MX on ICAT*

Alex de Maria Antolinos
Data Automation Unit
Software Group
ESRF
30/11/2024

Overview

- **Goals**
- **Comparison of the ingestion of experimental data**
  - With ISPyB
  - With generic approach

- **Dataset and relationships**
- **Microservices architecture**
  - Sample Tracking
  - Search API
  - Reprocessing

- **Adding new features**
  - **Merging datasets from a multi data collection with different kappa angles**
  - **New technique**

Overview

- **Goals**
- **Comparison of the ingestion of experimental data**
  - With ISPyB
  - With generic approach

- **Dataset and relationships**
- ~~**Microservices architecture**~~
  - ~~Sample Tracking~~
  - ~~Search API~~
  - ~~Reprocessing~~

- ~~**Adding new features**~~
  - ~~**Merging datasets from a multi-data-collection with different kappa angles**~~
  - ~~**New technique**~~

Goals

Looking for:

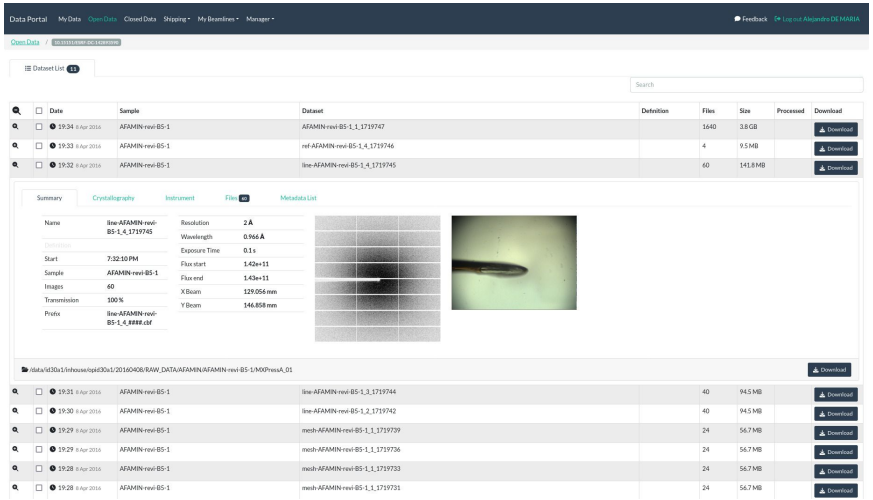
- **Sustainable in the long term**
- **Flexible**
  - Easier to adapt
  - Easier to extend
- Scalable
  - **New techniques**
    - Integrative structural biology
    - Others
  - **Short timescales**
- **Better data management**
  - No SPF
  - Easier to understand
  - Better organized
    - ++ Standardization

and:

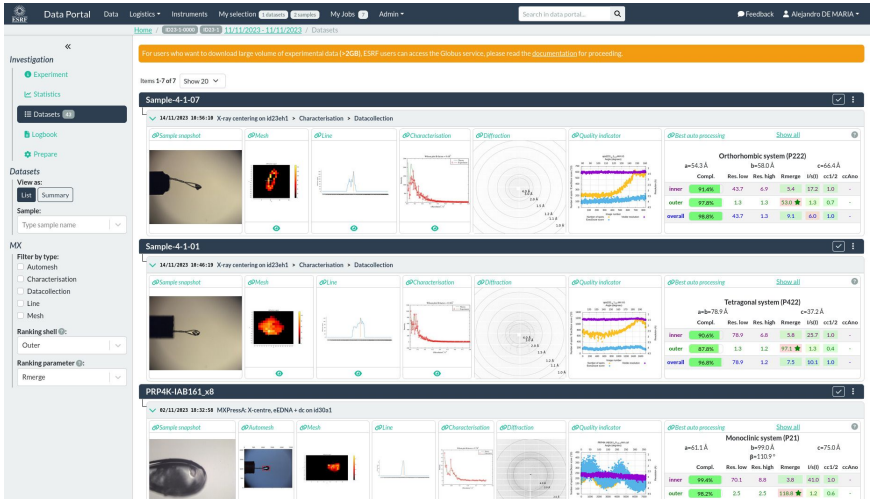
- **Modular** by design
  - Microservices
- **FAIR**
  - Ontologies
  - Public data and private data
  - Raw data and processed data
    - Tape interface
  - Logbook
- **Data publication**
    - DOI
    - PDB

Ingestion of experimental data

- **Work started early 2023**
- **Enrich the metadata catalog (ICAT) with MX processed data/metadata**
  - Define WHAT we want to store
  - The “How” already exists => Use existing software tools that are already in place
  - Development of the UI



Multiple raw dataset (data ingested in 2016)



Raw + Processed Datasets (2023)

