Metadata Practices at Synchrotron Facilities

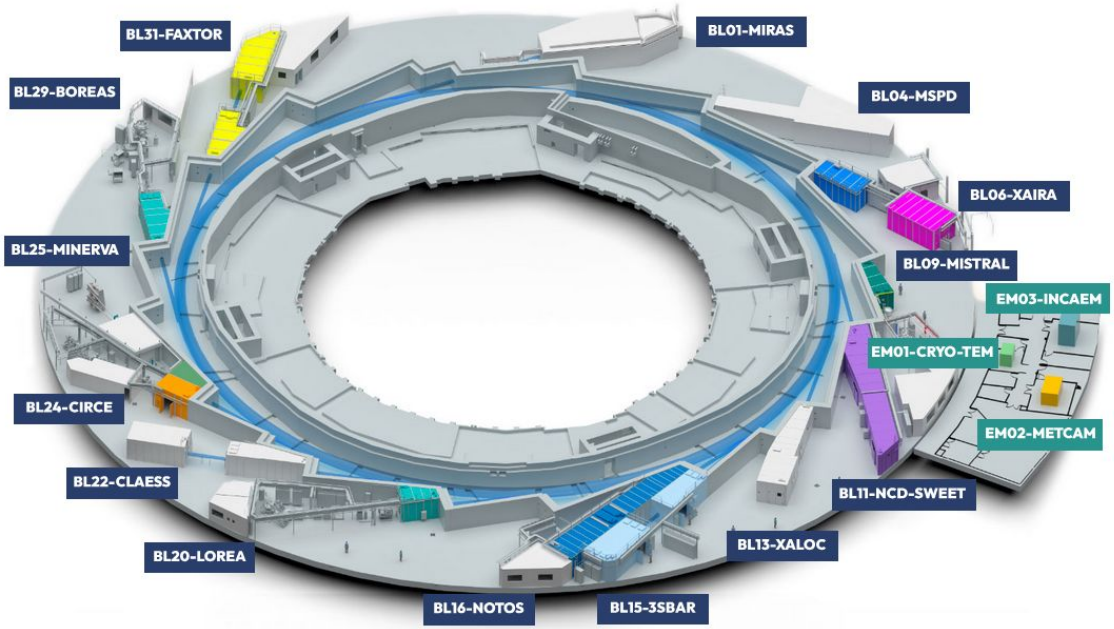
Nicolas Soler, head of SDM, ALBA synchrotron

NAPMIX Training Workshop 2026

19/05/2026

Beamlines and Electron Microscopes

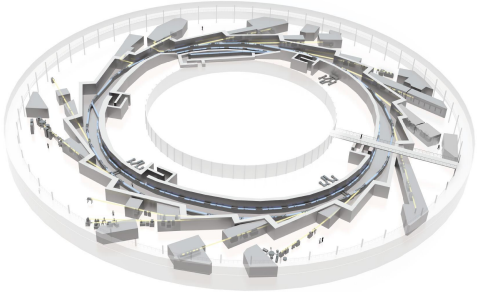
BL01 - MIRAS
BL04 - MSPD
BL06 - XAIRA
BL09 - MISTRAL
BL11 - NCD-SWEET
BL13 - XALOC
BL15 - 3Sbar
BL16 - NOTOS
BL20 - LOREA
BL22 - CLÆSS
BL24 - CIRCE
BL25 - MINERVA
BL29 - BOREAS
BL31 - FAXTOR
EM01-CRYO-TEM
EM02-METCAM
EM03-INCAEM

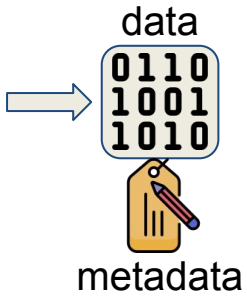


ALBA synchrotron at a glance: https://www.cells.es/en/instruments/map

The Scientific Data Management section

Our missions





Data processing
(mainly beamline support)

Data management
In synergy with other sections in Computing

- Orchestrating FAIR metadata capture (with CT)
- Providing provenance tools
- Persistent identifiers management (with MIS)
- Participate in data catalogue implementation (with MIS)
- Data analysis platform (with MIS)
- Data management plans

