



Towards a New Data Policy

Oscar Matilla
Computing Division Head
ALBA Synchrotron

VII ALBA Users' Meeting
04-09-2026



The new Data Policy is more than an update on the data storage rules. It reflects ALBA's commitment to our users, supports the success of ALBA's experiments, and serves as the contract connecting the entire scientific data lifecycle.

Timeline

- The first ALBA Data Policy was approved in 2017.
- Since then, several minor revisions have been introduced.
- Current version: https://www.cells.es/en/assets/pdfs/users/data_policy_cells
- Our goal is to **approve a new Data Policy by the end of 2026**, in time for the first call for proposals in 2027.
- Initial ideas were presented for feedback at the 42nd Scientific Advisory Committee Meeting; currently gathering input here at the VII ALBA Users' Meeting.

Key Drivers for the New Policy



Managed Lifecycle

Transitioning from a basic file custodian and delivery model to a comprehensive scientific data lifecycle with enhanced data services.



Infrastructure Sustainability

To address increasing data volumes and computing needs through intelligent data reduction and computing quotas without compromising scientific value.



Regulatory Landscape

To align with EU Legislation 2019/1024, Spanish RD 29/2021, and the Estrategia Nacional de Ciencia Abierta (ENC 2023-2027) guidelines, which mandate that publicly funded research data must be open by default.



The Regulatory Landscape

Open Data Directive

Publicly funded research data* must be open by default to support transparency and further scientific discovery.

European Legislation 2019/1024

Mandating open access to research data across the EU**.

Spanish Legislation RD 29/2021

National alignment for open science integration.

ENCA 2023-2027

National Strategy for Open Science directives.

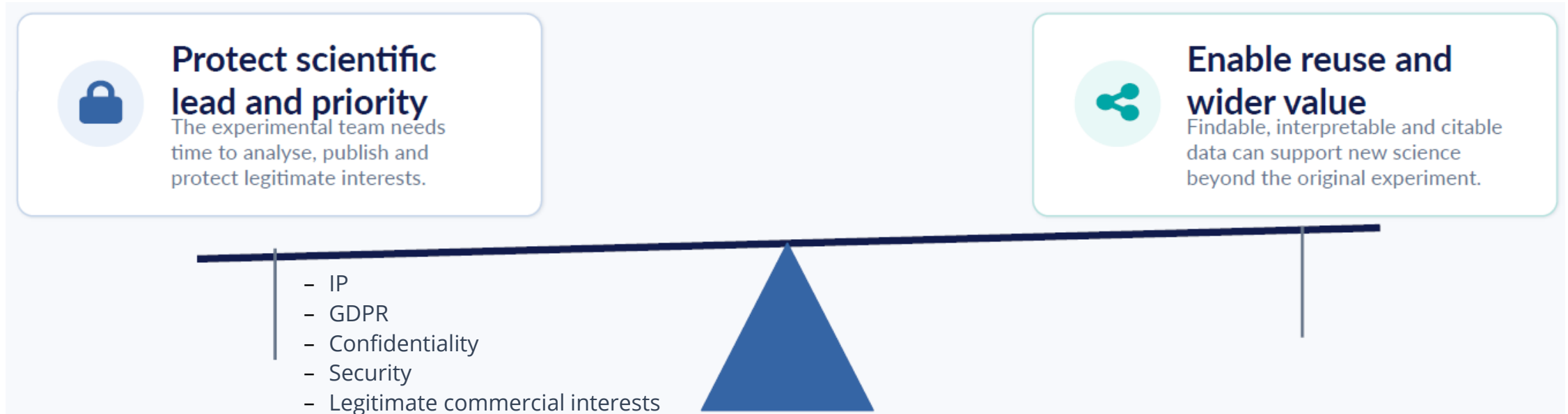
* Includes publication, data, code and methodologies

** Horizon Europe now strictly requires open metadata





Regulatory Landscape - The Balancing Act



“As open as possible, as closed as necessary”

Additional Regulatory Considerations

REGULATION (EU) 2024/1689 · AI ACT

Regulates AI-system risk

- **Risk-based duties.** Rules scale from prohibited practices to transparency, GPAI and high-risk systems.
- **Now enforceable.** Transparency requirements and enforcement started 2 Aug 2026; earlier AI-literacy, prohibition and GPAI rules also apply.
- **High-risk phase-in.** Annex III duties: 2 Dec 2027. Product-embedded systems: 2 Aug 2028.

REGULATION (EU) 2016/679 · GDPR

Regulates personal-data processing

- **Use data lawfully.** Identify a lawful basis and meet fairness, transparency and purpose-limitation duties.
- **Minimise and protect.** Use only necessary data; apply security, retention limits and accountability.
- **Assess impacts and rights.** Run a DPIA where processing is likely high-risk; safeguard rights in solely automated decisions (Art. 22).



Embracing FAIR Principles

Adopting FAIR principles (Findable, Accessible, Interoperable, Reusable) requires understanding what it is **not**:

- ✗ **FAIR is not strictly "Open"** Data can be FAIR while still being restricted to protect legitimate interests.
- ✗ **FAIR is not "Free Anonymous Access"** Access requires non-anonymous user registration and attribution (CC-BY).
- ✗ **FAIR does not mean "Store Forever"** It involves intelligent data triage, quotas, and defined retention periods to ensure long-term viability.



Types of metadata



Discovery metadata

Title • authors / PI • abstract • beamline

Public after the experiment so that the experiment can be found.

VISIBLE FOR DISCOVERY

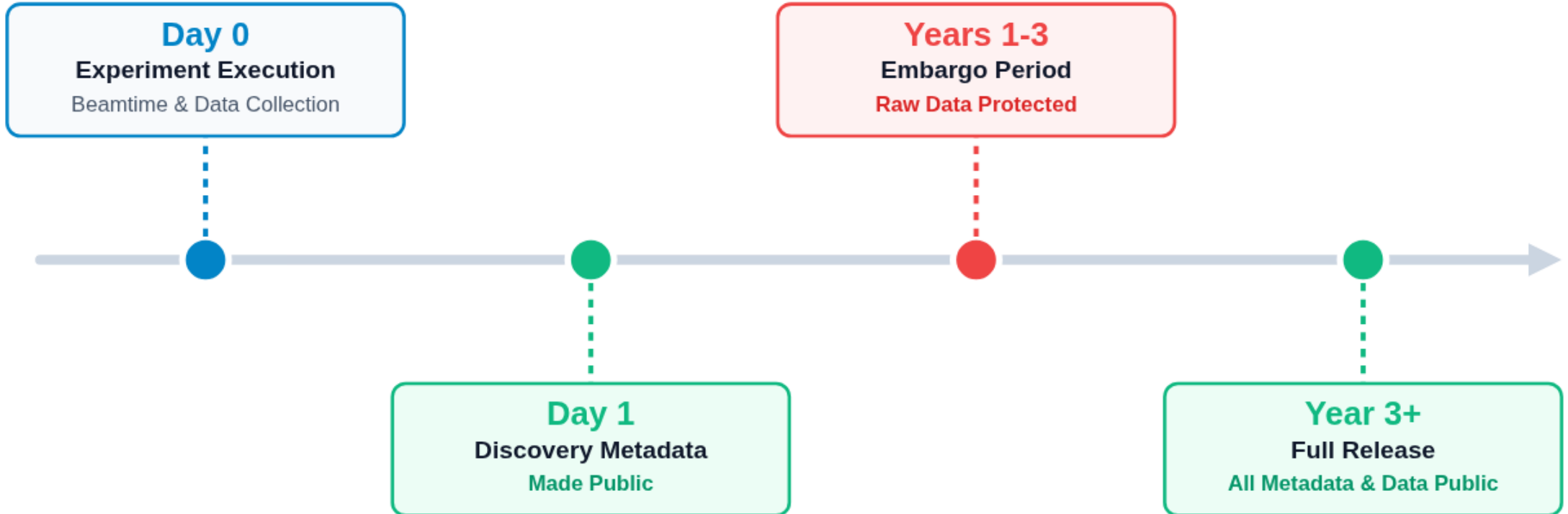


Detailed experimental + auxiliary metadata

Sample description and provenance • acquisition conditions • calibration • e-logbook • scripts • software tools and versions

Protected during the embargo; open afterwards.

