ESRF

10 years of Open Data management effort.
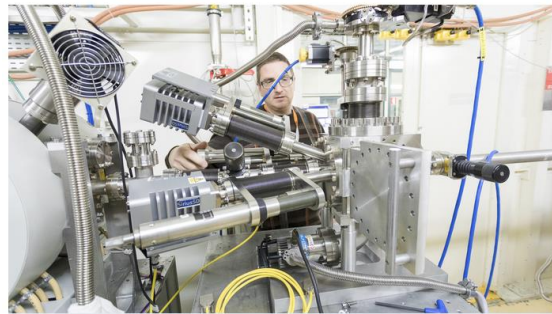
The European Synchrotron | ESRF

An international collaboration of 22 partners

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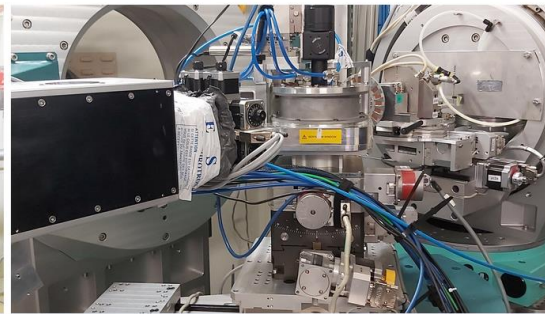
51 Instruments

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*

The European Synchrotron | ESRF

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**

ESRF EUROPEAN SYNCHROTRON RADIATION FACILITY

30 November 2015

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:

ESRF

Search Service

... for Big Data

Data Catalogue

Data Catalogue

https://data.esrf.fr
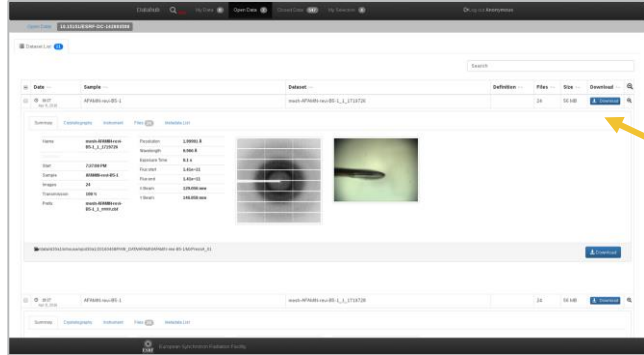
HDF5 Nexus

Download Data Service

Data Service

Explore and Download data

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**

DOI > 10.15151/ESRF-DC-186715732

Identifiers

...ta findable and searchable

Annotate your experiments

The European Synchrotron | ESRF

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 fllowing 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 European Synchrotron | ESRF