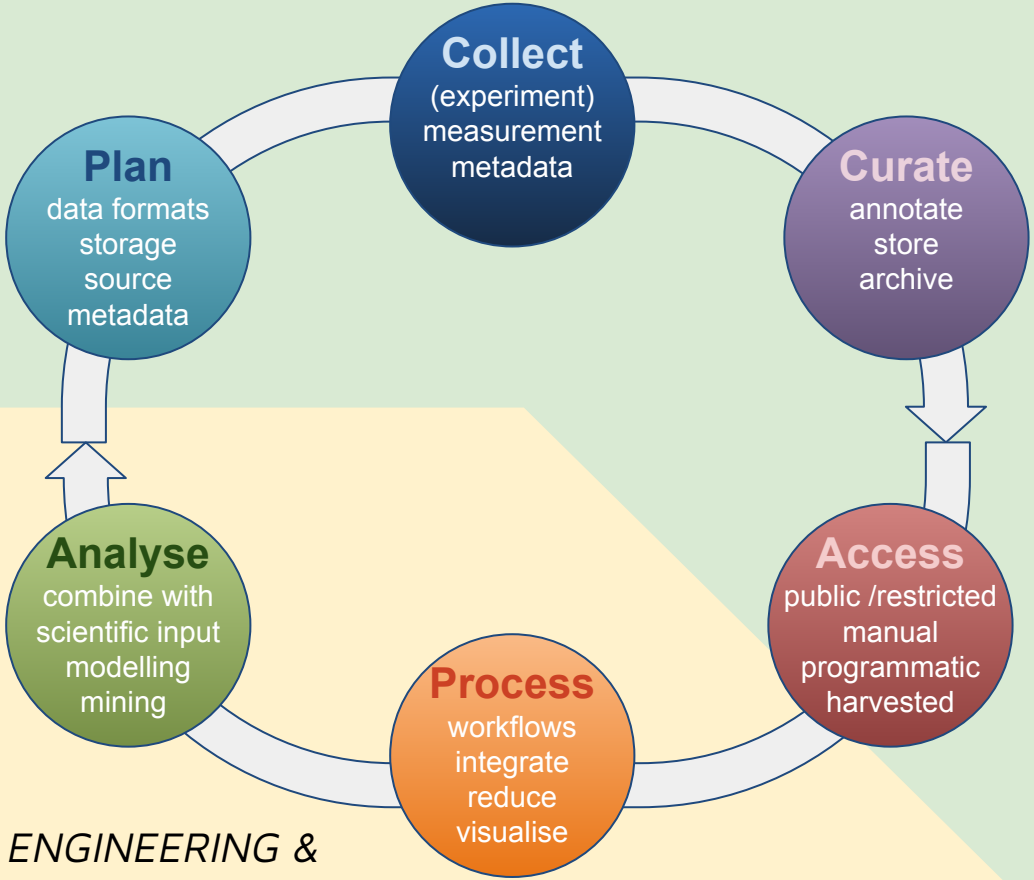
Section created in 2021, now 6 people
- 5 software engineers
- 1 section responsible

Data Management

DATA MANAGEMENT

SCIENTIFIC COMPUTING

SOFTWARE ENGINEERING & ENVIRONMENT



Changes and challenges

Upgrade to ALBA-II
New instruments and detectors

More brilliance & coherence

kHz fast detectors

MX, µCT, Ptycho, EM, etc

Faster data collection for conventional method

Previously marginal techniques will become routine (SSX, Ptychography)

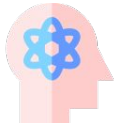
More data open, reusable

Needs

- Storage capacity
- Bandwidth
- HPC power
- Automation
- Scientific data processing expertise
- FAIR data policy
- Interfacility coordination and service federation (software and data)

European Commission, funders

Public research data must be browsable & **reusable** by others (FAIR)



Data challenges on the horizon

Data Challenges

- Data Volume
- Data Processing
- Data Complexity

High-quality data annotation systems are required.

EOSC: European Open Science Cloud

In May 2015, the European Commission proposed creating a European Open Science Cloud (EOSC) to the Competitiveness Council.

The aim was to federate existing research data infrastructures in Europe and realise a web of FAIR data and related services for science, making research data interoperable and machine actionable following the FAIR guiding principles.

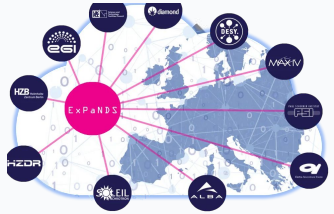
In the initial phase of development until 2020, the Commission invested around €320 million to start prototyping the EOSC through project calls in Horizon 2020 - the Commission's research and innovation funding programme.

The partnership aims to deploy and consolidate by 2030 an open, trusted virtual environment to enable the estimated 2 million European researchers to store, share and reuse research data across borders and disciplines.
The partnership will bring strategic coherence and complementary commitments at EU, national and institutional levels to bring together all advanced data infrastructures in Europe, modernise the ERA with a capability to produce "FAIR-by-design" (Findable, Accessible, Interoperable, Reusable) datasets, and populate EOSC with new FAIR data and related services, (iv) expand the FAIR data culture across Europe.

https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/european-partnerships-horizon-europe/candidates-across-themes_en

The PaNOSC and ExPaNDS projects

Photon **a**nd **N**eutron **O**pen **S**cience **C**loud
European Open Science Cloud (EOSC) **P**hoton **a**nd **N**eutron **D**ata **S**ervice



Were **4-year projects** started respectively in 2018 and 2019 to expand, accelerate and support the **data management** and **data services** provided through the EOSC for major national Photon and Neutron Research Infrastructures (PaN RIs)

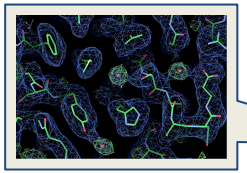
- Enable EOSC services and to **provide coherent FAIR data services** to the scientific users of national Photon and Neutron sources
- Connect national PaN RIs through **a platform of data analysis as a service** (DAaaS) for users from research institutes universities, industry etc.
- Develop and maintain a **catalogue of data and analysis software** for Photon and Neutron data
- Gather feedback and **cooperate with the EOSC** governance bodies to improve the EOSC and develop standard relationships between scientific publications, Photon and Neutron scientific dataset (raw data), experimental reports, instruments and authors (via ORCID)

https://www.panosc.eu/panosc-project/

https://expands.eu/

WP 2,3,4 overview

panosc
ExPaNDS



User: connection from local workstation or home laboratory

re-process

Data analysis as a service (WP4)