WHAT

ISPyB Tables

ISPyB Columns

					DATA COLLECTION			
Table	Column	Is Used	description	Empty since	To be kept	parameter Name	New proposed name	Description
DCGroup	experimentType	X	SAD MAD SAD - Inverse Beam MAD - Inverse Beam OSC Helical Mesh Characterization EM Collect - Multiwedge					
	startTime	?	Needed?					
	endTime	?	Needed?					
	crystalClass			2017				
	comments							
	detectorMode		Unbinned, Software binned	2012				
	actualSampleBarcode		Not sure why is needed here					
	actualSampleSlotInContainer	X						
	actualContainerBarcode	?	Makes certain sense nevertheless	2019				
	actualContainerSlotInSC	X						
	xtalSnapshotFullPath							
DataCollection	dataCollectionNumber	X						
	startTime	X						
	endTime	X						
	runStatus	?	It is free text and are like comments					
	axisStart	X						
	axisEnd	X						
	axisRange	X						
	overlap	X				MX_oscillationOverlap		
	numberOfImages	X				MX_numberOfImages		
	startImageNumber	?	Do we need it?			MX_startImageNumber		
	numberOfPasses	?	Do we need it?					
	exposureTime	X				~~MX_exposureTime~~	Detactor_real_time	https://manual.nexusfo
	imageDirectory	?	I do not think we need it					
	imagePrefix	?	I do not think we need it					
	imageSuffix	?	I do not think we need it					
	imageContainerSubPath	?	I do not think we need it					
	fileTemplate	?	I do not think we need it					
	wavelength	X				InstrumentMonochromator_wavelength		
	resolution	X						

Proposed Metadata to replace ISPyB Tables/Columns | Strategy/Characte. | Sample | Data Collection | Processing | Experiment Statistics | Primitives

- **Mapping ISPyB metadata in shape of columns into metadata parameters (dataset parameters)**
- **Comparing what needs to be kept/removed/added**
- **Lot of help from scientists (Many thanks!!!)**

WHAT

MX

```
MX_aperture                    # Aperture size in microns
MX_beamShape                     # Beam shape at sample position
MX_beamSizeAtSampleX                # Horizontal beam size in mm at sample position
MX_beamSizeAtSampleY                # Vertical beam size in mm at sample position
MX_dataCollectionId              # ISPyB data collection id
MX_detectorDistance               # Detector to sample distance in mm
MX_directory                   # Data collection directory
MX_exposureTime                  # Exposure time in s
MX_flux                  # Flux in photon/s before data collection
MX_fluxEnd               # Flux in photon/s before data collection
MX_motors_name                   # Motor names
MX_motors_value                # Motor positions in mm
MX_numberOfImages                  # Number of images
MX_oscillationOverlap            # Oscillation overlap per image
MX_oscillationRange              # Oscillation range per image
MX_oscillationStart             # Oscillation start of data collection
MX_resolution                  # Resolution in A
MX_resolution_at_corner             # Resolution in A
MX_scanType                   # mxCuBE experiment type
MX_startImageNumber                # Data collection image start number
MX_template                  # Image file name template
MX_transmission                # Transmission in %
MX_wavelength                  # Wavelength in A
MX_xBeam                 # Horizontal beam centre in mm
MX_yBeam                 # Vertical beam centre in mm
MX_rotation_axis               # Name of the rotation axis
MX_axis_range                 # Axis range
MX_axis_start                # Rotation start angle
MX_axis_end                  # Rotation end angle
```

AutoprocIntegration

```
MXAutoprocIntegration_start_image_number       # First image number of the integration
MXAutoprocIntegration_end_image_number         # Last image number of the integration
MXAutoprocIntegration_detector_distance      # Refined detector distance
MXAutoprocIntegration_beam_x               # Refined beam x
MXAutoprocIntegration_beam_y               # Refined beam y
MXAutoprocIntegration_rotation_axis_x        # X position of the rotation axis
MXAutoprocIntegration_rotation_axis_y        # Y position of the rotation axis
MXAutoprocIntegration_rotation_axis_z        # Z position of the rotation axis
MXAutoprocIntegration_beam_vector_x          # Vector X
MXAutoprocIntegration_beam_vector_y          # Vector Y
MXAutoprocIntegration_beam_vector_z          # Vector Z
MXAutoprocIntegration_space_group           # Space group
MXAutoprocIntegration_cell_a             # cell a
MXAutoprocIntegration_cell_b              # cell b
MXAutoprocIntegration_cell_c              # cell c
MXAutoprocIntegration_cell_alpha           # cell alpha
MXAutoprocIntegration_cell_beta            # cell beta
MXAutoprocIntegration_cell_gamma             # cell gamma
MXAutoprocIntegration_anomalous            # anomalous
```

