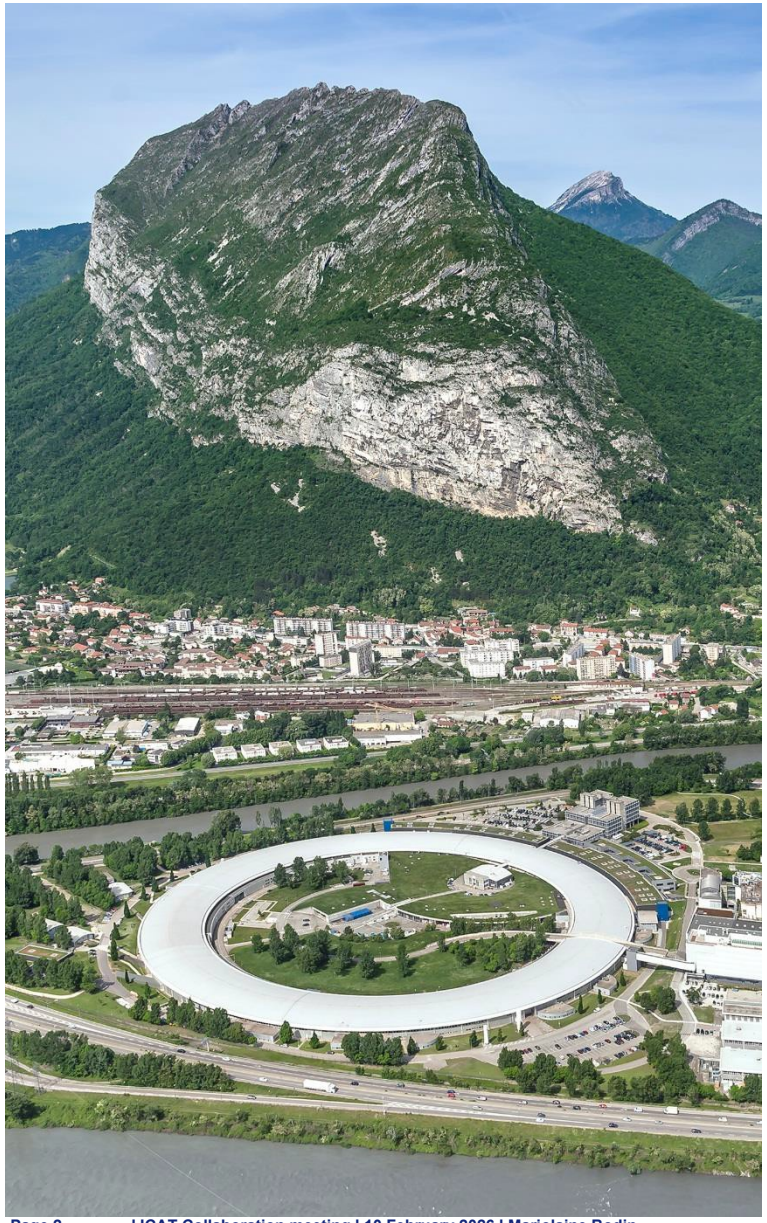
ESRF
The European Synchrotron



Data management at ESRF: what's new and what's next

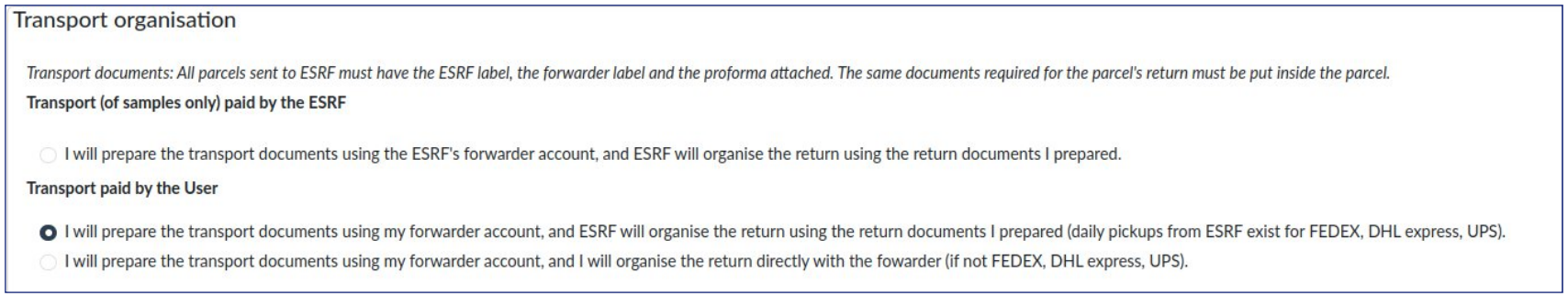
Marjolaine Bodin
(ISDD / Software Group / DAU)
marjolaine.bodin@esrf.fr

OUTLINE: FROM THE EXPERIMENT PREPARATION TO THE PUBLICATION

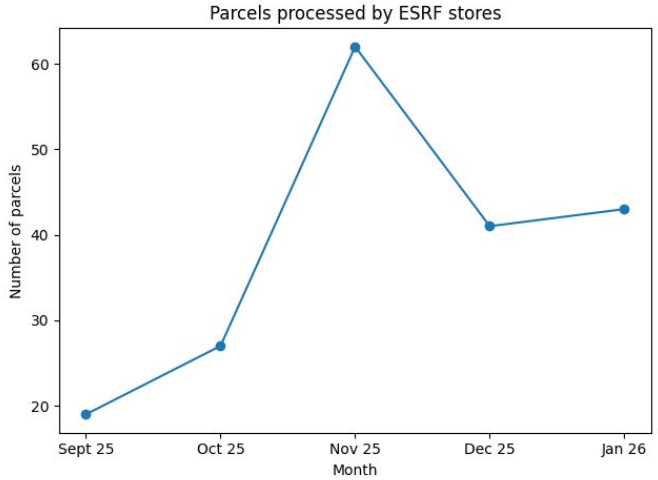
- Experiment preparation: sample tracking, experiment setup, sample metadata
- Tailored display: beamline, techniques
- ESRFET ontology
- DracBot: AI assistant
- Logbook/Notebook
- Invoicing, statistics
- DOIs
- Domain Portals
- Other ongoing & upcoming topics

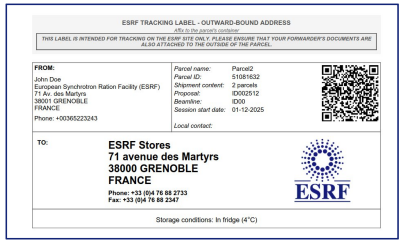
EXPERIMENT PREPARATION: SAMPLE TRACKING

Improvement of sample tracking (on-site parcel tracking), across all beamlines.



New sample transport organization section

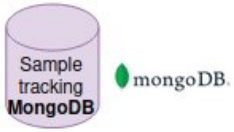




Labels

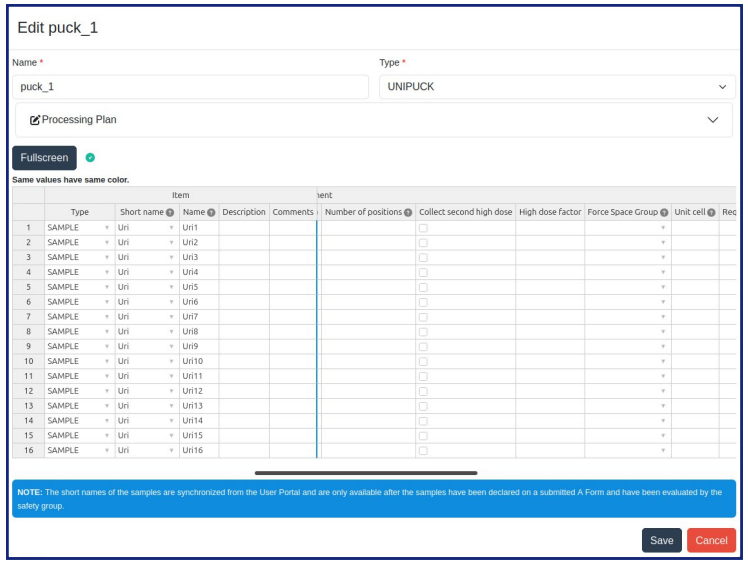
Parcel status

- **ESRF stores are better organized and standardized.**
- **Reduces loss or misplacement of parcels**
- **ISPyB sample tracking will be stopped in March**