- **EOSC Service portal**
- **PaN portal**
- **Containers**
  Virtual machines
  Remote desktops
  Jupyter notebooks
- **HPC**
- **Reference datasets**

Use re-use

Process raw data

Annotate & store processed data + provenance

FAIR-ready data (WP2)

- **Certified** (meta)data repository
- Standard metadata framework
- FAIR Data Management Plan (**DMP**) for each technique
- Persistent Identifier (**PID**) for data, instrument, software, sample)
- Common **data policy** framework

Data catalogues (WP3)

- **API / UI**
- **EOSC hub**
- **Metadata ontologies** (shared vocabulary)
- PaNOSC **federated catalogue**
- **Metadata formats** (NeXus)
- **Catalogues** (eg iCAT, SciCat)

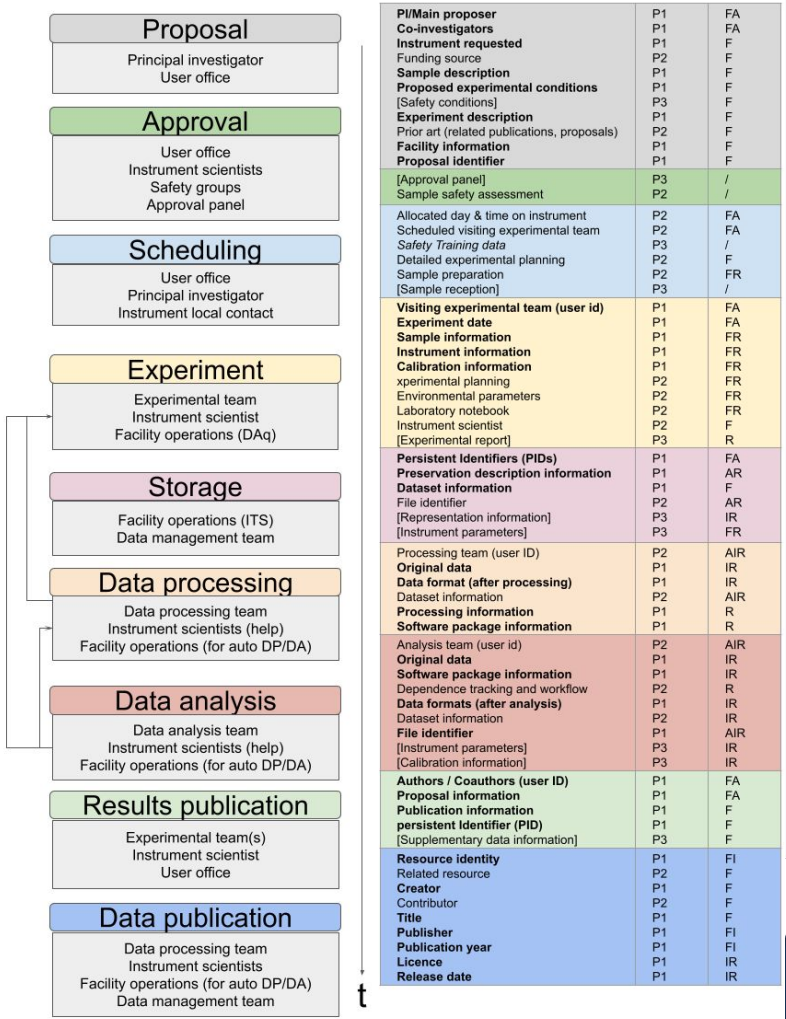
Data reduction / compression

- Metrics
- New algorithms
- Software vs hardware
- Technique-specific
- Lossy vs non lossy
- Meta-compressors (ML)

LEAPS INNOVATION (WP7)

Some of the outcomes

- A common FAIR data policy framework
- Common basis for Data Management Plans (DMP)
- Metadata recipes
- Common solutions of metadata catalogues: mainly Scicat and ICAT
- A federated interface for searching data
- Ontologies (NeXus and PaNET)
- A common (meta)data format: NeXus/HDF5
- A common solution for remote data processing and visualization: VISA