Scaling

```
MXAutoprocIntegrationScaling_overall_resolution_limit_low #
MXAutoprocIntegrationScaling_overall_resolution_limit_high #
MXAutoprocIntegrationScaling_overall_r_merge      #
MXAutoprocIntegrationScaling_overall_r_meas_within_IPlus_IMinus #
MXAutoprocIntegrationScaling_overall_r_meas_all_IPlus_IMinus #
MXAutoprocIntegrationScaling_overall_r_pim_within_IPlus_IMinus #
MXAutoprocIntegrationScaling_overall_r_pim_all_IPlus_IMinus #
MXAutoprocIntegrationScaling_overall_fractional_partial_bias #
MXAutoprocIntegrationScaling_overall_n_total_observations #
MXAutoprocIntegrationScaling_overall_n_total_unique_observations #
MXAutoprocIntegrationScaling_overall_mean_I_over_sigI #
MXAutoprocIntegrationScaling_overall_completeness #
MXAutoprocIntegrationScaling_overall_multiplicity #
MXAutoprocIntegrationScaling_overall_anomalous_completeness #
MXAutoprocIntegrationScaling_overall_anomalous_multiplicity #
MXAutoprocIntegrationScaling_overall_anomalous      #
MXAutoprocIntegrationScaling_overall_cc_half      #
MXAutoprocIntegrationScaling_overall_ccAno      #
MXAutoprocIntegrationScaling_overall_sigAno      #
MXAutoprocIntegrationScaling_overall_isa      #
MXAutoprocIntegrationScaling_overall_completeness_spherical #
MXAutoprocIntegrationScaling_overall_completeness_ellipsoidal #
MXAutoprocIntegrationScaling_overall_anomalous_completeness_spherical #
MXAutoprocIntegrationScaling_overall_anomalous_completeness_ellipsoidal #
MXAutoprocIntegrationScaling_inner_resolution_limit_low #
MXAutoprocIntegrationScaling_inner_resolution_limit_high #
MXAutoprocIntegrationScaling_inner_r_merge      #
MXAutoprocIntegrationScaling_inner_r_meas_within_IPlus_IMinus #
MXAutoprocIntegrationScaling_inner_r_meas_all_IPlus_IMinus #
MXAutoprocIntegrationScaling_inner_r_pim_within_IPlus_IMinus #
MXAutoprocIntegrationScaling_inner_r_pim_all_IPlus_IMinus #
MXAutoprocIntegrationScaling_inner_fractional_partial_bias #
MXAutoprocIntegrationScaling_inner_n_total_observations #
MXAutoprocIntegrationScaling_inner_n_total_unique_observations #
MXAutoprocIntegrationScaling_inner_mean_I_over_sigI #
MXAutoprocIntegrationScaling_inner_completeness    #
MXAutoprocIntegrationScaling_inner_multiplicity    #
MXAutoprocIntegrationScaling_inner_anomalous_completeness #
MXAutoprocIntegrationScaling_inner_anomalous_multiplicity #
MXAutoprocIntegrationScaling_inner_anomalous      #
MXAutoprocIntegrationScaling_inner_cc_half      #
MXAutoprocIntegrationScaling_inner_ccAno      #
MXAutoprocIntegrationScaling_inner_sigAno      #
MXAutoprocIntegrationScaling_inner_isa      #
MXAutoprocIntegrationScaling_inner_completeness_spherical #
MXAutoprocIntegrationScaling_inner_completeness_ellipsoidal #
MXAutoprocIntegrationScaling_inner_anomalous_completeness_spherical #
MXAutoprocIntegrationScaling_inner_anomalous_completeness_ellipsoidal #
MXAutoprocIntegrationScaling_outer_resolution_limit_low #
MXAutoprocIntegrationScaling_outer_resolution_limit_high #
MXAutoprocIntegrationScaling_outer_r_merge      #
MXAutoprocIntegrationScaling_outer_r_meas_within_IPlus_IMinus #
MXAutoprocIntegrationScaling_outer_r_meas_all_IPlus_IMinus #
MXAutoprocIntegrationScaling_outer_r_pim_within_IPlus_IMinus #
MXAutoprocIntegrationScaling_outer_r_pim_all_IPlus_IMinus #
MXAutoprocIntegrationScaling_outer_fractional_partial_bias #
MXAutoprocIntegrationScaling_outer_n_total_observations #
MXAutoprocIntegrationScaling_outer_n_total_unique_observations #
MXAutoprocIntegrationScaling_outer_mean_I_over_sigI #
MXAutoprocIntegrationScaling_outer_completeness    #
MXAutoprocIntegrationScaling_outer_multiplicity    #
MXAutoprocIntegrationScaling_outer_anomalous_completeness #
MXAutoprocIntegrationScaling_outer_anomalous_multiplicity #
MXAutoprocIntegrationScaling_outer_anomalous      #
MXAutoprocIntegrationScaling_outer_cc_half      #
MXAutoprocIntegrationScaling_outer_ccAno      #
MXAutoprocIntegrationScaling_outer_sigAno      #
MXAutoprocIntegrationScaling_outer_isa      #
MXAutoprocIntegrationScaling_outer_completeness_spherical #
MXAutoprocIntegrationScaling_outer_completeness_ellipsoidal #
MXAutoprocIntegrationScaling_outer_anomalous_completeness_spherical #
MXAutoprocIntegrationScaling_outer_anomalous_completeness_ellipsoidal #
```

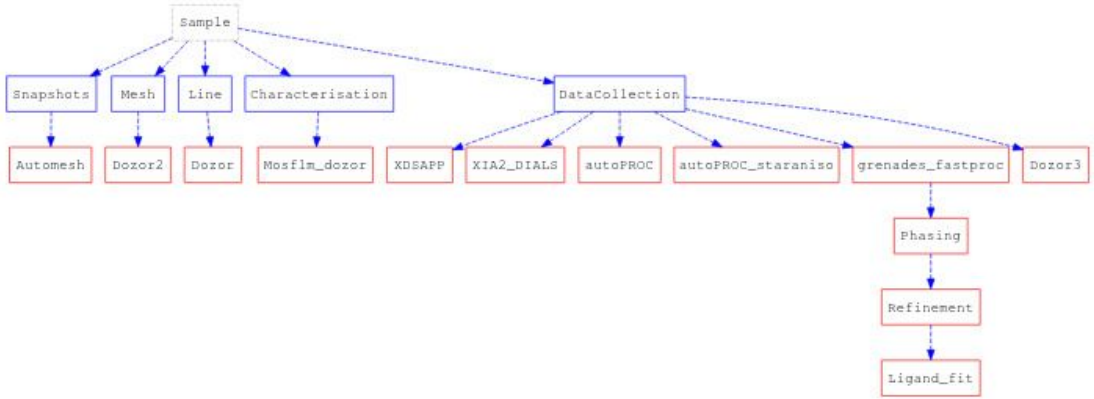
- List of metadata applicable to datasets
- For fully description of currently available parameters hdf5_config.xml

Credits: Olof S.

