Processed Data

Alex de Maria
Software Engineer
Data Manager@Data Automation Unit
Berlin 03/05/2023

Index

- Motivation
- Use cases
- Linking datasets in ICAT
- Implementation
  - Ingestion
  - Metadata
- Current status
- Example CryoET:
  - Dataset model
  - Metadata
  - UI
- Example MX
- Implications
- Conclusions

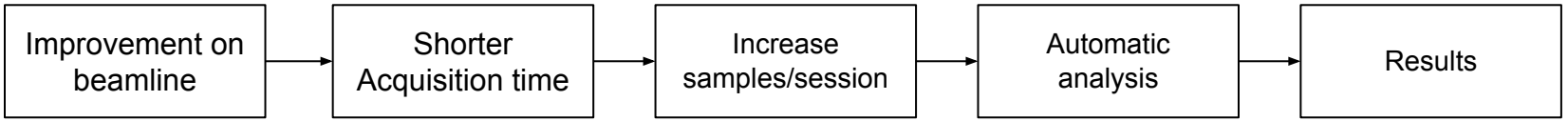
Motivation

- Data policy in process to be extended to include **processed data**
- The data portal is used to access to the online catalogue where data and metadata is available
- Processed data has been added to the catalogue **under demand**
  - Example: DOI needed for publication
- Experiment folder is been reorganized to better integrate the processed data in all beamlines
  - RAW_DATA and PROCESSED_DATA in all beamlines
- Beamlines are requesting processed data to be managed with the possibility of having a more or less sophisticated way to display the data
- We expect this demand to increase when:
  - Large number of sample to be processed
  - Automatic pipelines in place (ex: with ewoks)

Motivation

- **Standardization**
  - **beamlines:** standard procedures for acquisition and analysis = automation
  - **facility:** data is handled in a homogeneous way. Example:
    - same folder structure for each experiment
    - 1 dataset = 1 folder