Source: https://doi.org/10.5281/zenodo.6799106

A Common approach to manage scientific data

Data policy
Regulations
Rights
Duties

Data Management Plan (DMP)
Think ahead

Data storage
Ultra fast
Disc
Tape

Data format:
Wrapping data and metadata

NX

Data catalogue
Interface to find and browse data, logbook

ICAT

Automatic pipelines
Produce processed data

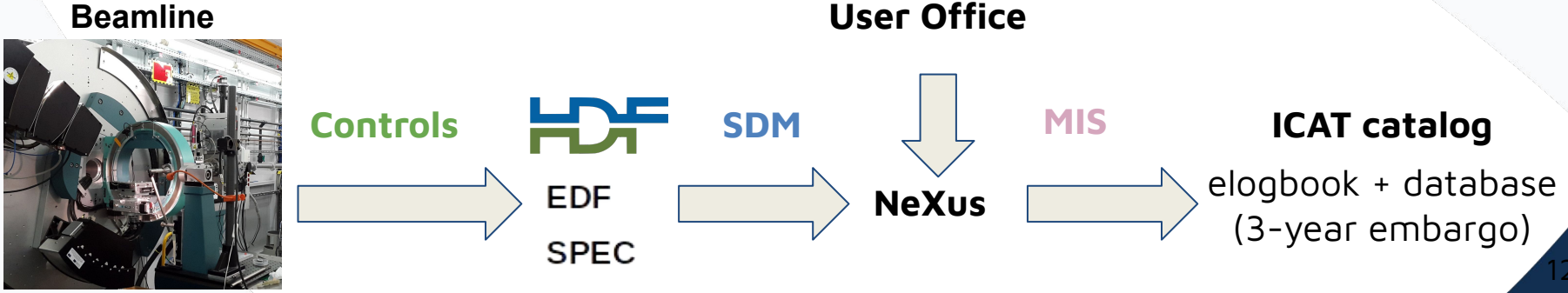
During and after experiment

(Re)processing platform
Provides remote interfaces for using ALBA's computing resources

VISA

(meta)data format: NeXus/HDF5

- ALBA synchrotron has pledged to follow the FAIR principles
- **NeXus** taken as the standard format to store (meta)data in ALBA to comply with FAIR principles
- Since 2023, Dr. Fernán Saiz (SDM) is acting as ALBA's representative in the NeXus International Advisory Committee (NIAC)
- **NIAC** develops and maintains NeXus as the common format for neutron, x-ray, and muon science facilities in Europe, Asia, Australia, and North America.
- Raw data is retained for a minimum period of 5 years (archived after 1 year)



(meta)data format: NeXus/HDF5

3.3.2. Application Definitions

A description of each NeXus application definition is given. NeXus application definitions define the *minimum* set of terms that *must* be used in an instance of that class. Application definitions also may define terms that are optional in the NeXus data file. The definition, in this case, reserves the exact term by declaring its spelling and description. Consider an application definition as a *contract* between a data provider (such as the beam line control system) and a data consumer (such as a data analysis program for a scientific technique) that describes the information is certain to be available in a data file.

Use NeXus links liberally in data files to reduce duplication of data. In application definitions involving raw data, write the raw data in the NXinstrument tree and then link to it from the location(s) defined in the relevant application definition.

Application definitions are grouped together based on the research fields where these are typically used. Definitions that address multiple research fields are listed in each category:

- Atom Probe Microscopy
- Diffraction & Scattering Techniques
- Electron Microscopy
- Imaging Techniques
- Multi-Dimensional Photoemission Spectroscopy
- Optical Spectroscopy
- Time-of-Flight Techniques
- Complete List

https://manual.nexusformat.org/classes/applications/index.html

(meta)data format: NeXus/HDF5

3.3.2.2. Diffraction & Scattering Techniques

Introduction

Application definitions for different diffraction and (small-angle) scattering techniques

Application Definitions

NXiqproc
: Application definition for any $I(Q)$ data.

NXlauetof
: This is the application definition for a TOF laue diffractometer.

NXmonopd
: Monochromatic Neutron and X-Ray Powder diffractometer.

NXxbase
: This definition covers the common parts of all monochromatic single crystal raw data application definitions.

NXxeuler
: Raw data from a four-circle diffractometer with an eulerian cradle, extends NXxbase.

NXxkappa
: Raw data from a kappa geometry (CAD4) single crystal diffractometer, extends NXxbase.

NXxlaue
: Raw data from a single crystal laue camera, extends NXxrot.

NXxlaueplate
: Raw data from a single crystal Laue camera, extends NXxlaue.

NXxnb
: Raw data from a single crystal diffractometer, extends NXxbase.

NXxrot
: Raw data from a rotation camera, extends NXxbase.

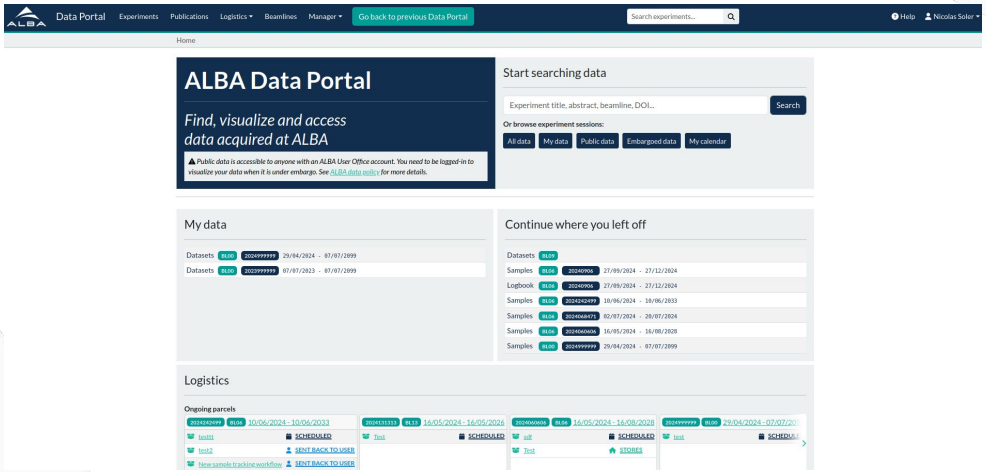
NXcanSAS
: Implementation of the canSAS standard to store reduced small-angle scattering data of any dimension.

NXsas
: Raw, monochromatic 2-D SAS data with an area detector.