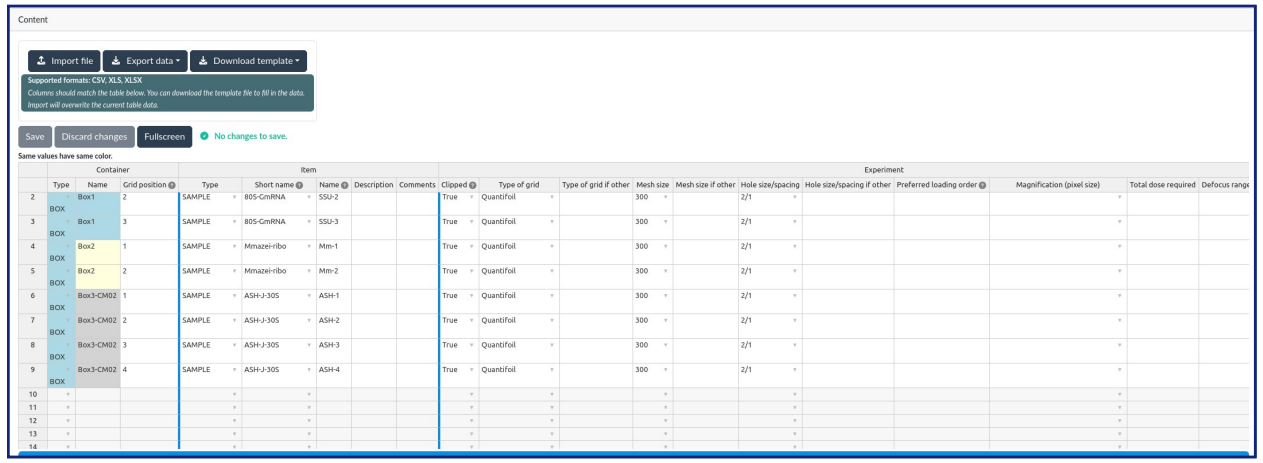
EXPERIMENT PREPARATION: EXPERIMENT SETUP

Experiment and processing plan managed in the Data Portal, configurable per beamline

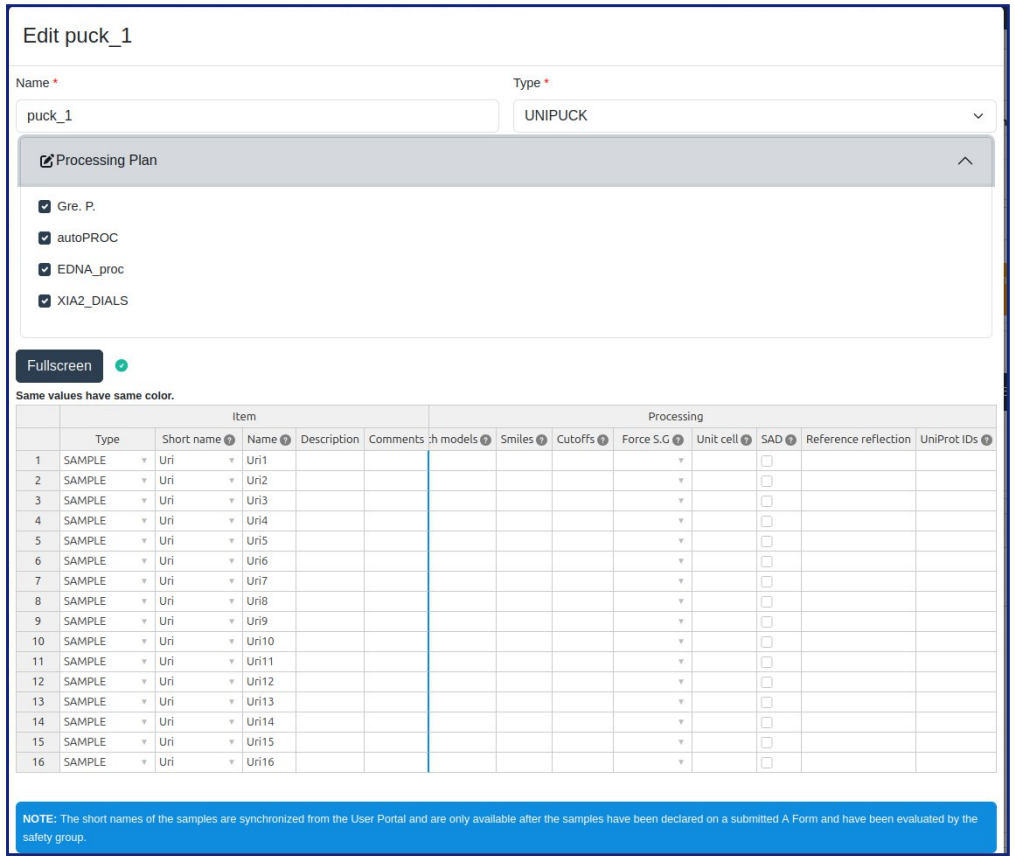


Macromolecular X-ray crystallography

Experiment plan



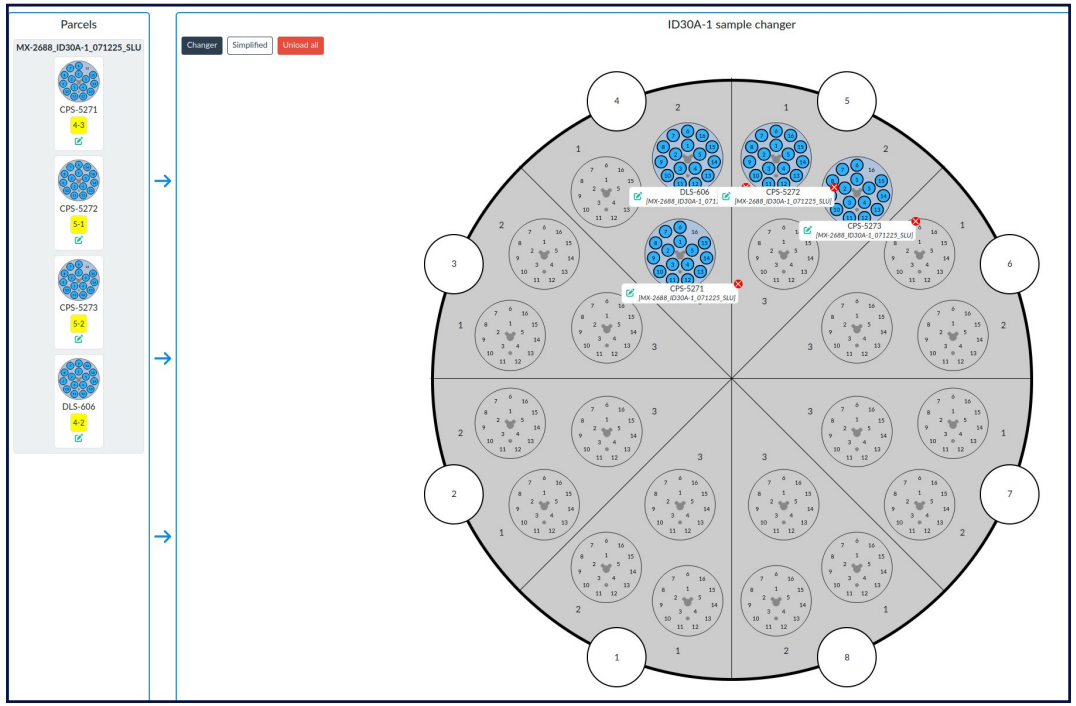
Cryo-electron microscopy



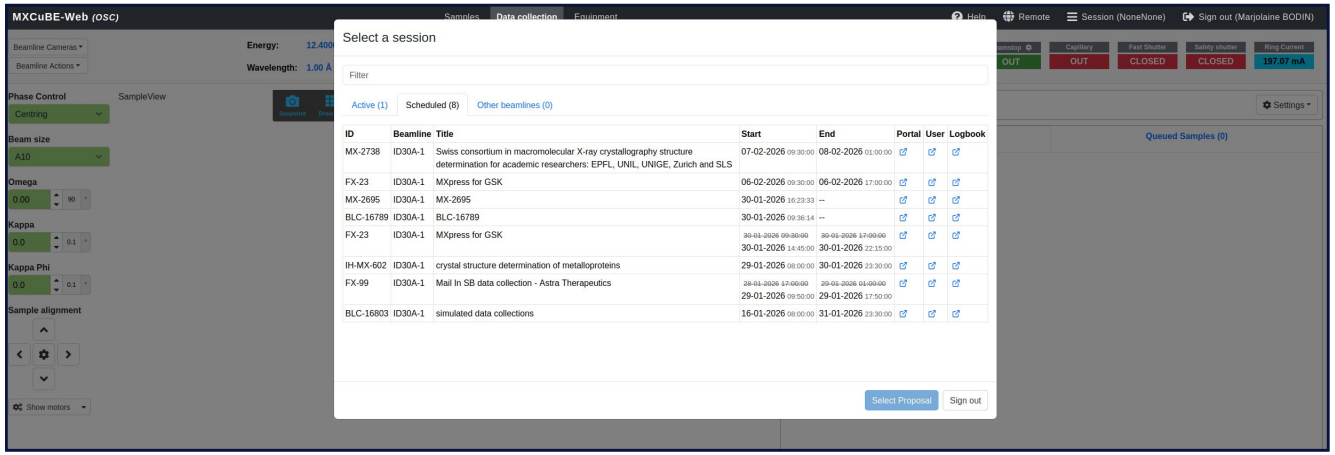
Processing plan (MX)

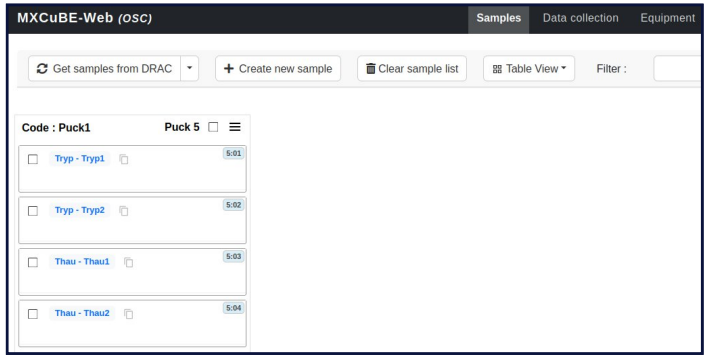
EXPERIMENT PREPARATION: EXPERIMENT SETUP