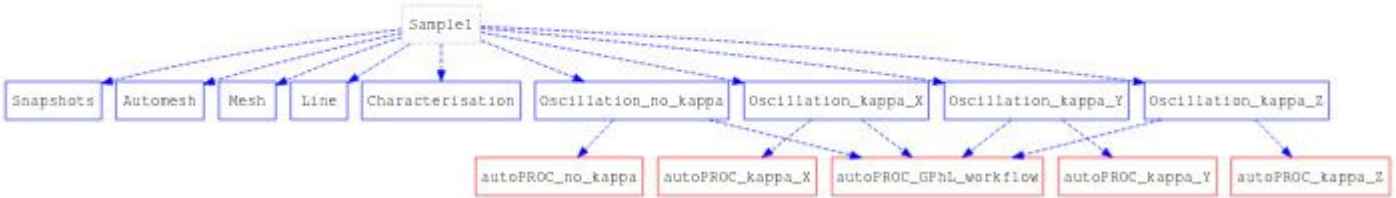
WHAT

Documented in hackmd

Workflow MxPress-E



Multiple oscillations with different kappa angles - Olof's version



ISPyB Web services and Database

MxCube

BsxCube

EMCube

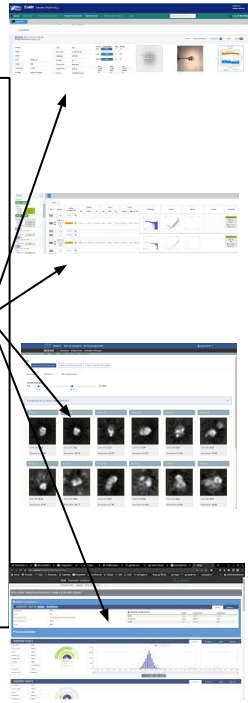
SOAP Web services

DATA IN

ISPyB
Database

REST Web services

DATA OUT

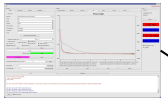


ISPyB

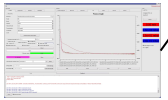
- ISPyB provides the functionality to ingested/read data via web services

ISPyB Web services and Database

MxCube

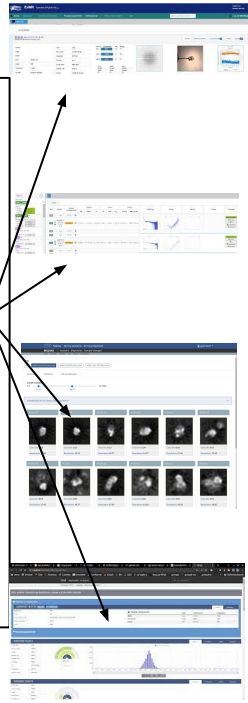


BsxCube



EMCube

?