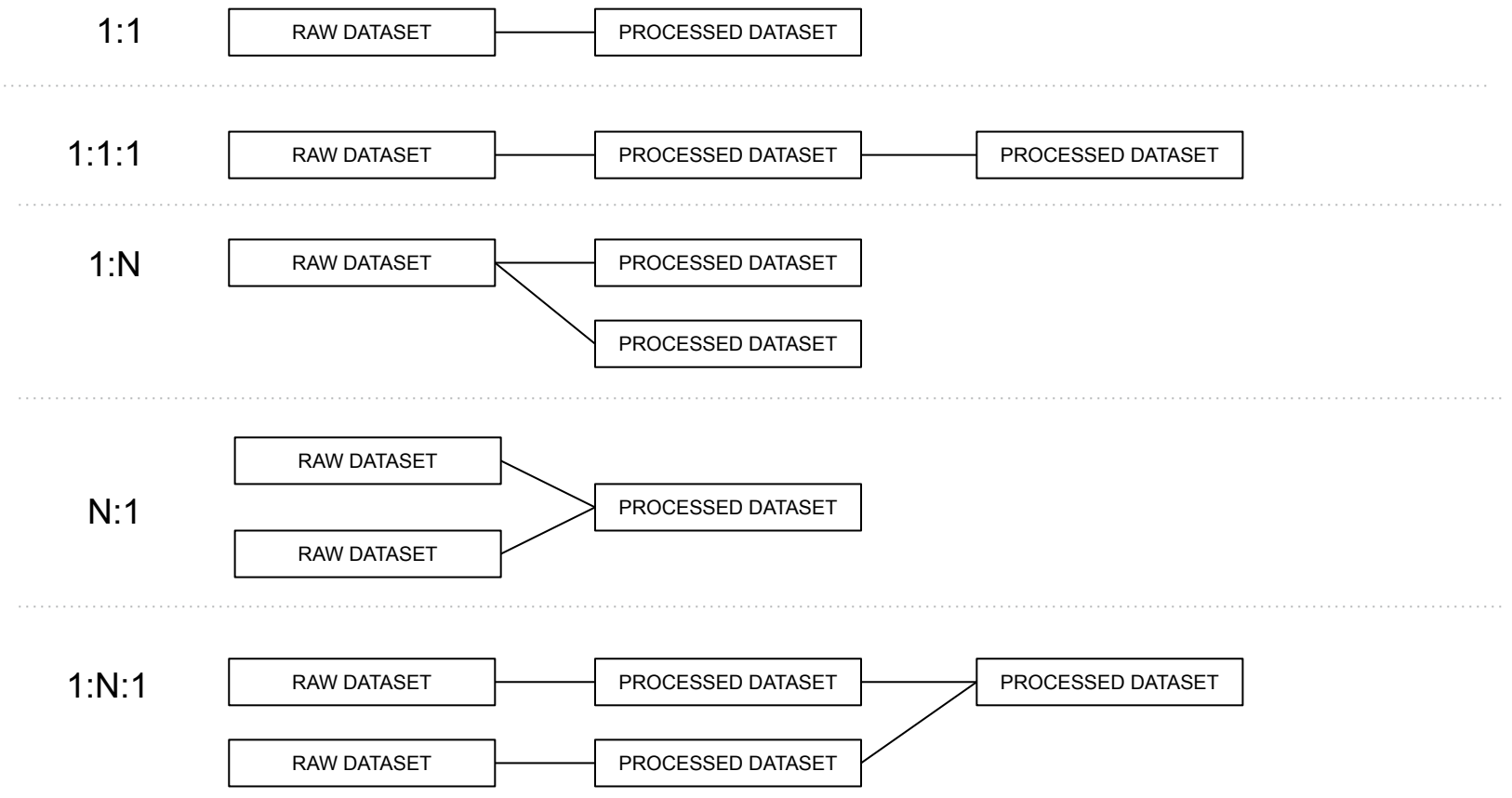
- **Automation**



Example:

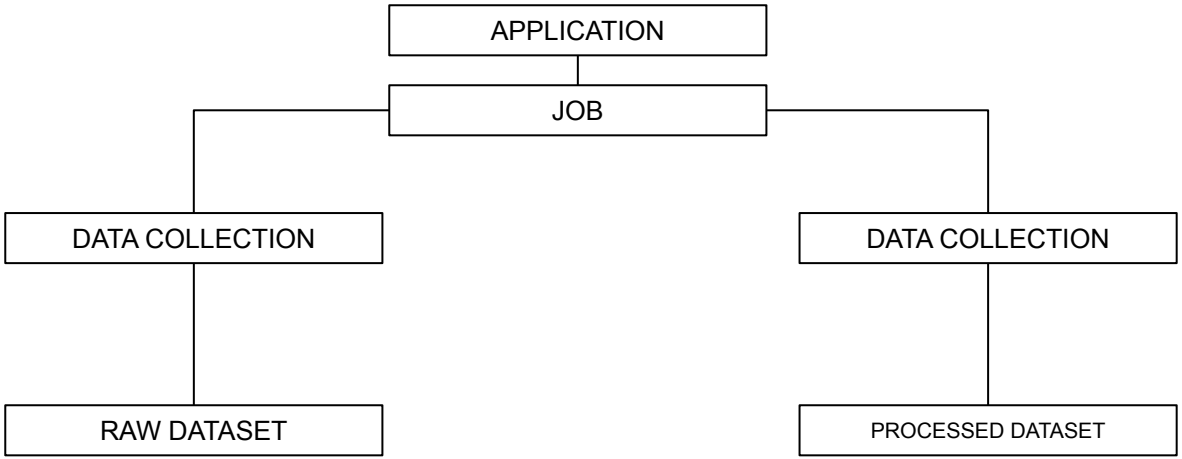
Faster detectors
Sample changers
Fully automatic beamlines

Use cases

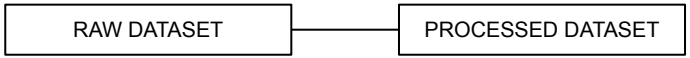

1:1
RAW DATASET
PROCESSED DATASET
1:1:1
RAW DATASET
PROCESSED DATASET
PROCESSED DATASET
1:N
RAW DATASET
PROCESSED DATASET
PROCESSED DATASET
N:1
RAW DATASET
RAW DATASET
PROCESSED DATASET
1:N:1
RAW DATASET
PROCESSED DATASET
PROCESSED DATASET
RAW DATASET
PROCESSED DATASET

How to link two datasets in ICAT?