E-LOGBOOK

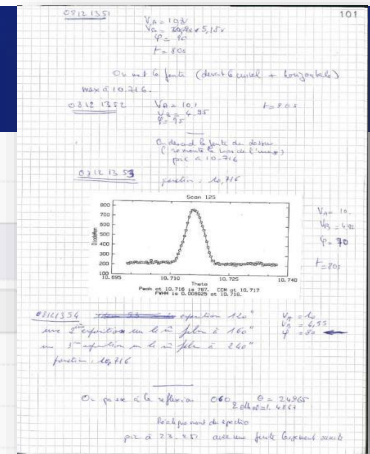Notify commands

Notify errors

Comment on a command

Annotations
- Created by humans
- Editable

Notifications
- Created by software:
  - Bliss
  - SPEC
  - Flint
- It can only be commented
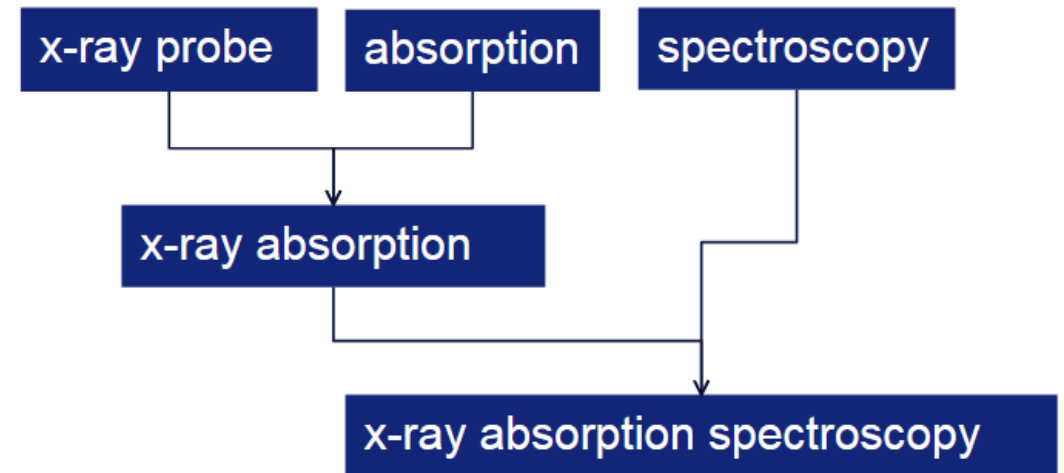
The European Synchrotron | ESRF

The PaNET ontology provides a taxonomy (classification) and thesaurus (synonyms and acronyms) of Photon and Neutron Experimental Techniques.
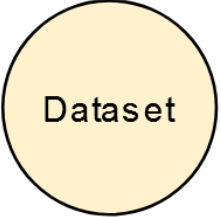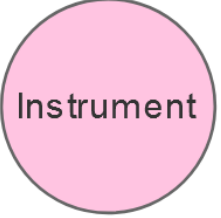
EXAFS	EXAFS	extended x-ray absorption fine structure	PaNET01198
FT-SSX	FT-SSX	fixed target serial synchrotron crystallography	PaNET01310
FTIR	FTIR	fourier transform infrared spectroscopy	PaNET01320
GISANS	GISANS	grazing incidence small angle neutron scattering	PaNET01276
GISAS	GISAS	grazing incidence 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

The European Synchrotron | ESRF

Institution

ROR

ESRF : https://ror.org/02550n020

Person

iD ORCID

G. Bouchard : 0000-0003-0705-6026

Publication

Crossref doi

https://doi.org/10.1038/s41592-021-01317-x

Journal

ISSN

Nature Communications : 2041-1723

Dataset

DataCite doi

ESRF data doi :
https://doi.org/10.15151/ESRF-DC-572196058

Grant/Proposal

Crossref doi

Instrument

DataCite doi

PaN Technique

PaNET

vocabulary/ontology
Photon **a**nd **N**eutrons **E**xperiment **T**echniques

SAXS - Small Angle X-ray Scattering :
http://purl.org/pan-science/PaNET/PaNET01188

RCSB PDB PROTEIN DATA BANK

GBIF Global Biodiversity Information Facility

PubChem

WIKIDATA

The European Synchrotron | ESRF

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

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

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.

raw

ES
da

Research Institutions
species geographic distributions

Paleobiology Database (PBDB)

Museum *collections*

The European Synchrotron | ESRF

Proposal
>94,000

ESRF User Portal

Universities, companies, RI
~110,000

Accepted proposal
>62,000

?

