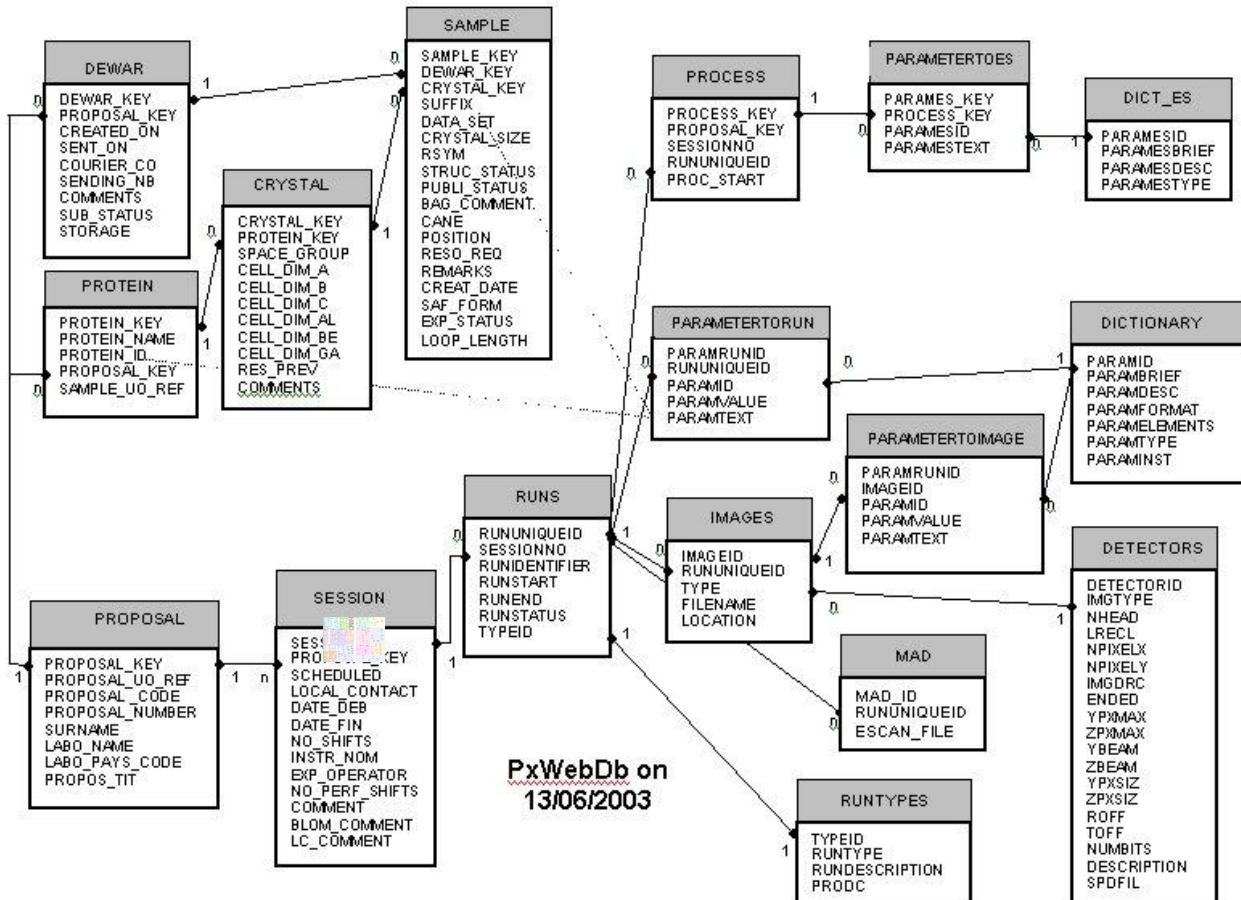


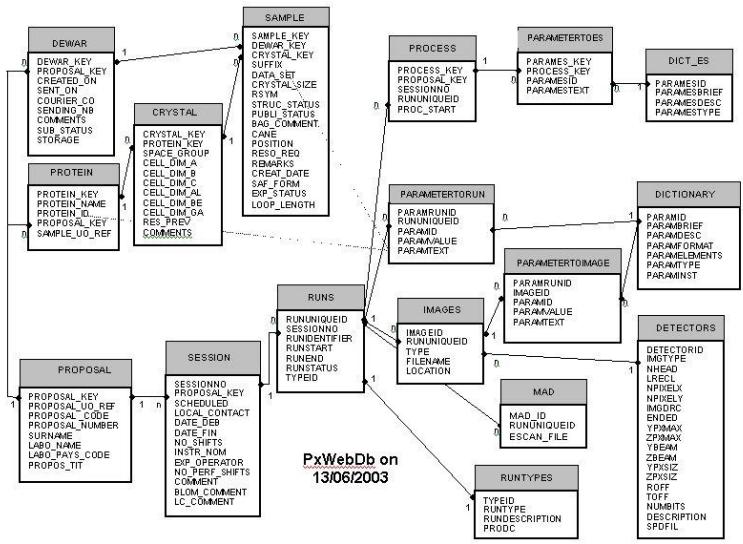
A proposal for a Next Generation of ISPyB Software