1) **With <u>JOB</u> table**
   a) Consistency and standard
   b) Needs changes everywhere
   c) Performance –

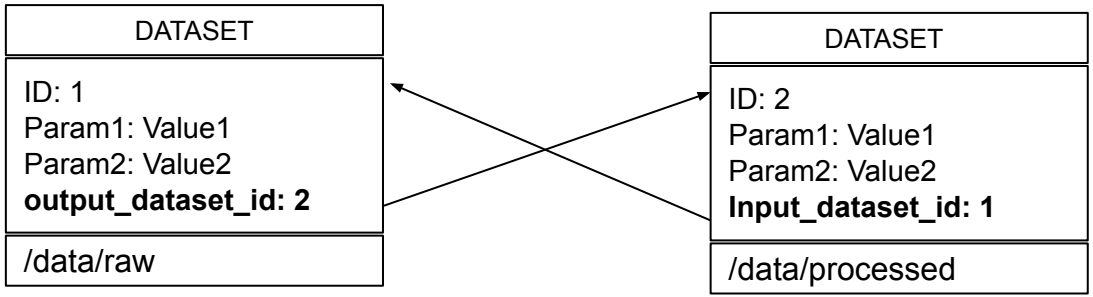


How to link two datasets in ICAT?



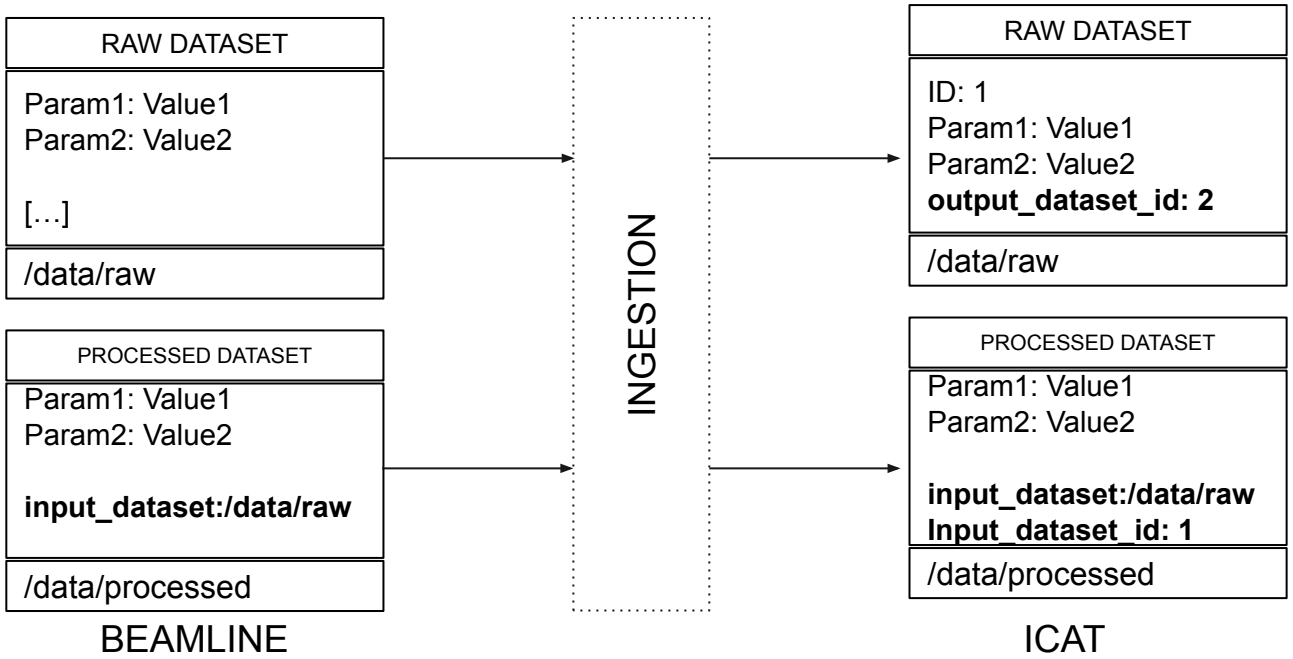
2) Soft links

a) Consistency is not ensured
b) Easy to implement (depending on your use case, e.g: no deletion of datasets is allowed)
c) Performance ++
d) Job and application parameters are stored as dataset parameters (see denormalization later)