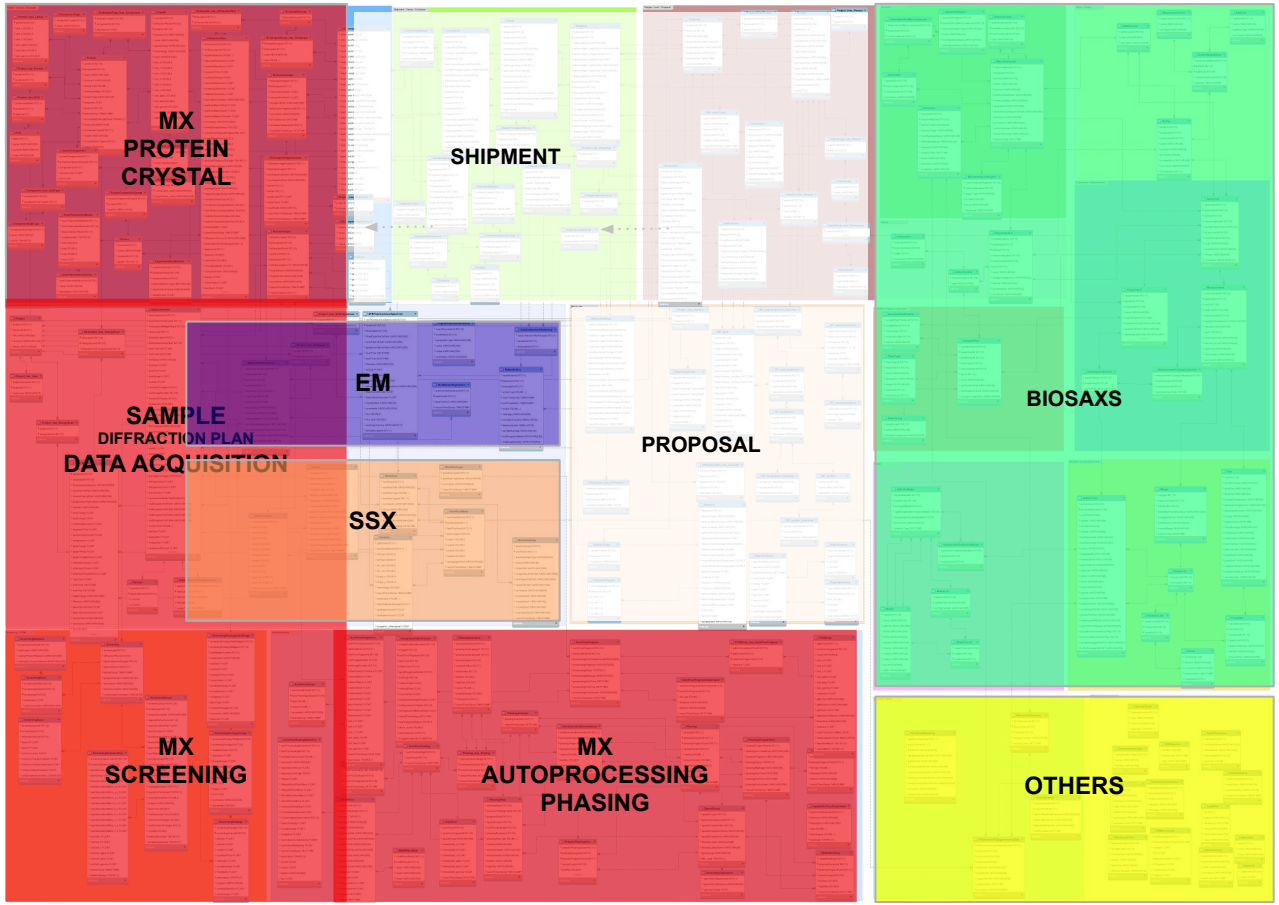


Can a ICAT's based solution might provide the same functionality and UX?

ISPyB Data Model

210 tables
40 Views

ISPyB Data Model



CRYSTALLOGRAPHY

BIOSAXS

EM

SSX

COMMON TABLES

OTHERS

210 tables
40 Views

ISPyB Web services and Database

SOAP Web services

MX
storeOrUpdateDataCollection()
storeOrUpdateAutoProcScalingStatistics()
storeOrUpdateDataCollectionGroup()
103 more…

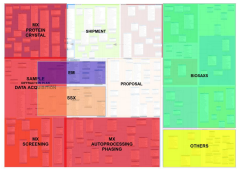
SAXS
createEmptyExperiment()
appendMeasurementToExperiment()
addAveraged()
46 more…

EM
addMovie()
addMotionCorrection()
addCTF()
5 more…

SSX

DATA IN

ISPyB
Database



- The set of webmethods to be called depends on the technique

ISPyB Web services and Database

SOAP Web services

MX
storeOrUpdateDataCollection()
storeOrUpdateAutoProcScalingStatistics()
storeOrUpdateDataCollectionGroup()
103 more…

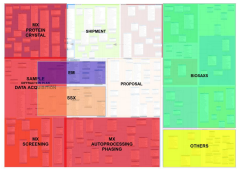
SAXS
createEmptyExperiment()
appendMeasurementToExperiment()
addAveraged()
46 more…

EM
addMovie()
addMotionCorrection()
addCTF()
5 more…

SSX

DATA IN

ISPyB Database



REST Web services

MX
/proposal/{proposal}/mx/session
/proposal/{proposal}/mx/crystal
94 more…

SAXS
/{proposal}/saxs/macromolecule
/{proposal}/saxs/buffer
68 more…

EM
/{proposal}/em/movie
/{proposal}/em/ctf
/{proposal}/em/motion
12 more…

SSX

DATA OUT

ISPyB Web services and Database

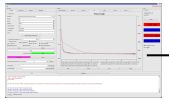
SOAP Web services

MxCube

MX

storeOrUpdateDataCollection()
storeOrUpdateAutoProcScalingStatistics()
storeOrUpdateDataCollectionGroup()

103 more…

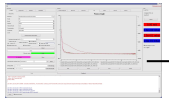


BsxCube

SAXS

createEmptyExperiment()
appendMeasurementToExperiment()
addAveraged()

46 more…



EMCube

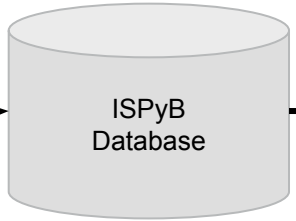
EM

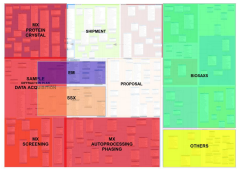
addMovie()
addMotionCorrection()
addCTF()

5 more…

MxCube

SSX





REST Web services

MX

/proposal/{proposal}/mx/session
/proposal/{proposal}/mx/crystal

94 more…

SAXS

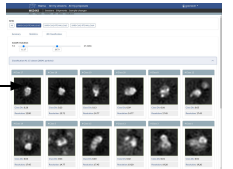
/{proposal}/saxs/macromolecule
/{proposal}/saxs/buffer