Embargo Period in the current Data Policy





Types of metadata

- Since the first Data Policy, **Discovery metadata** has been **public after beamtime ends**. Experimental and Auxiliary metadata remain restricted, together to raw data.
- Under the new Data Policy, users may request an **exception** to restrict Discovery Data until the end of the data embargo*.
- We plan to **extend** the option to request this exception to all users **retroactively**, including those covered by the current Data Policy.

* Valid exceptions

- i. ethical and legal constraints, in particular regarding personal data which cannot be anonymized to a sufficient degree and medical data,
- ii. security constraints, including possible code abuse.
- iii. ongoing preparation of a patent application,
- iv. contractual obligations e.g. to industry partners,
- v. intended commercial use of a software linked to a dataset,
- vi. intended use of data for commercial purposes or
- vii. other special circumstances validated by the CELLS



Discovery metadata

Title • authors / PI • abstract • beamline

Public after the experiment so that the experiment can be found.

VISIBLE FOR DISCOVERY



Detailed experimental + auxiliary metadata

Sample description and provenance • acquisition conditions • calibration • e-logbook • scripts • software tools and versions

Protected during the embargo; open afterwards. A minimum subset is defined in Appendix 1.



Types of metadata



Discovery metadata

Title • authors / PI • abstract • beamline

Public after the experiment so that the experiment can be found.

- Additionally, during the Call, **one additional metadata field** will be provided to users. This information will be **used exclusively by the Scientific Panel for evaluation** and will not be available afterward.



Detailed experimental + auxiliary metadata

Sample description and provenance • acquisition conditions • calibration • e-logbook • scripts • software tools and versions

Protected during the embargo; open afterwards. A minimum subset is defined in Appendix 1.

Embargo Period Management

- The **option** for users to **request an embargo extension** beyond three years was included in the previous Data Policy and will remain in the New Data Policy.
- Additionally, users will receive **automated notifications** when the embargo period is nearing its end.



Extending the Scientific Data Lifecycle



- Experiment is not finished acquiring raw data. **Data Policy will be extended to cover Collected Data which includes processed data.** ALBA will allocate resources and make efforts to ensure the successful processing of data and results.





From policy to production:

- Today, ALBA is actively deploying the tools and services required to manage the modern scientific data lifecycle.

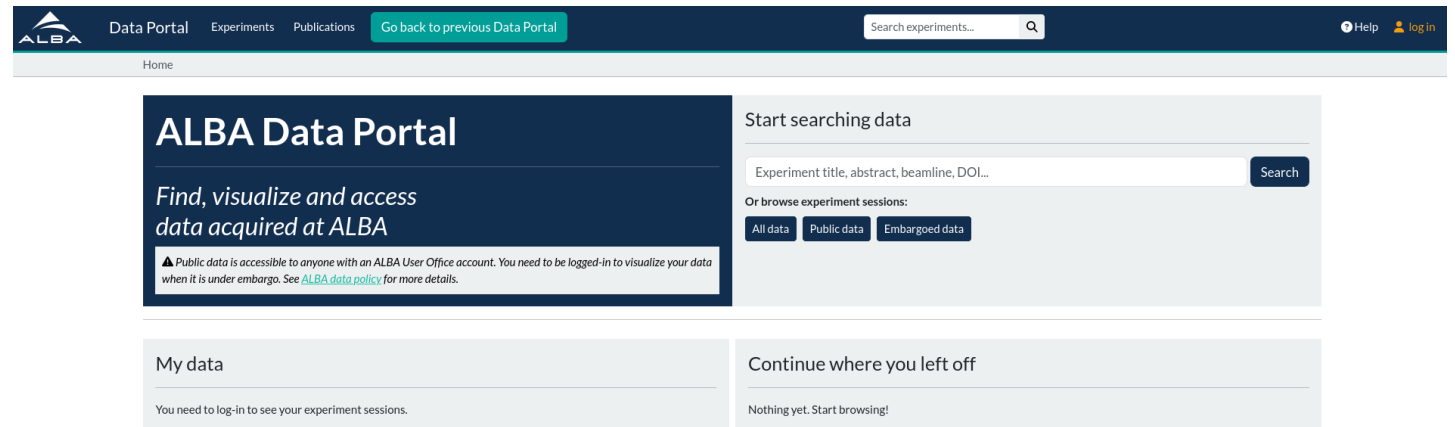
Scientific need	Capability already available
Structure data	NeXus/HDF5
Find and access data	ICAT Data Portal
Transfer large datasets	Globus and 100 Gbps connectivity
Cite and publish data	Automatic DOI minting
Process data remotely	VISA and other DaaS environments
Federate and reuse data	PaN Data Space integration

The ALBA Data Portal: ICAT

<https://data.cells.es>

- [ICAT](#) is a data catalogue, resulting from a collaboration between different European facilities (DLS, ESRF, HZB, SESAME, STFC and ALBA).
- It allows finding, browsing, viewing, downloading and annotating data during and after the experiment.

2023	BL20-LOREA,
2024	BL06-XAIRA BL09-MISTRAL BL24-CIRCE_PEEM
2025	BL11-NCD_SWEET BL22-CLAESS BL29-BOREAS
2026	BL01-MIRAS BL04-MSPD BL24-CIRCE-NAPP BL31-FAXTOR



The rest of the instruments will follow from 2027 onward

ICAT detailed Features

Experiment sessions

Search: name, title, abstract, DOI...

Filter by: Public data Embargoed data My data

Beamline: any

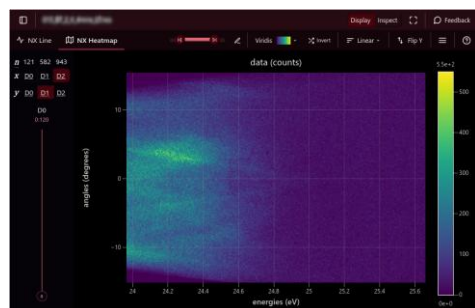
Start date: 24/08/2024

End date: 24/08/2024

User: Select a user...

Beamline	Start	Title	A-Form	Samples	Datasets	Files	Release	DOI	Logbook
2024028206	BL11	18/07/2024	Operando SAXS/WAXS Investigation of Silicon Anode Morphology evolution upon cycling, and effect of binder selection for High-Energy Li-Ion Batteries				21/07/2027		
2024028317	BL01	16/07/2024	Evaluation of lipid and protein alterations in the secondary brain lesion of a preclinical model of intracerebral hemorrhage by μ FTIR				20/07/2027		
2024028359	BL20	16/07/2024	Spin and Angle-Resolved photoemission investigations on the spin texture of magnetically modified topological surface states	14 265.58 MB	14 265.58 MB	14	20/07/2027	DOI: 10.57710/ALBA-ES-2024028359	
2024028174	BL11	11/07/2024	Unveiling the mechanism of hydrometallurgical recycling of neodymium from permanent magnets using phosphonium ionic liquids				11/07/2027		
2024028232	BL20	10/07/2024	Atomic mechanisms of electron doping in perovskite nickelates	254 3.98 GB	270 3.98 GB	270	14/07/2027	DOI: 10.57710/ALBA-ES-2024028232	
2024028286	BL01	09/07/2024	Structural changes in bioinspired thermoplastics during aging studies in the marine environment				11/07/2027		
2024066477	BL01	05/07/2024	Feasibility test for "Investigating the interaction between microstructure and mechanical performance in MEW Soft Robotic Actuators I (MIRAS part)"				13/07/2027		
2024028081	BL11	05/07/2024	Investigating coiled coil-based self-assembled structures				07/07/2027		
2024028164	BL11	03/07/2024	Correlation between nature, loading and spatial confinement of biomolecules within Metal-Organic-Frameworks				05/07/2027		

Browsing, visualizing, annotating and downloading data before and after the experiment



Integrated NeXus/HDF5 data viewer

Data Collection **Open access**

Test DC

Rodrigo Cabezas Quins

Test DC

Experimental data

Data can be accessed by clicking on the link below

Access data

Reference

Researchers must acknowledge the source of the data and cite its unique identifier as well as any publications linked to the same raw data. Below is the recommended format for citing this work in a research publication.