Implementation: ingestion

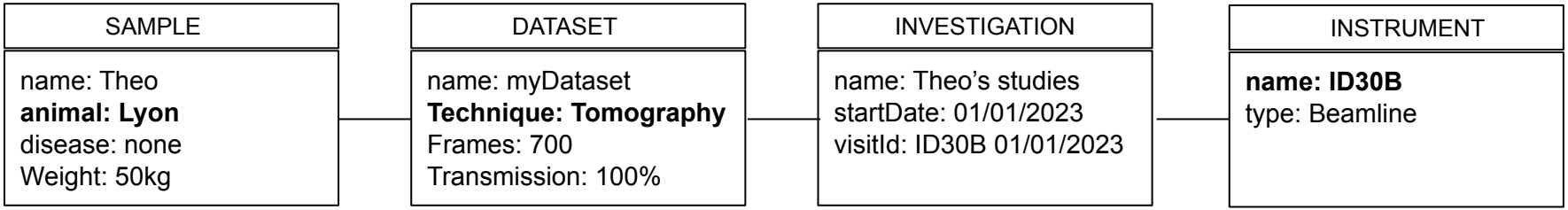
- **Implementation based on soft links**
  - Processed datasets are identified via parameters and linked based on the location
  - Instruments do not know anything about ICAT
  - Ingester enriches the metadata with the output/input ids. Pitfall: processed might be ingested before raw



Denormalization of dataset parameters

Denormalization is a strategy used on a previously-normalized database to increase performance. In computing, denormalization is the process of trying to improve the read performance of a database, at the expense of losing some write performance, by adding redundant copies of data or by grouping data.[1][2] It is often motivated by performance or scalability in relational database software needing to carry out very large numbers of read operations. Denormalization differs from the unnormalized form in that denormalization benefits can only be fully realized on a data model that is otherwise normalized.

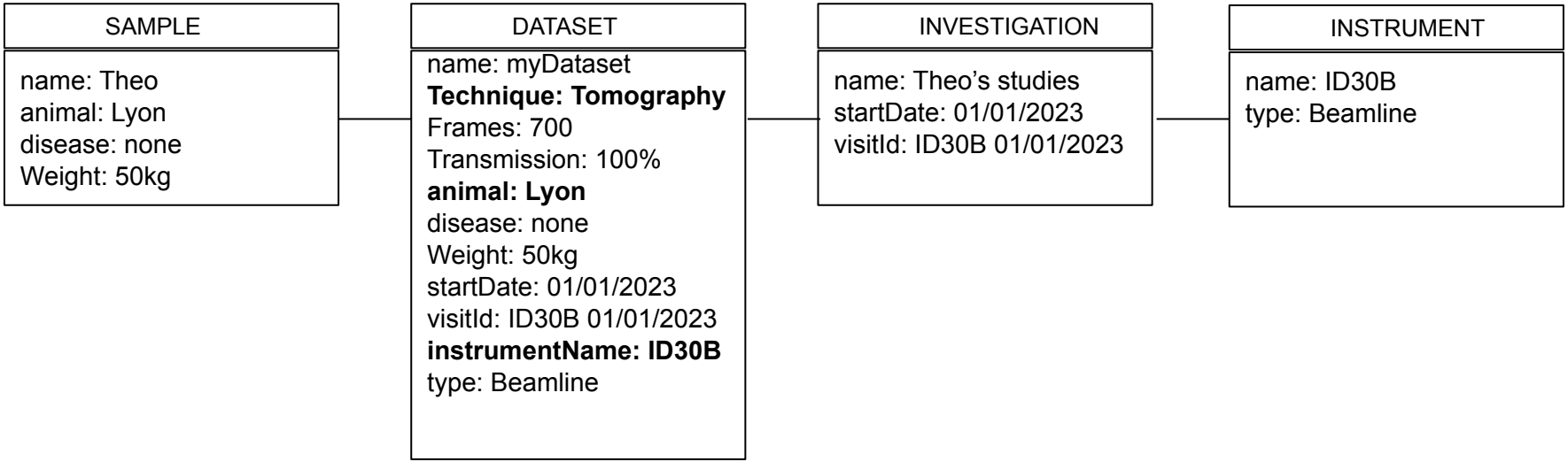
Example: normalized



SELECT dataset FROM Dataset
JOIN Sample **JOIN** Investigation **JOIN** Instrument
WHERE Sample.animal = 'Lyon' AND Instrument.Name = 'ID30B' and Data.technique='Tomography'