Configure and set up the sample changer via the Data Portal



Information synchronized in the data acquisition system MXCuBE





- Data collection is guided by the predefined parameters
- Analysis is performed according to user-defined processing strategies

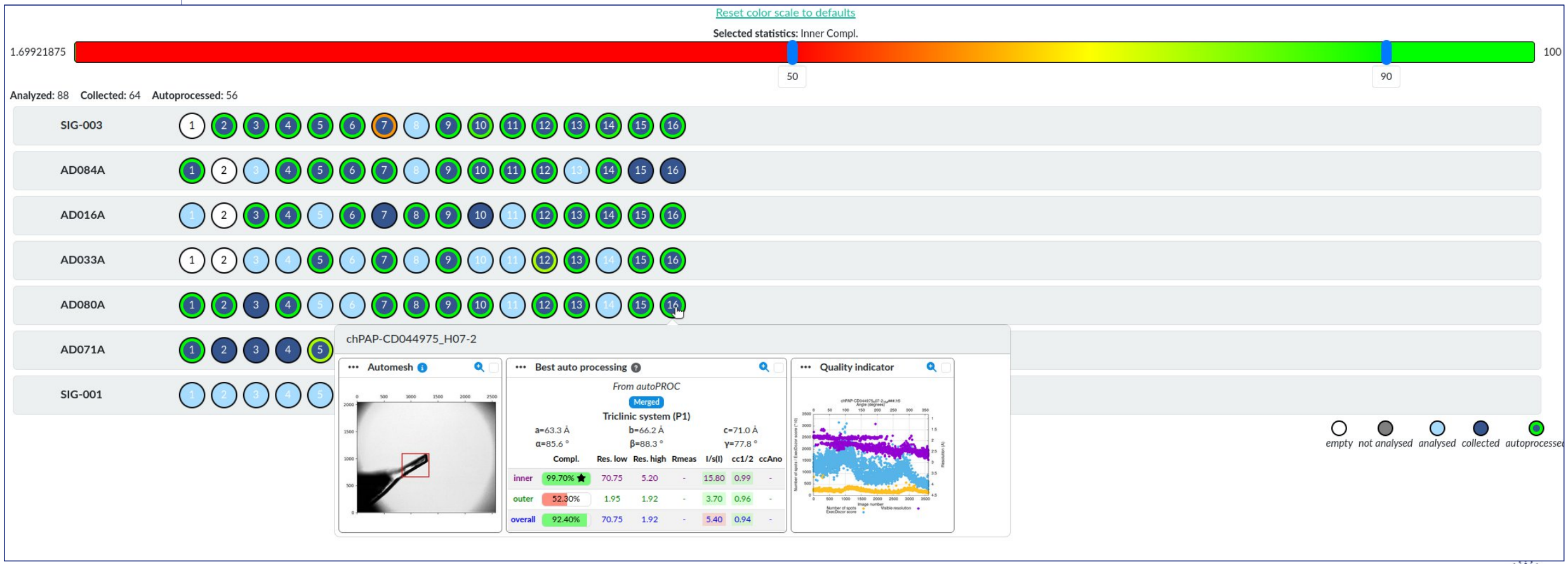
Oscarsson, M. et al. 2019. "MXCuBE2: The Dawn of MXCuBE Collaboration." Journal of Synchrotron Radiation 26 (Pt 2): 393–405.

METADATA-DRIVEN EXPERIMENT SUMMARY

From detailed experiment metadata => clear, concise summaries for easy access and review

Show progress in data processing

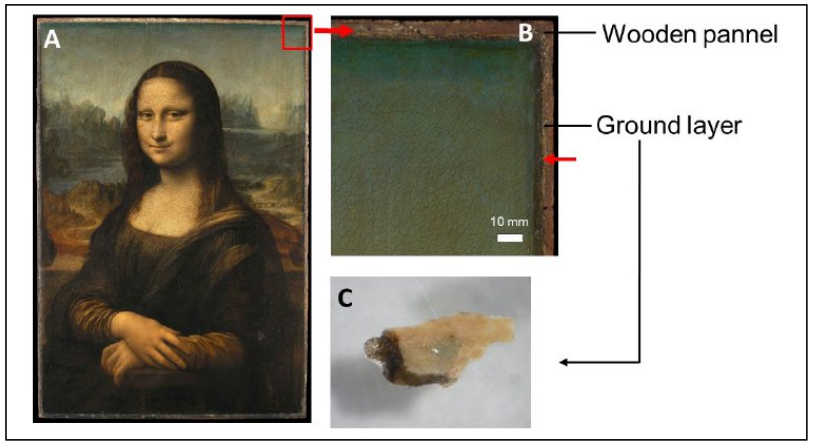
Could be extended to other techniques



EXPERIMENT PREPARATION: SAMPLE METADATA

Context: OSCARS ***SHARE*** *project : Synchrotron X-ray analysis of Heritage Accessible to and Reusable by Everyone (start date: 01/01/2025)*

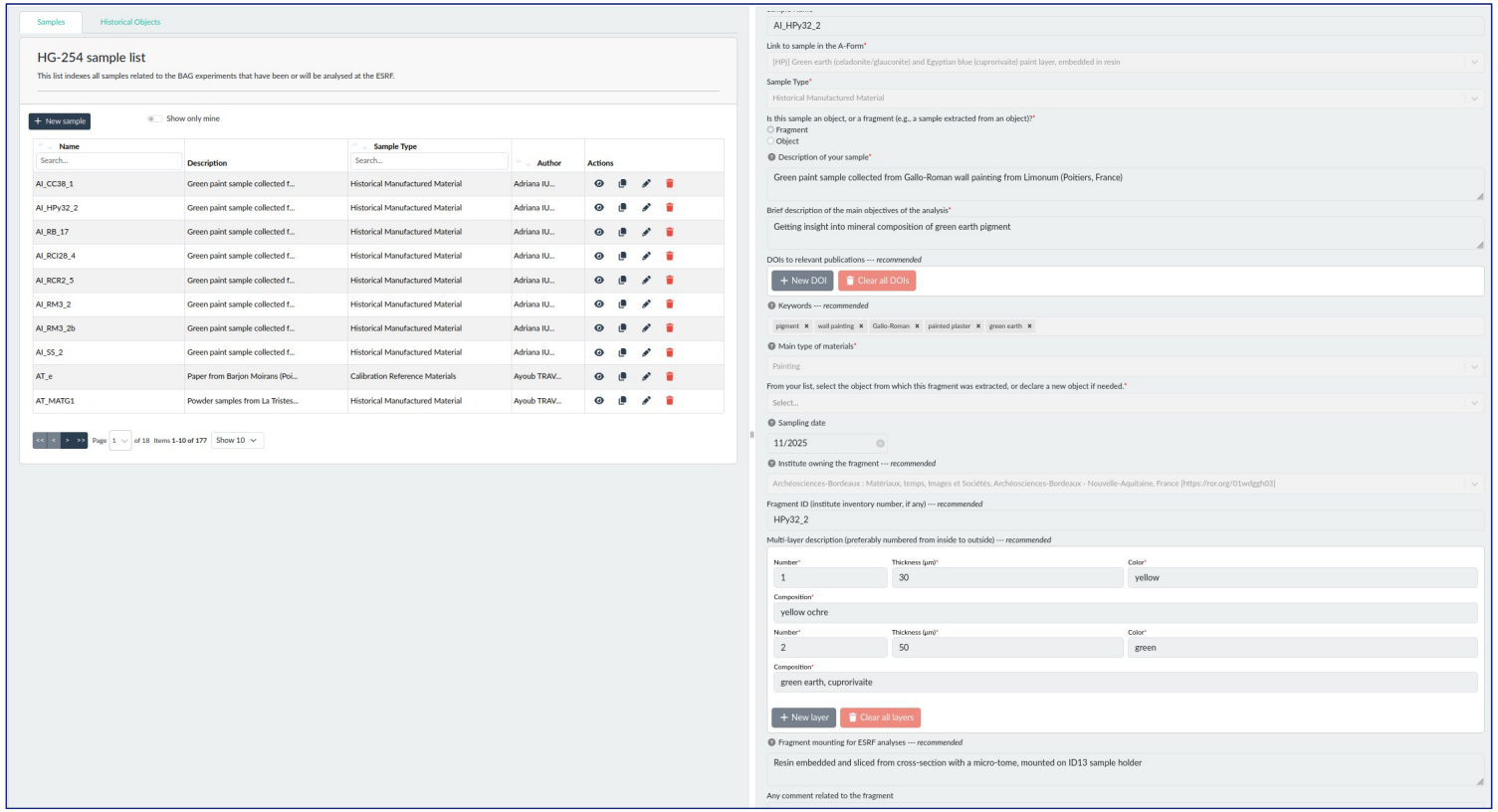
- Searchable database of art objects and model samples (Historical Materials BAG)
- Includes rare and precious samples
- Access to raw & processed data, with full metadata
- Links to museum databases
- Preserves valuable data, promotes FAIR/Open Science
- Enables global research, sharing, and reuse



