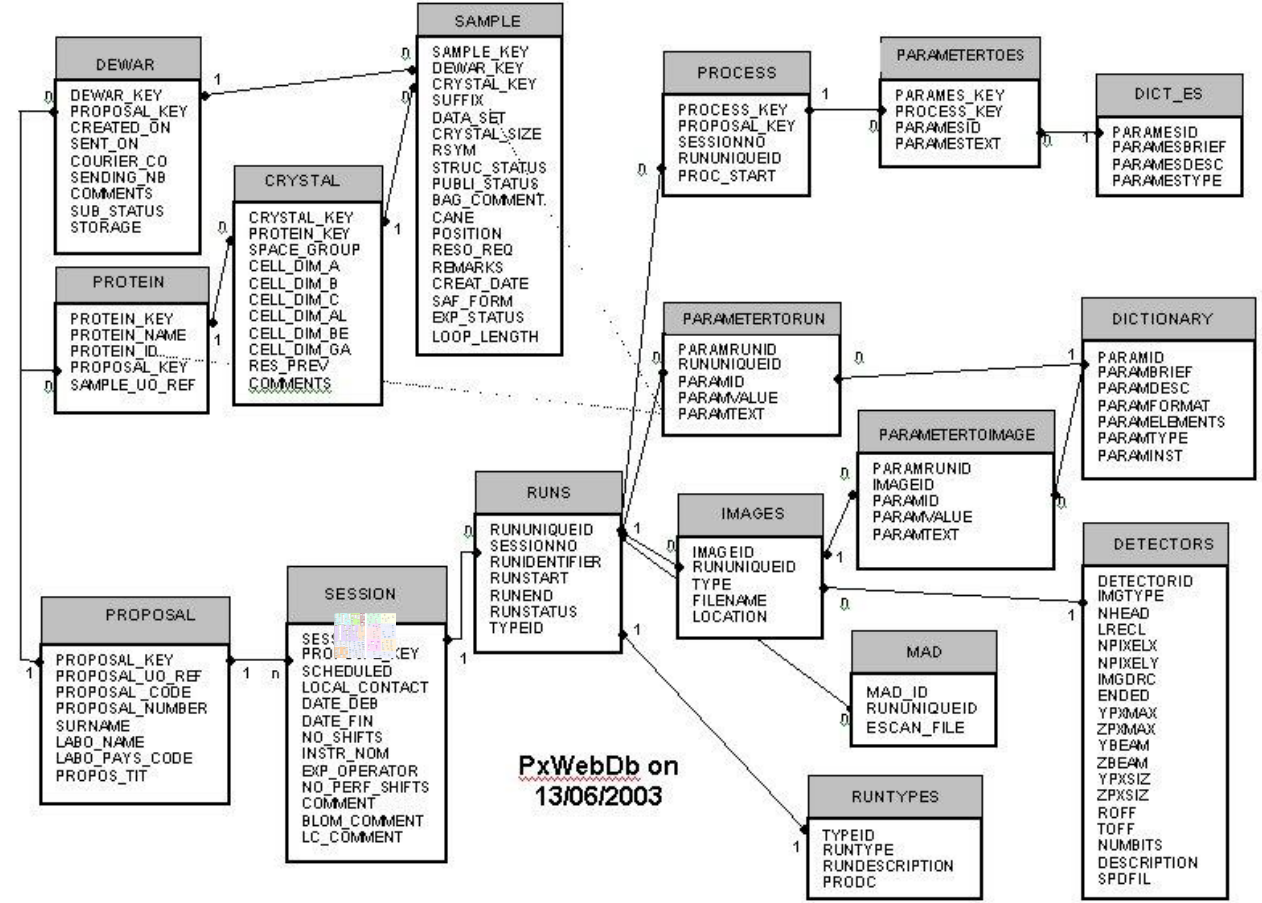
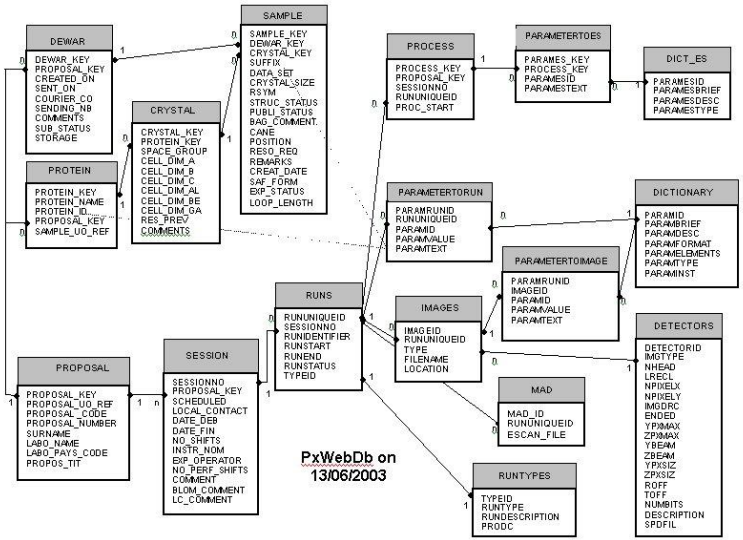
A proposal for a Next Generation of ISPyB Software

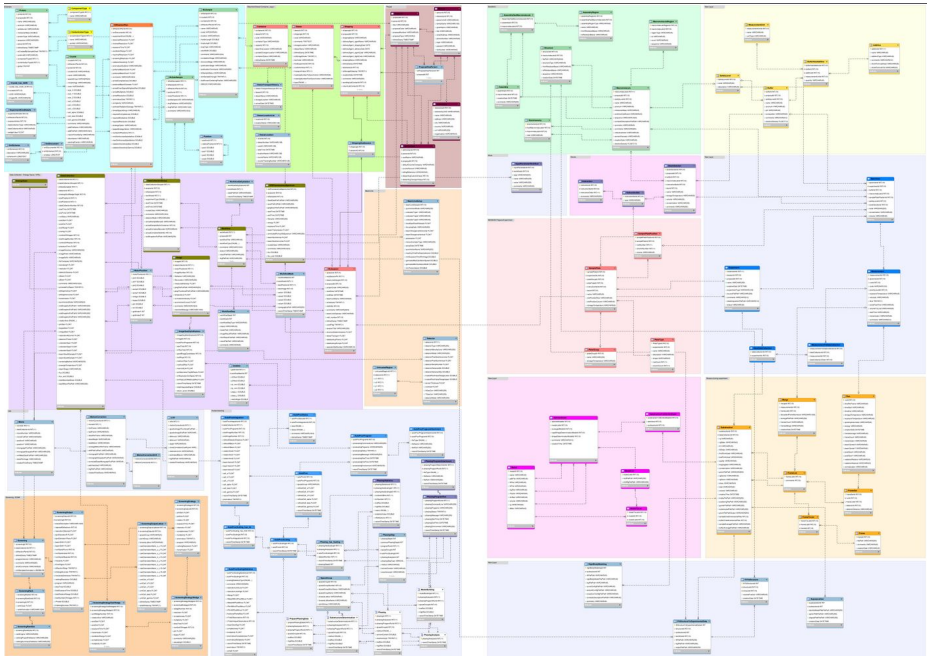
Alex de Maria Antolinos
On behalf of:
Software and Structural Biology Group

20th Anniversary



Database Evolution





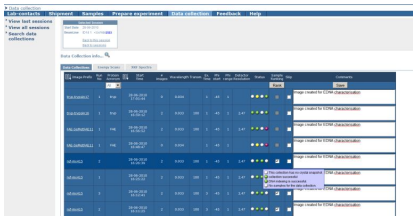
2003

17 tables

2023

209 tables
42 views

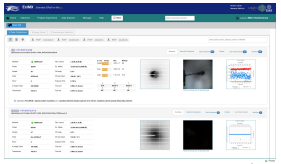
ISPyB Software Evolution

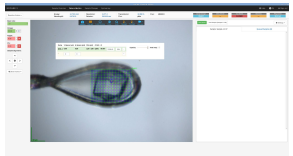


Java Application
With a built-in UI written in JSP

SQL Database

ISPyB (? - 2012)





UI | Web Services

Java Application
With a built-in UI written in JSP

SQL Database

ISPyB (2012 - now)

Web Services

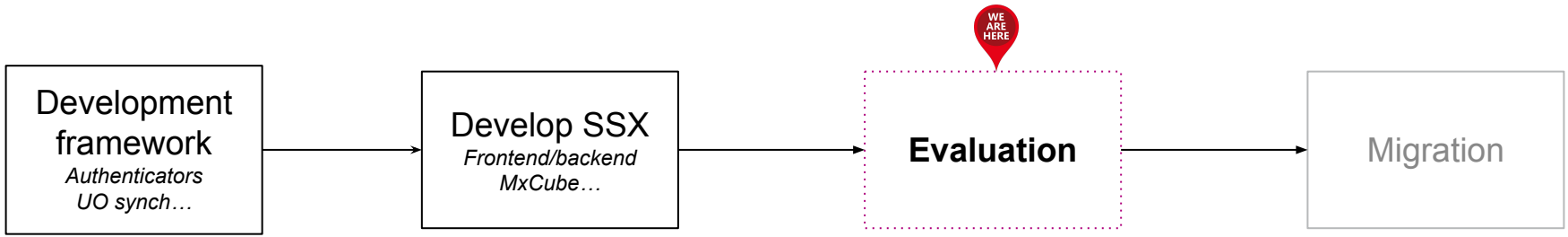
Python

SQL Database

PY-ISPyB

Roadmap achieved for last 4 years

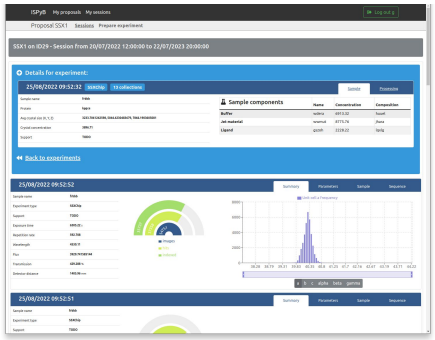
- Following what was agreed in the ISPyB Strategy Meeting @Hamburg 2020



GitHub

FastAPI

python



Main contributors: EMBL DESY. ESRF The European Synchrotron

GOALS

Wishlist for the next ISPyB-like software

Looking for:

- **Long term** solution (~20 years)
- Improve **sustainability**
- **Flexibility**
  - vs rigid data model (e.g: dc)
- Scalability
  - Implement **new techniques** and **extensions** to existing on **short timescales**
    - Standardization
    - Automation
  - Features
    - Reprocessing
    - Generic and Custom UIs
- High performance
- Reliability
- Make easier to contribute and collaborate

Wishlist for the next ISPyB-like software

Looking for:

- **Long term** solution (~20 years)
- Improve **sustainability**
- **Flexibility**
  - vs rigid data model (e.g: dc)
- Scalability
  - Implement **new techniques** and **extensions** to existing on **short timescales**
    - Standardization
    - Automation
  - Features
    - Reprocessing
    - Generic and Custom UIs
- High performance
- Reliability
- Make easier to contribute and collaborate

and:

- **Modular** by design
- **FAIR** principles
  - Ontologies
  - Global search
    - Public data and private data
- Compliant with the **Data policy**
  - Embargoed and public data
  - Logbook
  - **Data publication**
    - DOI
- **Raw and processed** data handling
  - Restore data from tape
  - Advanced file transfer
  - Sharing capabilities

Wishlist for the next ISPyB-like software

Looking for:

- **Long term** solution (~20 years)
- Improve **sustainability**
- **Flexibility**
  - vs rigid data model (e.g: dc)
- Scalability
  - Implement **new techniques** and **extensions** to existing on **short timescales**
    - Standardization
    - Automation
  - Features
    - Reprocessing
    - Generic and Custom UIs
- High performance
- Reliability
- Make easier to contribute and collaborate

and:

- **Modular** by design
- **FAIR** principles
  - Ontologies
  - Global search
    - data and private data
- Co… the **Data policy**
  - Embargoed and public data
  - Logbook
  - **Data publication**
    - DOI
- **Raw and processed** data handling
  - Restore data from tape
  - Advanced file transfer
  - Sharing capabilities

NEEDED FOR ALL TECHNIQUES not only Structural Biology

How to achieve that?