Cabezas Quins, R. (2024). Test DC (Version 1) [Dataset]. ALBA Synchrotron. doi.org/10.57710/ALBA-DC-15

Cited by

Metadata

Identifier

DOI: 10.57710/ALBA-DC-15

Proposal

2023999999

Public release year

2024

Publisher

ALBA Synchrotron

License (for files)

Creative Commons Attribution 4.0

Related DOIs

No related DOIs were found.

Experiment's landing page

Home / 2023097869 BL20 24/01/2024 - 29/01/2024 / Logbook

+ New comment Edit Tags Take Picture Download Help

Items 1-2 of 2 Show 100

Massimo Tolleria

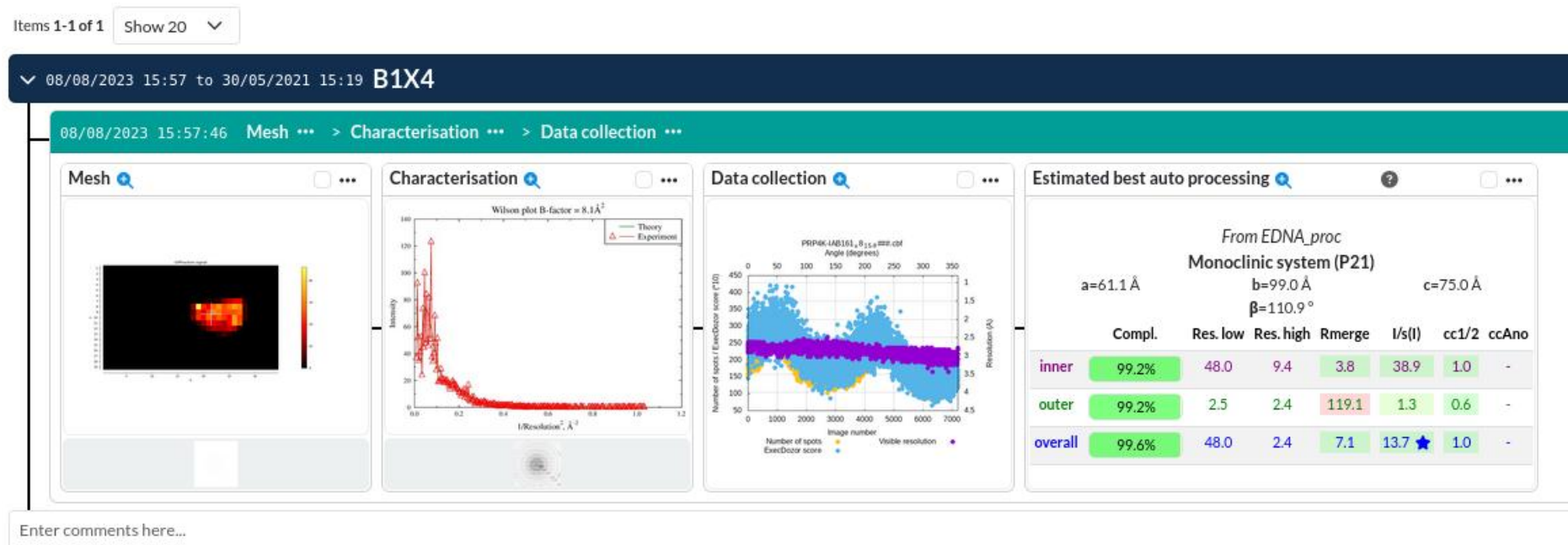
09:09:32

hello world

eelogbook

Features: DOI minting, integrated eelogbook, NeXus/HDF5 data viewer and more.

ICAT detailed Features



- MX techniques have an integrated **Laboratory Management Information System** (ICAT-DRAC/ESRF).
- More techniques will be included in the future (EM, SAXS, SSX,...).

Scientific Data Formats

- Embedding data and metadata: the [NeXus/HDF5](#) format

In production:

BL01-MIRAS ([NXoptical_spectroscopy](#))

BL04-MSPD ([NXmonopd](#), [NXxrd](#))

BL06-XAIRA ([NXmx](#))

BL09-MISTRAL ([NXtomo](#))

BL11-NCD ([NXsas](#))

BL20-LOREA ([NXarpes](#))

BL24 CIRCE-PEEM (new [NXmpes_peem](#))

BL24 CIRCE-NAPP ([NXsensor_scan](#), [NXxas](#), [NXmpes](#))

BL29-BOREAS ([NXsas](#) based)

WIP:

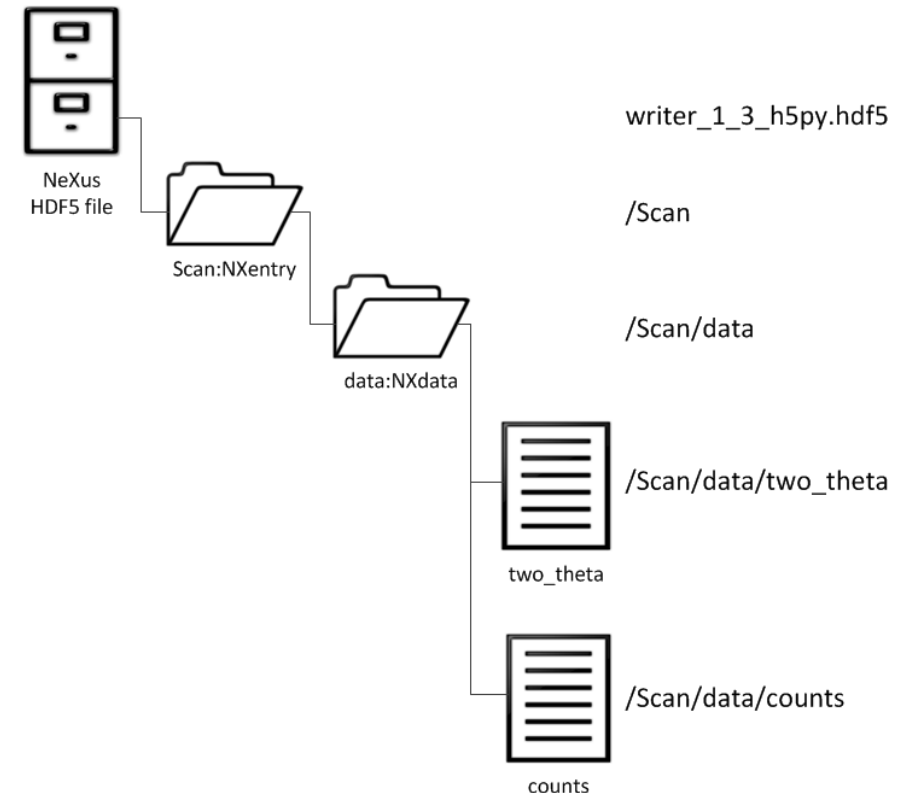
BL13-XALOC

BL15 3Sbar

BL16-NOTOS

BL22-CLAESS

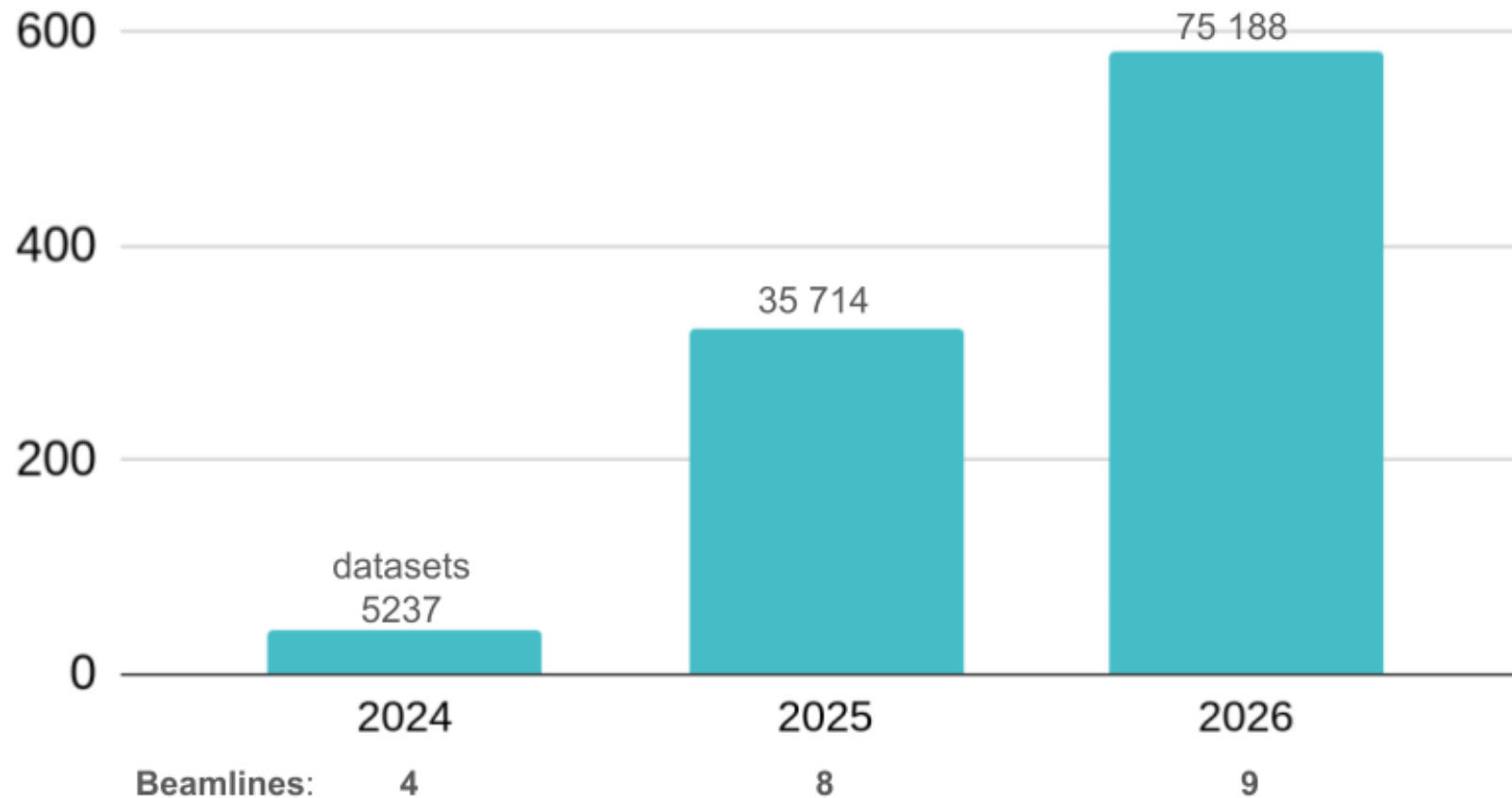
BL31-FAXTOR (using JPEG2K compression)



[Application definitions link](#)

ICAT Usage at ALBA

Cumulated number of experiments registered in ICAT



ICAT - Access and Data Transfer

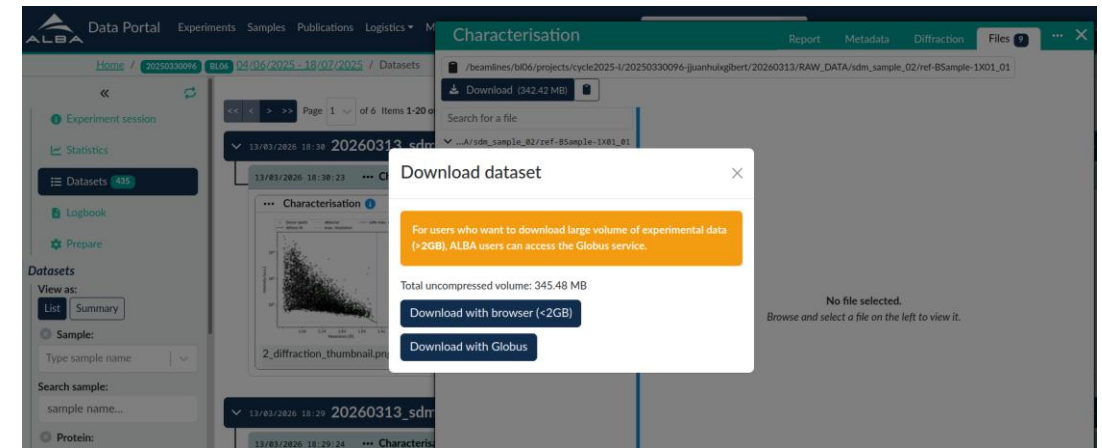


ALBA Data Portal does not allow anonymous access:

- Access is restricted to registered users.
- Authentication currently relies on ALBA User Office accounts.
- We plan to integrate with the broader European federated research identity ecosystem, using GÉANT eduGAIN and MyAccessID. (Early 2027)

Faster Data Transfers for ALBA Users:

- Internet connectivity has been upgraded from 10 Gbps to 100 Gbps, with 25 Gbps reserved for Globus.
- Globus is now available for high-volume data transfers from BL06-XAIRA, BL13-XALOC, and BL31-FAXTOR, all integrated with ICAT.





ICAT - DOI minting

DOIs are minted automatically for every experiment

2024098932	BL06	19/09/2025	Proteins, Protein-Nucleic Acid Complexes and Carbohydrate modifying enzymes.	39	346 176.78 GB	1 866	21/11/2028	DOI 10.57710/ALBA-ES-2024098932	
2024098934	BL06	18/09/2025	Towards personalized molecular medicine: Structural basis of neurodegenerative diseases and cancer (X)	20	317 155.95 GB	1 559	25/09/2028	DOI 10.57710/ALBA-ES-2024098934	
2025003333	BL06	15/09/2025	SDM TEST INVESTIGATION	129	644 361.41 GB	3 863	15/09/2028		
2024098938	BL06	24/07/2025	Structural studies of macromolecular interactions relevant to infectious and metabolic diseases. Standard Proposal	140	1463 1.57 TB	9 132	07/11/2028	DOI 10.57710/ALBA-ES-2024098938	
2024098929	BL06	02/07/2025	BAG-CRI347EM Proposal	30	214 214.85 GB	1 302	27/11/2028	DOI 10.57710/ALBA-ES-2024098929	
2024078923	BL06	20/06/2025	Molecular mechanisms of proteins and complexes involved in signaling, microbial pathogenic processes	102	584 415.78 GB	3 384	26/11/2028	DOI 10.57710/ALBA-ES-2024078923	

Users can manually mint a DOI to an ensemble of chosen datasets

ALBA Data Portal Experiments Samples Publications Logistics Beamlines My selection Manager

Home / Mint DOI

Clear selection Manage selection

Items 1-3 of 3 Show 20

2023999999 (09/11/2023 on BL00)

08/08/2023 15:57:46 Sprint 2023-9 review--TESTESTEST Summary Instrument Files 2 Metadata

Sample Peanut butter

Description

Technique

/beamlines/bl01/projects/cycle2019-1/2018092997-herandez Explore Download

08/08/2019 15:57:46 Test parameter analyser_entrance slit_shape Summary Files 2 Metadata

Sample SAMPLE 1

Description

Technique

/beamlines/bl01/projects/cycle2019-1/2018092997-herandez Explore Download

08/08/2019 15:57:46 Test only datafile6 Summary Files 1 Metadata

Sample null

Description

Technique

/asd Explore Download

You can only mint a DOI for datasets which are linked to experiment sessions where you are a user.