Instrument experiment
48

PUMA

Data

Publication, book, thesis
~48,000
ESRF Library database

Experiment report
>21,000

The European Synchrotron | ESRF
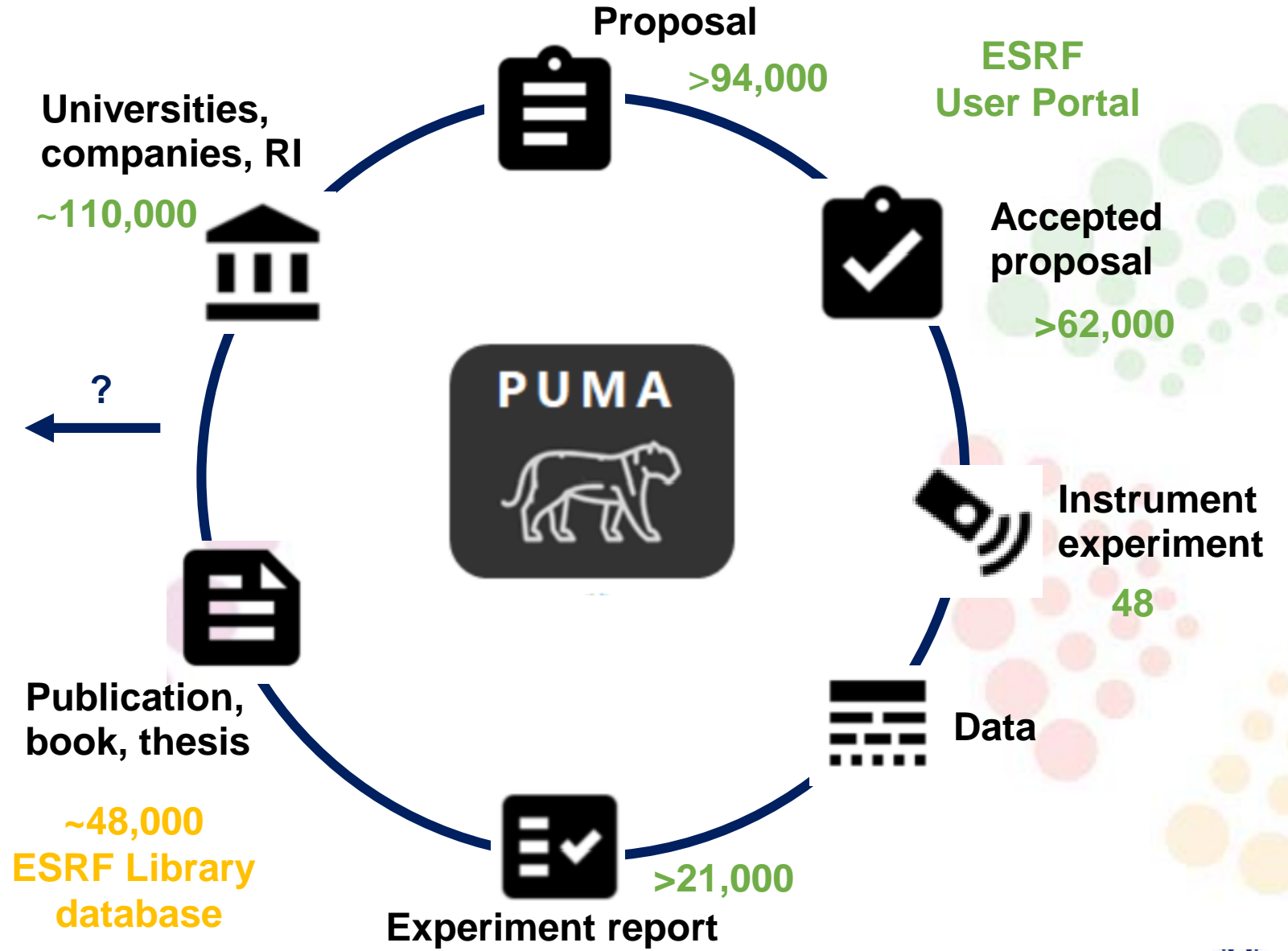
DataCite Metadata Properties