3 options have been considered by ESRF:

1. **ISPyB**
   a. Stay with ISPyB's java backend
   b. Nothing to implement (but goals not achieved)

2. **py-ISPyB**
   a. Keep database schema but change backend
   b. Needs:
      i. Implementation of backend (partially done)
      ii. Implementation of frontend (partially done)
      iii. Change data exchange between beamline and ISPyB

3. **ICAT-based solution**
   a. Replace database schema and backend by ICAT
   b. Needs:
      i. Extend data exchange between beamline and ISPyB
      ii. Implementation of frontend (partially done)
      iii. Sample shipping/tracking (partially done)

How to achieve that?

3 options have been considered by ESRF:

1. ~~ISPyB~~
   a. ~~Stay ISPyB's java backend~~
   b. ~~Nothing to reimplement~~

2. **py-ISPyB**
   a. Needs:
      i. Implementation of backend (partially done)
      ii. Implementation of frontend (partially done)
      iii. Change data exchange between beamline and ISPyB

3. **ICAT-based solution**
   a. Replaces backend and database by ICAT
   b. Needs:
      i. Extend the ingestion of data from beamlines
      ii. Implementation of frontend (partially done)
      iii. Logistics

Evolution

Revolution

py-ISPyB

Repository:
https://github.com/ispyb/py-ispyb
Documentations:
https://ispyb.github.io/py-ispyb/

Similar to ISPyB in features and architecture but:

- Python
- Modern and cleaner
- No legacy
- Easier to configure and deploy

Evaluation

- No doubt this solution **can indeed work** as replacement of ISPyB
- The main medium/long term **problem is the data model**.
  - Built around MX technique, it does **not scale** well
- Adding new techniques or extensions implies changes in the data model.
  - Refactoring the data models or making truly generic might take longer than starting from scratch or adopting an existing generic solution
- **py-ispyb is not a long term solution for both MX and non-MX use**

What is ICAT?

What is ICAT?

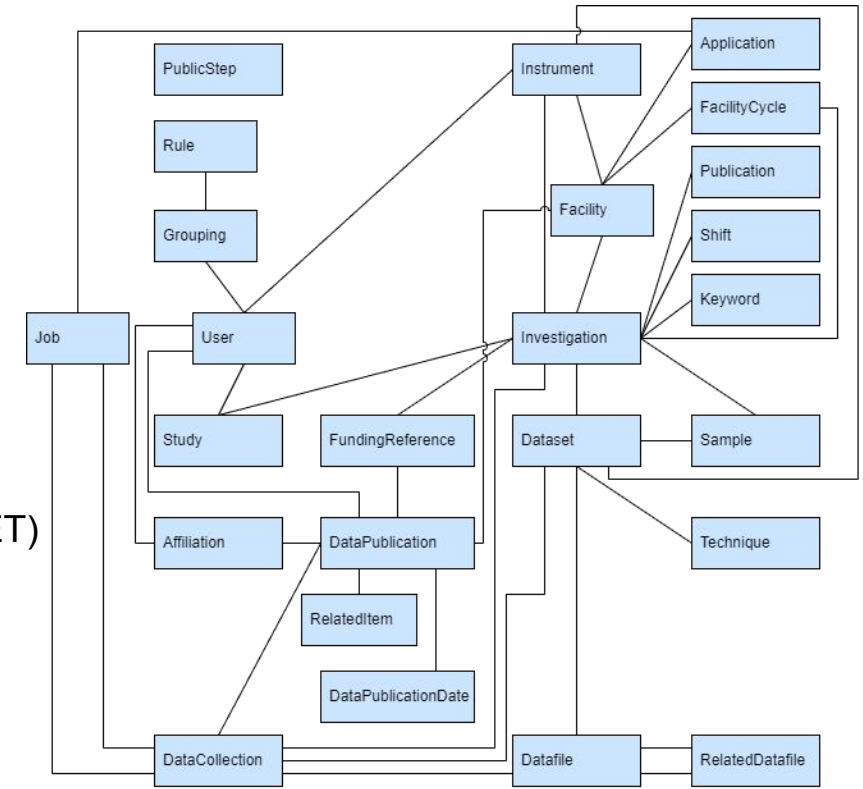
- ICAT is a generic **metadata catalogue** developed at STFC
- It supports experimental data management for large-scale facilities
- It is a set of components that includes:
    - ICAT Server:
        - Supports ORACLE and MariaDB
        - Rest/SOAP API
    - Authenticators: openId, SSO, DB, custom…
    - Fine-grained authorisation model based on roles
    - OAI-PMH
    - Search component based on Apache Lucene
    - Python-icat: python client components
    - PANOSC/ExPands API
    - Etc…

UKRI Science and Technology Facilities Council

- Used @ESRF since 2015 as **metadata repository** for the implementation of the **data policy**
    - All beamlines connected including MX via MxCube
    - Primarily for **raw data** but recently POC have been carried out for **processed data**

ICAT Data Model

- ICAT has around 40 tables
- It covers:
  - BLSessions
  - Users
  - Data collections
  - Data processing
- It does not cover:
  - Dewars
  - Containers
- Additionally:
  - Data publication
  - DOI
  - Techniques (PANET)



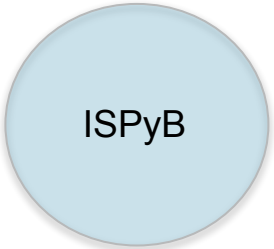
Main Entities

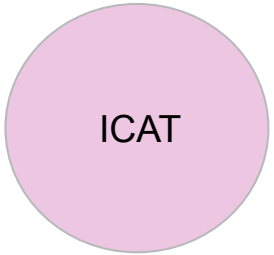
Science and Technology Facilities Council

ICAT and ISPyB at the ESRF

ICAT and ISPyB functionality over the time

~2014 started the implementation of data policy





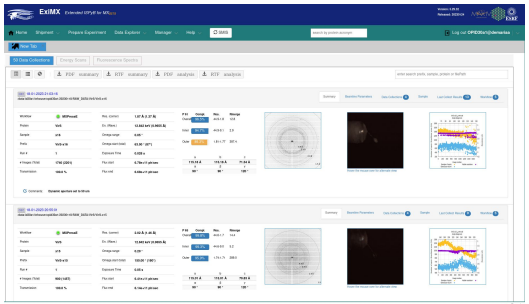
For MX/SAXS:

- LIMS
- Sample tracking
  - Diffraction plan
  - Processing plan
- Display of processed results
- Real time feedback
- Rich visualizations

For all techniques:

- Metadata Catalogue
- Implementation of data policy
  - Public and close data
  - FAIR
  - Etc..
- Raw data only
- Real time access

ICAT and ISPyB functionality over the time

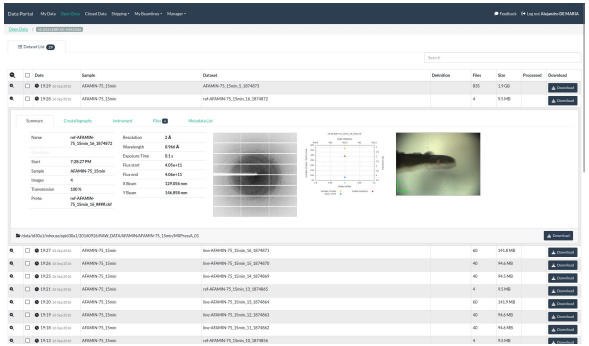


EXI/ISPyB

~2018

ISPyB

ICAT



Data Portal/ICAT

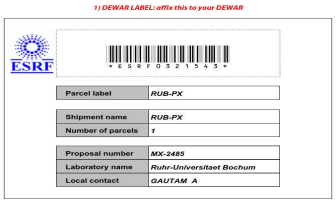
For MX/SAXS/EM:

- LIMS
- Sample tracking
  - Diffraction plan
  - Processing plan
- Display of processed results
- Real time feedback
- Rich visualizations

For all techniques:

- Metadata Catalogue
- Implementation of data policy
  - Public and close data
  - FAIR
  - Etc..
- Raw data only
- Logbook
- DOI minting
- Restoration from tape
- Richer visualization

ICAT and ISPyB functionality

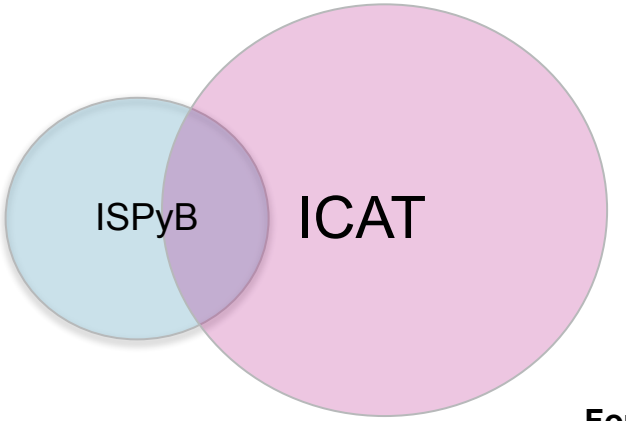

1) DEWAR LABEL: affix this to your DEWAR

ESRF

ESRF032154

Parcel label	RUB-PX
Shipment name	RUB-PX
Number of parcels	1
Proposal number	MX-2485
Laboratory name	Ruhr-Universitaet Bochum
Local contact	GAUTAM A

ISPyB's parcel label



~2020


OUTWARD-BOUND ADDRESS LABEL

Affix to the parcel's container

FROM:
Sonya Girodon
User Office
ESRF 71 Avenue des Martyrs
38000 GRENOBLE
FRANCE
Phone: 2880

Parcel name: Joanne's Parcel
Shipment content: 16 parcels
Proposal: ID002020
Beamline: ID00
Session start date: 14-10-2020
Local contact: Stéphanie MONACO

TO:
ESRF Stores
71 avenue des Martyrs
38000 GRENOBLE
FRANCE
Phone: +33 (0)4 76 88 2733
Fax: +33 (0)4 76 88 2347

ESRF

Storage conditions: At room temperature - None specified

ICAT's parcel label

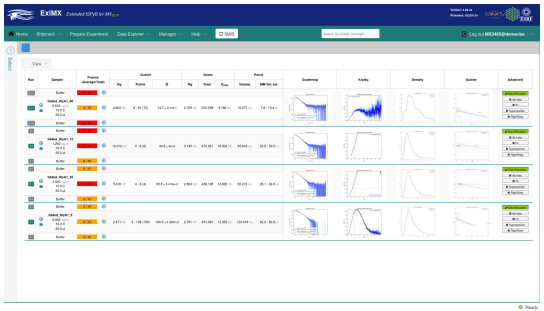
For MX/SAXS/EM:

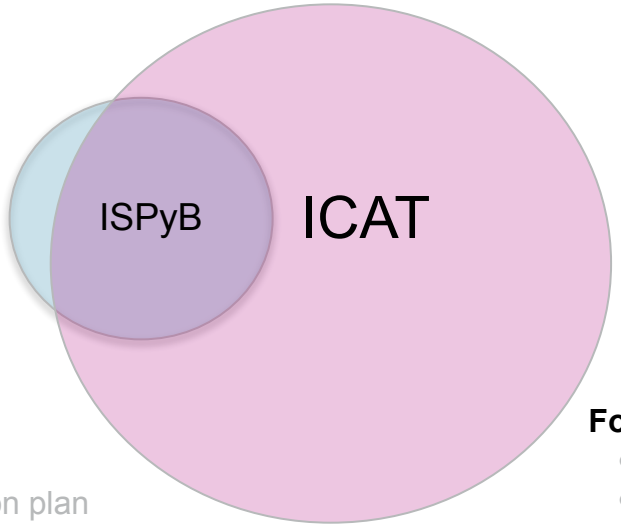
- LIMS
- Sample tracking
  - Diffraction plan
  - Processing plan
- Display of processed results
- Real time feedback
- Rich visualizations

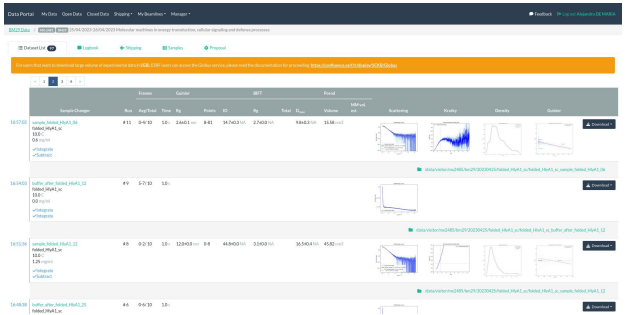
For all techniques:

- Metadata Catalogue
- Implementation of data policy
  - Public and close data
  - FAIR
  - Etc..
- Raw data only
- Logbook
- DOI minting
- Restoration from tape
- Richer visualization
- Improved sample tracking
  - 2nd generation

ICAT and ISPyB functionality







For MX/SAXS/EM:

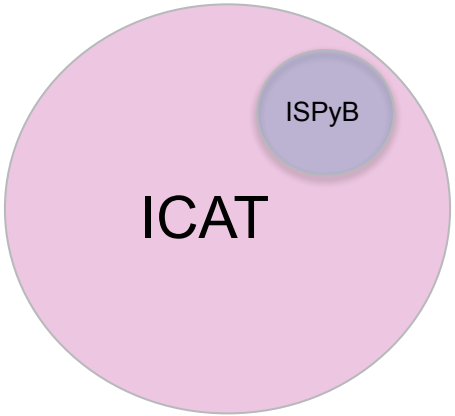
- LIMS
- Sample tracking
  - Diffraction plan
  - Processing plan
- Display of processed results
- Real time feedback
- Rich visualizations

~2022

For all techniques:

- Metadata Catalogue
- Implementation of data policy
  - Public and close data
  - FAIR
  - Etc..
- Raw data only
- Logbook
- DOI minting
- Restoration from tape
- Richer visualization
- Generic sample tracking
- Processed data

Hypothetical scenario with full overlapping

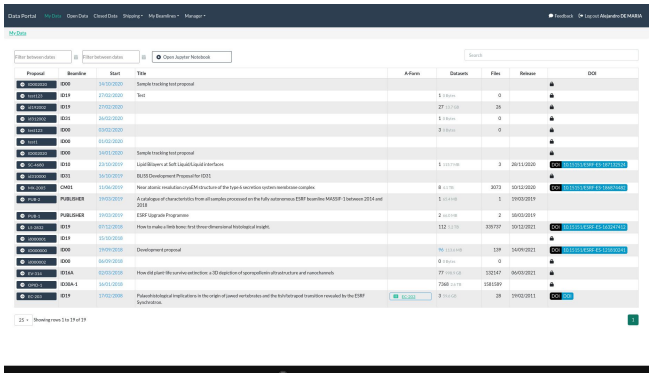


- Adopting a single solution based on a generic approach has advantages:
  - Facility level:
    - Offering to other BL the current **SB capabilities**
    - Reduce maintenance
  - Beamline and MX's level
    - **Rapid** development
    - More **autonomy**
      - No need to new tables or columns
      - Metadata nomenclature is defined by scientist
    - Generic backend
      - Focus in:
        - Do science
        - Define procedures and methods
    - Modular architecture
- Why ICAT and not ISPyB?

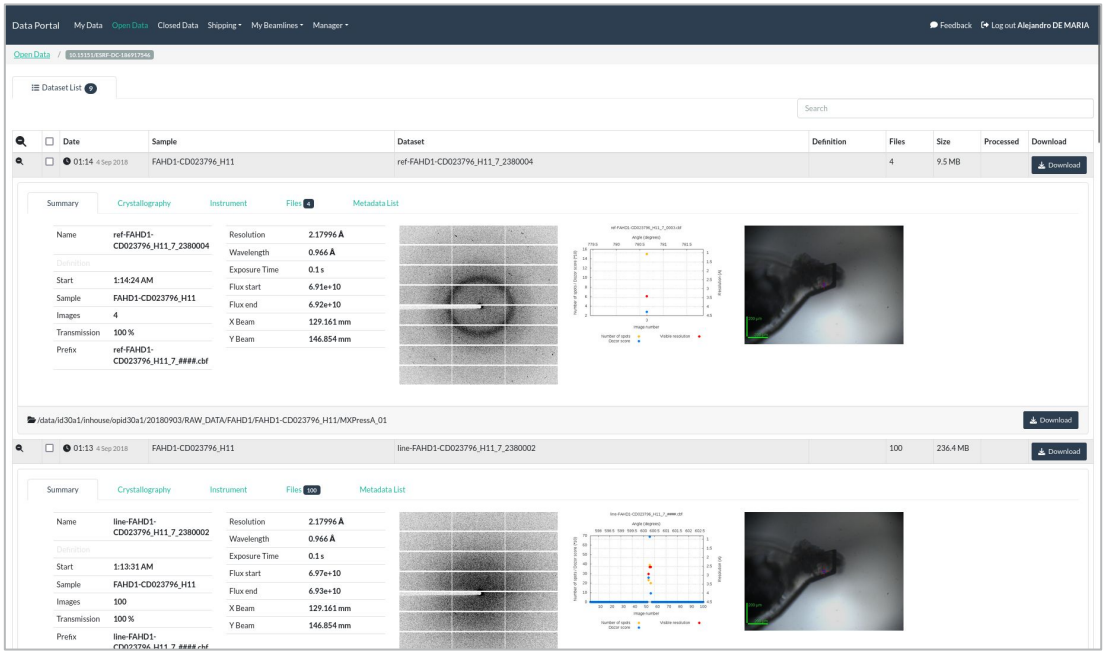
Features available today

Some features currently available

- Data portal: UI displays the information stored in ICAT. Public and embargoed data



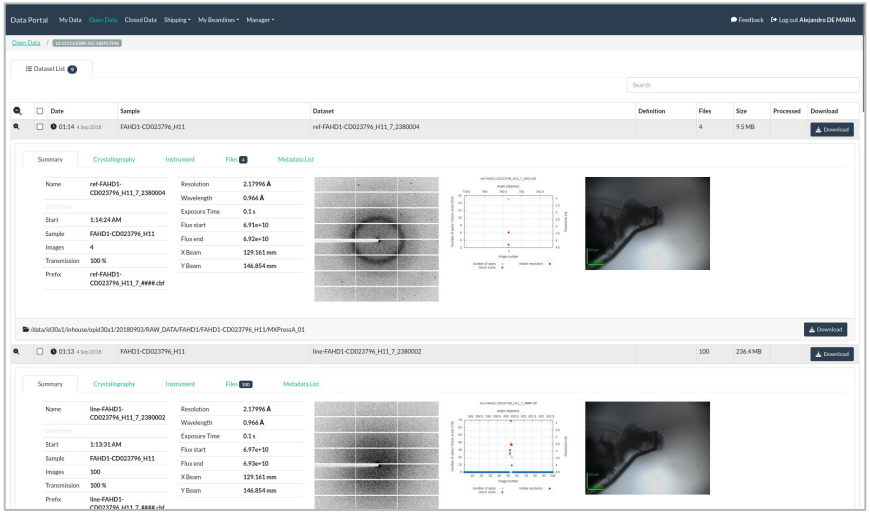
Browse proposal/visits

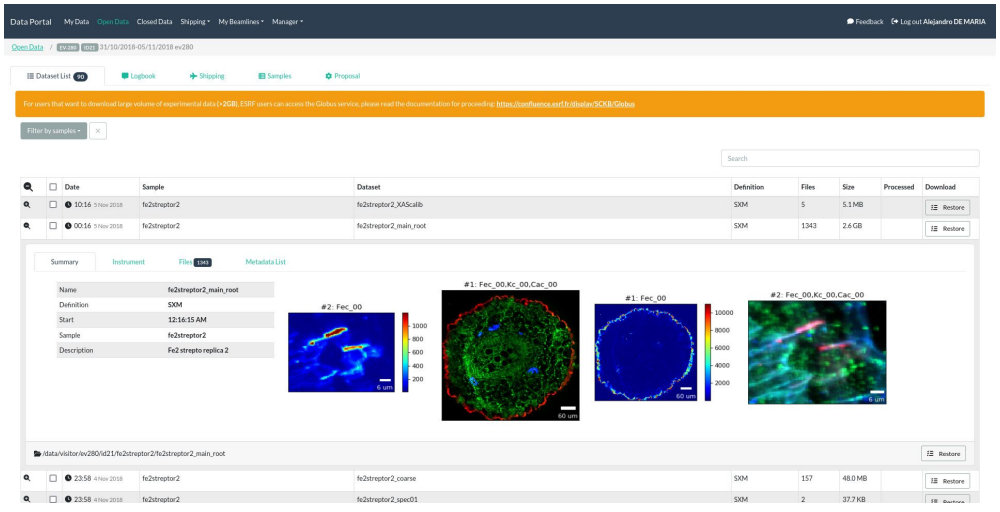


Generic Data Collection view

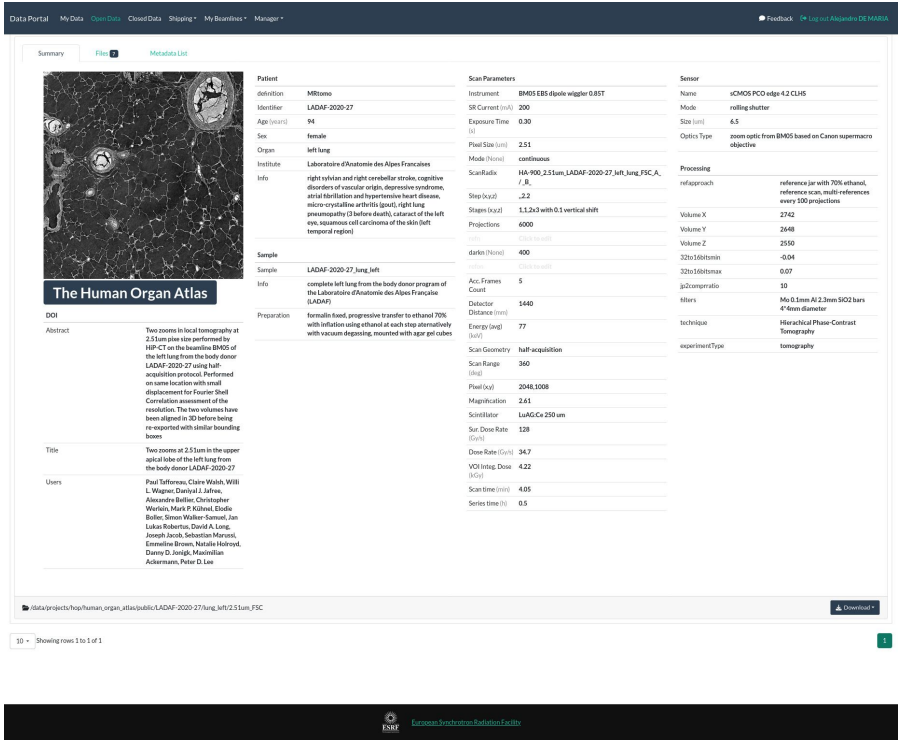
Some features currently available

- Data portal: UI displays the information stored in ICAT. Public and embargoed data