68 more…

EM

/{proposal}/em/movie
/{proposal}/em/ctf
/{proposal}/em/motion

12 more…

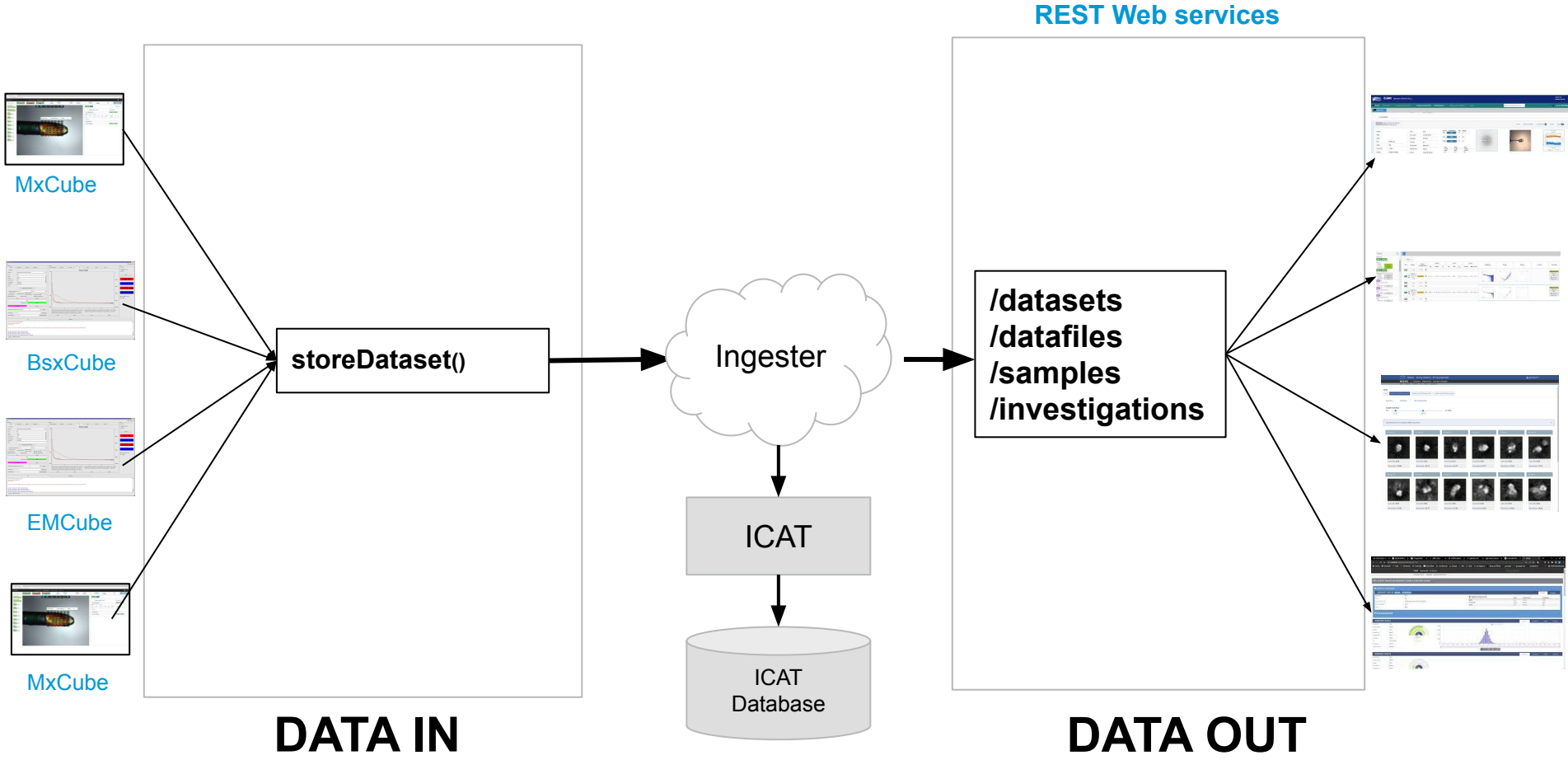


SSX

DATA IN

DATA OUT

Ingestion and exposing data via generic approach



What is a dataset?

A dataset is a data structure composed by:

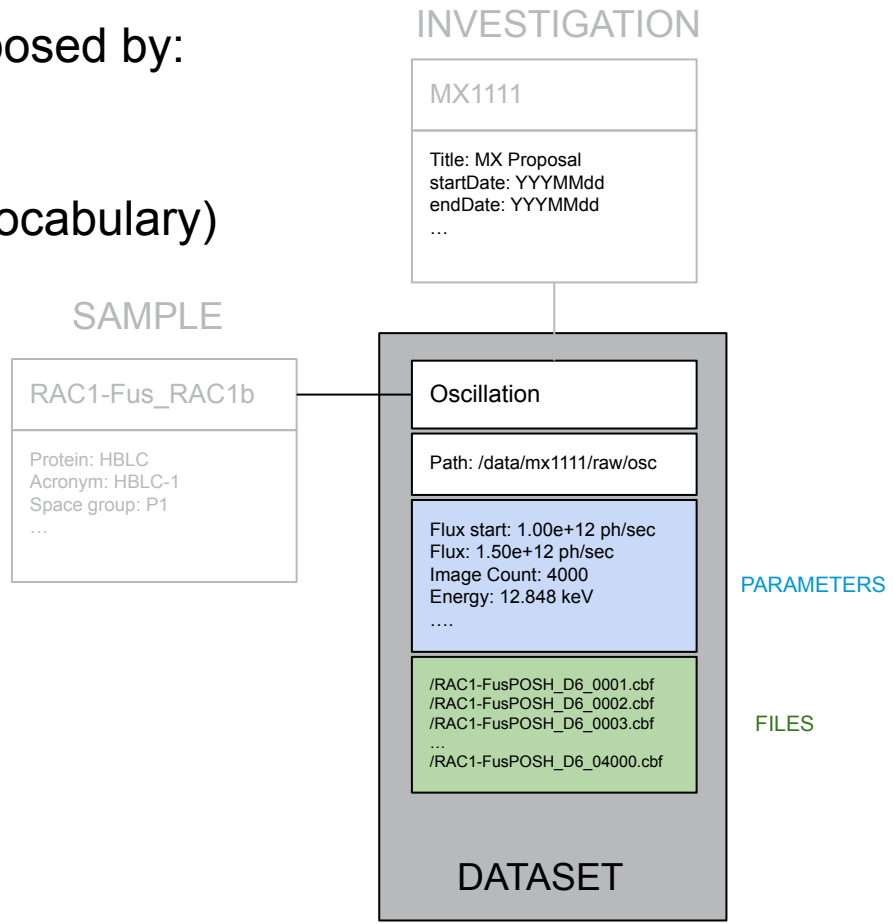
- Name
- Path
- Parameters (controlled vocabulary)
- Files