Minting the DOI makes the datasets public immediately.

Title*
A minimum of 40 characters is required.

Abstract*
A minimum of 1000 characters is required.

Authors*
Select an author to add... + New author Clear all authors

Keywords
+ New keyword Clear all keywords

Publication DOI

External reference URLs (PDB...)
+ New external reference urls (pdb...) Clear all external reference urls (pdb...)

Next Clear

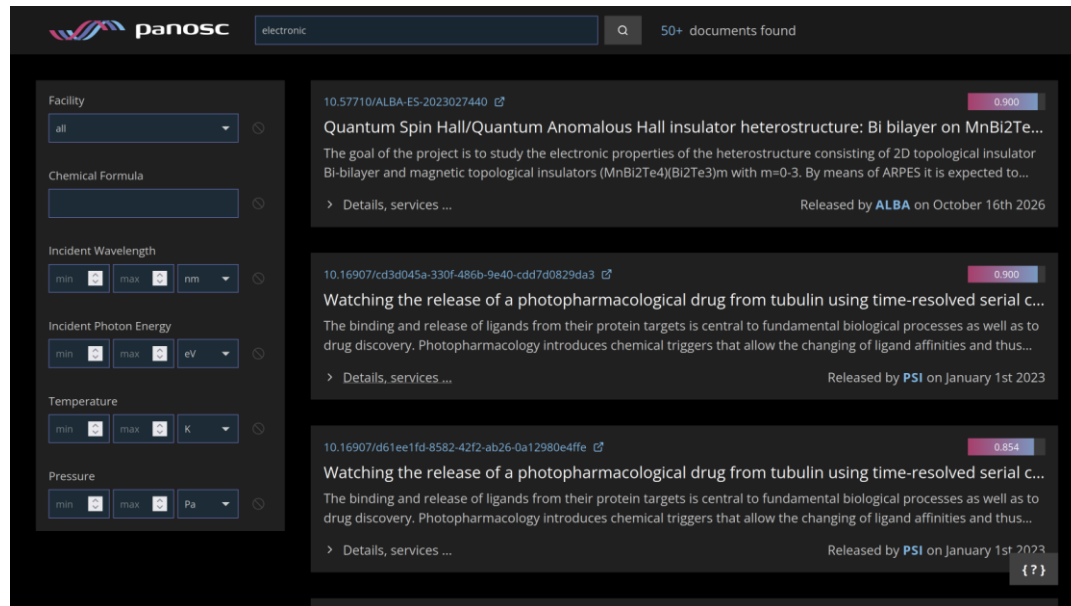
Each DOI minting leads to the creation of a public landing page

Federated Data Repositories



Federating European facilities

- We participate into the **creation of a European PaN Data Space** and to **facilitate Spanish users** access to it.
- Our data catalogue seamlessly **integrates with the PanOSC data catalogue**, allowing for easy discovery through general search engines.



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

PaNOSC Node and the EOSC Federation

• Node buildup – 2025-now

1. PaNOSC was selected as one of the first wave of EOSC Nodes
2. A Project Charter was developed in 2024 followed by a [Project Plan](#) with 6 Work Packages
3. PaNOSC is involved in the buildup phase working groups (Handbook, Training, MoU, Service Catalogue)

• PaNOSC Node is building a platform for Open Science in Photon and Neutrons:

- AAI – [UmbrellaID AAI](#) has been validated on the technical level with the EOSC AAI (MyAccessId)
- Data catalogues – [PaN-Finder](#) (OSCARS), an AI-driven search engine for FAIR data
- VRE – work ongoing on the [VISA](#) (OSCARS) and extending it to EOSC users
- Training – developing [PaN-Training platform](#) based on outcome of mTeSS-X (OSCARS)
- Website – developing the [PaNOSC EOSC Node home page](#) + service catalogue
- PaN-Space – developing a [PaN user portal](#) for users to discover and manage FAIR data as a project
- Metadata – defining metadata and extending existing standards ([Nexus](#))
- Software – sharing workflows ([Ewoks](#)) and package research software ([CVFMS](#))
- Visualisation – developers of the cross-domain visualization service [H5Web](#)

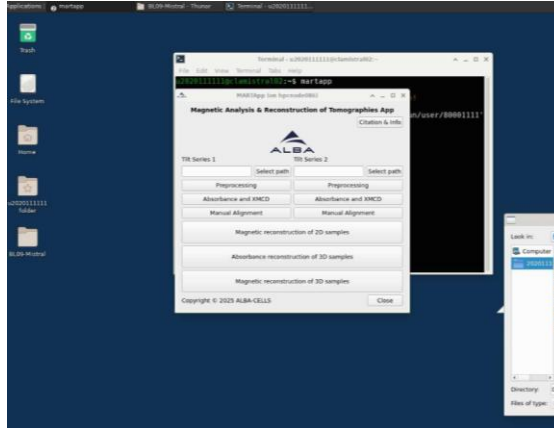


<https://panosc.eu/>



DaaS : VISA & others

- **VISA deployed at ALBA:** images for **tomography** and **MX** in test phase, to be deployed in the next weeks.
- MISTRAL uses **NoMachine** for the moment
- Possibility to outsource computation to the **Barcelona Supercomputing Center** (BSC, work in progress).



Remote processing by
MISTRAL users via NoMachine

<https://visa.cells.es>

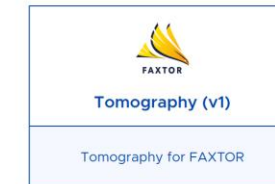


New compute instance

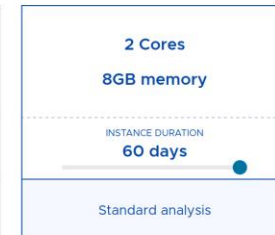
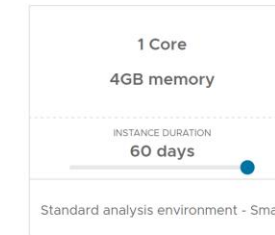
Please fill in the details below to create a new compute instance. For [information](#) about VISA