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

20th Anniversary

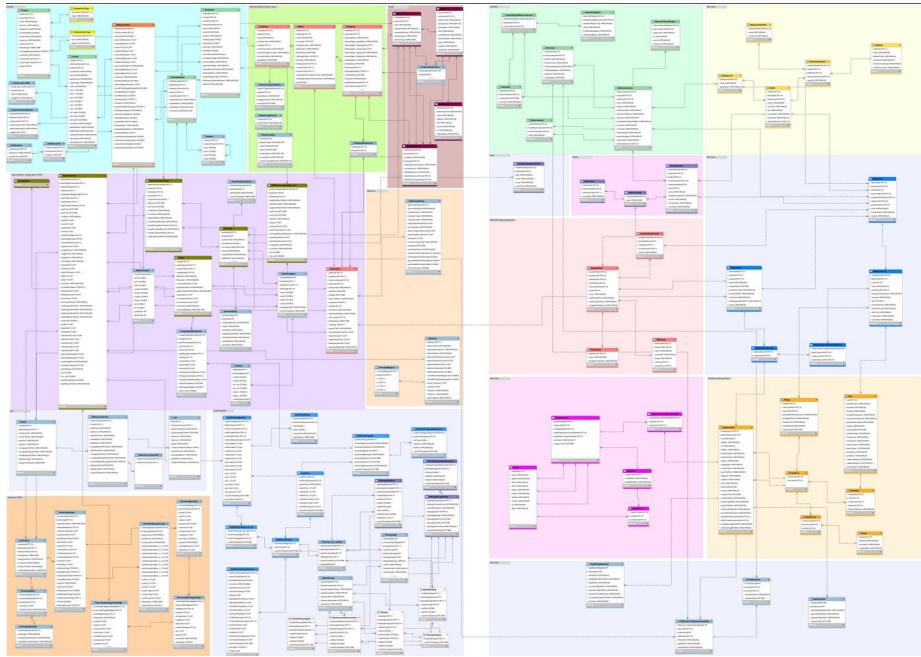


Database Evolution



2003

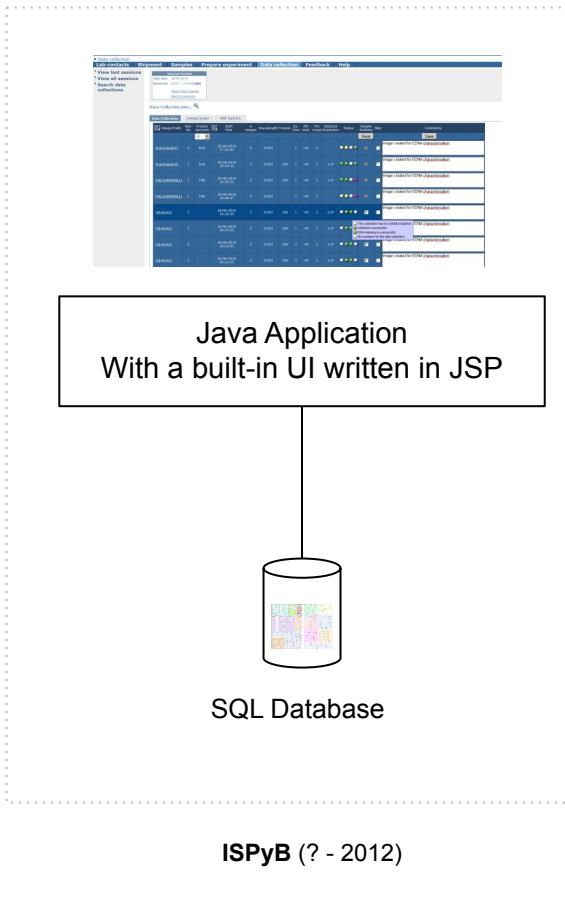
17 tables



2023

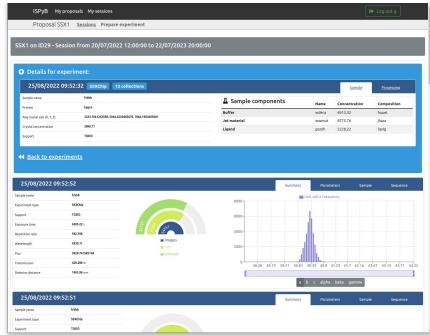
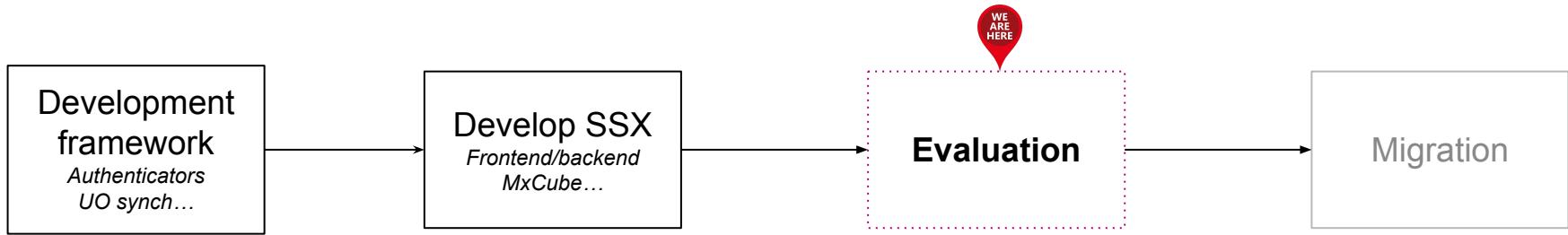
209 tables
42 views

ISPyB Software Evolution



Roadmap achieved for last 4 years

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



GOALS

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

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

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
- **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**

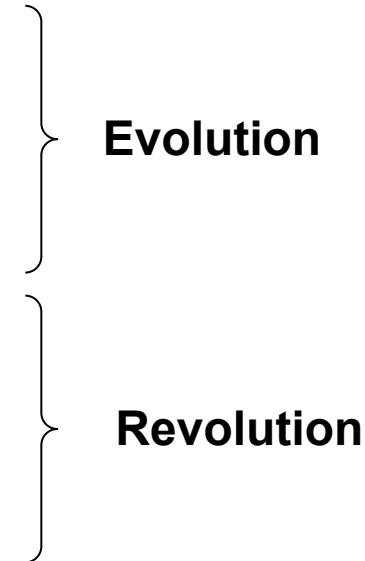
- a. ~~Stay ISPyB's java backend~~
- b. ~~Nothing to reimplement~~

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



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?

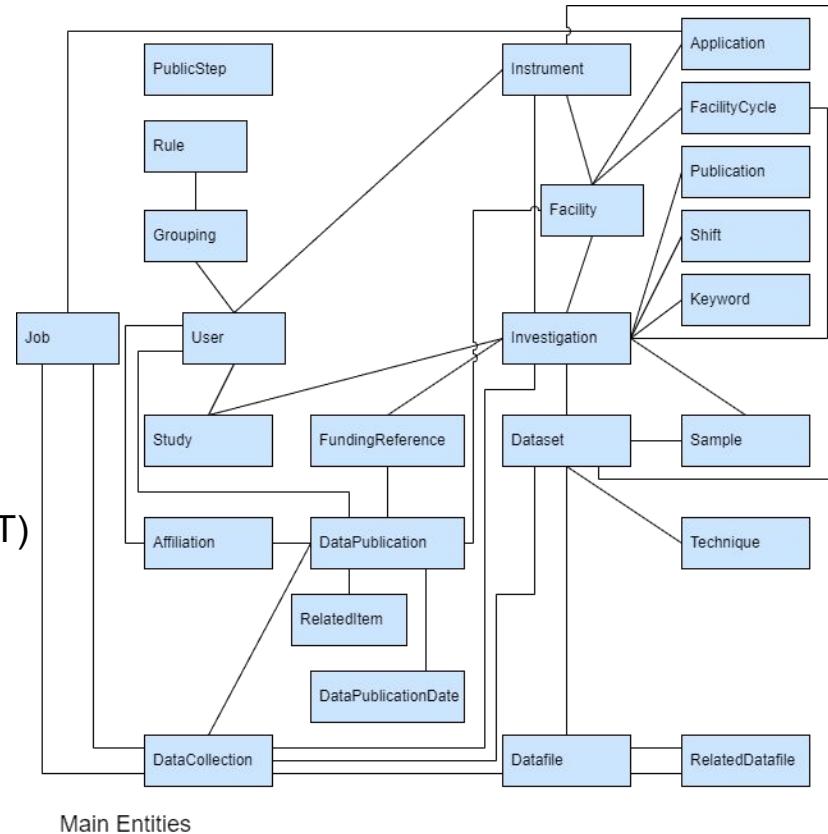
What is ICAT?

- 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...
- 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**



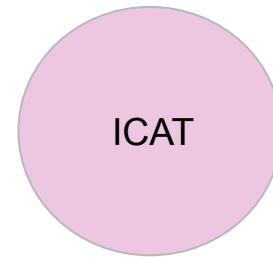
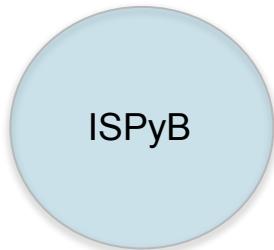
ICAT Data Model

- 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)



ICAT and ISPyB at the ESRF

~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

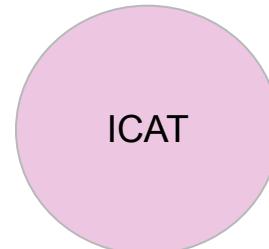
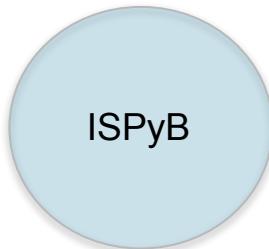
For all techniques:

- 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

The screenshot shows the EXI/ISPyB interface. At the top, there are tabs for Home, Sample, Project Expert, Data Explorer, Manager, Help, and Tools. Below the tabs, there are sections for 'Data Catalogue' and 'Data Processing'. The main area displays experimental parameters such as Beamline (EXI), Wavelength (0.94 Å), Energy (100 eV), and various detector settings. It also shows processed data including a 2D intensity map, a 1D intensity profile, and a corresponding plot.

~2018



The screenshot shows the Data Portal/ICAT interface. At the top, there are tabs for Home, Sample, Dataset, Manager, and Help. Below the tabs, there is a 'Data Catalogue' section. This section lists datasets with details like Name (e.g., 'ref-450000-1'), Resolution (1 Å), Wavelength (0.94 Å), Exposure Time (0.1 s), and Sample (ADMXN-75_25m). It also includes a preview of a 2D intensity map and a 1D intensity profile. At the bottom, there is a 'Data Download' section with a table showing file details such as File, Size, Processed, and Downloaded.