NXsastof
: Raw 2-D SAS data with an area detector with a time-of-flight source.

https://manual.nexusformat.org/classes/applications/diff-structure.html#appdef-diff-structure

Implementation of the data catalogue (ICAT)

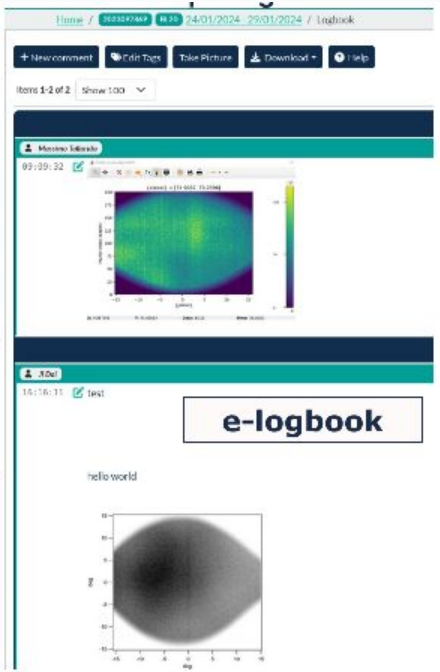


https://data.cells.es

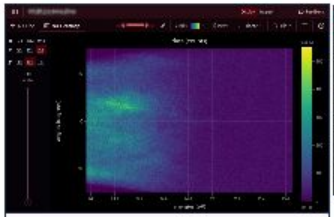
- ICAT is a data catalogue solution, resulting from a collaboration between different european facilities (ESRF, SFTC, HZB, Diamond, etc)
- It allows finding, browsing, viewing, downloading and annotating data during and after the experiment
- Now in **deployment phase** at ALBA

Implementation of the data catalogue (ICAT)

The **data catalogue** allows browsing, downloading annotating and visualising your data during and after experiment