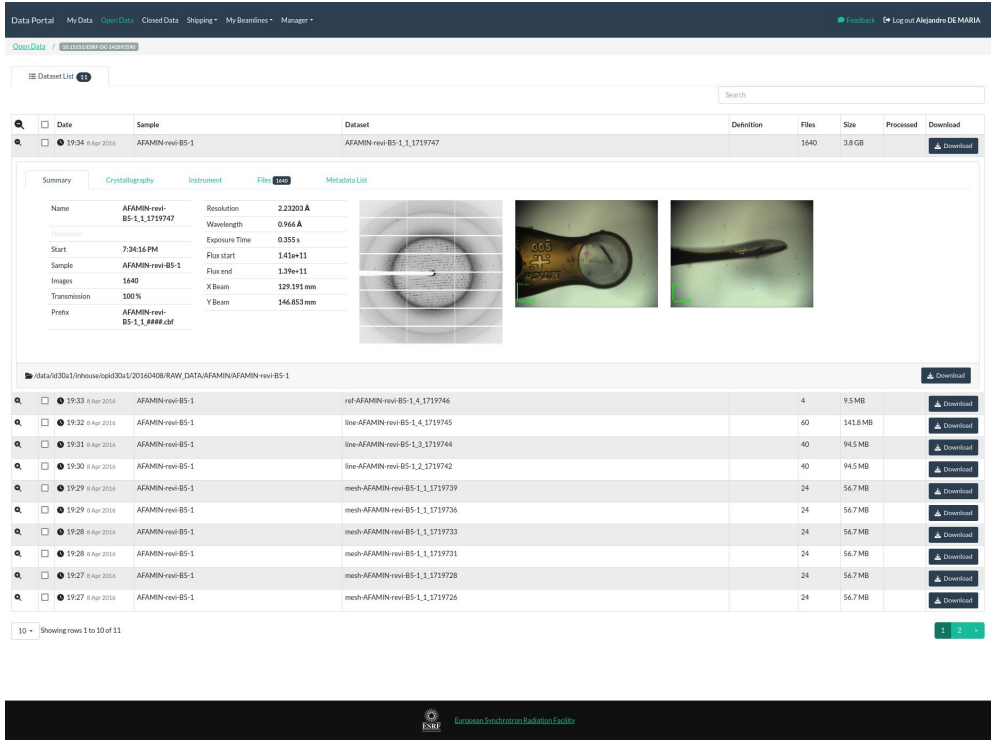




Custom views

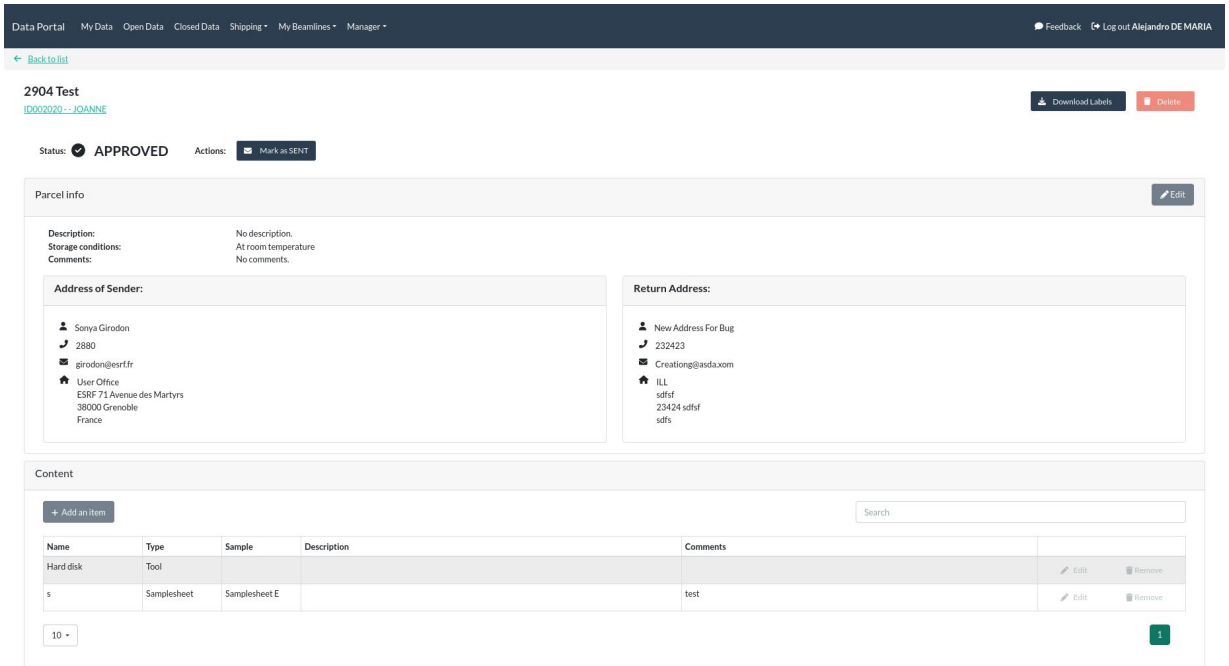


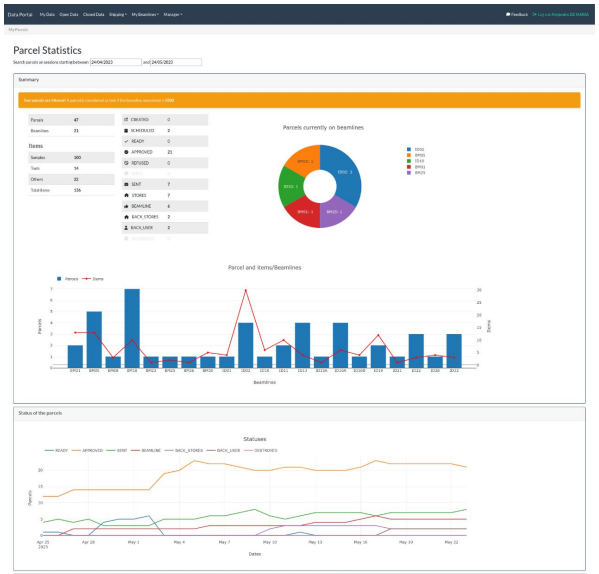
Human organ atlas dataset view with
Edit mode