Data Portal/ICAT

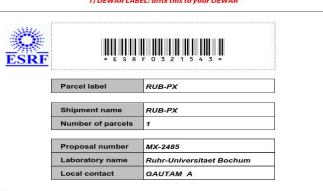
For MX/SAXS/EM:

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

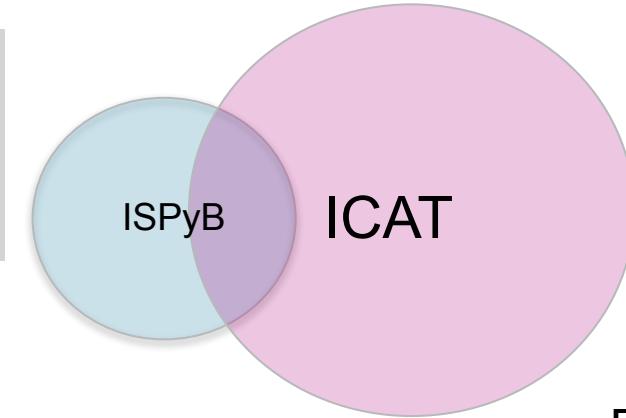
For all techniques:

- 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



ISPyB's parcel label



For MX/SAXS/EM:

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

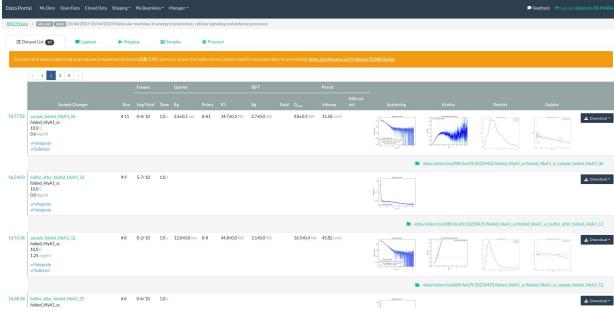
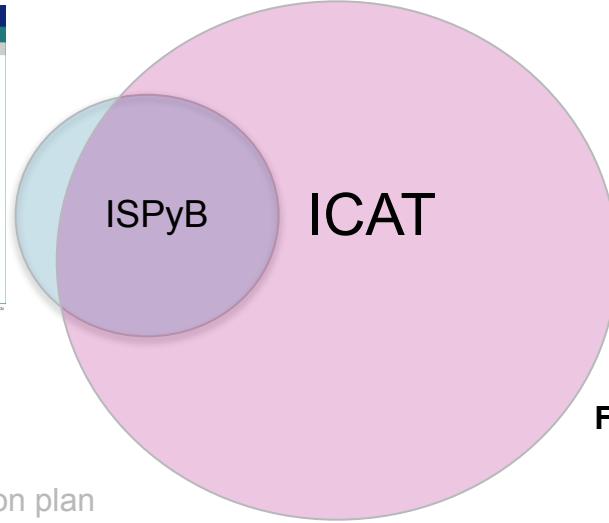
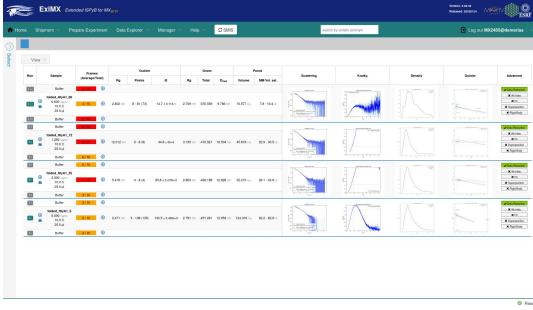


ICAT's parcel label

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:

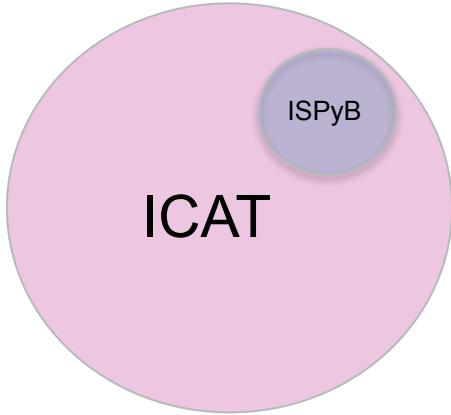
- LIMS
- Sample tracking
 - Diffraction plan
 - Processing plan
- Display of processed results
- Real time feedback
- Rich visualizations

~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

This screenshot shows the Data Portal interface. At the top, there are tabs for Data Portal, Open Data, Closed Data, Shipping, My Beamlines, Manager, Feedback, and Logout. A user profile for 'Alejandro DE MARIA' is visible. Below the header, a search bar and a 'New' button are present. The main area displays a table of proposals and visits. The columns include: Proposal, Beamline, Start Date, Title, Affairs, Datasets, Files, Release, and DOI. The table lists several entries, such as 'Sample tracking test proposal' (DOI: 10.13115/1458F-DC-18697294) and 'Sample tracking test proposal' (DOI: 10.13115/1458F-DC-18697295). Each entry includes a small thumbnail image and a 'DOI' link.

This screenshot shows a detailed view of a dataset from the Data Portal. The top navigation bar includes Data Portal, My Data, Open Data, Closed Data, Shipping, My Beamlines, Manager, Feedback, and Logout. The URL in the address bar is 'Open Data / 10.13115/1458F-DC-18697294'. The main content area is titled 'Dataset List' and shows a single dataset entry: '01:14 4 Sep 2018 FAHD1-CD023796_H11 ref-FAHD1-CD023796_H11_7_2380004'. Below this, a 'Summary' tab is selected, displaying various parameters: Name (ref-FAHD1-CD023796_H11_7_2380004), Resolution (2.17996 Å), Wavelength (0.966 Å), Exposure Time (0.1 s), Start (1:14:24 AM), Sample (FAHD1-CD023796_H11), Flux start (6.91e+10), Flux end (6.92e+10), Images (4), Transmission (100 %), X Beam (129.161 mm), Y Beam (146.854 mm), and Prefix (ref-FAHD1-CD023796_H11_7_####.cbf). To the right, there are two large images: a grayscale diffraction pattern and a 3D surface plot of a protein structure labeled 'Hfq'.

Browse proposal/visits

This screenshot shows another detailed view of a dataset from the Data Portal. The top navigation bar and URL are identical to the previous screenshot. The main content area shows a dataset entry: '01:13 4 Sep 2018 FAHD1-CD023796_H11 line-FAHD1-CD023796_H11_7_2380002'. Below this, a 'Summary' tab is selected, displaying parameters: Name (line-FAHD1-CD023796_H11_7_2380002), Resolution (2.17996 Å), Wavelength (0.966 Å), Exposure Time (0.1 s), Start (1:13:31 AM), Sample (FAHD1-CD023796_H11), Flux start (6.97e+10), Flux end (6.98e+10), Images (100), Transmission (100 %), X Beam (129.161 mm), Y Beam (146.854 mm), and Prefix (line-FAHD1-CD023796_H11_7_####.cbf). To the right, there are two large images: a grayscale diffraction pattern and a 3D surface plot of a protein structure labeled 'Hfq'.

Generic Data Collection view

Some features currently available

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

Data Portal My Data Open Data Closed Data Shipping My Beamlines Manager

Open Data / [Dataset List](#) [Log out Alejandro DEMARIA](#)

Dataset List

Search

Date	Sample	Dataset	Definition	Files	Size	Processed	Download
01/14 4 Sep 2018	FAHDI-CD023796_H11	ref-FAHDI-CD023796_H11_7_2380004		4	9.5 MB		Download

Summary Crystallography Instrument Filter Metadata List

Name: ref-FAHDI-CD023796_H11_7_2380004 Resolution: 2.17996 Å
Wavelength: 0.966 Å
Exposure Time: 0.1 s
Start: 1:52:42 AM
Sample: FAHDI-CD023796_H11 Flux start: 6.97e+10
Images: 4 Flux end: 6.93e+10
Transmission: 100 % X Beam: 129.361 mm
Prefix: ref-FAHDI-CD023796_H11_7_####.cif Y Beam: 146.854 mm

[/data/d30/1/linhouse/opj30/1/20189103/RAW DATA/FAHDI/FAHDI-CD023796_H11/MXPyessA_01](#) [Download](#)

01/13 4 Sep 2018 FAHDI-CD023796_H11 Ims-FAHDI-CD023796_H11_7_2380002 100 236.4 MB [Download](#)

Summary Crystallography Instrument Filter Metadata List

Name: Ims-FAHDI-CD023796_H11_7_2380002 Resolution: 2.17996 Å
Wavelength: 0.966 Å
Exposure Time: 0.1 s
Start: 1:19:31 AM
Sample: FAHDI-CD023796_H11 Flux start: 6.97e+10
Images: 100 Flux end: 6.93e+10
Transmission: 100 % X Beam: 129.361 mm
Prefix: Ims-FAHDI-CD023796_H11_7_####.cif Y Beam: 146.854 mm

Data Portal My Data Open Data Closed Data Shipping My Beamlines Manager

Open Data / [Dataset List](#) [Log out Alejandro DEMARIA](#)

Dataset List

For users that want to download large volumes of experimental data (>2GB), ESRF users can access the Globus service, please read the documentation for procedure: <https://confluence.esrf.fr/display/SCB/Globus>

Open Data / [Dataset List](#) [Logbook](#) [Shipping](#) [Samples](#) [Proposal](#)

Filter by samples: [X](#)