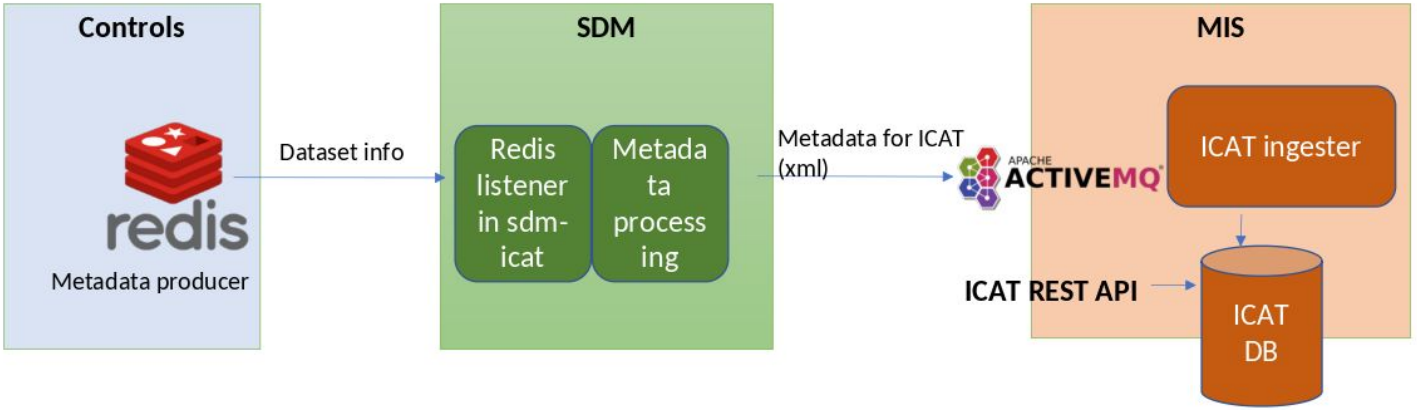
e-logbook



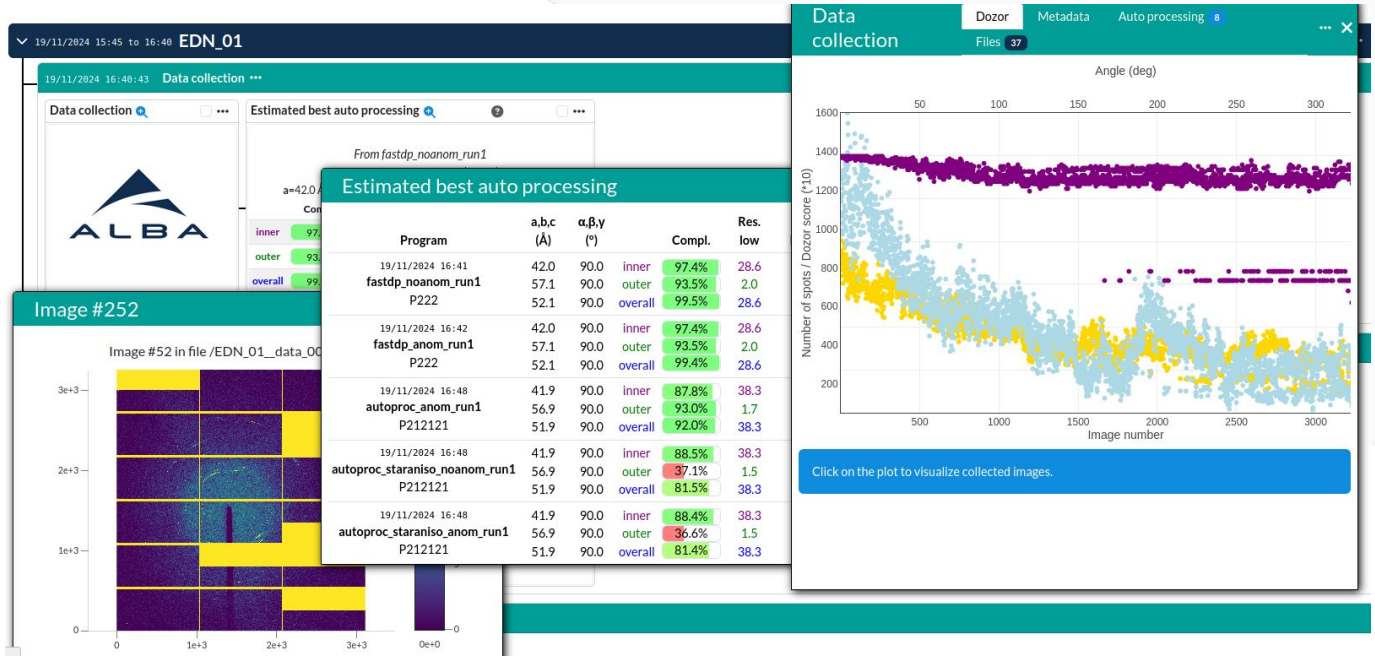
Integrated NeXus / HDF5 data **viewer**

Publication **landing page** with data DOI

ICAT communication setup



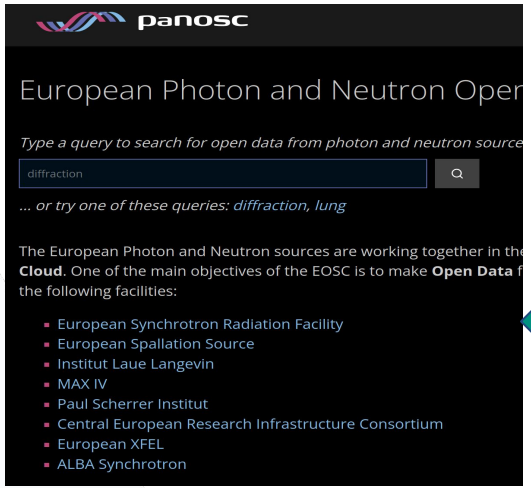
ICAT-DRAC (for XAIRA, developed by ESRF)



An integrated Laboratory Information Management System (LIMS)

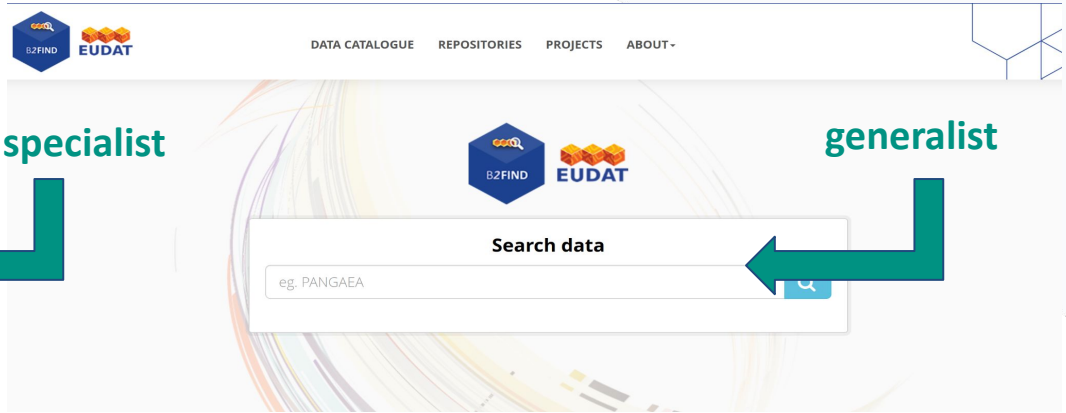
Harvesting PaN Data catalogues

- We aim to 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.