Computing Environment

Choose an environment



Choose hardware requirements



Display Settings

Choose the resolution of your display



Choose layout

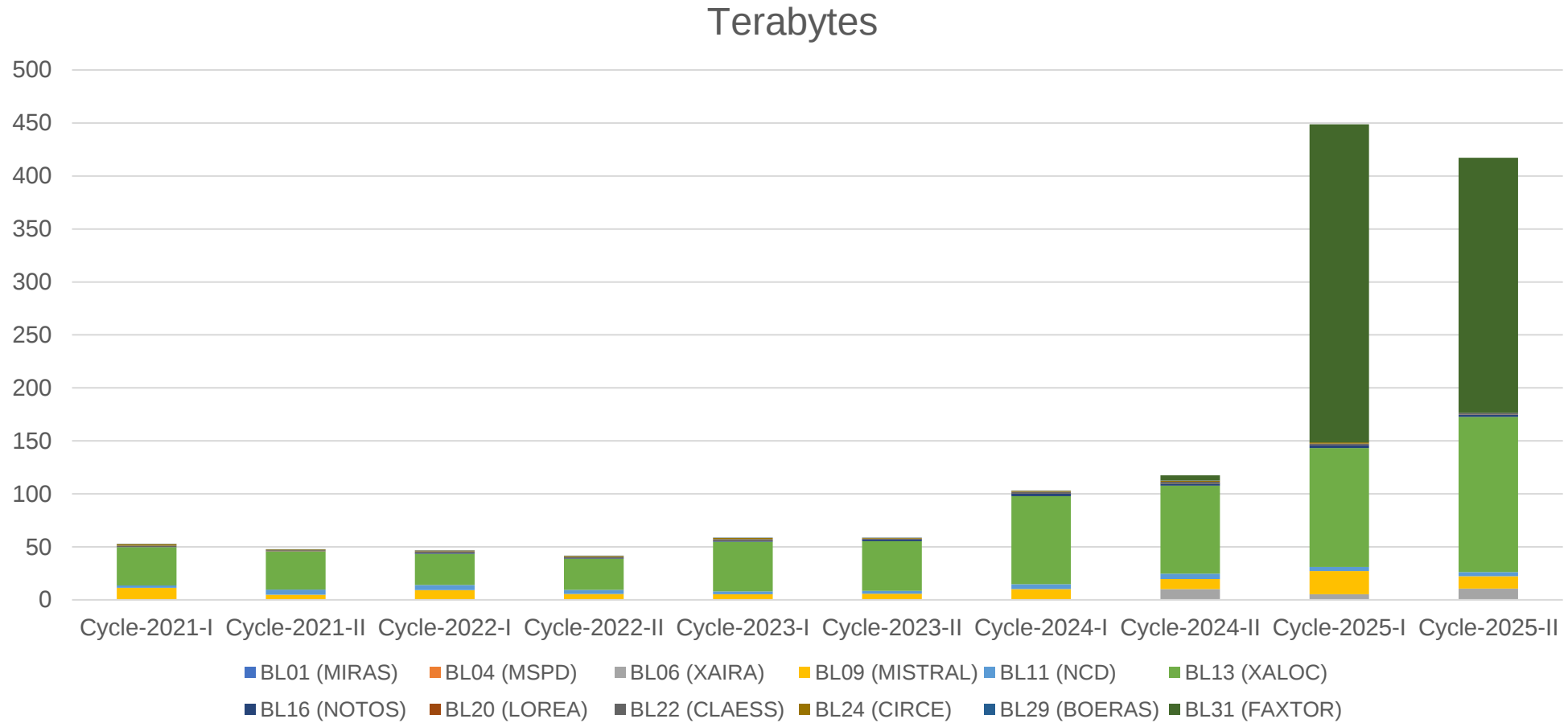


Data Services Sustainability

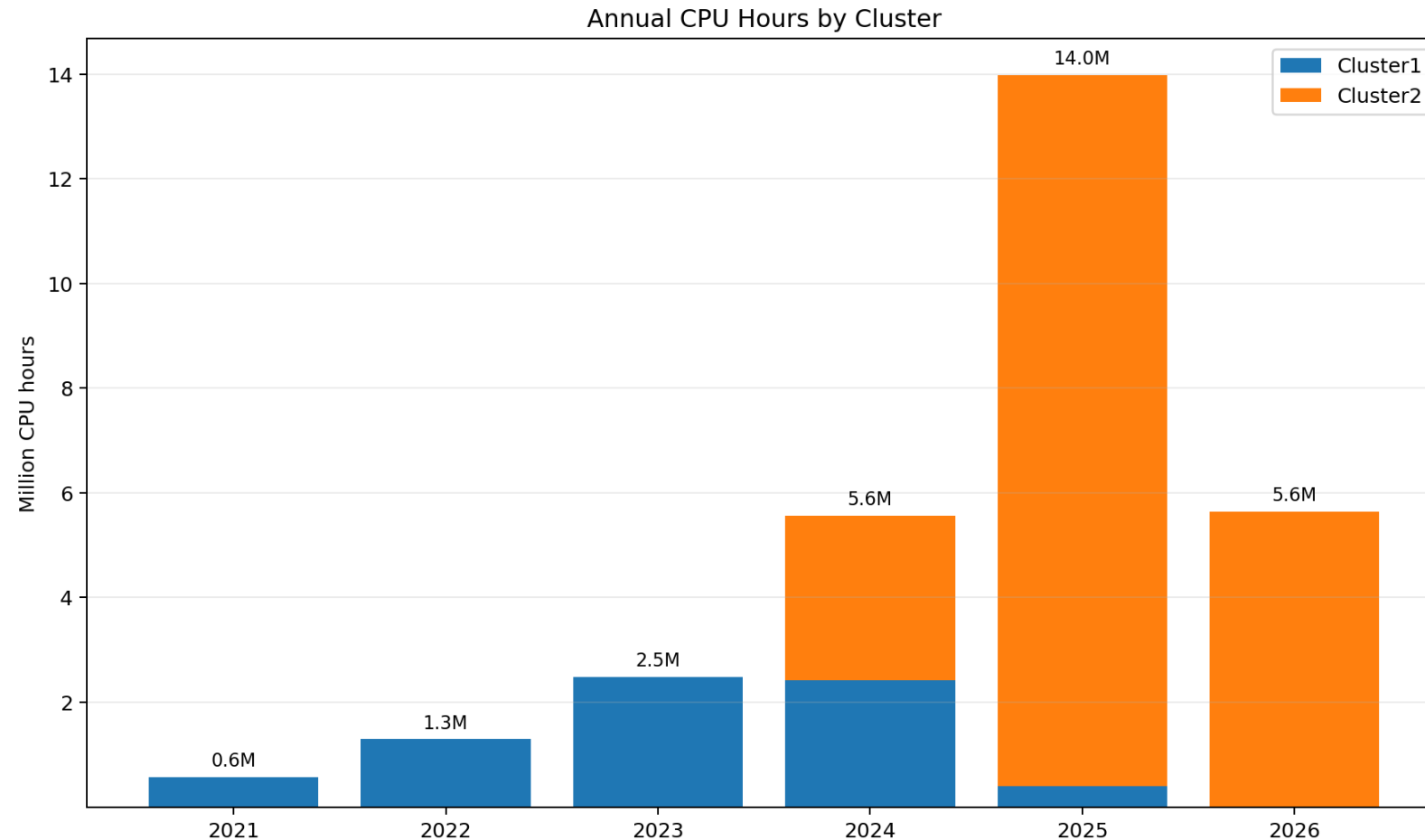
Data services can grow in levels



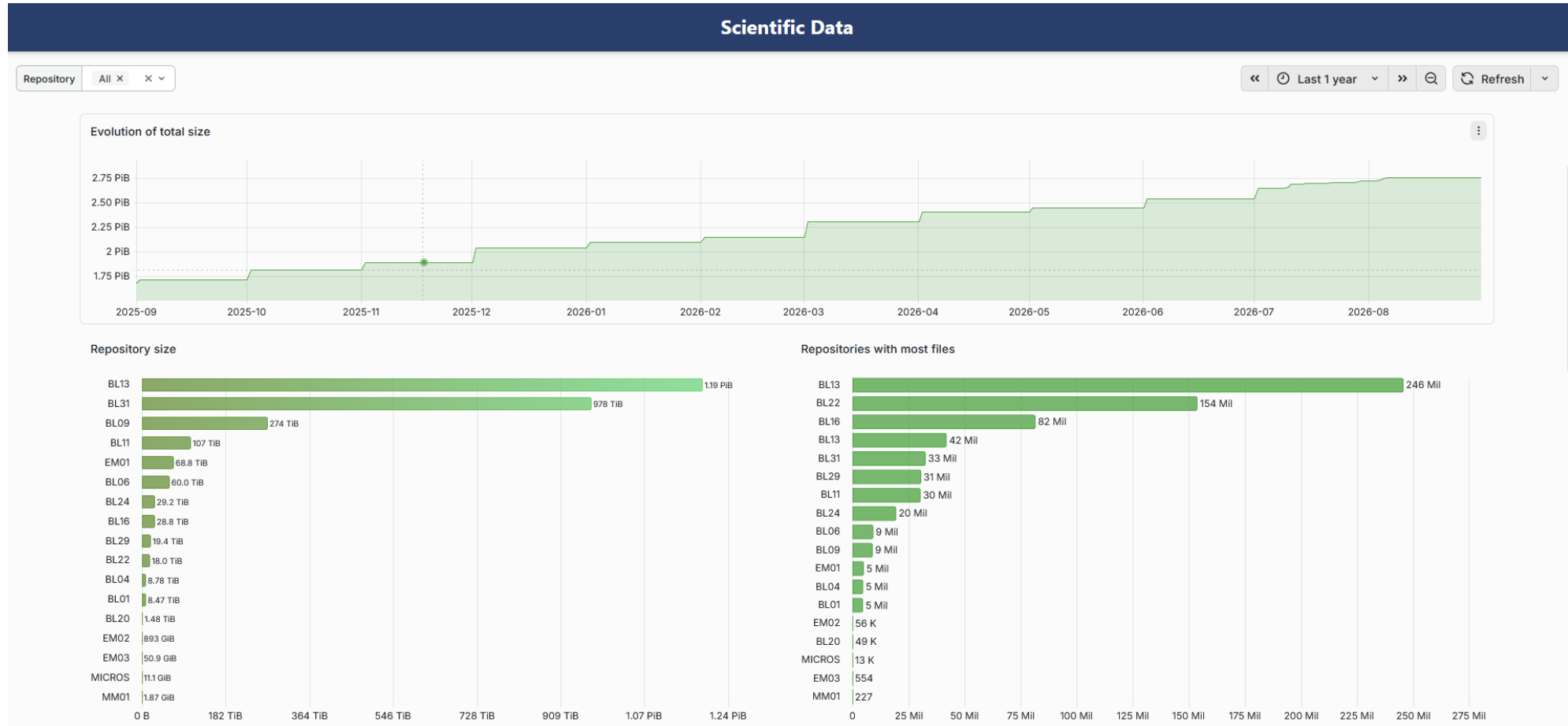
Evolution of data production at ALBA



Associated with increased processing needs...

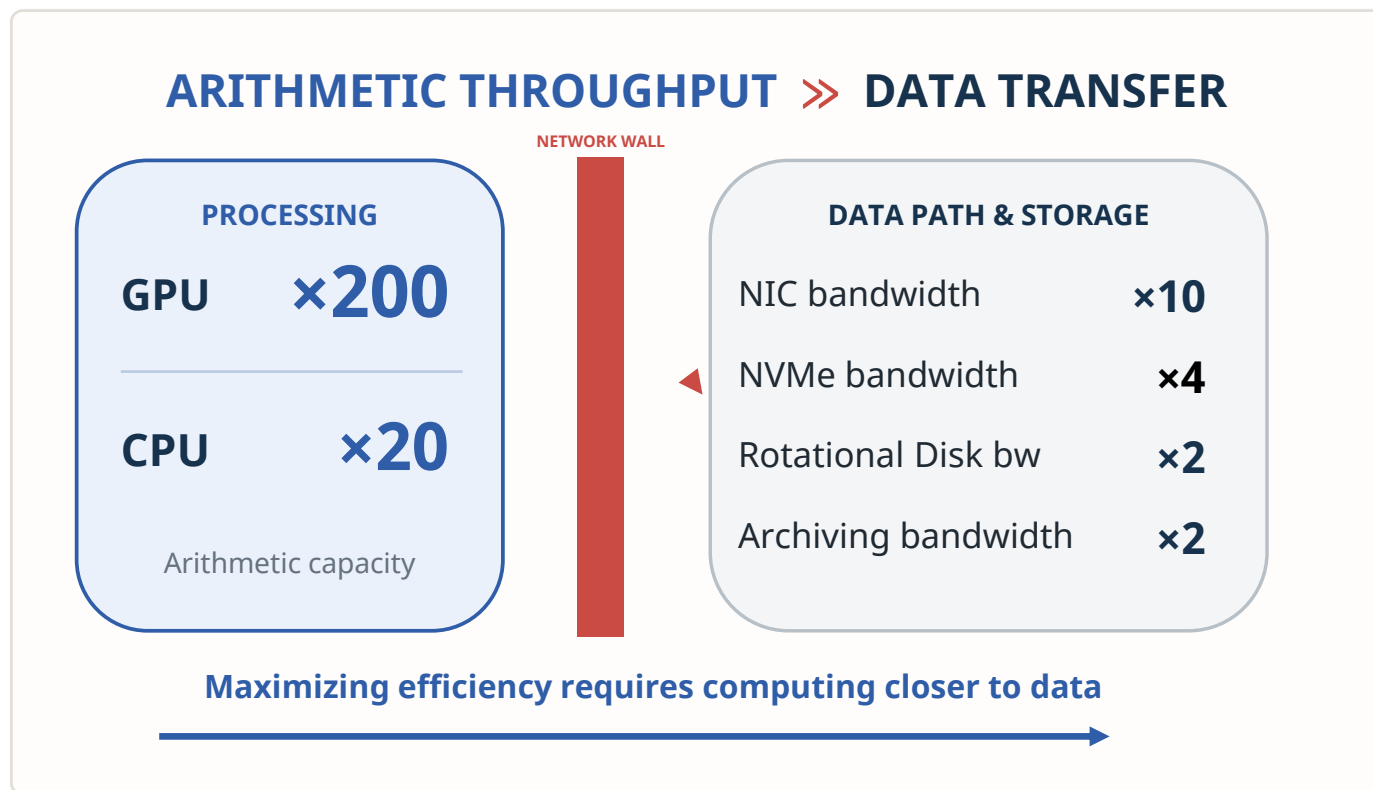


- We are implementing strict mechanisms to monitor the use of our IT infrastructure:





IT Evolution Over the Last Decade



IT is becoming structurally unbalanced: computing power is growing faster than our ability to move data.

The gap is architectural, not a single-component issue.

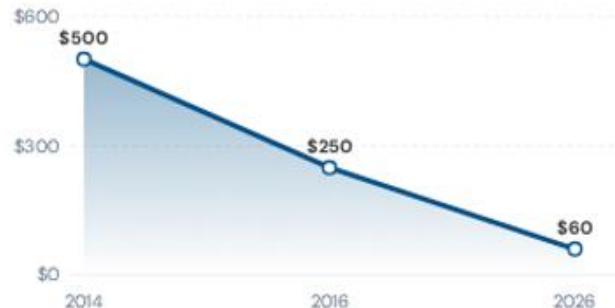
IT Evolution Over the Last Decade



HDD Cost per TB



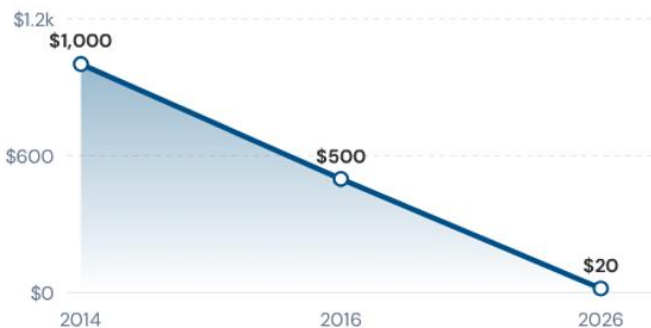
NVMe Cost per TB



Tape Archive Cost per TB



CPU Cost per TFLOP/s



GPU Cost per TFLOP/s (FP32)



- Data volumes are growing exponentially.
- Data processing costs are falling rapidly, while storage costs are falling much more gradually.