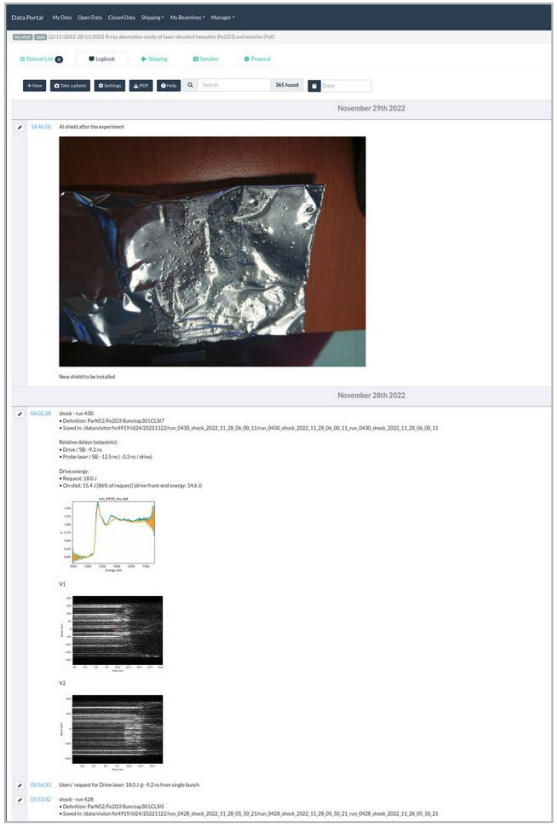
Standard view based in the gallery

Generic Sample tracking

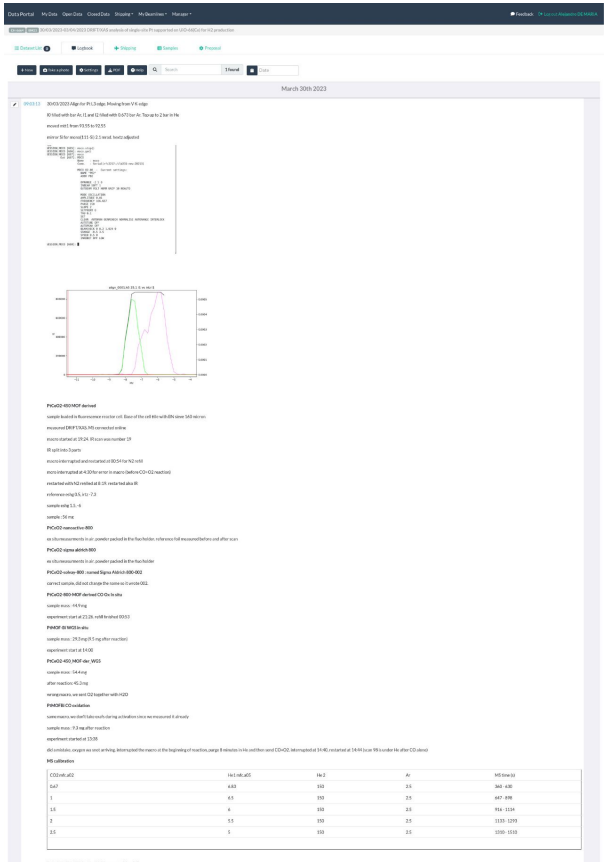




Features: Electronic logbook

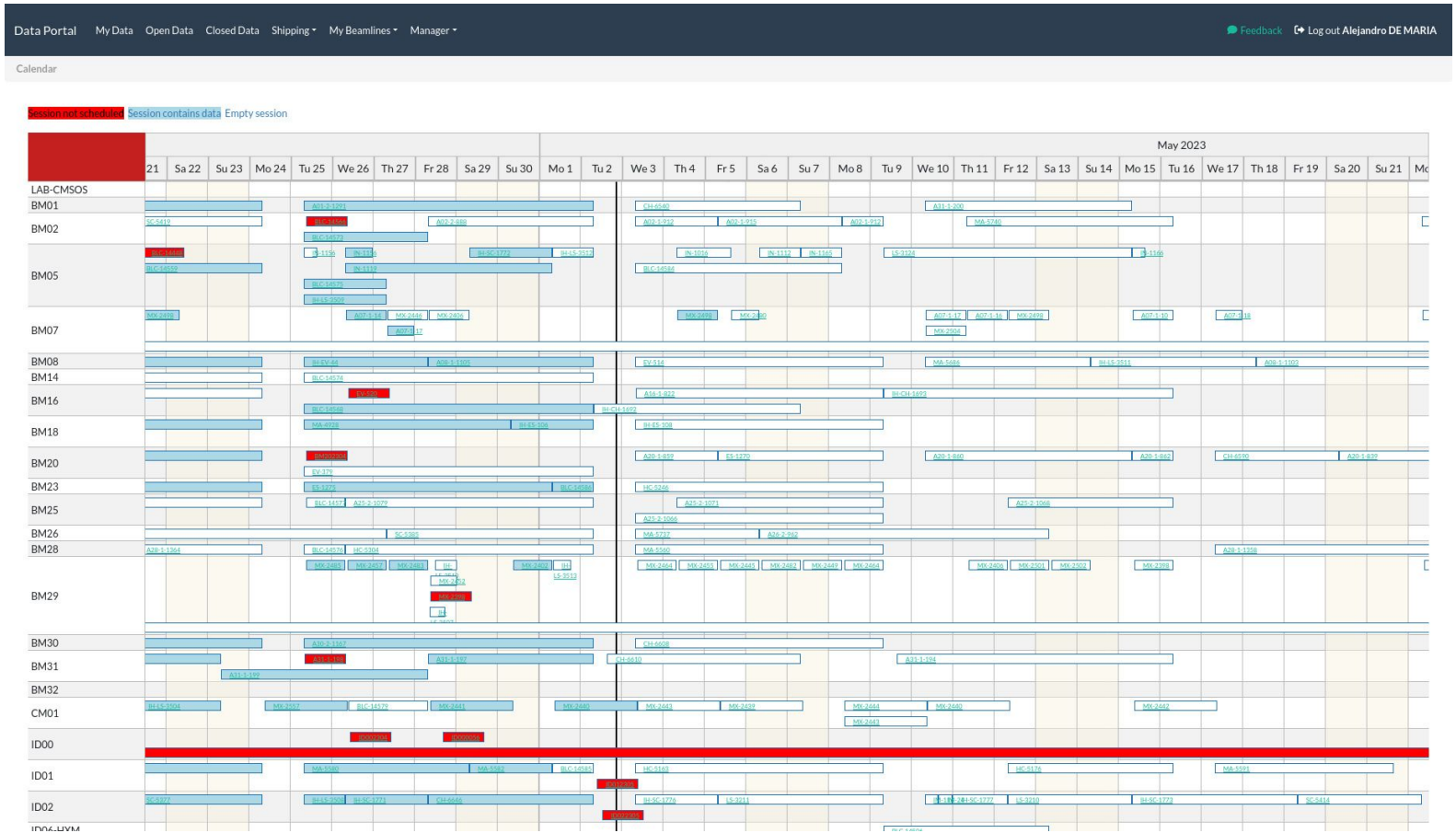


Electronic logbook

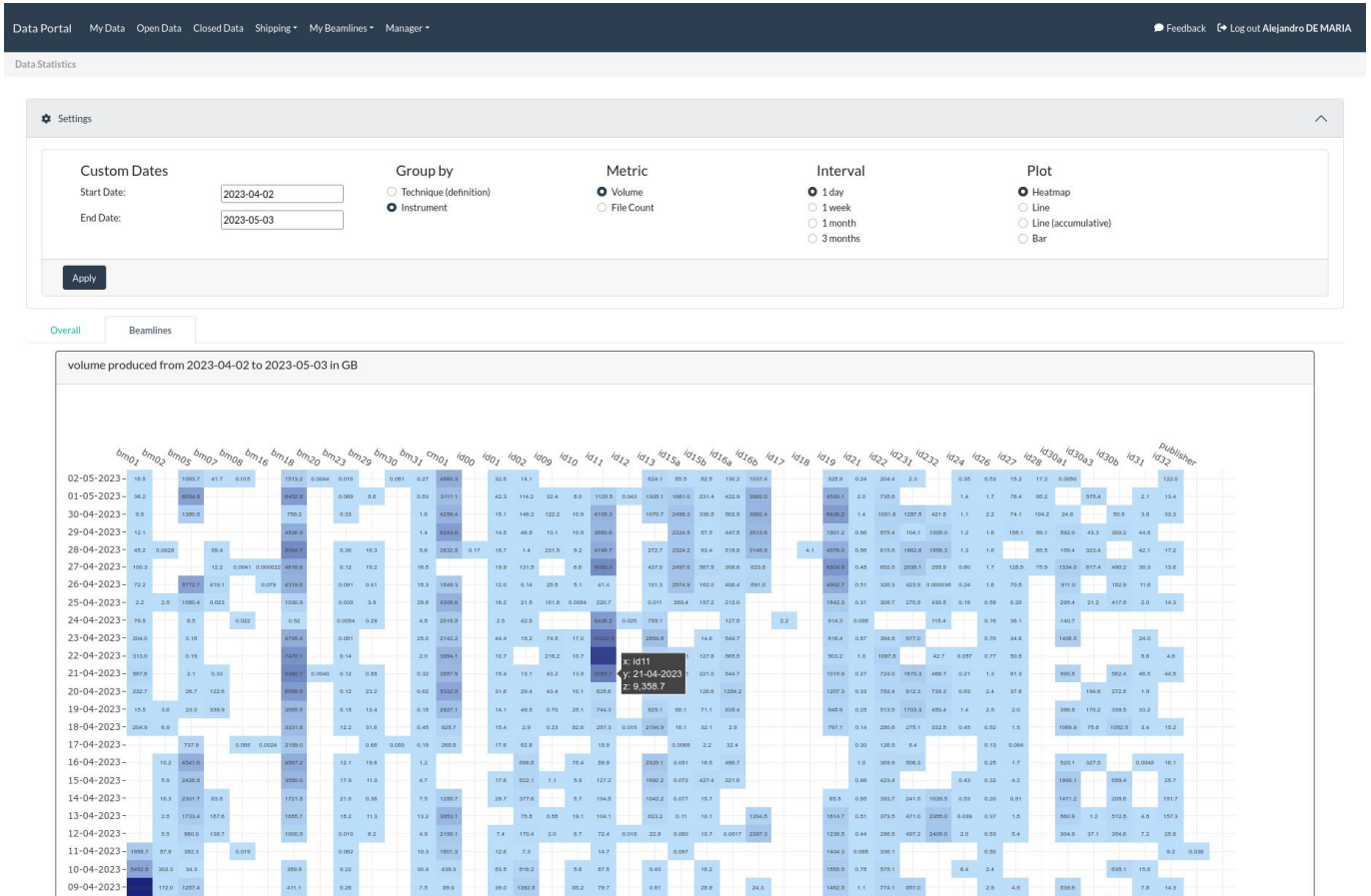


- Allows to annotate you experiments
  - Automatically
  - Manually
- Logbook types
  - Experiment
  - Beamline
  - Facility
- It is included and come out of the box for free