Search

Date	Sample	Dataset	Definition	Files	Size	Processed	Download
10:16 5 Nov 2018	fe2streptor2	fe2streptor2_XAScalib	SXM	5	5.1 MB		Restore
00:16 5 Nov 2018	fe2streptor2	fe2streptor2_main_root	SXM	1343	2.6 GB		Restore

Summary Instrument Filter Metadata List

Name: fe2streptor2_main_root Definition: SXM
Start: 12:16:15 AM
Sample: fe2streptor2 Description: Fe2streptor replica 2

#2: Fec_00 #1: Fec_00_Kc_00_Cac_00

#1: Fec_00 #2: Fec_00_Kc_00_Cac_00

[/data/visitor/v280/d21/fe2streptor2/fe2streptor2_main_root](#)

Date	Sample	Dataset	Definition	Files	Size	Processed	Download
23:58 4 Nov 2018	fe2streptor2	fe2streptor2_coarse	SXM	157	48.0 MB		Restore
23:58 4 Nov 2018	fe2streptor2	fe2streptor2_spec01	SXM	2	37.7 KB		Download

Custom views

[Data Portal](#) [My Data](#) [Open Data](#) [Closed Data](#) [Shipping](#) [My Beamlines](#) [Manager](#) [Feedback](#) [Logout Alejandro DE MARIA](#)

[Summary](#) [File](#) [Metadata List](#)

Patient

definition MRromo
Identifier LADAF-2020-27
Age (years) 94
Sex female
Organ left lung
Institute Laboratoire d'Anatomie des Alpes Françaises
Info right Sylvian and right cerebellar stroke, cognitive disorders of vascular origin, degenerative arachnoiditis, bilateral carpal tunnel syndrome, micro-crystalline arthritis (gout), right lung pneumopathy (?) before death; catarract of the left eye; hyperemia and edema of the skin (left temporal region)

Sample

Sample LADAF-2020-27_left_lung_left_info complete lung tissue from body donor program of the Laboratoire d'Anatomie des Alpes Françaises (LADAF)
Preparation formalin fixed, progressive transfer to ethanol 70% with inflation using ethanol at each step alternately with vacuum degassing, mounted with agar gel rates

Scan Parameters

Instrument BMOS EB5 dipole wiggler 0.85T
SR Current (mA) 200
Exposure Time 0.30
Pixel Size (um) 2.51
Mode (None) continuous
ScanRadius 1400_251um_LADAF_2020-27_left_lung_FSC_A
Stage (x,y,z) 1.1,0.3 with 0.1 vertical shift
Projections 4000
dark (None) 400
Detector Distance (mm) 1440
Energy (keV) 77
Scanning Geometry half acquisition
Scan Range (deg) 360
Pixel (xy) 2048,1008
Magnification 2.61
Scintillator LuAG:Ce 250 um
Sur.Dose Rate (Gy/h) 128
Dose Rate (Gy/h) 34.7
VOL Integ. Dose (kGy) 4.22
Scan time (min) 4.05
Series time (s) 0.5

[Download](#)

DOI

Abstract Two zooms in local tomography at 2.51 um resolution performed by the LADAF-2020-27 using half acquisition mode. Both volumes are centered on same location with small modifications for Fourier Shell Correlation analysis of the resolution. The two volumes have been aligned in 3D volume stack re-exported with similar bounding boxes.

Title

Two zooms at 2.51 um in the upper spatial half of the left lung from the body donor LADAF-2020-27

Users

Paul Tafferau, Clare Wahl, Will Wagner, Daniel V. Juras, Alessandro Caviglia, Stephan Weinkauf, Mark P. Kuhnel, Eddie Bauer, Michael A. Hildebrand, Lukas Robertis, David A. Lososki, Joseph Jacob, Sebastian Moroz, Emmanuel G. S. de Souza Moreyto, Darren D. Jostig, Martinien Ackermann, Peter D. Lee

[Download](#)

[/data/projects/hip/human_organ_atlas/edits/LADAF-2020-27/left/2.51um_FSC](#)

Showing rows 1 to 1 of 1

European Synchrotron Radiation Facility

*Human organ atlas dataset view with
Edit mode*

[Data Portal](#) [My Data](#) [Open Data](#) [Closed Data](#) [Shipping](#) [My Beamlines](#) [Manager](#) [Feedback](#) [Logout Alejandro DE MARIA](#)

[Open Data](#) / [/data/d30a1/inhouse/opid30a1/20160408/RAW_DATA/AFAMIN/AFAMIN-rev1-B5-1](#)

Dataset List [11]

[Search](#)

	Date	Sample	Dataset	Definition	Files	Size	Processed	Download
<input type="checkbox"/>	● 19:34 4 Apr 2016	AFAMIN-rev1-B5-1	AFAMIN-rev1-B5-1_1_1719747		1640	3.8 GB		Download

Summary **Crystallography** **Instrument** **File** [Metadata List](#)

Name AFAMIN-rev1-B5-1_1_1719747 **Resolution** 2.23303 Å **Wavelength** 0.966 Å
Start 7:34:16 PM **Exposure Time** 0.355 s **Detector** 141e+11
Sample AFAMIN-rev1-B5-1 **Flux start** 1.139e+11 **Flux end** 1.39e+11
Images 1640 **XBeam** 129.191 mm **Transmission** 100 % **Y Beam** 146.853 mm
Prefix AFAMIN-rev1-B5-1_1_###.cif

Scans

[/data/d30a1/inhouse/opid30a1/20160408/RAW_DATA/AFAMIN/AFAMIN-rev1-B5-1](#)

[Download](#)

<input type="checkbox"/>	● 19:33 4 Apr 2016	AFAMIN-rev1-B5-1	ref- AFAMIN-rev1-B5-1_1_1719746	4	9.5 MB	Download
<input type="checkbox"/>	● 19:32 4 Apr 2016	AFAMIN-rev1-B5-1	line- AFAMIN-rev1-B5-1_1_1719745	60	141.8 MB	Download
<input type="checkbox"/>	● 19:31 4 Apr 2016	AFAMIN-rev1-B5-1	line- AFAMIN-rev1-B5-1_1_1719744	40	94.5 MB	Download
<input type="checkbox"/>	● 19:30 4 Apr 2016	AFAMIN-rev1-B5-1	line- AFAMIN-rev1-B5-1_1_1719742	40	94.5 MB	Download
<input type="checkbox"/>	● 19:29 4 Apr 2016	AFAMIN-rev1-B5-1	mesh- AFAMIN-rev1-B5-1_1_1719739	24	56.7 MB	Download
<input type="checkbox"/>	● 19:29 4 Apr 2016	AFAMIN-rev1-B5-1	mesh- AFAMIN-rev1-B5-1_1_1719736	24	56.7 MB	Download
<input type="checkbox"/>	● 19:28 4 Apr 2016	AFAMIN-rev1-B5-1	mesh- AFAMIN-rev1-B5-1_1_1719733	24	56.7 MB	Download
<input type="checkbox"/>	● 19:28 4 Apr 2016	AFAMIN-rev1-B5-1	mesh- AFAMIN-rev1-B5-1_1_1719731	24	56.7 MB	Download
<input type="checkbox"/>	● 19:27 4 Apr 2016	AFAMIN-rev1-B5-1	mesh- AFAMIN-rev1-B5-1_1_1719728	24	56.7 MB	Download
<input type="checkbox"/>	● 19:27 4 Apr 2016	AFAMIN-rev1-B5-1	mesh- AFAMIN-rev1-B5-1_1_1719726	24	56.7 MB	Download

Showing rows 1 to 10 of 11

European Synchrotron Radiation Facility

Standard view based in the gallery

Generic Sample tracking

Data Portal My Data Open Data Closed Data Shipping My Beamlines Manager

Feedback Log out Alejandro DE MARIA

[Back to list](#)

2904 Test
ID002020 - JOANNE

Status: ● APPROVED Actions: [Mark as SENT](#)

[Download Labels](#) [Delete](#)

Parcel info

[Edit](#)

Description: No description.
Storage conditions: At room temperature
Comments: No comments.