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/yyyyy",
        "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"
    }
]
```

Crossref

PID

GBIF Global Biodiversity Information Facility

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
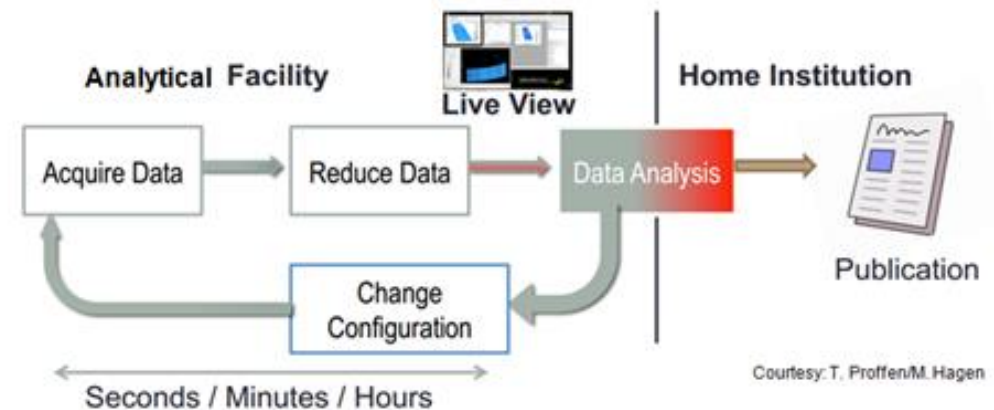
…

The European Synchrotron | ESRF

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**
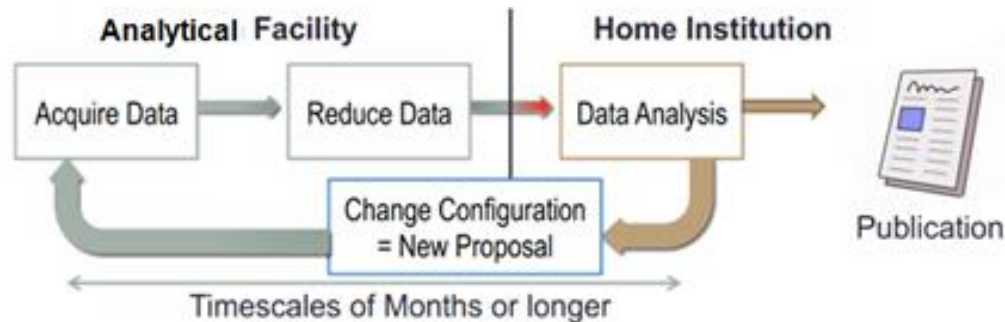
To go from this...   ...to this

Courtesy: T. Proffen/M. Hagen

The European Synchrotron | ESRF

VISA - REMOTE DATA PROCESSING/ANALYSES PLATFORM
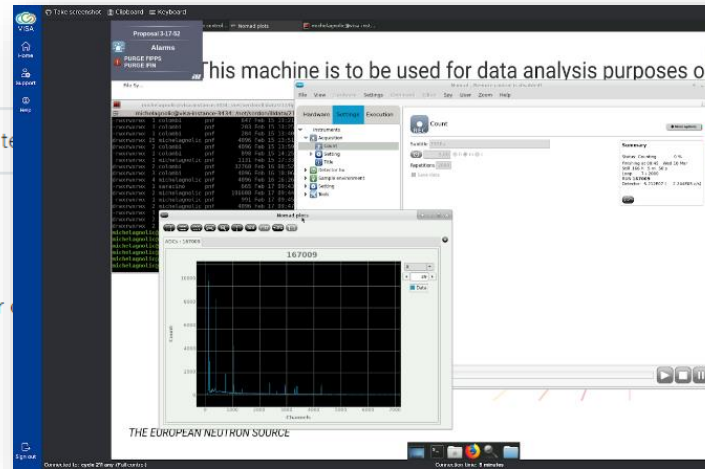
Infrastructure for remote data processing / analysis

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

The European Synchrotron | ESRF

FAIR data

Configurable

The European Synchrotron | ESRF

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
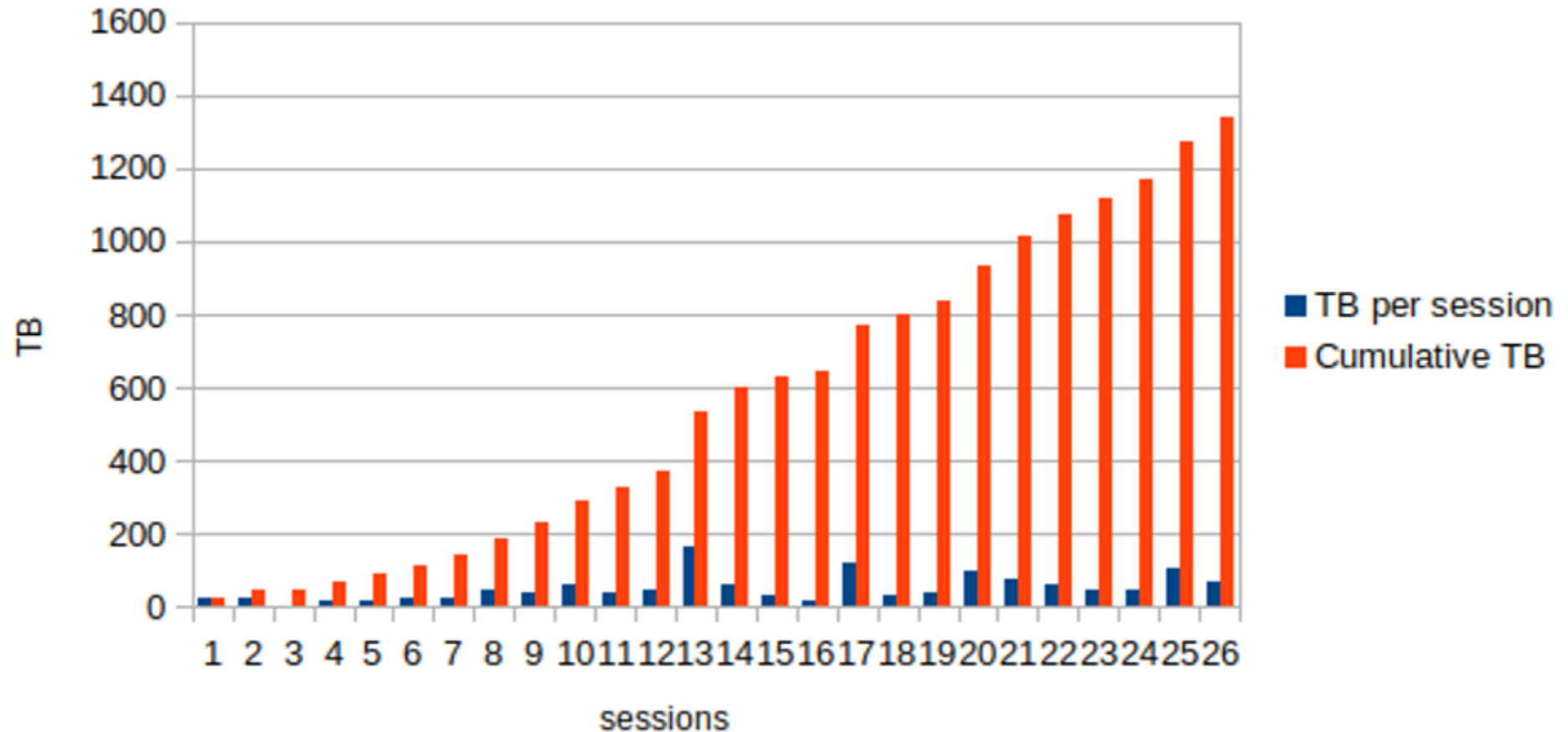
EBS DATA PRODUCTION MEASURED

	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 8.1 TB	8.4 GB 7.8 TB	6.9 GB 8.5 TB	6.8 GB 11.1 TB
Ave / Max files	428 39k	474 43k	293 51k	213 195k
Main Beamline	CM01 40%	ID19 20%	BM18 CM01 21%	BM18 19%
Archived (tape)	--	3.86 PB	6.52 PB	7.75 PB

Mediaflux software

Disk storage