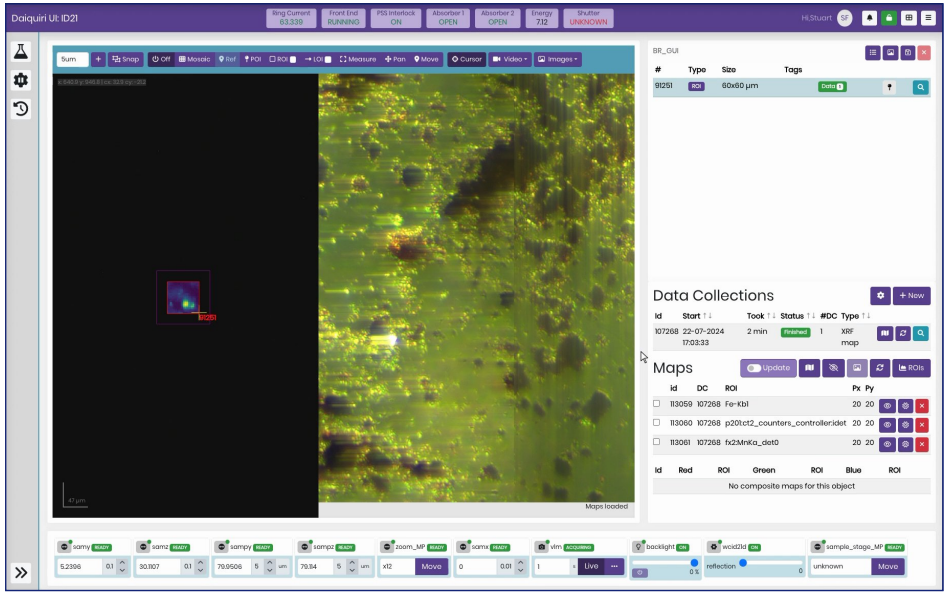
Study of the ground of Mona Lisa
Credits: M. Cotte

Gonzalez, V.; Wallez, G.; Ravaud, E.; Eveno, M.; Fazlic, I.; Fabris, T.; Nevin, A.; Calligaro, T.; Menu, M.; Delieuvin, V.; Cotte, M. X-ray and Infrared Microanalyses of Mona Lisa's Ground Layer and Significance Regarding Leonardo da Vinci's Palette. J. Am. Chem. Soc. ***2023****, 145 (42), 23205– 23213, DOI:10.1021/jacs.3c07000*

EXPERIMENT PREPARATION: SAMPLE METADATA



Sample metadata form for Cultural Heritage



Daiquiri
Credits: S. Fischer

https://doi.org/10.1107/S1600577521009851

Information synchronized in the data acquisition system Daiquiri

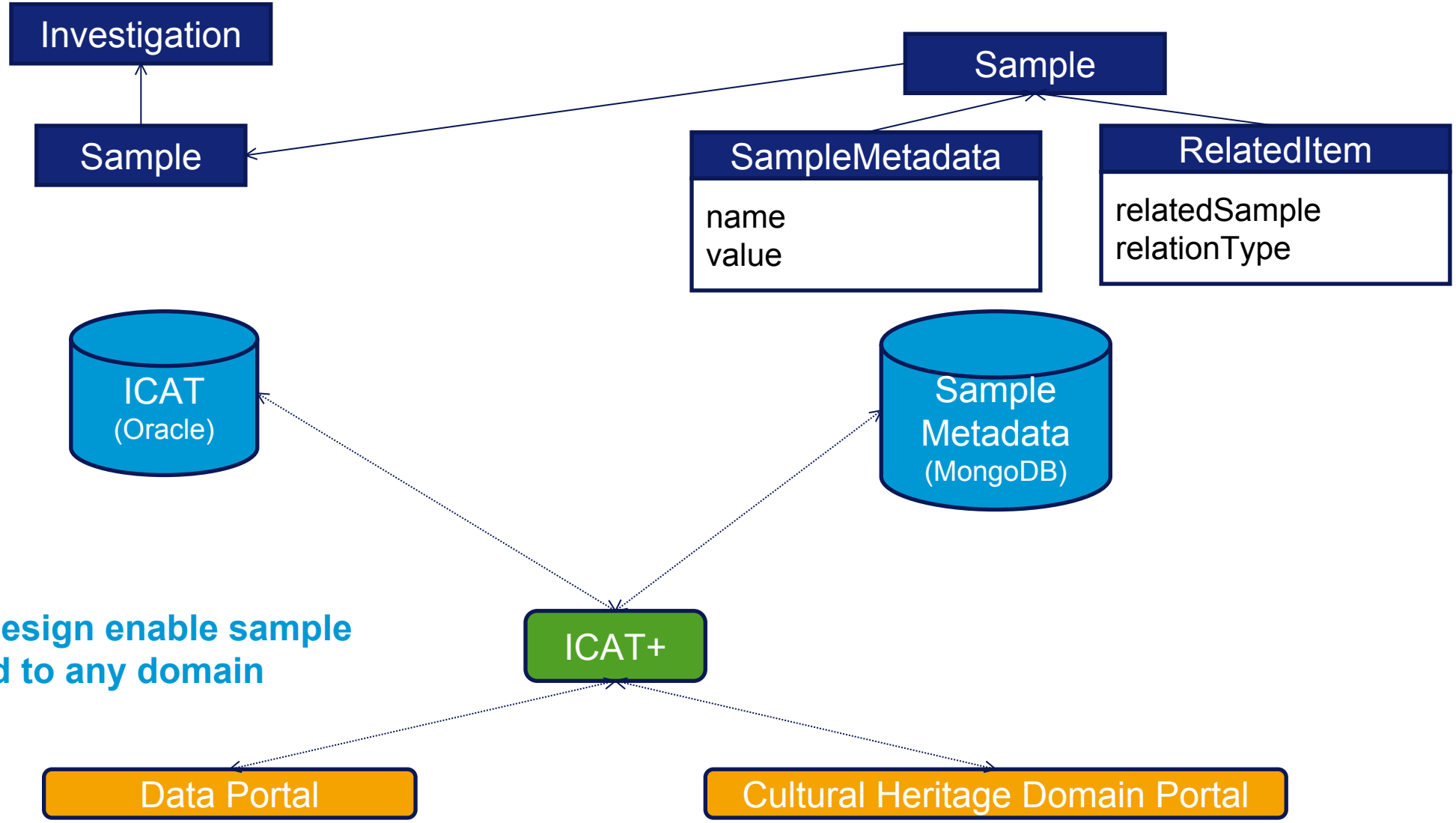
- **Consistent, high-quality acquisition metadata**

Used in production in the last run of 2025

- **3 experiment sessions**
- **> 200 samples created**

EXPERIMENT PREPARATION: SAMPLE METADATA - ARCHITECTURE OVERVIEW

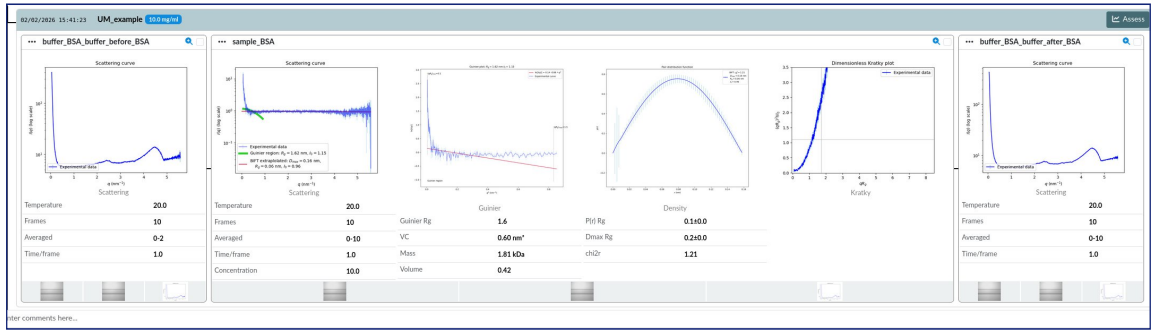
High interest: many-to-many relation Sample-Investigation



Architecture and design enable sample metadata to extend to any domain

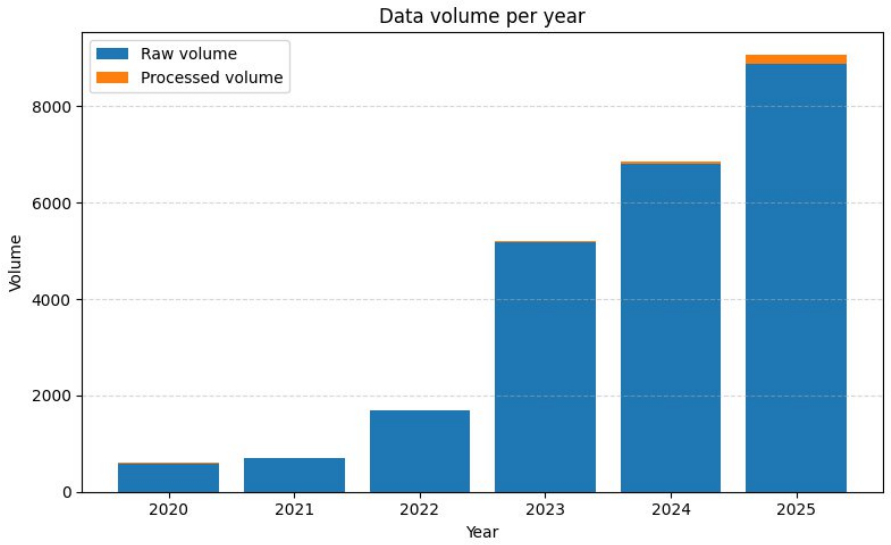
DATA VISUALIZATION: TAILORED DATA DISPLAY