PaN specialist

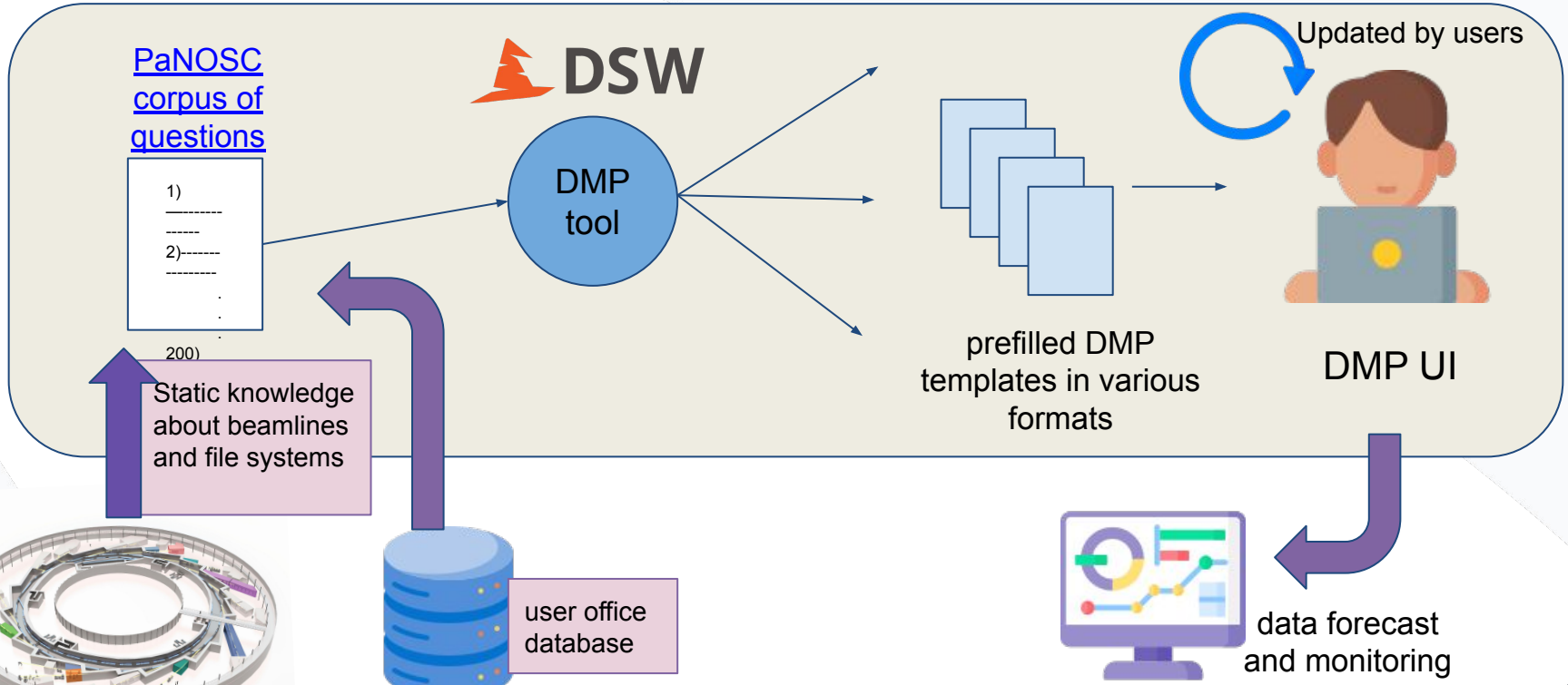
generalist



https://data.panosc.eu/

https://b2find.eudat.eu

Data Management Plans (DMP)



Current data policy for public experiments
A few keypoints

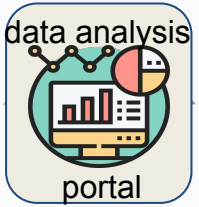
- **No need to copy data home**, ALBA custodies it
- Raw data is **retained** for a minimum period of **5 years** (archived after 1 year)
- A **3-year embargo period** is applied to raw data, afterwards it becomes **publicly available**
- High-level metadata, such as proposal title, authors and abstracts, are visible immediately after the experiment is completed.

We must update our current data policy

- Based on the template and guidelines from the PaNOSC and ExPaNDS projects
- Explicitely states our commitment to embrace the FAIR principles
- E.g. commits to rich metadata annotation, persistent identifiers, interoperable formats, etc
- Must include derived/processed data (and software)
- Encompasses our new Microscopes (InCAEM)
- Must take into account operational sustainability of the long term access and custody of the data

3) Keystones

Data Analysis as a Service: VISA

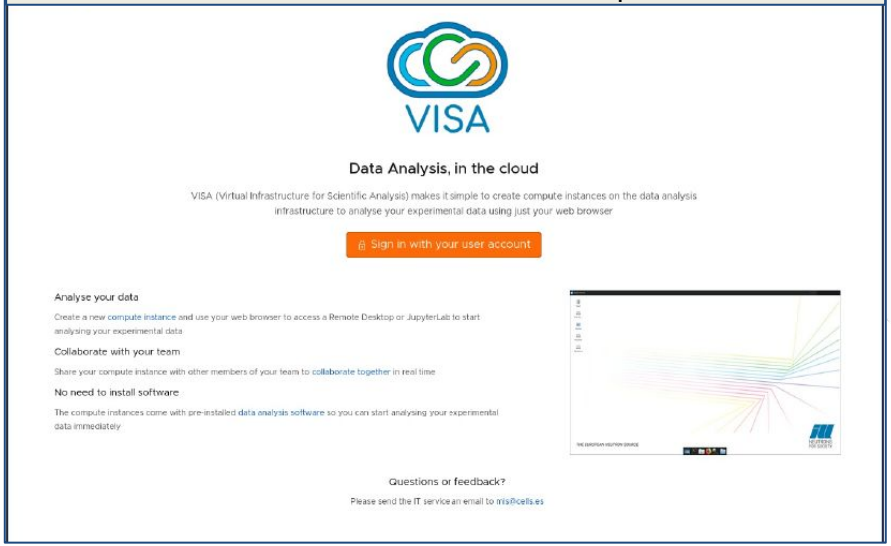


Portal for (re)processing / remote data analysis using HPC resources. (leading group: **MIS**)

- Active collaboration (MoU) between ALBA and other European list institutes
- can be run on premises or even on external backends (e.g Google Cloud)
- MIS team is exploring **new ways of deploying it** (Kubernetes, Windows)
- ALBA is candidate for the first in-person workshop in 2025

Why VISA?