Address of Sender:

Sonya Girodon
2880
girodon@esrf.fr
User Office
ESRF 71 Avenue des Martyrs
38000 Grenoble
France

Return Address:

New Address For Bug
232423
Creation@asda.xom
ILL
sdfs
23424 sdfs
sdfs

Content

+ Add an item

Name	Type	Sample	Description	Comments	Edit	Remove
Hard disk	Tool				Edit	Remove
s	Samplesheet	Samplesheet E		test	Edit	Remove

10 ▾

Data Portal My Data Open Data Closed Data Shipping My Beamlines Manager

Parcel Statistics

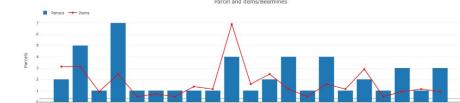
Search parcels or beamlines (29042020) [id] (24552020)

Summary

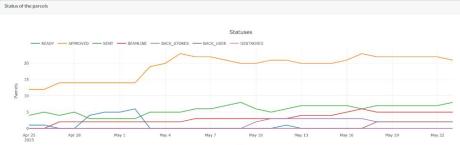
Parcels	Branches	IP	SHIPPING
47	20	0	0
Branches	20	0	0
ESRF	20	0	0
Beamline	0	0	0
ILL	0	0	0
Services	0	0	0
None	0	0	0
Others	0	0	0
Total	54	0	0
Locations	554	0	0



Parcels and items/beamlines



Status of the parcels



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: Proprietary
Barcode ID: ID002020
Session start date: 14-10-2020
Local contact: Stéphanie MONACO

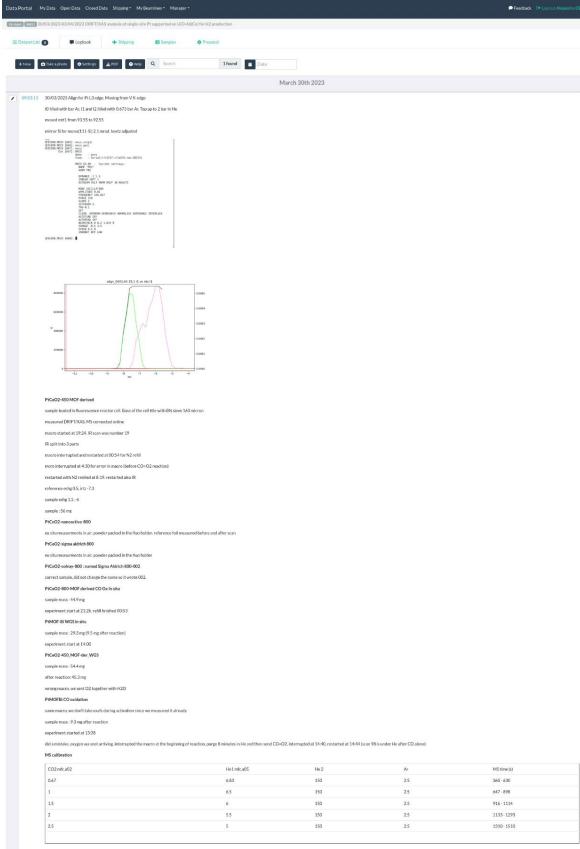
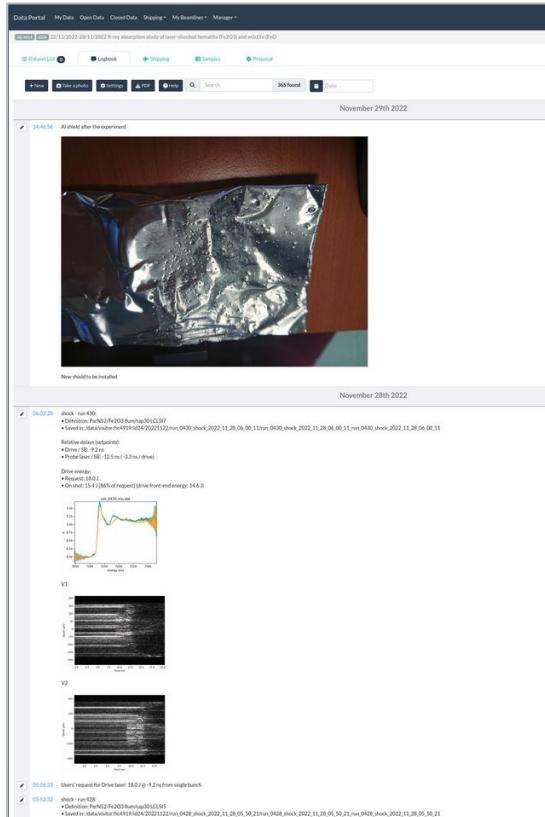


TO:
ESRF Stores
71 avenue des Martyrs
38000 GRENOBLE
FRANCE
Phone: +33 (0)4 76 88 2733
Fax: +33 (0)4 76 88 2347



Storage conditions: At room temperature - None specified

Features: Electronic logbook

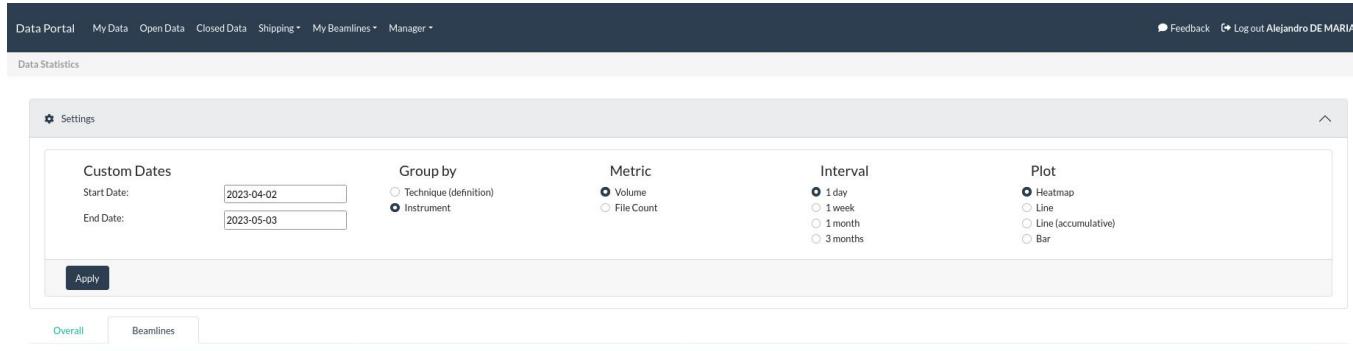


- Allows to annotate your experiments
 - Automatically
 - Manually
 - Logbook types
 - Experiment
 - Beamline
 - Facility
 - It is included and comes out of the box for free