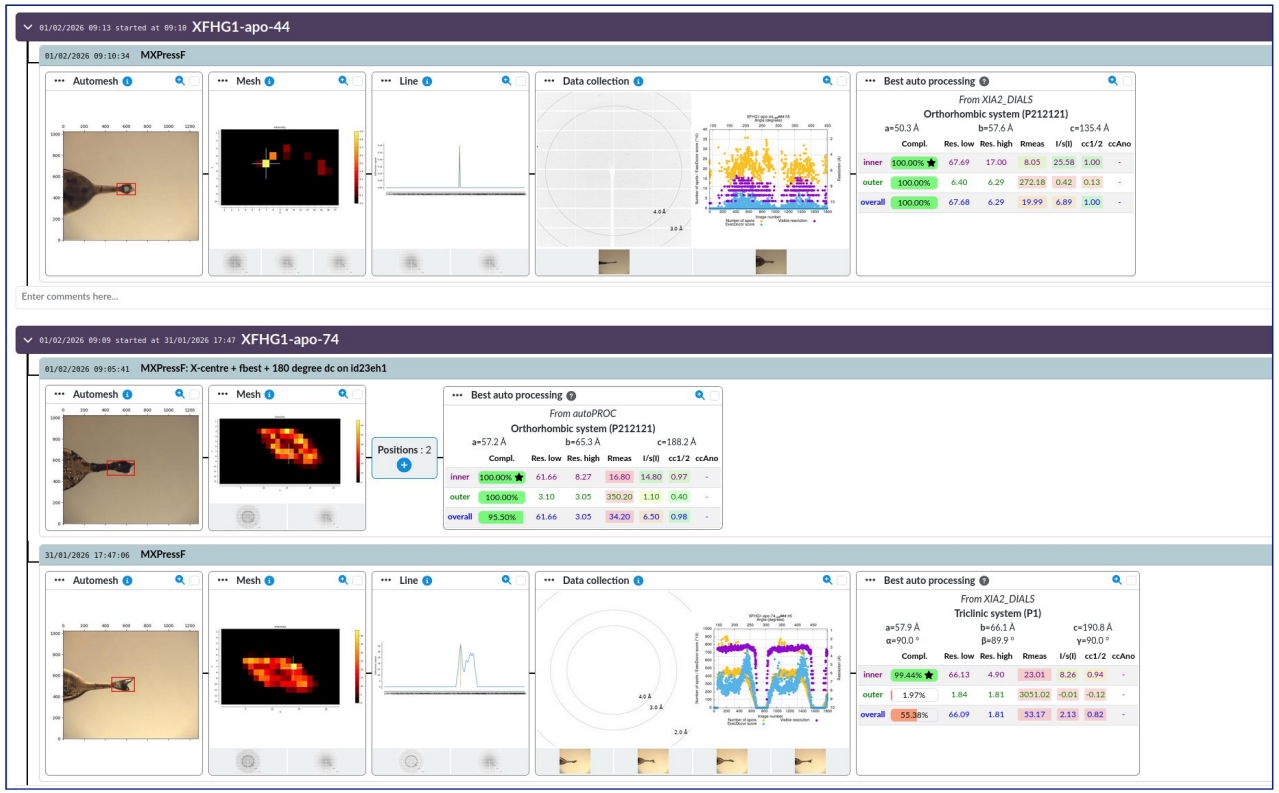
Beamline-specific



BioSAXS beamline - BM29

Processed data is progressively increasing in ICAT.



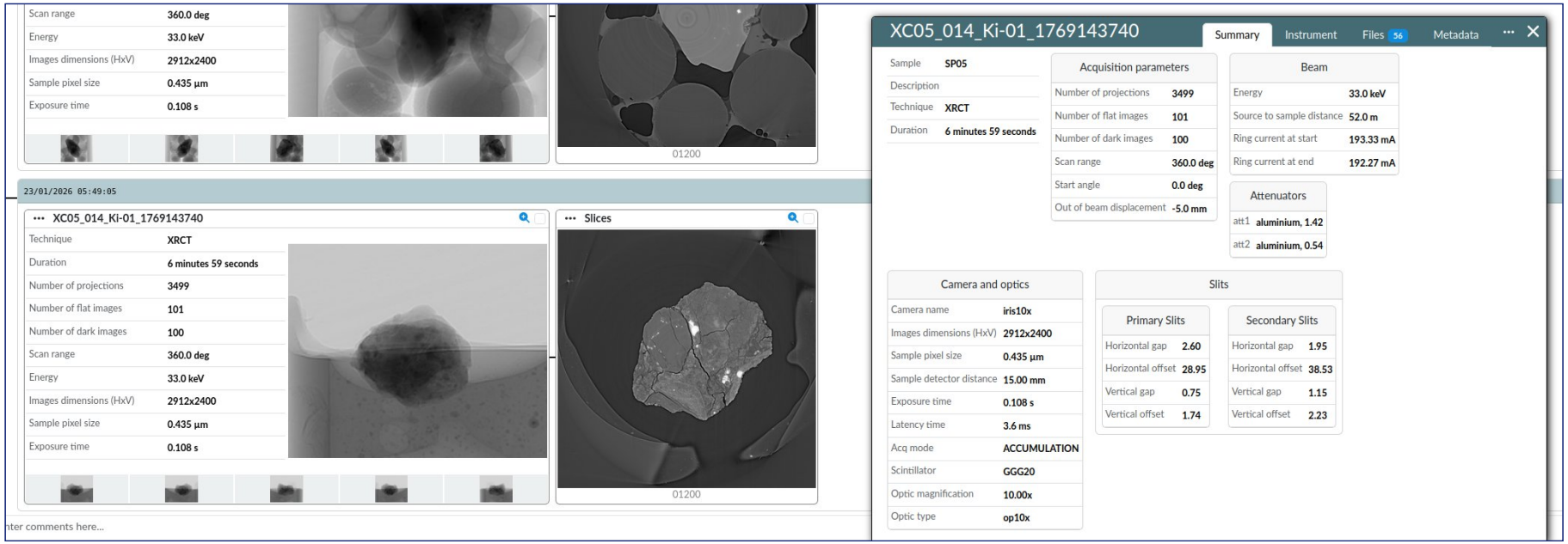


MX beamlines

DATA VISUALIZATION: TAILORED DATA DISPLAY

Technique-specific

- **Immediate data quality assessment**
- **Early detection of issues (e.g. scintillator problems)**



X-ray Computed Tomography (XRCT)

X-ray Powder Diffraction (XRPD)

Single Crystal X-ray Diffraction (SCXRD)

Next steps

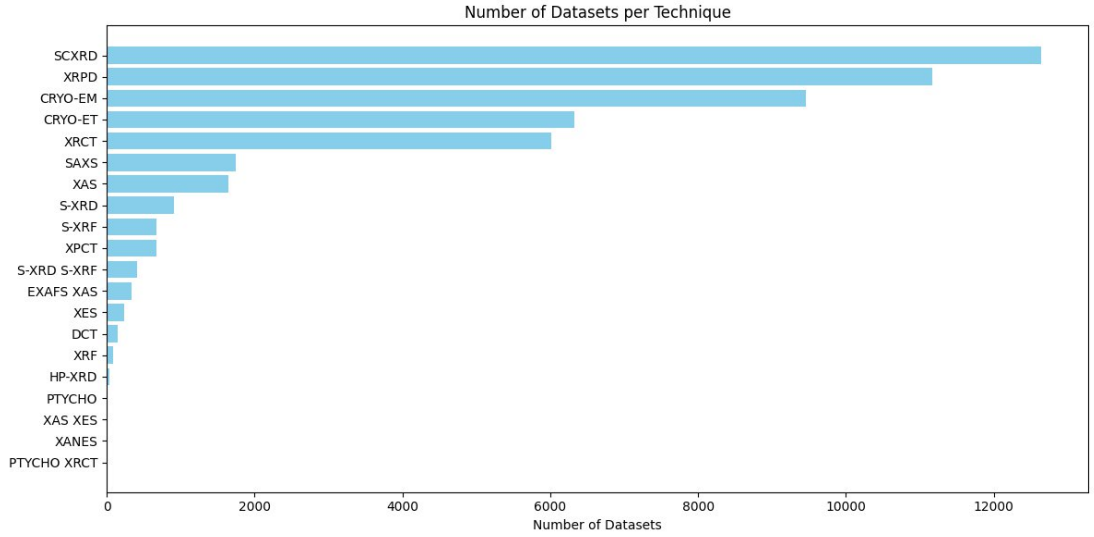
New techniques to be included

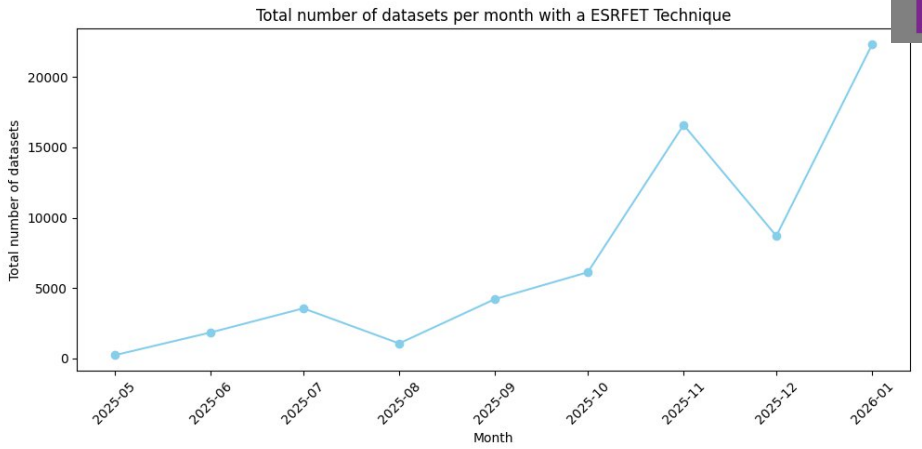
- **Cryo-Electron Tomography (Cryo-ET)**
- **Cryo-Electron Microscopy (Cryo-EM)**
- **Serial Macromolecular Crystallography (SSX)**