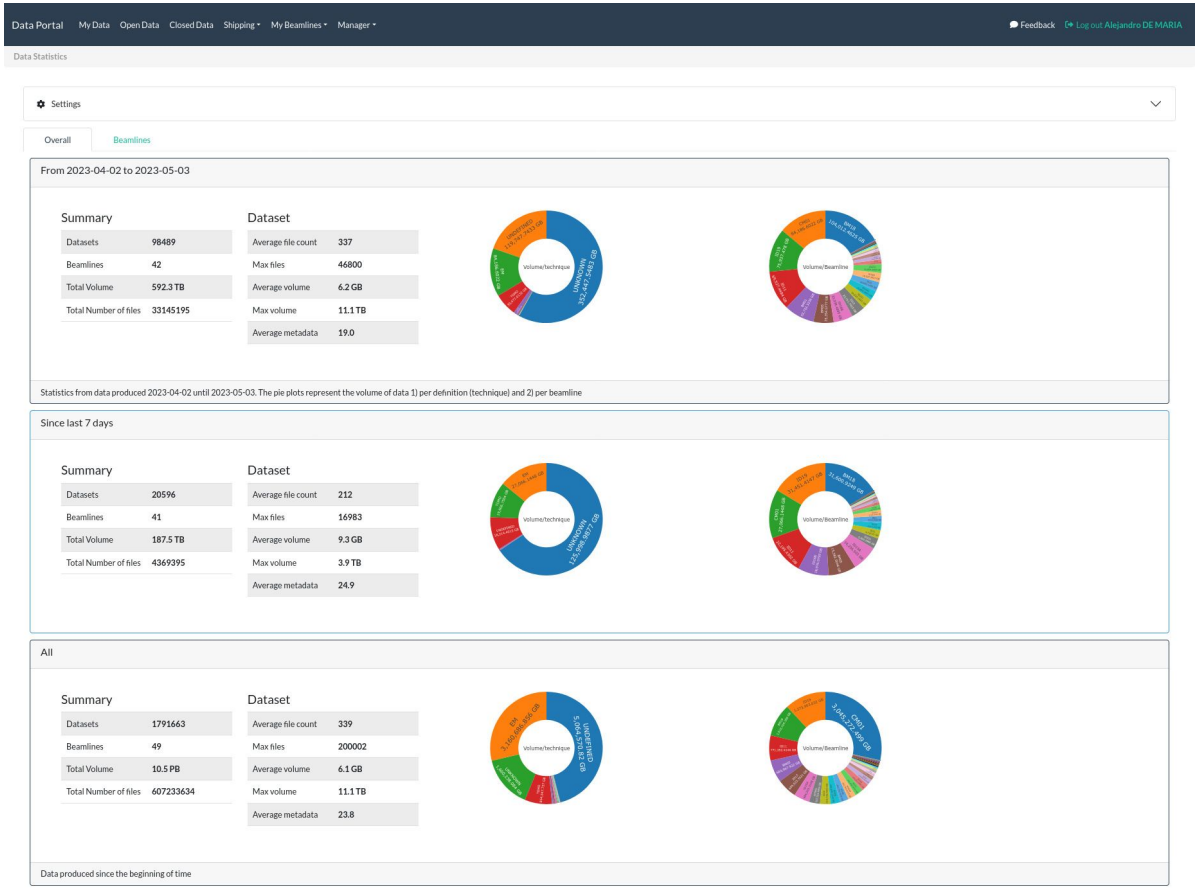
Features: Investigation calendar



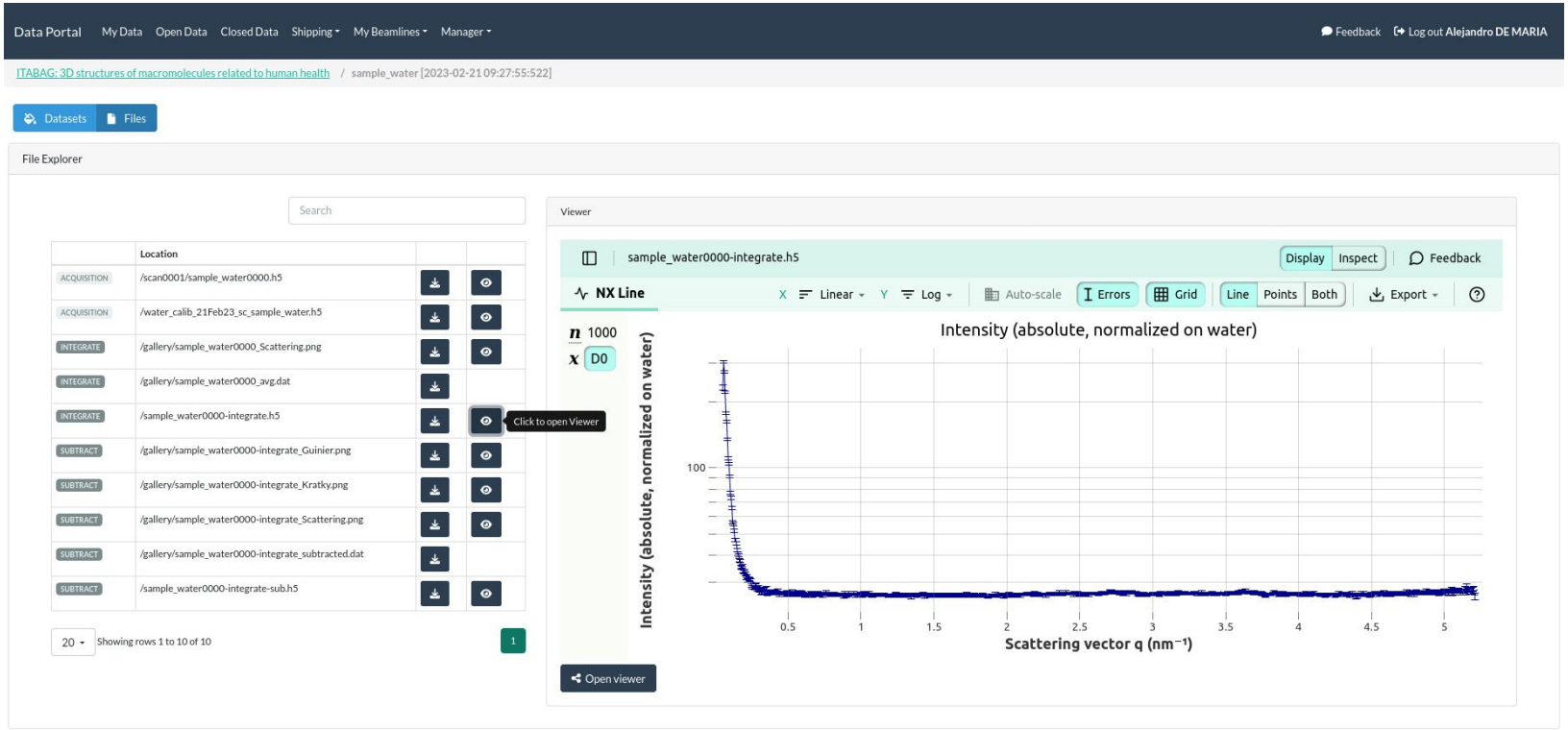
Features: Statistics



Features: Statistics

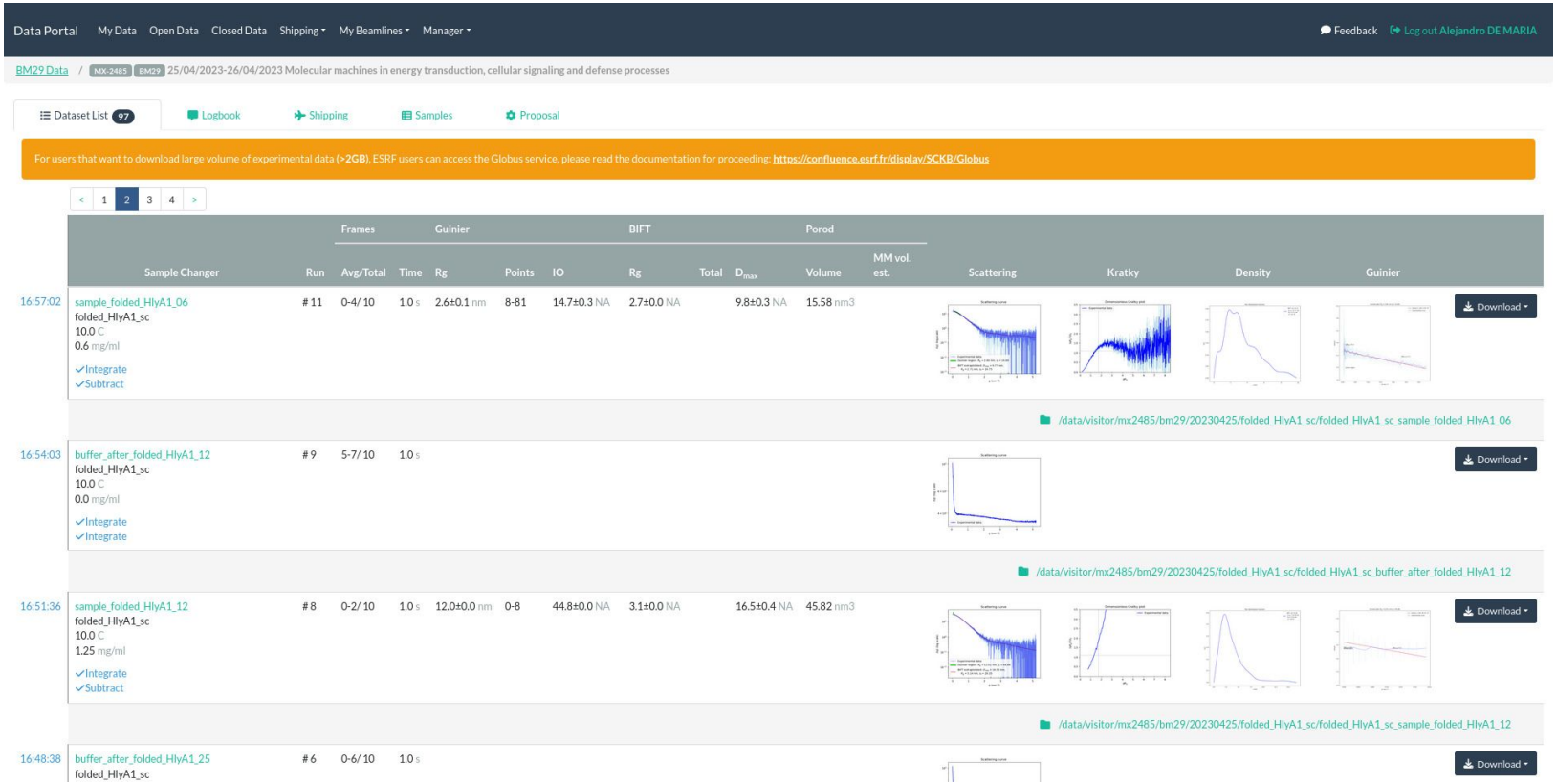


H5Viewer

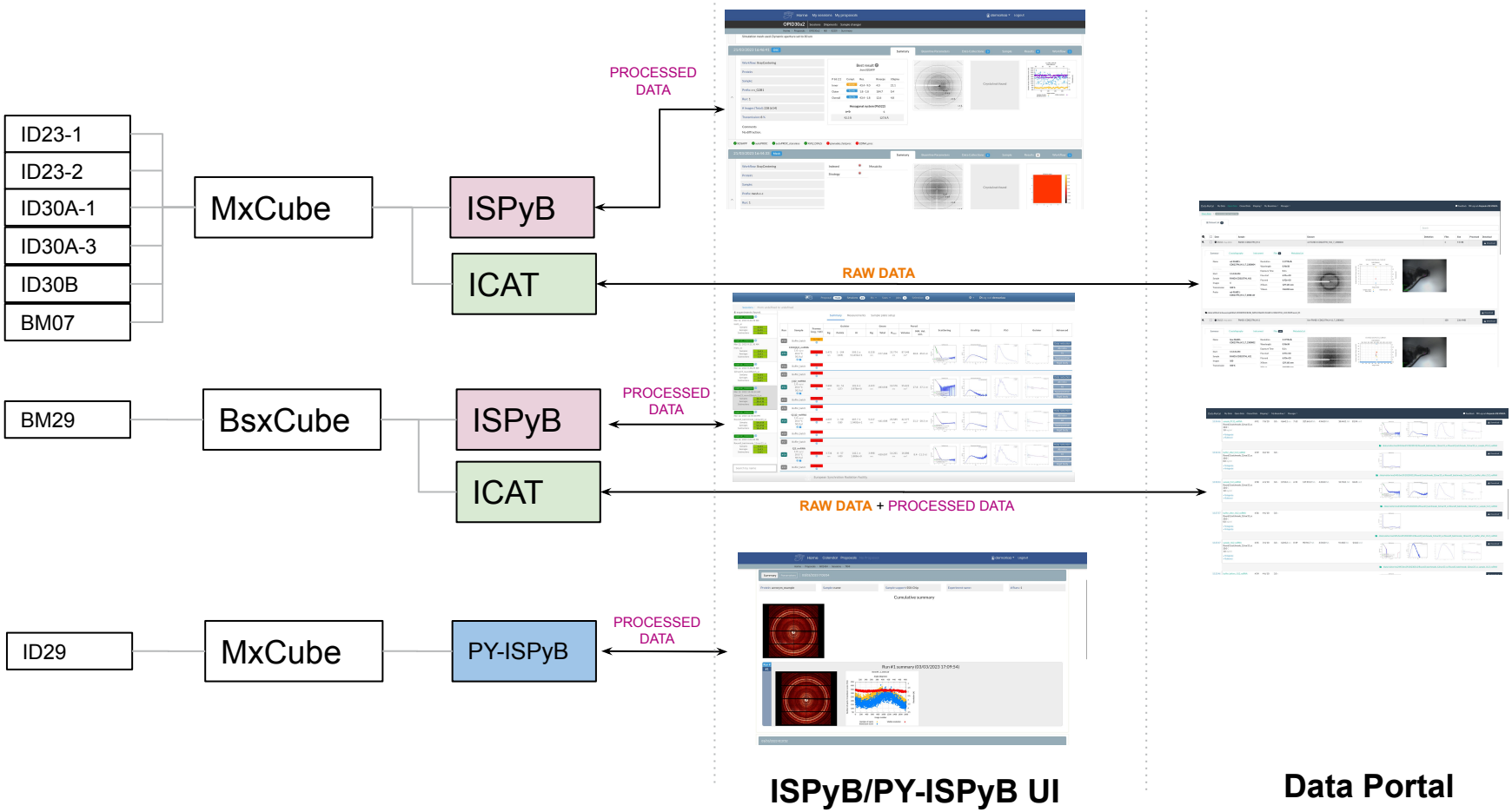

Data Portal
My Data
Open Data
Closed Data
Shipping
My Beamlines
Manager
Feedback
Log out Alejandro DE MARIA
ITABAG: 3D structures of macromolecules related to human health / sample_water [2023-02-21 09:27:55:522]
Datasets
Files
File Explorer
Search
Location
ACQUISITION /scan0001/sample_water0000.h5
ACQUISITION /water_calib_21Feb23_sc_sample_water.h5
INTEGRATE /gallery/sample_water0000_Scattering.png
INTEGRATE /gallery/sample_water0000_avg.dat
INTEGRATE /sample_water0000-integrate.h5
Click to open Viewer
SUBTRACT /gallery/sample_water0000-integrate_Guinier.png
SUBTRACT /gallery/sample_water0000-integrate_Kratky.png
SUBTRACT /gallery/sample_water0000-integrate_Scattering.png
SUBTRACT /gallery/sample_water0000-integrate_subtracted.dat
SUBTRACT /sample_water0000-integrate-sub.h5
20
Showing rows 1 to 10 of 10
1
Viewer
sample_water0000-integrate.h5
Display
Inspect
Feedback
NX Line
X Linear
Y Log
Auto-scale
Errors
Grid
Line
Points
Both
Export
n 1000
x D0
Intensity (absolute, normalized on water)
Intensity (absolute, normalized on water)
100
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Scattering vector q (nm⁻¹)
Open viewer