- Some data are becoming too big to be transferred to home institution
- Users take advantage of the facility HPC
- Exhaustive software environment and provenance



FACILITIES INVOLVED:
ALBA, DESY, ILL, ESRF, SOLEIL, ESS, EuXFEL

Tools to manage scientific data

Data policy
Regulations
Rights
Duties

Data storage
Ultra fast
Disc
Tape

Data Management Plan (DMP)
Think ahead

Data format: NX
Wrapping data and metadata

Data catalogue ICAT
Interface to find and browse data, logbook

Automatic pipelines
Produce processed data

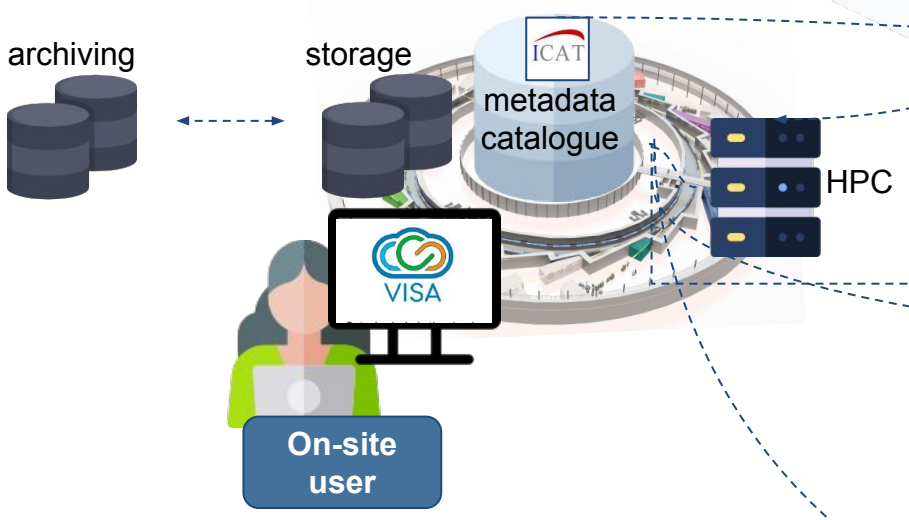
During and after experiment

(Re)processing platform
Provides remote interfaces for using ALBA's computing resources

VISA

Entire setup ready: foreseen for 2026

Vision of future data services for users



Outside World

Remote (re)user

Cloud / HPC

External databases/ interfaces

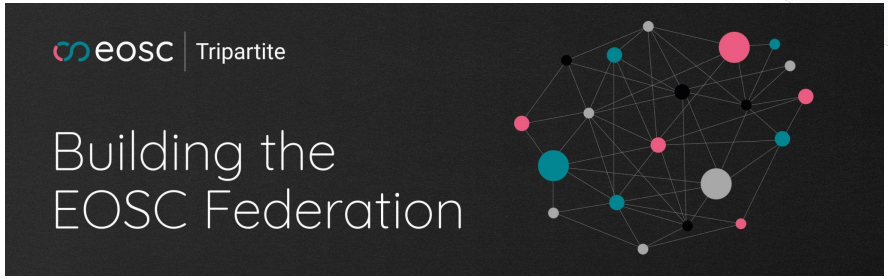
B2FIND
OpenAIRE

EOSC data discovery platforms

Non-specialist

Finding, Accessing, Understanding, Reprocessing data

PaN EOSC node



The European Open Science Cloud (EOSC) is federating European data and services under various thematic nodes.

The first NODE (EU) was launched in October 2024!

ALBA member of the future PaNOSC node

Conclusions

- As a prerequisite for new developments in data processing & automation, ALBA needs to fully implement a FAIR-compliant data management setup. This includes:
  - An update of our **data policy**
  - Interoperable data **formats** (mainly NeXus) and standards
  - The deployment of a data **catalogue**, which includes
    - findability tools, (also via the PaN data portal and OpenAIRE/B2find)
    - persistent identifiers (DOIs)
    - data landing pages
    - electronic lab. Notebooks
    - data visualization tools (e.g. NeXusviewer, ICAT-DRAC)
  - A **Data Analysis as a Service** platform (VISA), available on-site and from home institution
  - Alignment with / participation in joint PaN **Open Science** initiatives at European level

The first beneficiaries of this FAIR setup are the **users** themselves!

What's next?

- *Harmonized and exportable data processing workflows, providing rich metadata (provenance)*
- *More metadata standards, in particular for the **samples***
- *Combining data from different experiments*
- *AI to evaluate data quality*
- *Real time feedback (experiment steering)*
- *Tailored data compression*
- *AI- assisted DMP generation*
- *Software sharing among PaN facilities*
- *etc*

For all this, we need a good data management basis !

Acknowledgements

Controls
Oriol Vallcorba
Fulvio Becheri

ITS
Gemma Rosas
Sergi Pusó
Sergio Vicente

MIS
Rodrigo Cabezas
Marc Armenter
Toni Fernández

Computing Head: Óscar Matilla

SDM
Emilio Centeno, Albert Castellví, Fernán Saiz
Przemyslaw Karczmarczyk, Joaquín Gómez, Gabriel Jover

ESRF: Alex de Maria, Marjolaine Bodin, Andy Goetz

Questions? nsoler@cells.es