TECHNIQUES: PANET AND ESRFET ONTOLOGIES

ESRFET **is an ontology of experimental techniques used at the ESRF connected to the PaNET ontology.**

- **Python API's: generate technique metadata for ESRF data producers (BLISS, etc.) to save in NeXus-compliant HDF5 and ICAT.**
- **> 120 techniques defined**





ESRFET Techniques in the Data Portal

Next steps

- **Tag more datasets with techniques**
- **purl replaced by w3id**
- **Add techniques metadata in DOI**
- **Search and filter by technique**
- **Monitoring tool**

ICAT

Dataset ↔ DatasetTechnique ↔ Technique

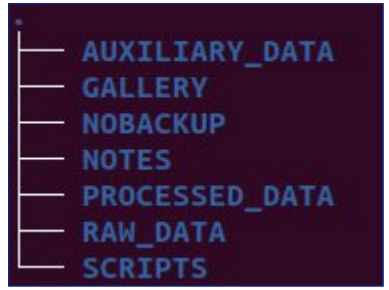
Koumoutsos, G., De Nolf, W. The ESRF Experimental Techniques Ontology.

DATA: CAPTURE THE FULL EXPERIMENT CONTEXT

Integrate

- **Auxiliary data** (ESRF data policy: *refers to data that provide contextual information regarding the experiment, such as information regarding the sample (images, provenance and preparation) or the software used for data processing*),
- **scripts**,
- **experiment gallery** (setup photos).

How should these data be catalogued in ICAT, especially when they are not linked to a specific sample?

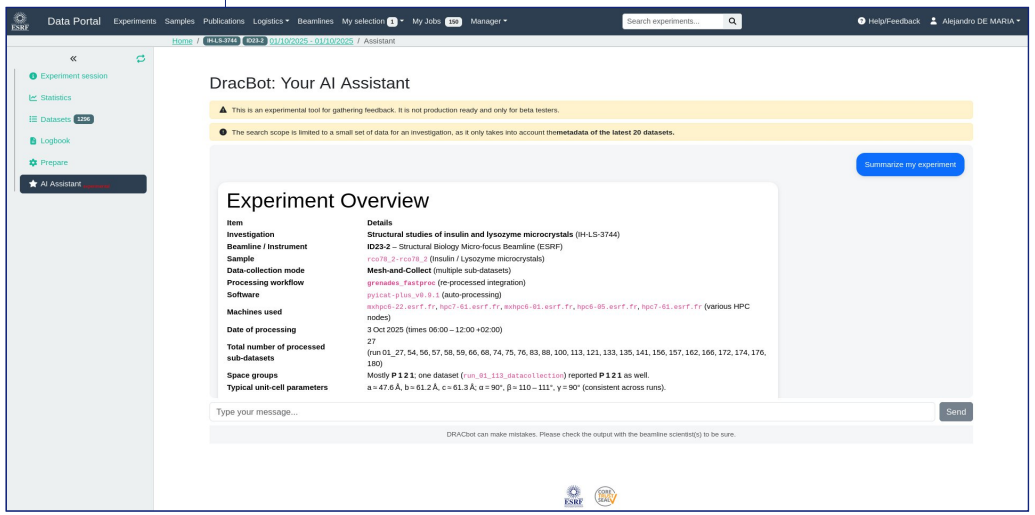


ESRF Data folder structure

DRACBOT: AI ASSISTANT

POC Accessible to ESRF beamline staff only

Limited context sent to the LLM (gpt-oss model): metadata of the last 20 datasets



AI Assistant in the Data Portal
Credits: A. De Maria

Data-quality Summary (representative values)

Metric	Typical range across runs
Overall completeness	**7 % – 15 %** (most runs 12-15 %)
Overall ⟨I/σ(I)⟩	**0.8 – 2.5** (inner shells often much higher, up to ~40 for the best runs)
Overall CC½	**0.96 – 0.998** (inner shells often > 0.99)
Overall R_merge	**0 – 30 %** (many runs report 0, a few high values up to ~95 % for low-resolution shells)
Overall R_meas	**0 – 65 %** (some anomalously large negative values indicate statistical noise in very low-count data)
Overall R_pim	**0 – 27 %** (most ≤ 12 %)
Overall multiplicity	**1.0 – 1.3** (generally low, as expected for single-crystal mesh scans)
Resolution limits	**High-resolution shell** typically **1.63 – 1.66 Å**; **low-resolution cut-off** varies from **~18 Å** to **> 40 Å** depending on crystal quality.
Anomalous signal	Generally **absent** (overall anomalous multiplicity ≈ 0, anomalous completeness ≈ 0 – 0.3 %).

Note: Many parameters are reported as **0** for outer shells (e.g., R_merge, R_meas) because the corresponding reflections are either absent or have intensities below the detection threshold.

Key Observations

1. **Consistent crystal lattice** – All processed sub-datasets share nearly identical unit-cell dimensions, confirming that the same crystal (or identical crystals) were used throughout the mesh-and-collect experiment.
2. **High inner-shell data quality** – The inner resolution shells routinely show very high ⟨I/σ(I)⟩ (up to ~40) and CC½ > 0.99, indicating that the diffraction spots are strong and well measured.
3. **Low overall completeness** – Because each mesh-and-collect sub-dataset only samples a small portion of reciprocal space, overall completeness per sub-dataset is modest (7-15 %). The full experiment (combining all sub-datasets) would be expected to reach a much higher completeness.
4. **Resolution** – The majority of runs reach a high-resolution limit of **~1.64 Å**, sufficient for detailed atomic modelling of insulin and lysozyme.
5. **Anomalous signal** – No significant anomalous signal was detected, which is typical for native protein crystals without heavy-atom derivatization.
6. **Processing consistency** – All datasets were processed with the same pipeline (`grenades_fastproc` → `pyicat-plus_v0.9.1`) on ESRF HPC resources, ensuring comparable statistical treatment.

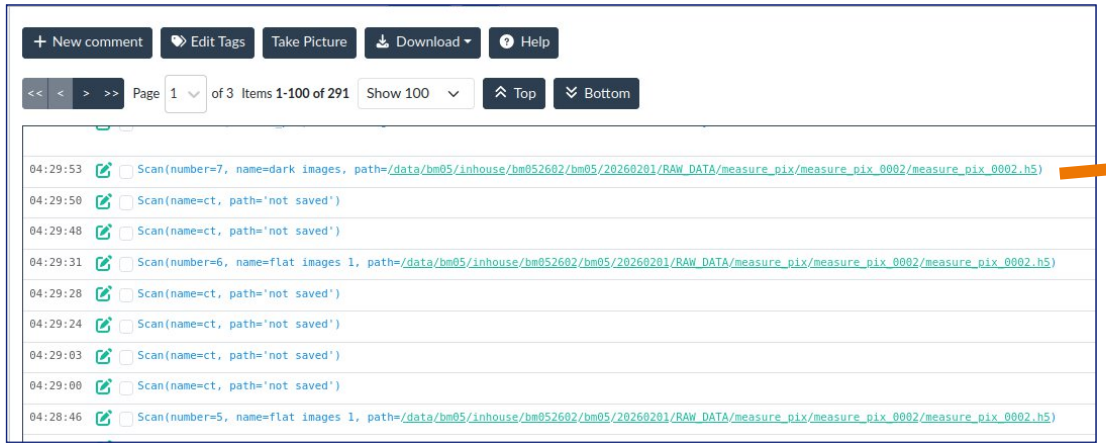
Conclusion

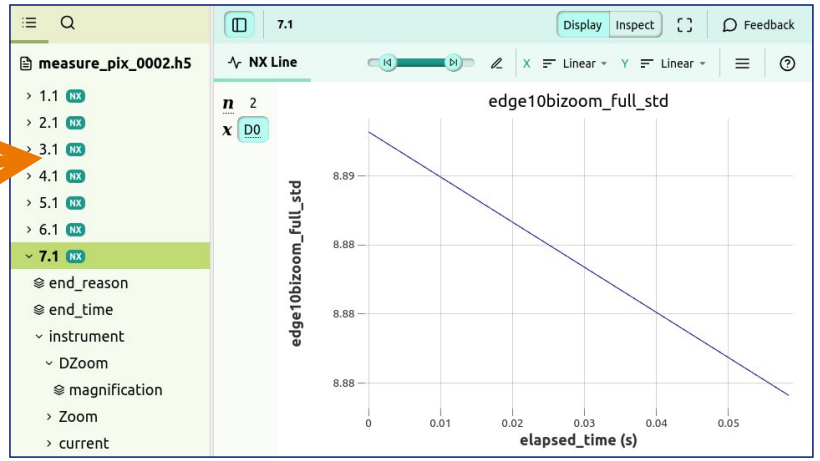
The mesh-and-collect experiment on the ID23-2 beamline generated a large set of high-resolution diffraction sub-datasets from insulin/lysozyme microcrystals. While each individual sub-dataset has limited completeness, the data quality within the measured resolution shells is excellent (high CC½, strong I/σ(I), and low R_merge/R_meas). When merged, these data should provide a robust structural model of the target proteins at ~1.6 Å resolution.