SpectraLogic BlackPearl object storage & tape library gateways

SpectraLogic T-Finity tape libraries & drives

The European Synchrotron | ESRF

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

Findable
Accessible
Interoperable
Reusable

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

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

The European Synchrotron | ESRF

Pub citing ESRF data doi, by year

Collection : *2014-2024. Publications with ESRF experiment, or with ESRF author (esrf_pub)*
Filter : JOURNAL_ARTICLE with instrument only.
'Y' : pub. with data doi. 'N' : pub. without data doi.
(total 'Publications' : 16186 – display as percentage)

Data doi id citation in 'data availability', 'references', 'acknowledgment' or 'supplement' sections. JOURNAL_ARTICLE and with instrument only.
Source PUMA 2024/09/26 17:02. - PUMA_CORPUS_PUB_BY_YEAR_CIT_DATA_DOI_STACKED

Time between publication and experiment

2019-2024

Last experiment date of proposal(s) associated to publication is used to calculate delay with publication date. commissioning and industrial experiments excluded.
(total 'Stacked' : 60 – display as percentage)

- Pub 7th year or more after experiment
- Pub 6th year after experiment
- Pub 5th year after experiment
- Pub 4th year after experiment
- Pub 3rd year after experiment
- Pub 2nd year after experiment
- Pub 1st year after experiment

Percentage of publications with link to experiment

2019-2024

Last experiment date of proposal(s) associated to publication is used to calculate delay with publication date. commissioning and industrial experiments excluded.
(total 'Stacked' : 148 – display as percentage)

- Nb Publications with experiment/publication link
- Nb Publications without experiment/publication link
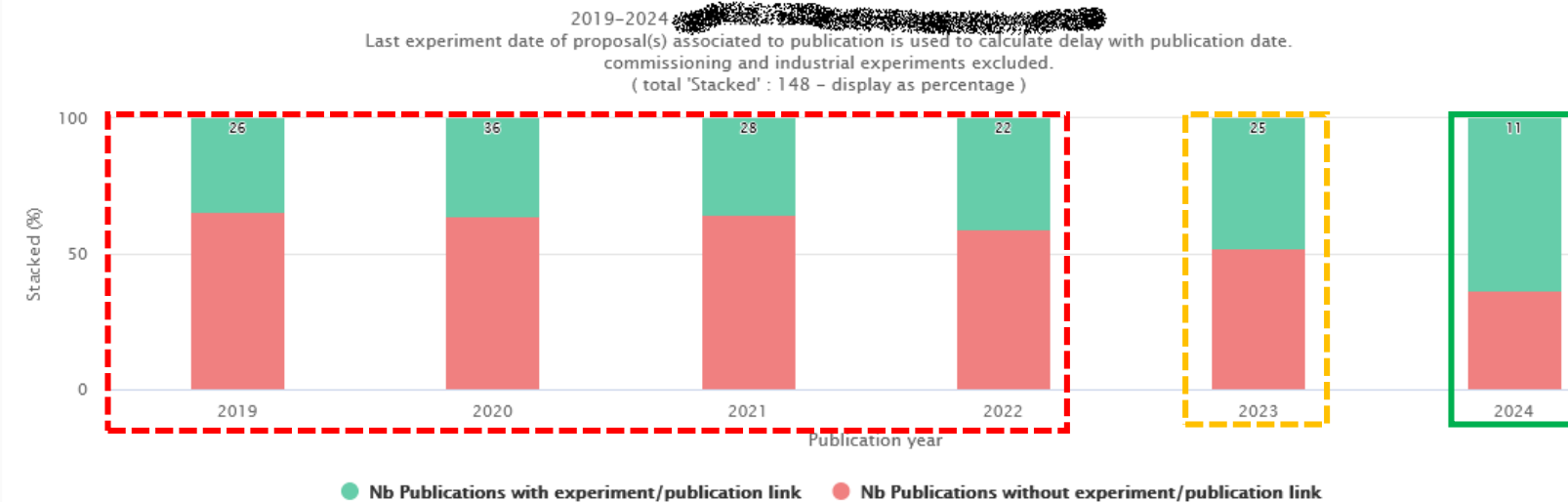
- Library email campaign : we now ask the users for missing proposal reference in publication ex: MX-12345. Flora **library** database is updated.
- Email sent 08/2024 to publication corresponding authors (for 2023 and 2024 publications).
- Good user responses : we see impact on data quality for these years.
- First ESRF DOIs created in 2018 (source Datacite search api)

Source PUMA 2024/09/26 16:28. - PUMA_CORPUS_PUB_BY_YEAR_EXP_PUB_DATA_STACKED

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)**

The European Synchrotron | ESRF

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 ?

The European Synchrotron | ESRF

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, ...

The European Synchrotron | ESRF