Denormalization of dataset parameters

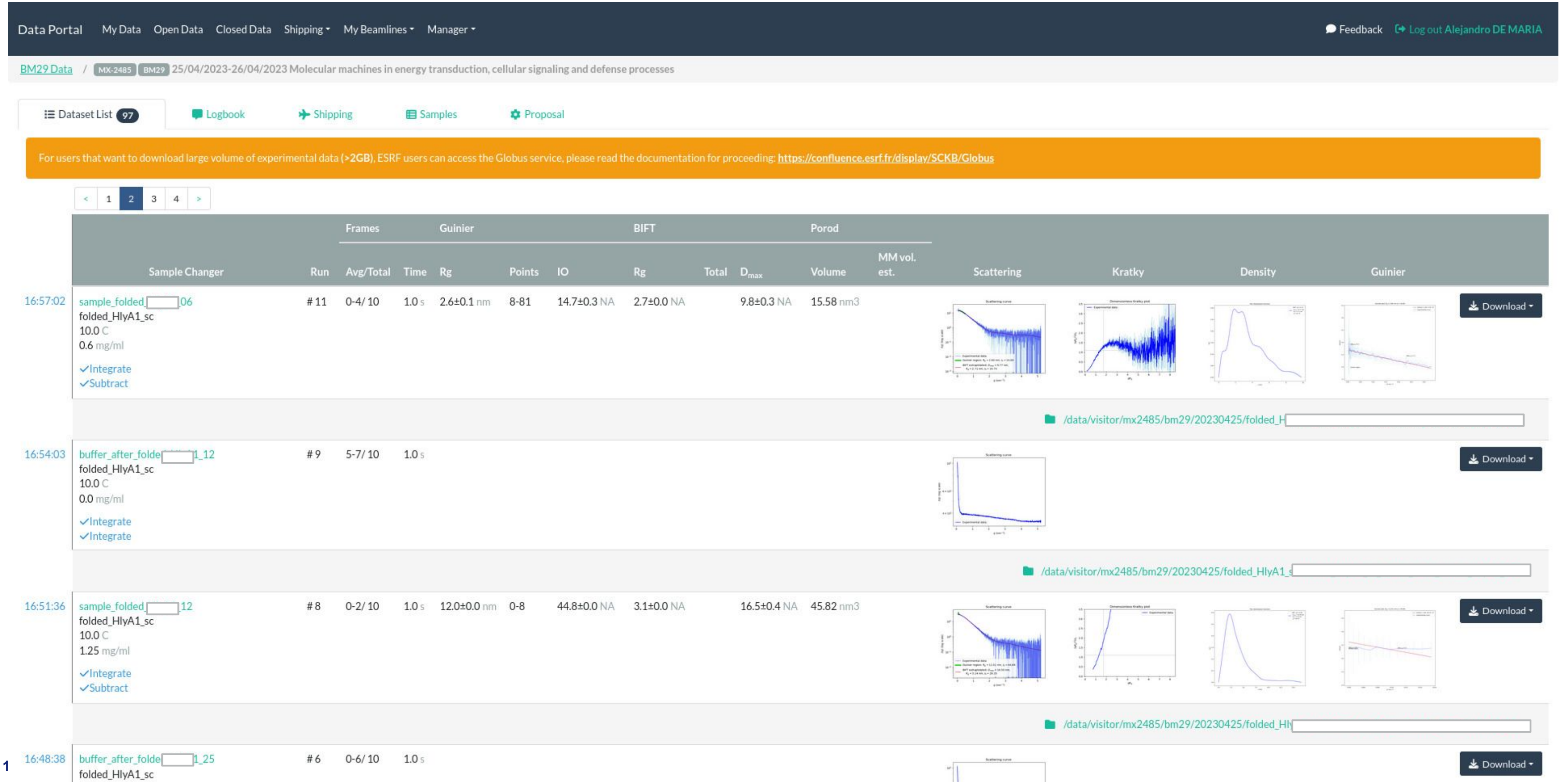
Example: denormalized



SELECT dataset FROM Dataset WHERE animal = 'Lyon' AND name = 'ID30B' and technique='Tomography'