All information extracted from the dataset metadata supplied (27 re-processed integration datasets, IDs 2260453301 – 2260552388).

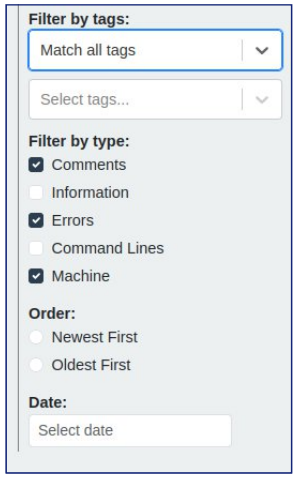
LOGBOOK

Log scan entries with links to HDF5: easy tracking of data acquisition for users





Advanced tag filtering for efficient logbook navigation: faster access to relevant logbook entries

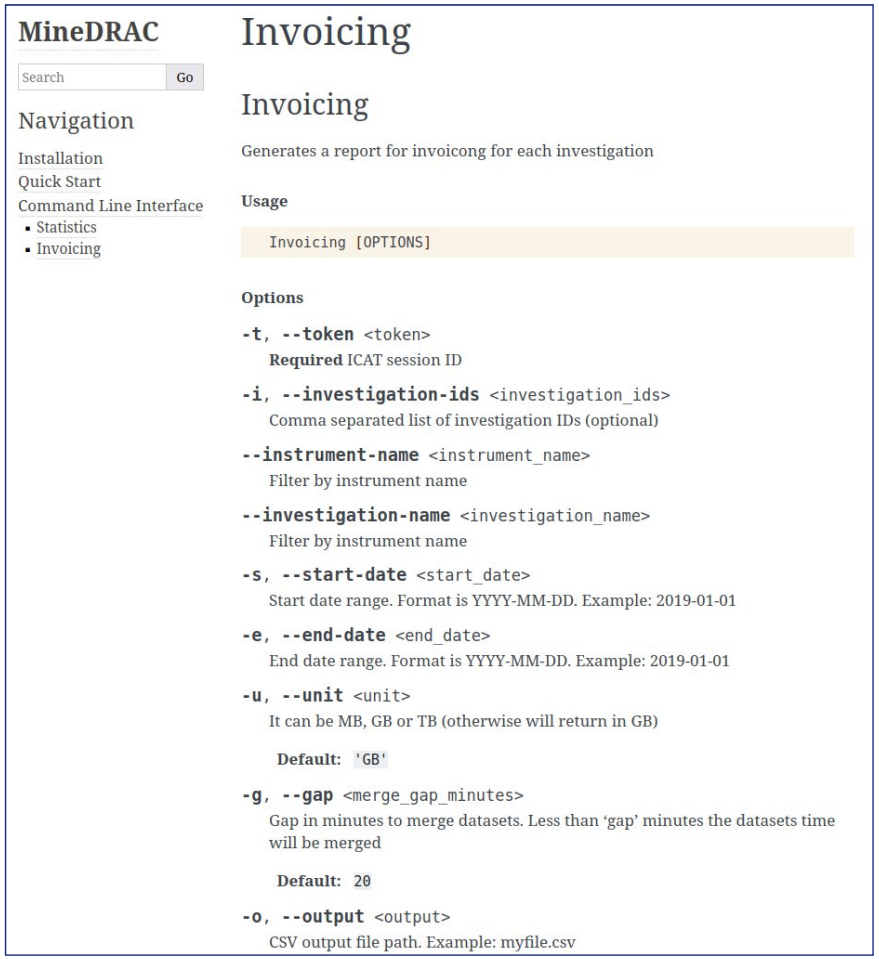


Next steps

- Notebook: “Google Doc”-like document
- E-logbook: view contextualized by dataset or sample

INVOICING, STATISTICS

MineDRAC tool to extract statistics and figures for invoicing (MX industrial experiments)



MineDRAC

Search Go

Navigation

Installation
Quick Start
Command Line Interface
- Statistics
- Invoicing

Invoicing

Invoicing

Generates a report for invoicong for each investigation

Usage

```
Invoicing [OPTIONS]
```

Options

-t, --token <token>
Required ICAT session ID

-i, --investigation-ids <investigation_ids>
Comma separated list of investigation IDs (optional)

--instrument-name <instrument_name>
Filter by instrument name

--investigation-name <investigation_name>
Filter by instrument name

-s, --start-date <start_date>
Start date range. Format is YYYY-MM-DD. Example: 2019-01-01

-e, --end-date <end_date>
End date range. Format is YYYY-MM-DD. Example: 2019-01-01

-u, --unit <unit>
It can be MB, GB or TB (otherwise will return in GB)

Default: `'GB'`

-g, --gap <merge_gap_minutes>
Gap in minutes to merge datasets. Less than 'gap' minutes the datasets time will be merged

Default: `20`

-o, --output <output>
CSV output file path. Example: myfile.csv

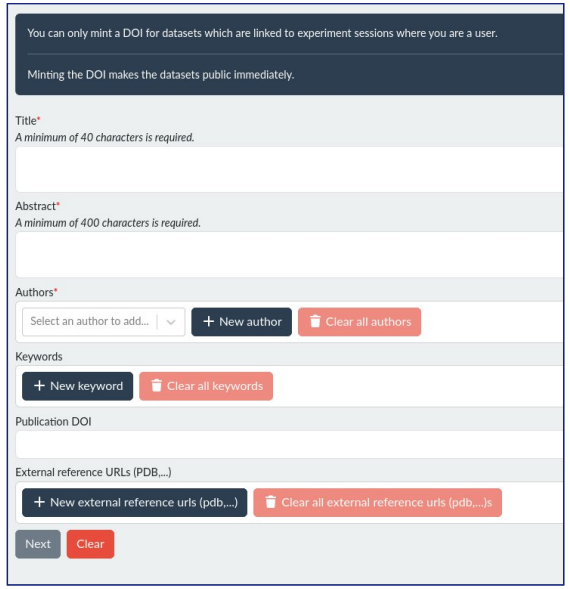
proposal	investigation_id	beamline	start_date	end_date	booked_time	samples_collected	samples_declared	parcels	volume	raw_volume	processed_volume	effective_time	total_time
MX-##	22920936	ID30A-3	2025-12-06 09:30:00	2025-12-06 17:00:00	7:30:00	87	90	1	160 GB	140 GB	20 GB	3:55:58.806000	19:36:25.457000
FX-##	22995581	ID30A-1	2025-12-06 09:30:00	2025-12-06 17:00:00	7:30:00	37	37	1	192 GB	189 GB	3 GB	2:41:49.020000	7:40:34.842000
MX-##	23004685	ID30B	2025-12-06 09:30:00	2025-12-06 17:00:00	7:30:00	193	16	1	424 GB	420 GB	4 GB	5:54:02.654000	0:00:39.868000
MX-##	23039195	ID30A-1	2025-12-06 09:30:00	2025-12-06 17:00:00	7:30:00	43	0	0	242 GB	224 GB	17 GB	4:30:37.512000	12:20:27.169000
MX-##	22233712	ID23-2	2025-12-05 09:30:00	2025-12-06 08:00:00	22:30:00	40	46	1	685 GB	681 GB	3 GB	3:08:28.800000	18:26:56.858000
QA-##	22296784	ID23-1	2025-12-05 09:30:00	2025-12-05 17:00:00	7:30:00	58	0	0	693 GB	644 GB	49 GB	1:38:22.530000	10:12:49.566000
FX-##	22692252	ID30B	2025-12-05 09:30:00	2025-12-05 17:00:00	7:30:00	111	0	0	202 GB	198 GB	4 GB	2:42:17.204000	11:51:31.742000
MX-##	22954669	ID30A-3	2025-12-05 09:30:00	2025-12-05 17:00:00	7:30:00	64	0	0	272 GB	243 GB	28 GB	3:24:33.339000	21:28:42.532000
MX-##	22322136	ID30A-3	2025-12-04 17:00:00	2025-12-05 08:00:00	15:00:00	43	45	1	200 GB	173 GB	27 GB	1:52:39.490000	10:58:38.632000

Invoicing report

– **Saves time for beamline staff and administrative teams**

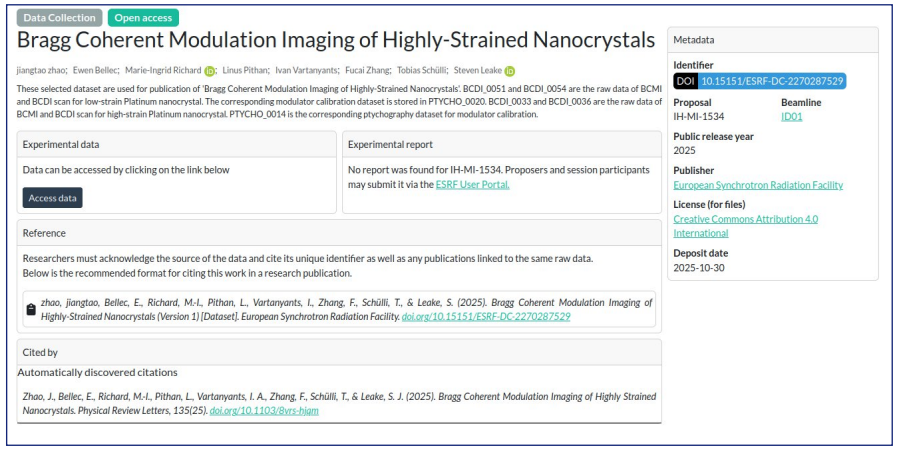
DOIS

- Enhance DOI minting/reservation: keywords, publication DOI, external references added
- Autodiscovering citations using **OpenAlex**