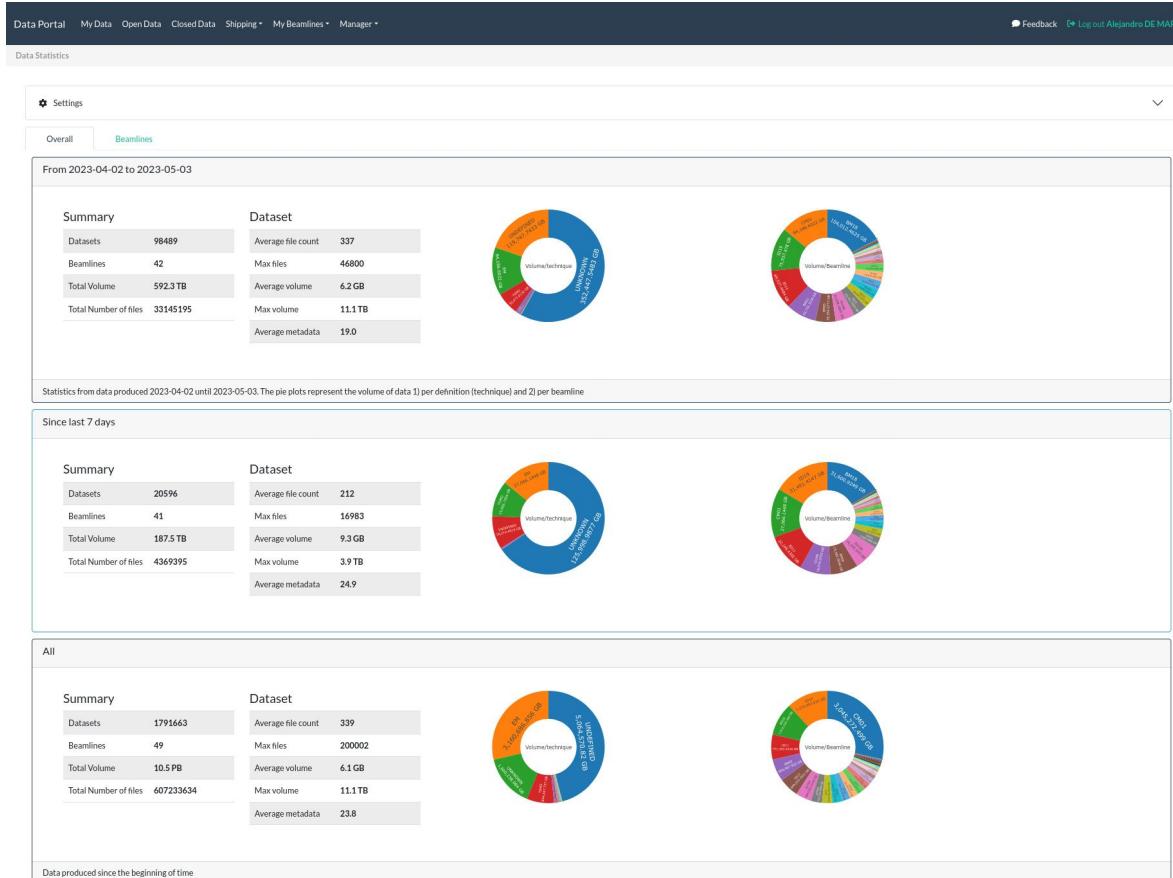
Electronic logbook

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

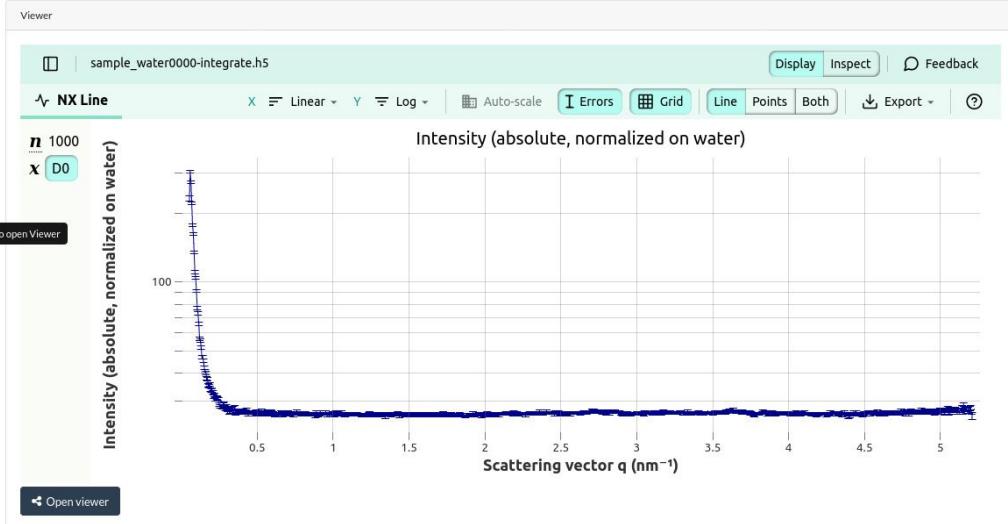
File Explorer

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



Features: dataset view

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

Data Portal My Data Open Data Closed Data Shipping My Beamlines Manager Feedback Log out Alejandro DE MARIA

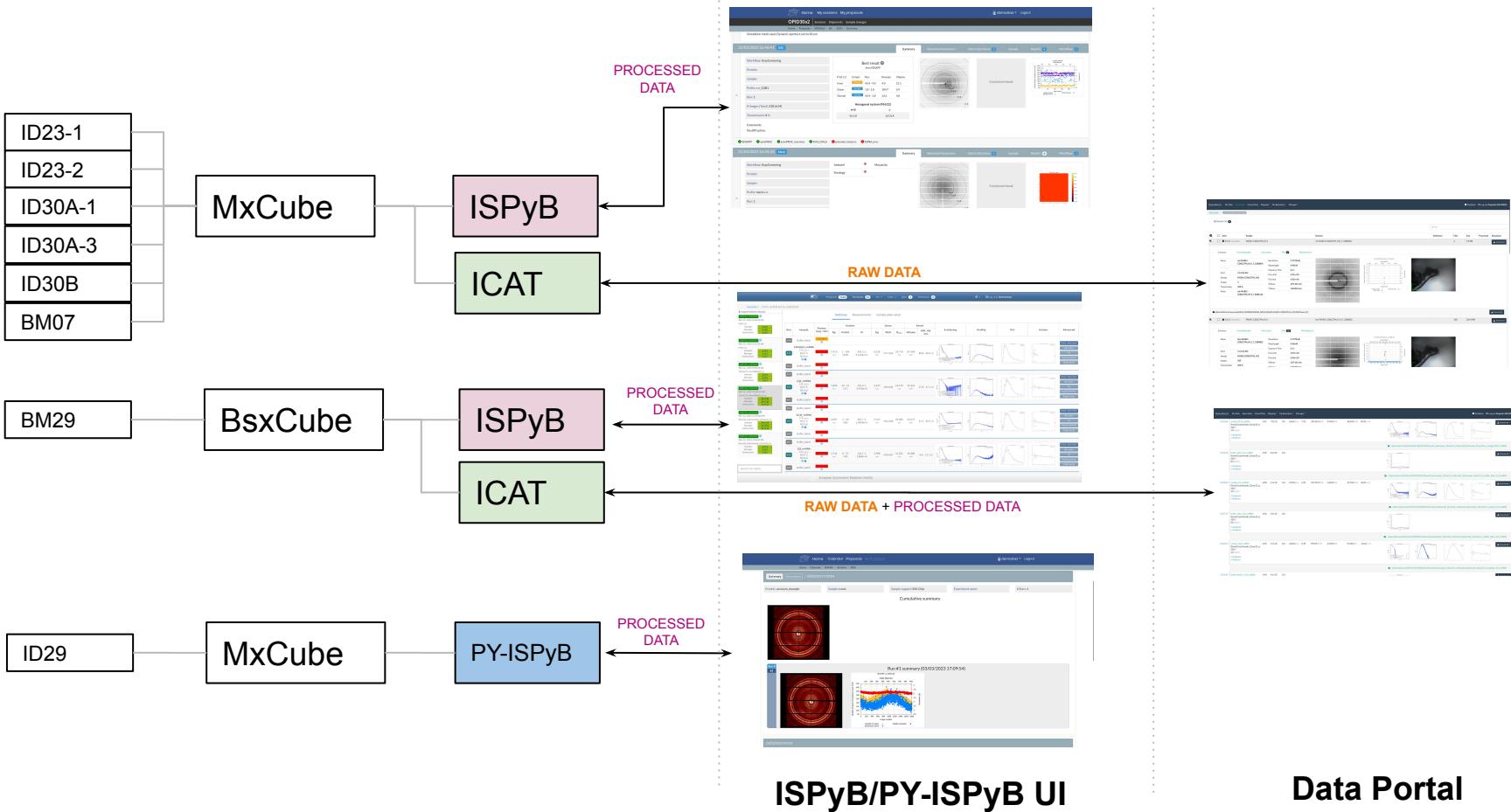
BM29 Data / MX-2485 BM29 25/04/2023-26/04/2023 Molecular machines in energy transduction, cellular signaling and defense processes

Dataset List 97 Logbook Shipping Samples Proposal

For users that want to download large volume of experimental data (>2GB), ESRF users can access the Globus service, please read the documentation for proceeding: <https://confluence.esrf.fr/display/SCKB/Globus>

	Frames	Guinier	BIFT	Porod		Scattering	Kratky	Density	Guinier							
	Sample Changer	Run	Avg/Total	Time	Rg	Points	IO	Rg	Total	D _{max}	Volume	MM vol. est.				
16:57:02	sample_folded_HlyA1_06 folded_HlyA1_sc 10.0 C 0.6 mg/ml ✓Integrate ✓Subtract	# 11	0-4/ 10	1.0 s	2.6±0.1 nm	8-81	14.7±0.3 NA	2.7±0.0 NA	9.8±0.3 NA	15.58 nm ³						Download
16:54:03	buffer_after_folded_HlyA1_12 folded_HlyA1_sc 10.0 C 0.0 mg/ml ✓Integrate ✓Integrate	# 9	5-7/ 10	1.0 s												Download
16:51:36	sample_folded_HlyA1_12 folded_HlyA1_sc 10.0 C 1.25 mg/ml ✓Integrate ✓Subtract	# 8	0-2/ 10	1.0 s	12.0±0.0 nm	0-8	44.8±0.0 NA	3.1±0.0 NA	16.5±0.4 NA	45.82 nm ³						Download
16:48:38	buffer_after_folded_HlyA1_25 folded_HlyA1_sc	# 6	0-6/ 10	1.0 s												Download