And linked to:

- Investigation
- Sample

Types:

- Raw
- Processed



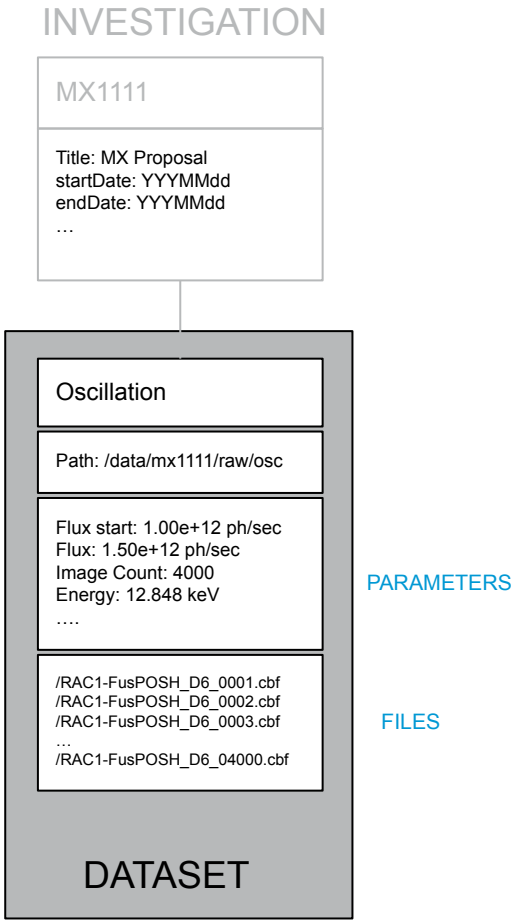
How is a dataset ingested?

- Generic signature for all datasets/techniques

```
storeDataset(datasetName,proposalName, sampleIdentifier, path, parameters)
```

- Example

```
storeDataset(“oscillation”, “MX1111”, “Fus_RAC1b”, “/data/mx1111/raw/osc",
        {
                Definition: MX,
                scanType : oscillation
                Flux start : 1.00e+12
                Image Count : 4000
                Energy :  12.848
                ….})
```



Linking datasets

- Information (datasets) need to be linked together
- Done in ISPyB through relations between tables

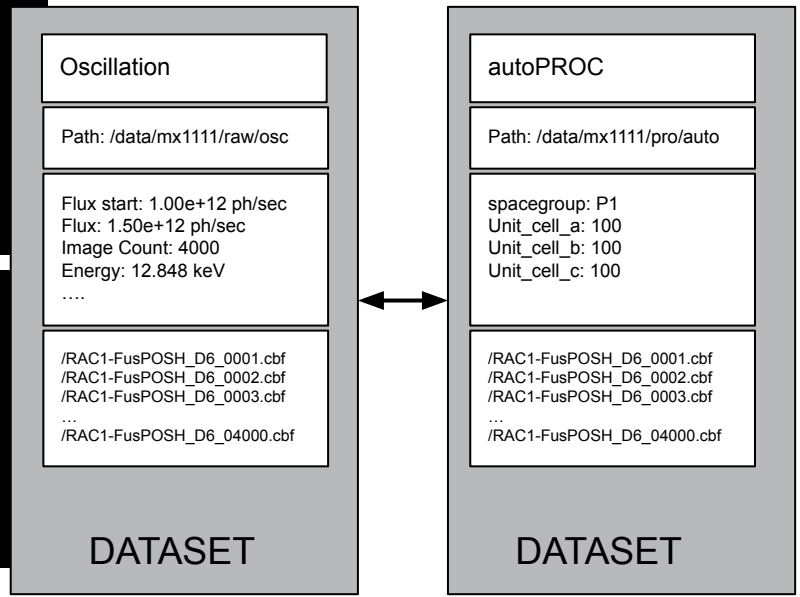
How is a dataset can be linked?