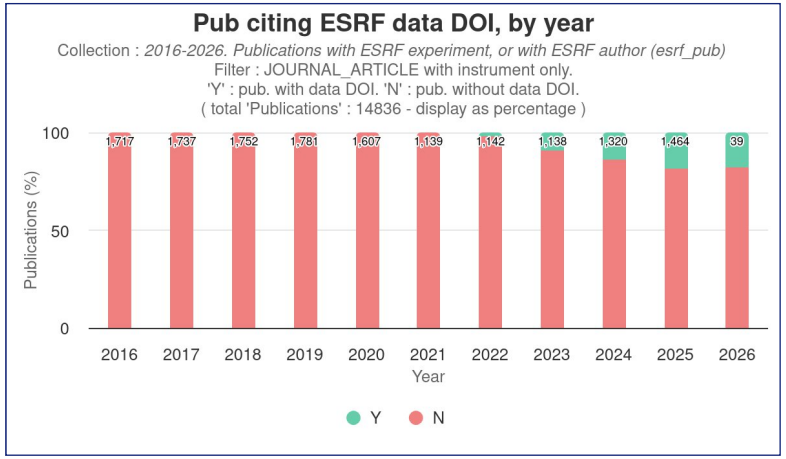


Mint bespoke DOI in the Data Portal

DOI landing page



Autodiscover citation

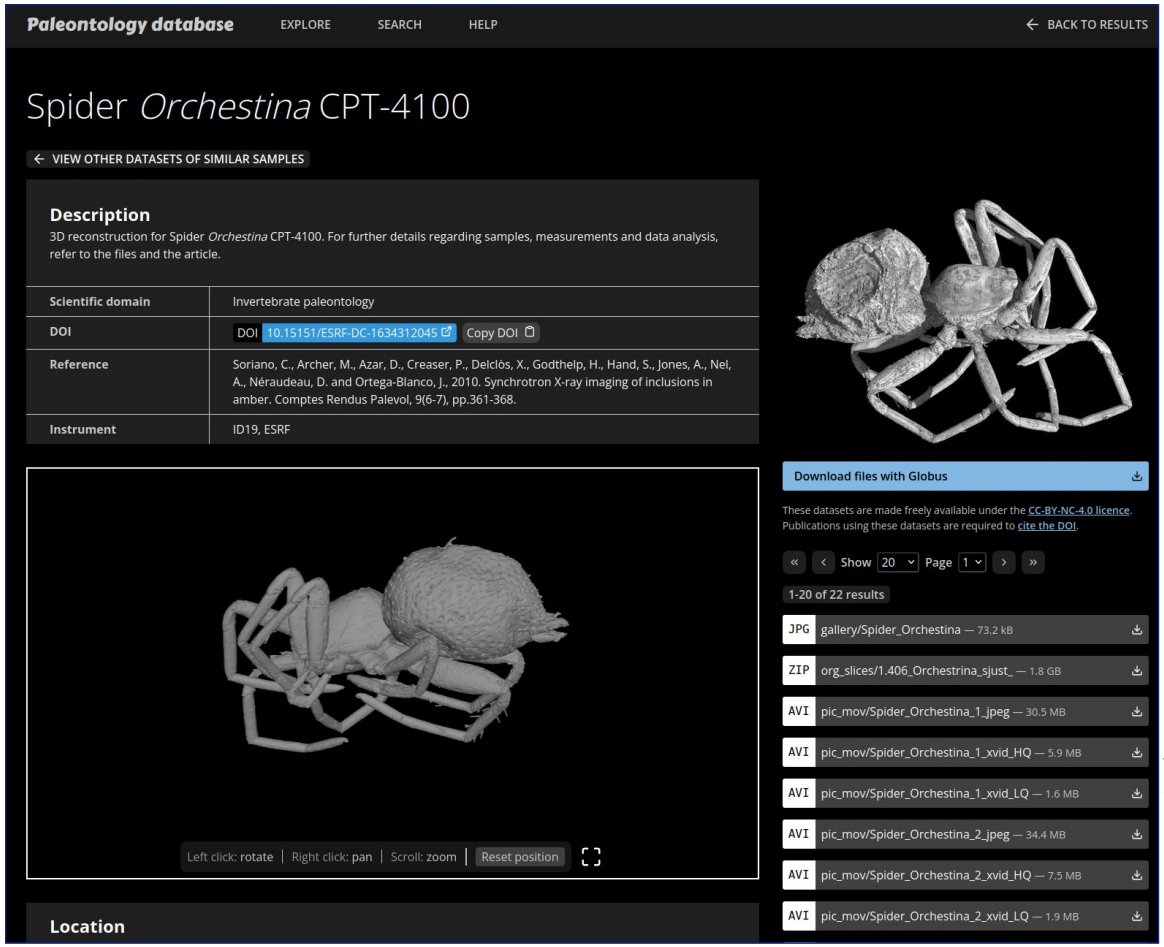


PUMA

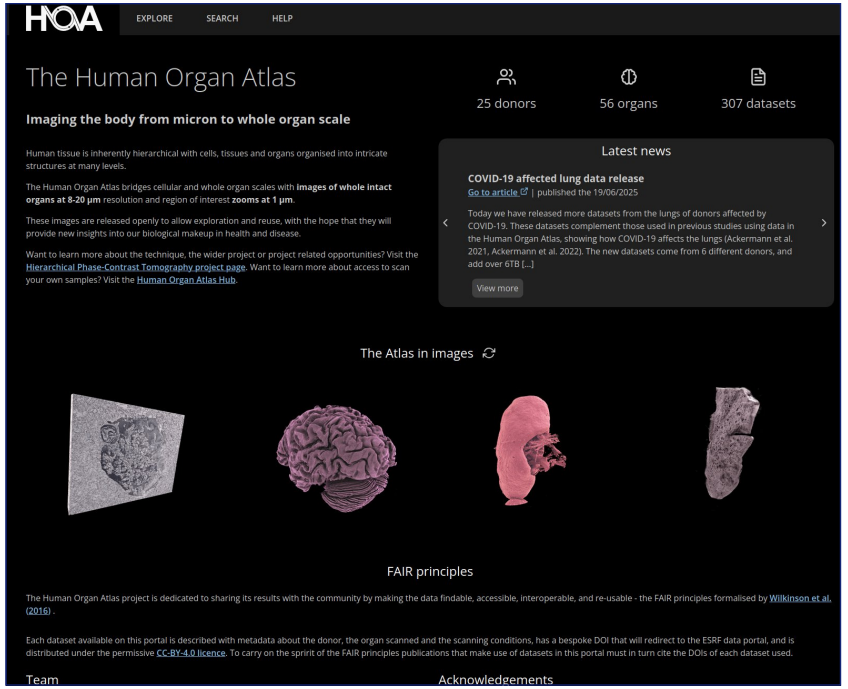
DOMAIN PORTALS

New data in both paleo and human organ atlas projects

Tool to extract and monitor statistics from Matomo

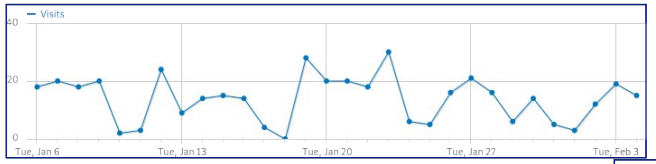


3D viewer in the Paleo portal

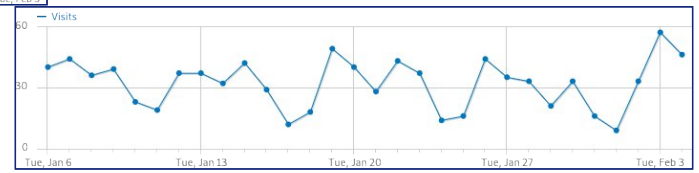


Human Organ Atlas Portal

Number of visits per day



paleo



hoa

Next steps

- Cultural Heritage portals (violins, painting, mummies)
- Connection to other databases

ADDITIONAL TOPICS

- PaN-Finder (Nadine talk)
- ICAT upgrade to be planned
- High interest in ICAT-Lucene
- Download and data restoration
- Data retention: long-term retention, criteria to be defined
- Performance issues:
  - users may experience up to 4× slower resource access than admins
  - Large-scale dataset ingestion/update may impact DB performance
- Instrument PIDs
- Sample PIDs
- D3A: Direct DOI Data Access (Tycho, Nadine and Andy)

ACKNOWLEDGMENTS

DAU team, especially

Alejandro De Maria

Axel Bocciarelli

Guillaume Gaisné

Simon Delcamp

Nadine Fischer

Andy Götz (unit leader)

Ludovic Broche

Olof Svensson

Pierre-Olivier Autran

Wout De Nolf

+ BCU, ITS group + beamline scientists

ICAT Collaboration team

QUESTIONS AND DISCUSSION

Thank you for your attention!

Discussion, questions?