Data policy implementation

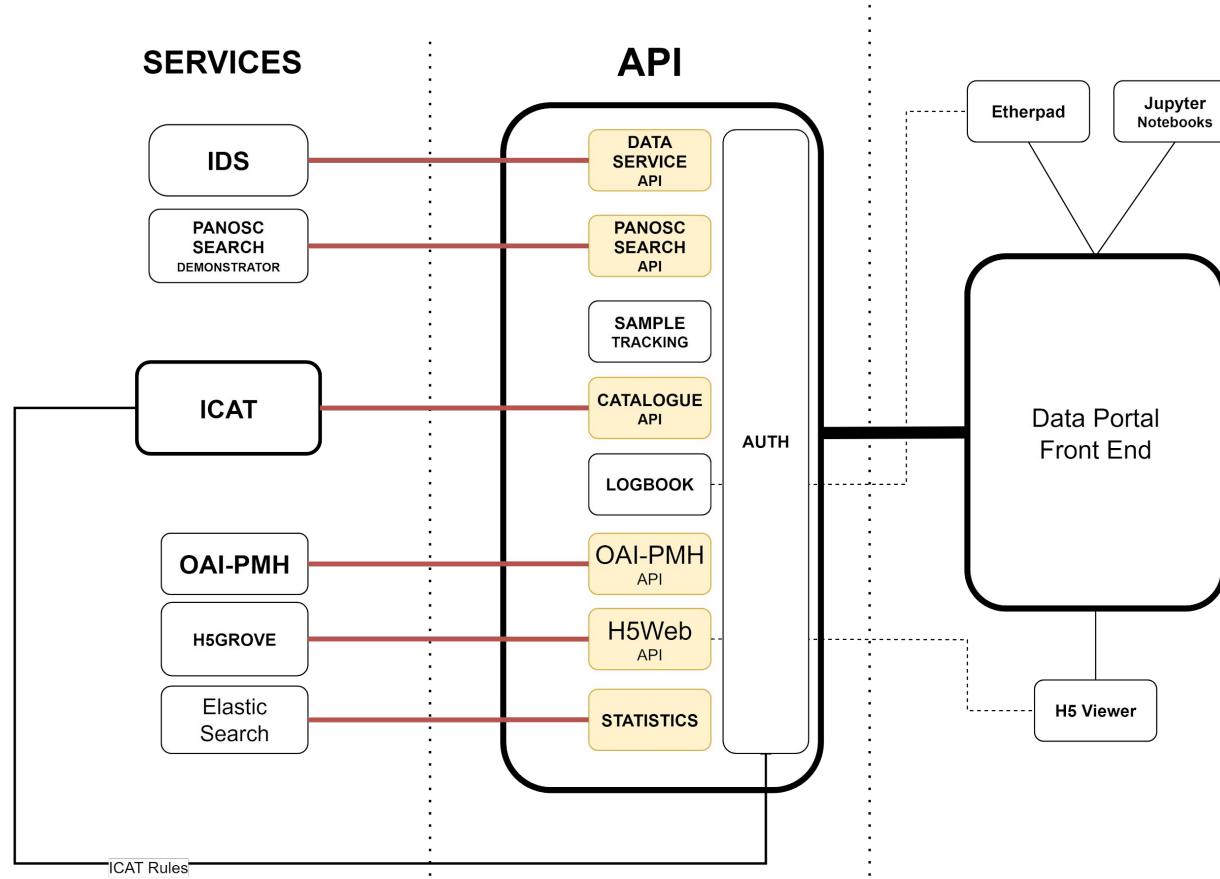


ISPyB/PY-ISPyB UI

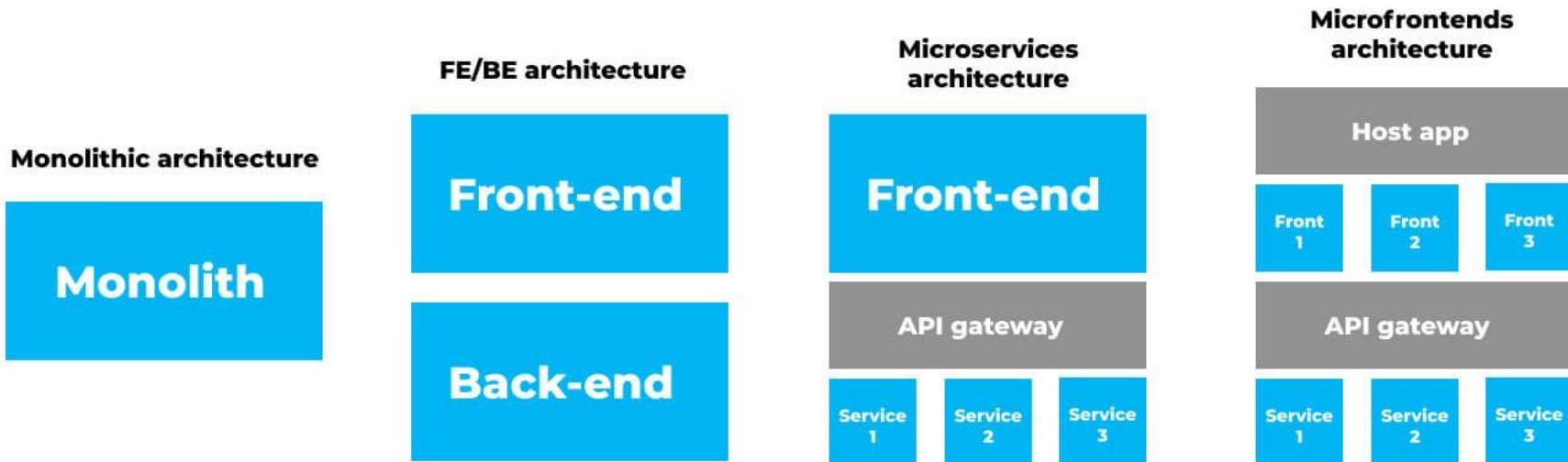
Data Portal

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?

Ingestion of the data

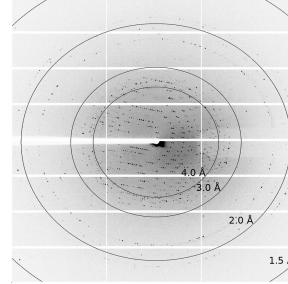
SAMPLE

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**

Ingestion of the data

SAMPLE

RAC1-Fus_RAC1b
Protein: HBLC
Acronym: HBLC-1
Space group: P1
...



DATASET

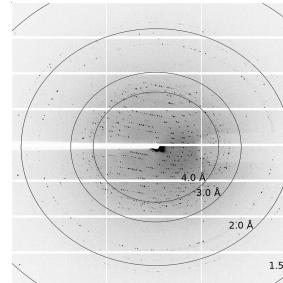
RAC1-Fus_RAC1b_1
Flux start: 1.00e+12 ph/sec
Flux end: 1.50e+12 ph/sec