- Datasets are linked dynamically via metadata
  - 1 dataset = 1 folder => (dataset path === identifier of a dataset)
  - The input parameter (is a list) and allows to link to multiple datasets

- Example of storing autoPROC linked to OSC

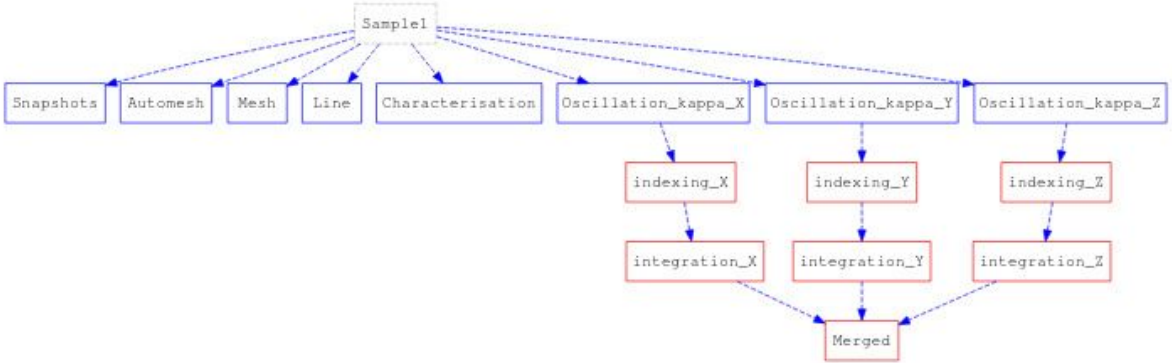
```
storeDataset(“oscillation”, “MX1111”, “Fus_RAC1b”, “/data/mx1111/raw/osc",
        {
                Definition: MX,
                scanType : oscillation
                Flux start : 1.00e+12
                Image Count : 4000
                Energy :  12.848
                ….})
```

```
storeDataset(“autoPROC”, “MX1111”, “Fus_RAC1b”, “/data/mx1111/pro/auto",
        {
                Definition: MX,
                scanType : integration
                spacegroup : P1
                Unit_cell_a : 100
        }
        input=[/data/mx1111/raw/osc])
```



Dynamic linking of datasets

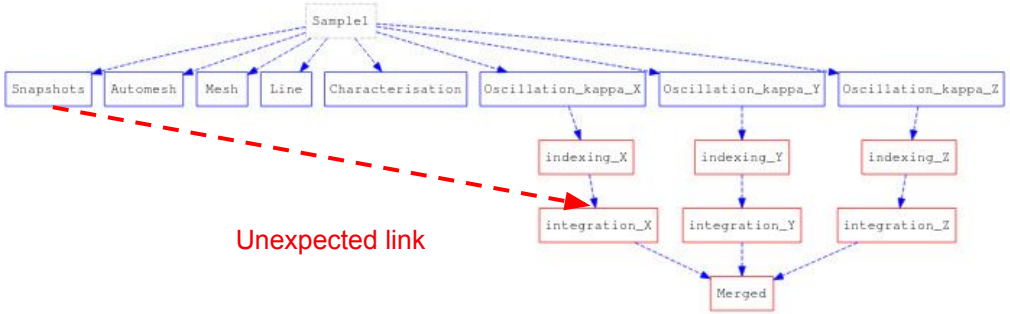
- Powerful and flexible way to link entities
- Changes in the relationship between datasets does not need changes in the backend



BUT currently:

- There is no any formal description of the dataset parameters and relationships
- Relationships are not enforced anymore
- A very high standardization is required if used multi-site

Possible Mitigations



- The risk might be considered low because ingestion of the data is done by data acquisition software (or processing pipelines)
  - Code rarely changes
  - No users allowed
  - Fixed rapidly
  - The risk already exists in ISPyB (snapshot = Datacollection -> AutoProcIntegration)
- In order to mitigate them several approaches can be envisaged:
  - High level API (on top of “storeDataset”)
  - Checking mechanism for metadata and relationships before/after the ingestion
  - Description of the metadata and relationships with a standard format (mmCIF?)
  - Others… (?)

Conclusions

- **A more generic approach**
  - simplifies the software making a huge impact on development and maintenance
  - can facilitate the progress of developments that would be challenging to execute using the current ISPyB
  - allows for the incorporation of new techniques in a seamless manner
  - can make easier to application developers to accommodate theirs needs on both metadata and UI
  - federates resources that otherwise would be jeopardized by standalone technique specific developments
  - removes duplication of efforts

- **But more work needs to be done because it**
  - requires a high standardization of metadata
  - and a mechanism to ensure consistency

Final words

As developer:

My personal opinion, along with the feedback I've received so far from people working in various areas of development (MxCube, processing, UI) with experience in both ISPyB and ICAT, has been very positive.

My perception is that scientists are pleased with the output achieved in so little time, which, in any case, can be considered a final product but a starting point.

The recent work needs to be consolidated; however, more experience and feedback from users are needed.

ACKNOWLEDGEMENTS

- **ISPyB Collaborators for your constructive feedback**
- **STFC ICAT Developers**

- **Mael Gaonach**
- **Marjolain Bodin**
- **Olof Svensson**
- **Marcus Oscarsson**
- **Andy Gotz**

- **Max Nanao**
- **Romain Talon**
- **Matthew Bowler**
- **Didier Nurizzo**
- **Estelle Mossou**

ACKNOWLEDGEMENTS

- **Thanks to the organizers of the ISPyB Meeting@ALBA!!**