Features: dataset view

- Automatic upload of processed data has been implemented in some beamlines

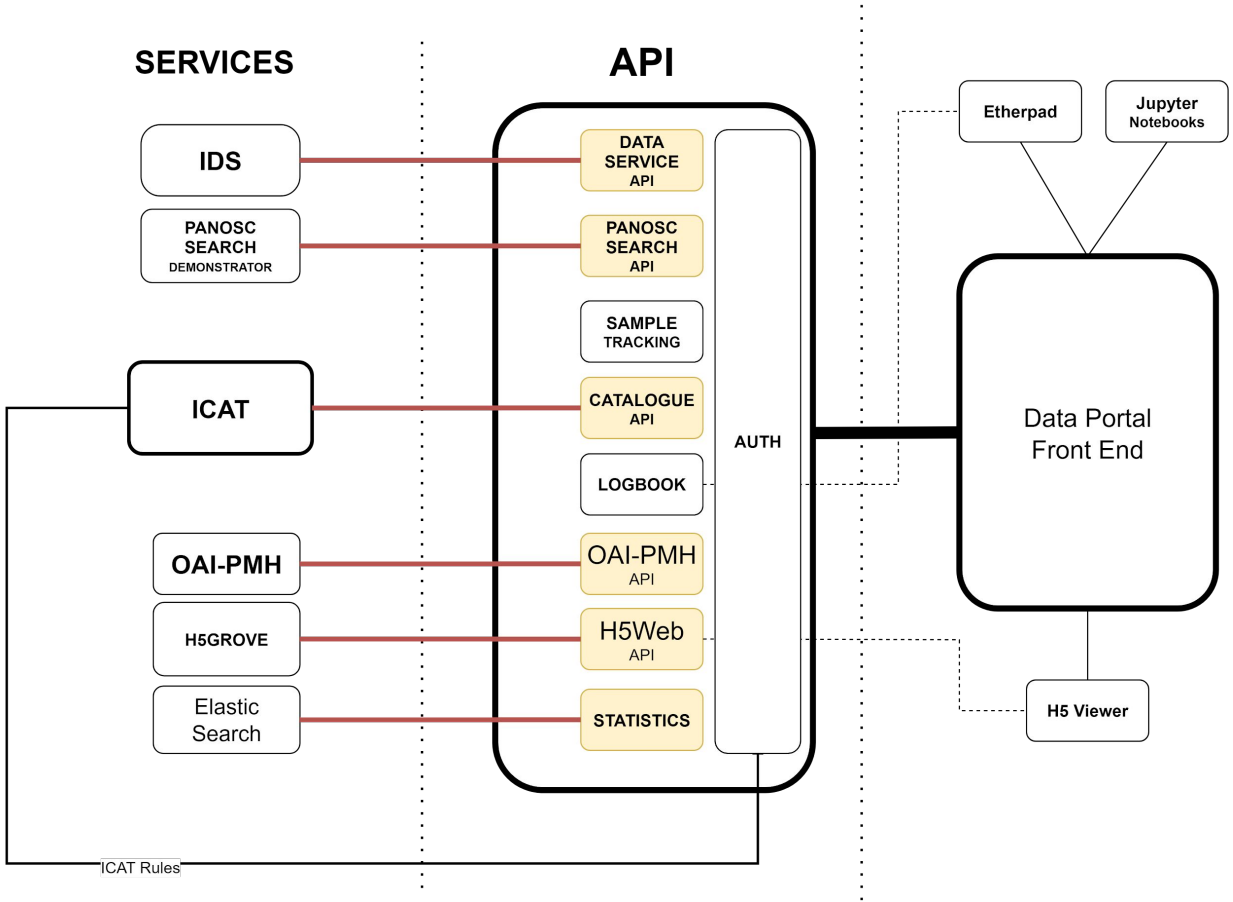


Data policy implementation

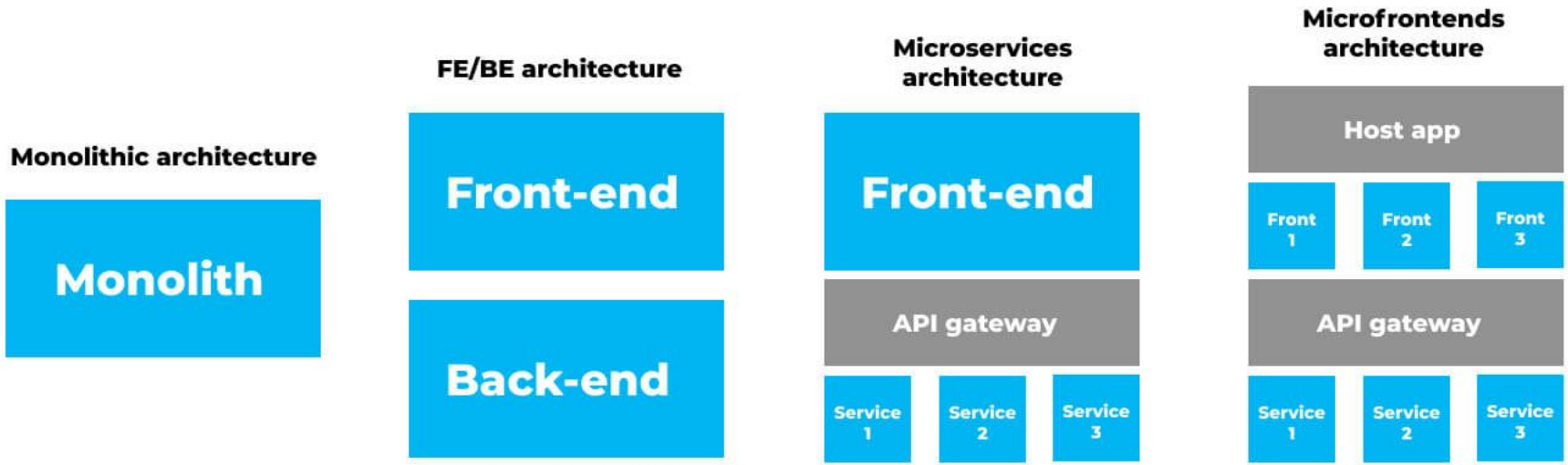


Architecture

Architecture based on microservices + microfrontends



Architecture based on microservices + micro frontends



Evaluation

Feasibility study

- Can ISPyB be eventually be replaced by ICAT?
  - **Ingestion** of the data:
    - Can MX data be ingested in ICAT? How?
  - **Data visualization**:
    - Can MX data be visualized in a ISPyB-like viewer from ICAT?
  - **Scalability**:
    - Can we manage many techniques with a single software? How?
  - **New technique:**
    - Can we evaluate how much time it takes to add a new technique?

STREAMLINE

Ingestion of the data

SAMPLE

RAC1-Fus_RAC1b

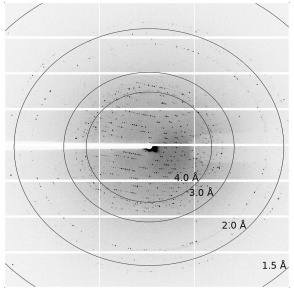
Protein: HBLC
Acronym: HBLC-1
Space group: P1
…

DATASET

RAC1-Fus_RAC1b_1

Flux start: 1.00e+12 ph/sec
Flux end: 1.50e+12 ph/sec
Image Count: 4000
Energy: 12.848 keV
….

/RAC1-FusPOSH_D6_0001.cbf
/RAC1-FusPOSH_D6_0002.cbf
/RAC1-FusPOSH_D6_0003.cbf
…
/RAC1-FusPOSH_D6_04000.cbf



- Data persistence is technique agnostic
- Dataset oriented
- Dataset contains:
  - **Parameters**: key-value pairs
  - **Data files**

STREAMLINE

Ingestion of the data

SAMPLE

RAC1-Fus_RAC1b

Protein: HBLC
Acronym: HBLC-1
Space group: P1
…

DATASET

RAC1-Fus_RAC1b_1

Flux start: 1.00e+12 ph/sec
Flux end: 1.50e+12 ph/sec
Image Count: 4000
Energy: 12.848 keV
….

/RAC1-FusPOSH_D6_0001.cbf
/RAC1-FusPOSH_D6_0002.cbf
/RAC1-FusPOSH_D6_0003.cbf
…
/RAC1-FusPOSH_D6_04000.cbf

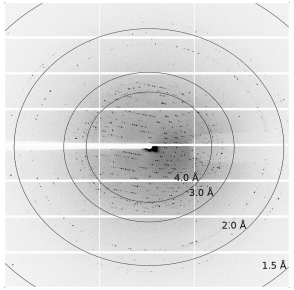
DATASET

Integration

Completeness: 70%
multiplicity: …
Space group: …

….

/processed/x.ext
/processed/x1.ext
…
/processed/xn.ext



STREAMLINE

Ingestion of the data

SAMPLE

RAC1-Fus_RAC1b

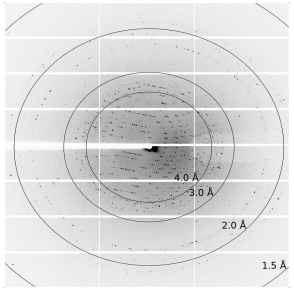
Protein: HBLC
Acronym: HBLC-1
Space group: P1
…

DATASET

RAC1-Fus_RAC1b_1

Flux start: 1.00e+12 ph/sec
Flux end: 1.50e+12 ph/sec
Image Count: 4000
Energy: 12.848 keV
….

/RAC1-FusPOSH_D6_0001.cbf
/RAC1-FusPOSH_D6_0002.cbf
/RAC1-FusPOSH_D6_0003.cbf
…
/RAC1-FusPOSH_D6_04000.cbf



DATASET

Integration

Completeness: 70%
multiplicity: …
Space group: …

….