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



Ingestion of the data

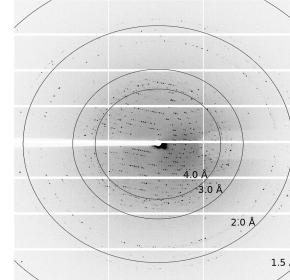
SAMPLE

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

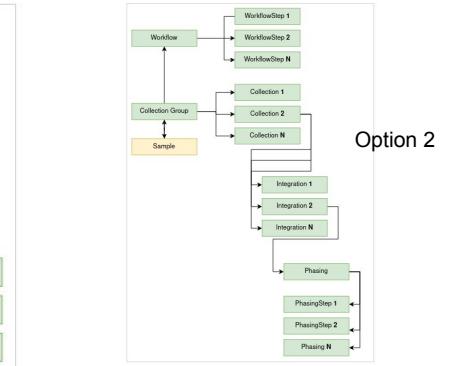
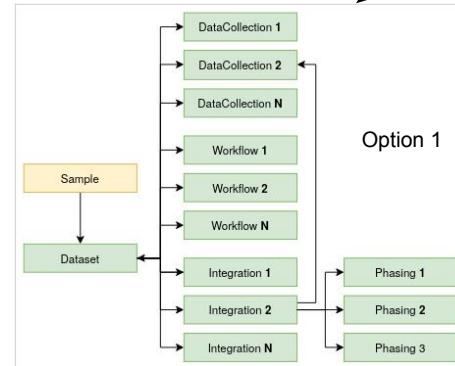
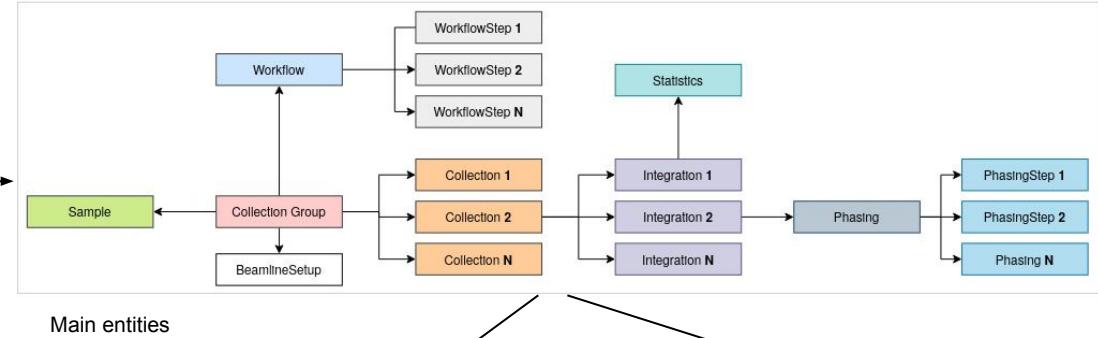
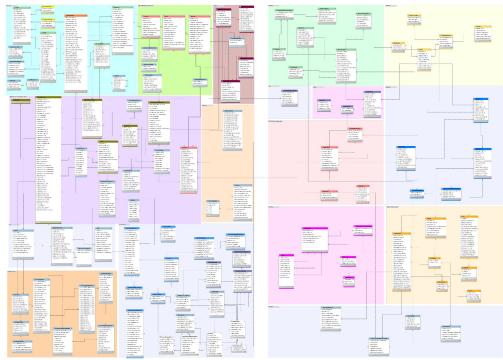
DATASET

SAD
Parameter1: ...
Parameter2: ...
....
/processed/x.ext
/processed/x1.ext
...
/processed/xn.ext

MR
....
/processed/x.ext
/processed/x1.ext
...
/processed/xn.ext

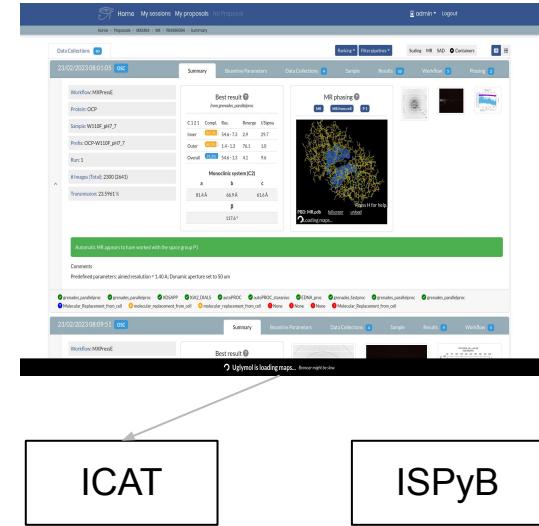
Ingestion of the data

- 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 visualization

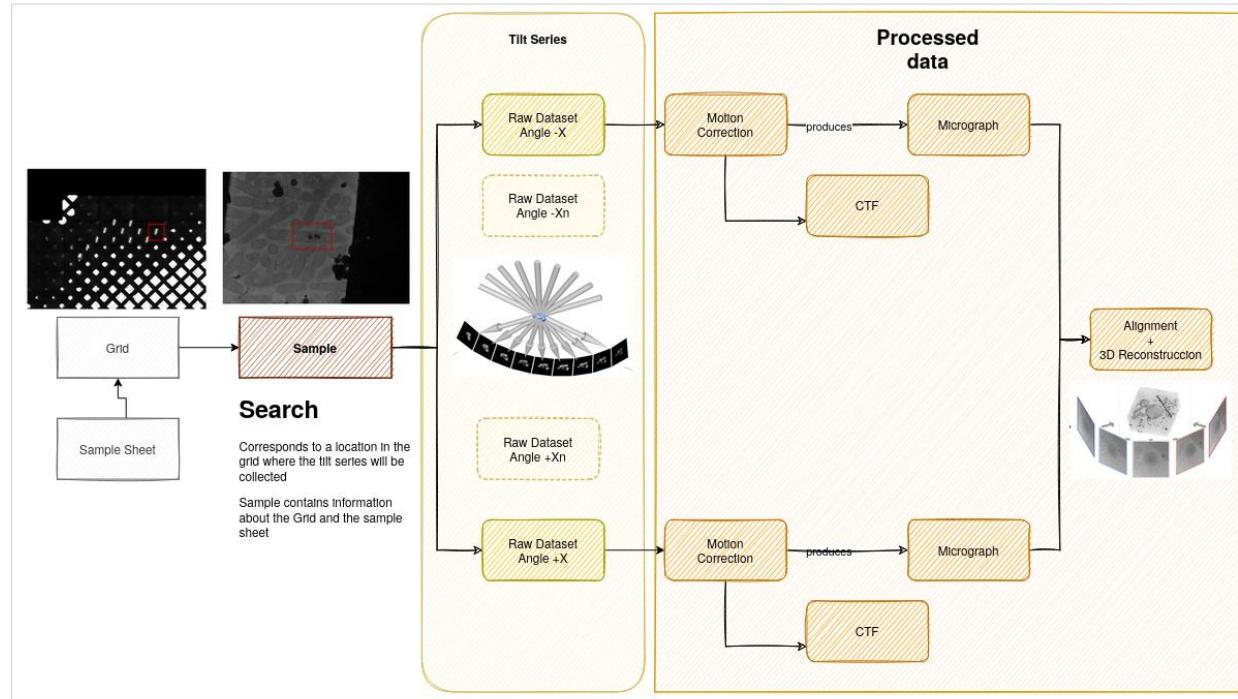
- Our approach was to directly connect py-ispyb-ui to ICAT
- The modification was done in less than 1 day
- Evaluation
 - Very happy with the results
 - **Out of the box** the performance is slightly poorer than ISPyB
 - Scientists and ourselves consider it is **acceptable**
- Mitigation
 - It is possible to get a better performance by using lazy loading
 - Improving the queries done to ICAT



- 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



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

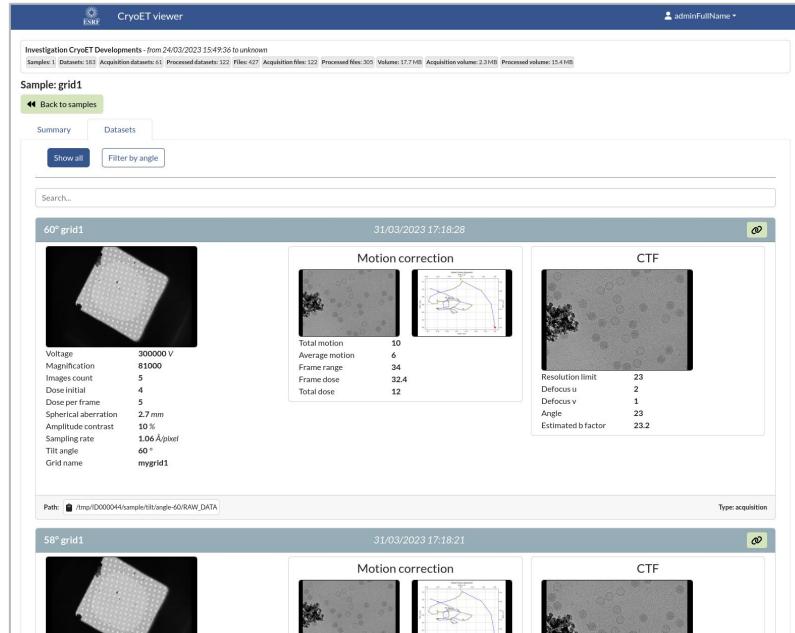
Metadata

Tilt Angle dataset	Motion Correction	CTF
sampleName	EMMotionCorrection_total_motion (NA)	EMCTF_resolution_limit (NA)
EM_voltage (V)	EMMotionCorrection_average_motion (NA)	EMCTF_correlation (NA)
EM_magnification (NA)	EMMotionCorrection_frame_range (NA)	EMCTF_defocus_u (NA)
EM_images_count (NA)	EMMotionCorrection_frame_dose (NA)	EMCTF_defocus_v (NA)
EM_position_x (NA)	EMMotionCorrection_total_dose (NA)	EMCTF_angle (NA)
EM_position_y (NA)		EMCTF_estimated_b_factor (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)		



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 ingestors
 - 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 logically 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