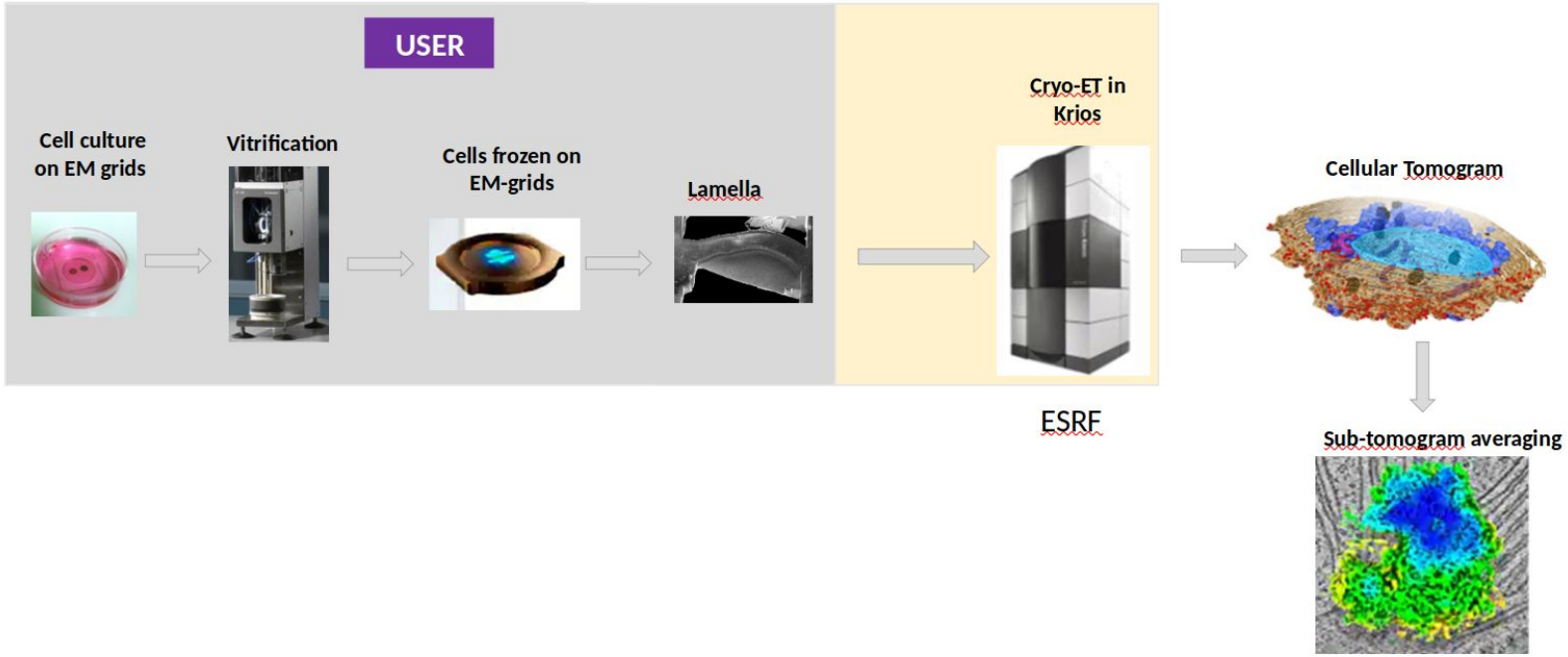
Current status

- Two techniques were implemented as proof of concept with <u>embedded viewers</u>