/processed/x.ext
/processed/x1.ext
…
/processed/xn.ext

DATASET

SAD

Parameter1: …
Parameter2: …

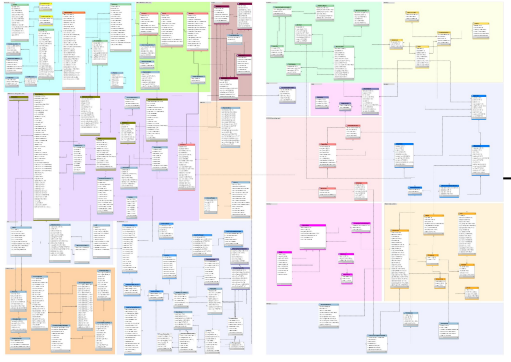
/processed/x.ext
/processed/x1.ext
…
/processed/xn.ext

MR

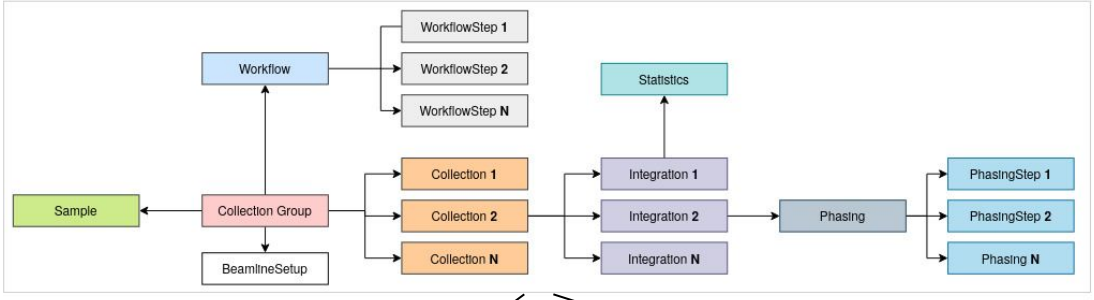
/processed/x.ext
/processed/x1.ext
…
/processed/xn.ext

Ingestion of the data

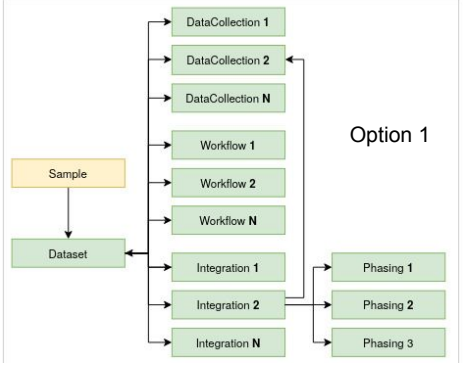
- ispyb2icat.py script transfers data and metadata from ISPyB to disk
- Then data is ingested then with another script simulating the beamline behaviour
- It is crucial to define the datasets and the relationship between them



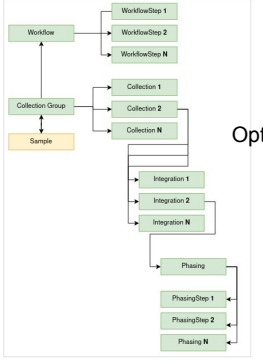
Data model



Main entities



Option 1



Option 2

Data visualization

- Our approach was to directly connect py-ispyb-ui to ICAT
- The modification was done in less than 1 day

- Evaluation
  - Very happy with the results
  - **Out of the box** the performance is slightly poorer than ISPyB
  - Scientists and ourselves consider it is **acceptable**

- Mitigation
  - It is possible to get a better performance by using lazy loading
  - Improving the queries done to ICAT

ICAT

ISPyB

STREAMLINE

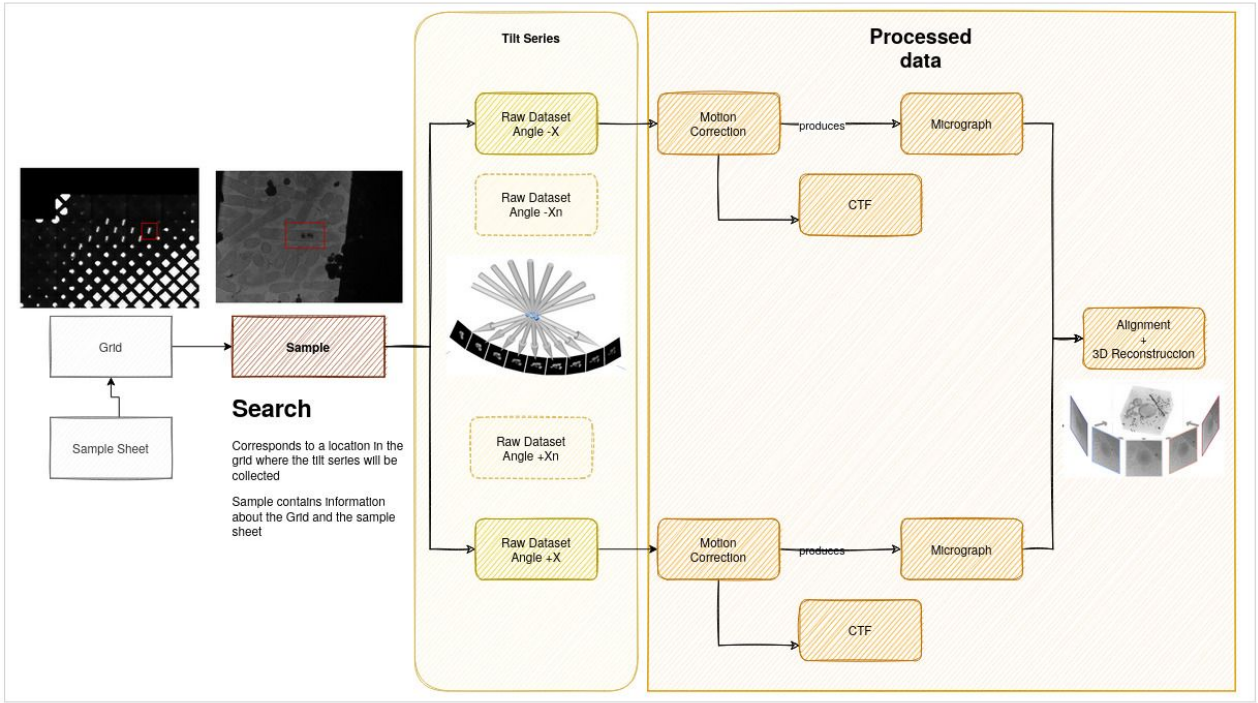
Scalability

- By choosing ICAT as backend the complexity in the DB is vanished

- However, multiple techniques implies to handle:
  - Set of parameters per technique -> Nexus convention-like
  - Specific visualizations -> Micro front-ends

Adding a new technique: cryoET

- A new technique was added from scratch (cryoET) following the next steps:
  - Identify the entities that will be stored as datasets



Adding a new technique: cryoET

- A new technique was added from scratch (cryoET) following the next steps:
  - Identify the entities that will be stored as datasets and their relationship
  - Identify the metadata associated to each dataset

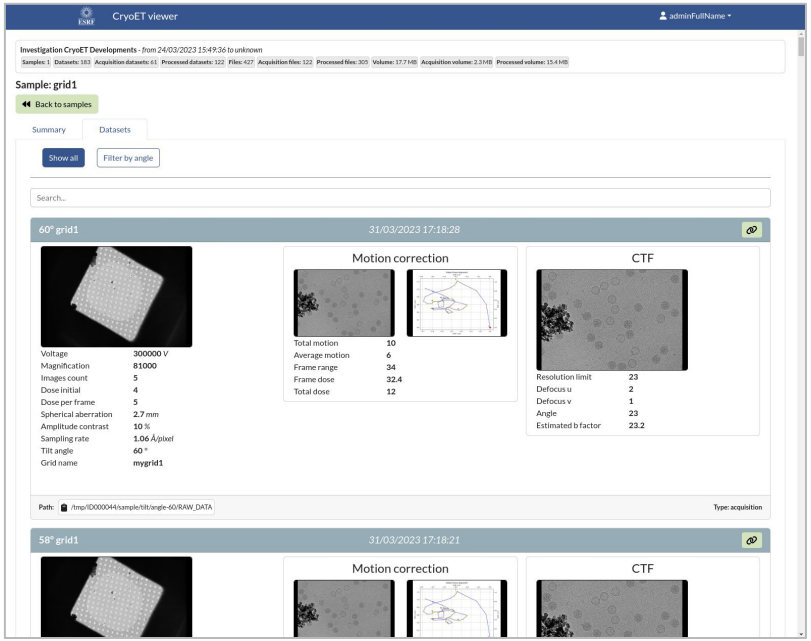
Metadata

Tilt Angle dataset	Motion Correction	CTF
sampleName EM_voltage (V) EM_magnification (NA) EM_images_count (NA) EM_position_x (NA) EM_position_y (NA) EM_dose_initial (NA) EM_dose_per_frame (NA) EM_spherical_aberration (mm) EM_amplitude_contrast (%) EM_sampling_rate (Å/pixel) EM_tilt_angle (°) EM_grid_name (NA)	EMMotionCorrection_total_motion (NA) EMMotionCorrection_average_motion (NA) EMMotionCorrection_frame_range (NA) EMMotionCorrection_frame_dose (NA) EMMotionCorrection_total_dose (NA)	EMCTF_resolution_limit (NA) EMCTF_correlation (NA) EMCTF_defocus_u (NA) EMCTF_defocus_v (NA) EMCTF_angle (NA) EMCTF_estimated_b_factor (NA)

STREAMLINE

Adding a new technique: cryoET

- A new technique was added from scratch (cryoET) following the next steps:
  - Identify the entities that will be stored as datasets and their relationship
  - Identify the metadata associated to each dataset
  - Build the UI



What is missing?

- Working in a version 2 of the generic sample logistics to define:
  - Configurable experiment plan
  - Configurable processing plan
- Currently testing the micro frontend architecture with module federation in React
- MxCube to send the processed data to the ingesters
  - Currently MXCube sends the raw dataset to ICAT
  - We need analysis pipeline to send the datasets too
- Reprocessing is a high priority and needs to be developed urgently

ESRF Answers to ISPyB Survey Questions #1

- What strategy will you/we use to manage data with the European data policy?

  ESRF started implementing the PaNdata/FAIR data policy in 2015, completed in 2020

- Which data catalogues are used (Icat, SciCat, ...) in the community to register experimental metadata (Proposal, Sample, Raw Data, Data Processing, ….)?
  ICAT v5

- How will the ISPyB Application be linked to these databases? **ESRF is currently running both backends in parallel, copying metadata to ICAT and testing ICAT as ISPyB backend**

- Which techniques do you plan to use ISPyB for?

- Which version of ISPyB are you currently using (tables+backend+frontend)?
  Latest (py-ispyb)

- When do you plan to upgrade to the latest version(s) of ISPyB? **DONE**

ESRF Answers to ISPyB Survey Questions #2

- What do you expect from the ISPyB collaboration?

  Active participation in the development as an open source project

  Contributions in code, documentation, tests, bug reports, metadata definitions

  Installation scripts for different platforms

- What can you contribute to the ISPyB collaboration?

  ESRF is implementing backends (py-ispyb + icat) and frontends for MX, Cryo + BioSAXS

  Open source code and project coordination

  A working solution to be adapted to your site

Conclusions

- Current ISPyB architecture has several limitations
  - Database changes are technically easy, but logistically difficult
  - ISPyB Java middleware is end of live and would need heavy refactoring, and ESRF does not have the resources
  - Some fundamental characteristics of the data model, such as a Data collection focus make proper implementation of some features (e.g. groups) non optimal
  - Table complexity becoming difficult to manage
  - Legacy decisions on e.g. shipping require major rewrites

- ICAT seems to be a good candidate to overcome these data model limitations

- Further evaluation and testing on the possibility to migrate to ICAT is highly recommended

Acknowledgement

- Mael Gaonach
- Romain Talon
- Marjolaine Bodin
- Olof Svensson
- Andy Götz
- Max Nanao
- Software and Structural Biology Group