Cloud Computing Alternative



AWS S3 Cost per GB/Month



AWS Data Egress Cost per GB



Public-cloud prices have not followed the declining cost of the underlying hardware.

- Cloud prices have remained relatively **flat over the past decade**, and some services are now experiencing price increases — sometimes referred to as “cloudflation.”
- Example: storing 100 TB in the public cloud.

Storage: ~\$27.6k/year

Data transfer: ~\$9k/transfer

- **Cloud repatriation** is becoming increasingly common: more than 40% of U.S. companies relying on cloud solutions have moved workloads back on-premises.



The challenge is shifting from “Can we compute it?” to “Where is the data, and how often must it move?”

➡ **Data Gravity**



Compute Moves to Data

Instead of pulling massive datasets into centralized public clouds, we must deploy execution environments directly at the edge or within institutional data centers where the data natively resides.



The Federated Model

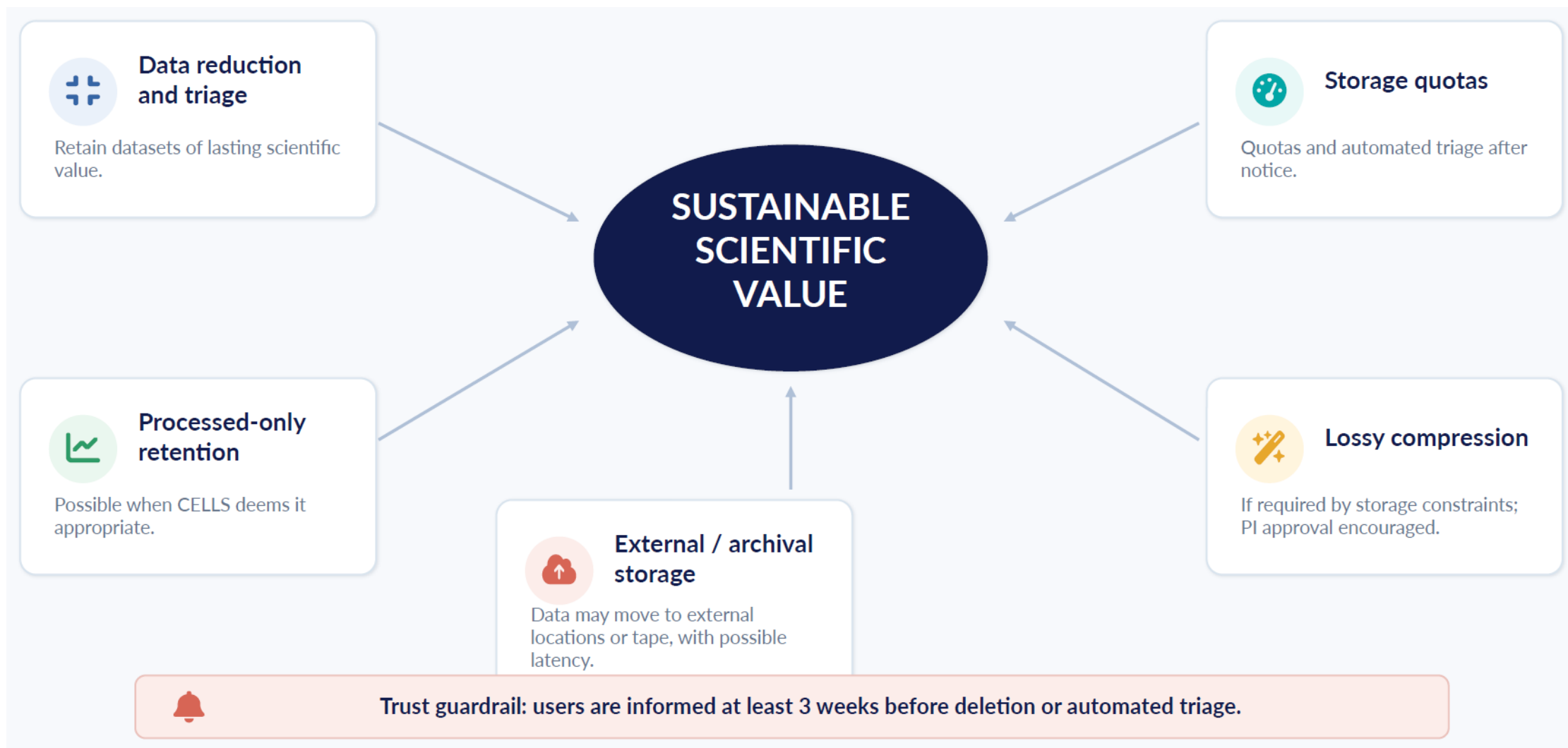
Embracing ecosystems like the European Open Science Cloud (EOSC) allows research entities to maintain distributed infrastructures that interoperate, sharing resources without unnecessary data transit.



Strategic Cloud Use

Public cloud is reserved for compute-heavy, data-light workloads, dynamic public web services, or specialized machine learning bursts where local capital expenditure is not justifiable.

Sustainability Mechanisms in New Data Policy



Conclusions



- The **new Data Policy**, planned for **early 2027**, is based on the following principles:
- **Open by default, protected where necessary**
Balancing openness with privacy, intellectual property, confidentiality and security.
- **Covering the complete scientific data lifecycle**
From acquisition to processing, publication, reuse and long-term preservation.
- **Built on services already in production**
ICAT, NeXus, Globus, DOI minting, VISA (or other DaaS) and European data federation.
- **Sustainable by design**
Defined retention periods, data reduction and resource monitoring.

Your input is a key part of this process.

Appendix- detailed policy changes

AREA	PREVIOUS POLICY	NEW POLICY	PRACTICAL EFFECT
Scope	Raw data and metadata.	"Collected data": metadata, raw, auxiliary, processed, simulated data and results.	Govern the full lifecycle with different commitments.
Principles	LOPD reference; open access after embargo.	LOPDGDD + GDPR; FAIR/open science; "as open as possible, as closed as necessary".	Openness designed with legal, security, IP and ethical safeguards.
Discovery metadata	High-level metadata public after the experiment.	Same default, with proposal-stage non-disclosure for legal, security, patent, contractual or commercial cases.	Findable by default; documented exceptions.
Portal and identifiers	Online catalogue; identifiers described as planned.	ALBA data portal; experiment PID operational; users can mint dataset PIDs; CC-BY archive.	Operational discovery, citation and publication workflows.
Processed data	Minimum six-month retention; deletion possible after notice; results stored without full curation.	Best-effort curation of pipeline outputs; raw-data access rules; processed-only retention; DAaaS.	Processing becomes a managed service layer.
Sustainability	Five-year custody plus instrument exceptions.	Quotas, automated triage, lossy compression, external storage, processed-only retention and deletion notice.	Scalable storage with transparent guardrails.
Instrument rules	JEMCA and FAXTOR exceptions stated in prose.	Appendix matrix for standard ALBA, FAXTOR and JEMCA instruments.	Exact rules are visible by instrument.
User responsibilities	Correct paths/experiment ID; metadata encouraged; cite data.	Minimum metadata, auxiliary data, DMP guidance, triage, citation and publication deposit.	Stewardship becomes part of experimental practice.