOIs created in 2018 (source Datacite search api)

Archiving is not only ensuring that files are safely stored on tapes

Open data/FAIR data and archiving are structuring data management operation.

In 2024, it has become difficult to identify the cost of Open/FAIR data, this has become the “normal” way of managing data. This is fully integrated into our user service portfolio.

FAIR Data = Quality of Service for RI users.

Costs :

- **Infrastructure and consumables are provided by ESRF budget (tape library, storage, ...) – DP validated by council**
- **4 FTE – Strong support from EU Grants**
- **Specific tools and standards have been developed by the PaN community + adopting open source solutions (ex: DS Wizzard)**

Archiving less

- **Data reduction during acquisition: lossless and lossy compression, data triage ...**
- **Data Processing: identify useless data.**

Deleting archives:

- **10 years max engagement**
- **Some data will deserve a longer preservation**
- **Contact with users before deletion**
- **1B datasets, need for automation**
- **How to identify the value of data (relation to scientific articles, nb of downloads, Scientific Prizes ...)**

Select the data to be preserved longer based on scientific articles value ?

It could be better to replay the experiment in some case ?

How to identify the scientific value of the data?

Could we delete data, if no publication X years after the experiment ?

Thank you for your attention.

This presentation has received support from many colleagues from scientific and technical ESRF departments and PaN (Photon and Neutron RI) community.

Andrew Goetz, Thomas Vincent, Antoine Roux, Cédric Vinet, Marcus Oskarsson, Aidan Campbell, Stuart Fisher, Vincent Favre-Nicolin, Axel Bocciarelli, Loic Huder, Alex de Maria, Marjolaine Bodin, Emmanuel Eyer, François Sturma, Stuart Caunt, ...