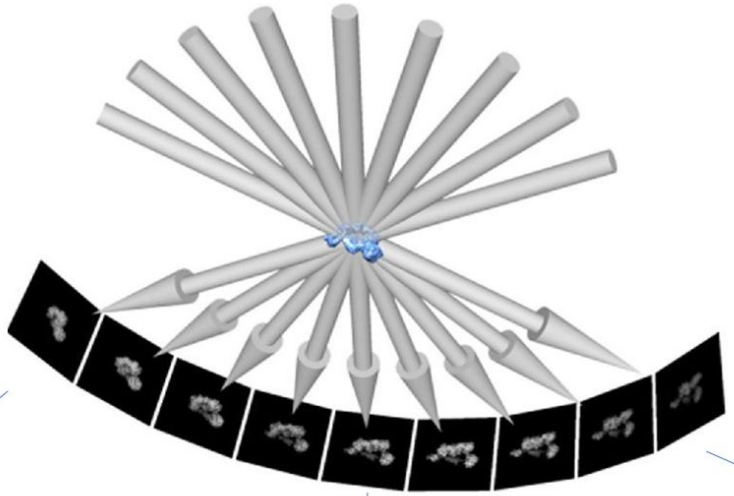


Example CryoET

Tomography Workflow

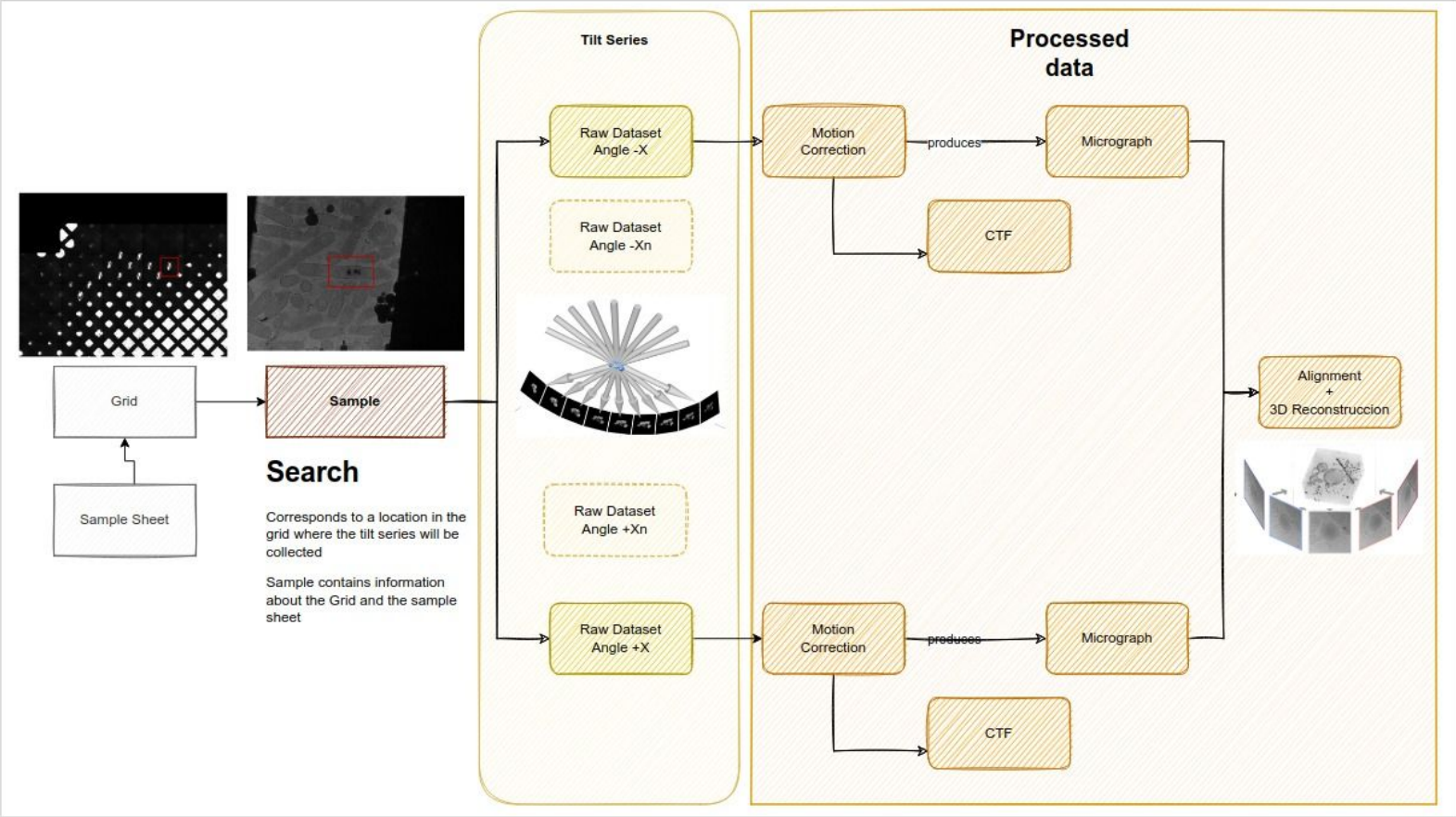


Example CryoET



	Movie tilt angle -60° 5 frames	Movie at tilt angle 0° 5 frames	Movie at tilt angle +60° 5 frames
Essential	Motion correction	Motion correction	Motion correction
Could be useful	CTF estimation	CTF estimation	CTF estimation
RESULT 1	Micrograph at tilt angle -60°	Micrograph at tilt angle 0°	Micrograph at tilt angle +60°

Example CryoET



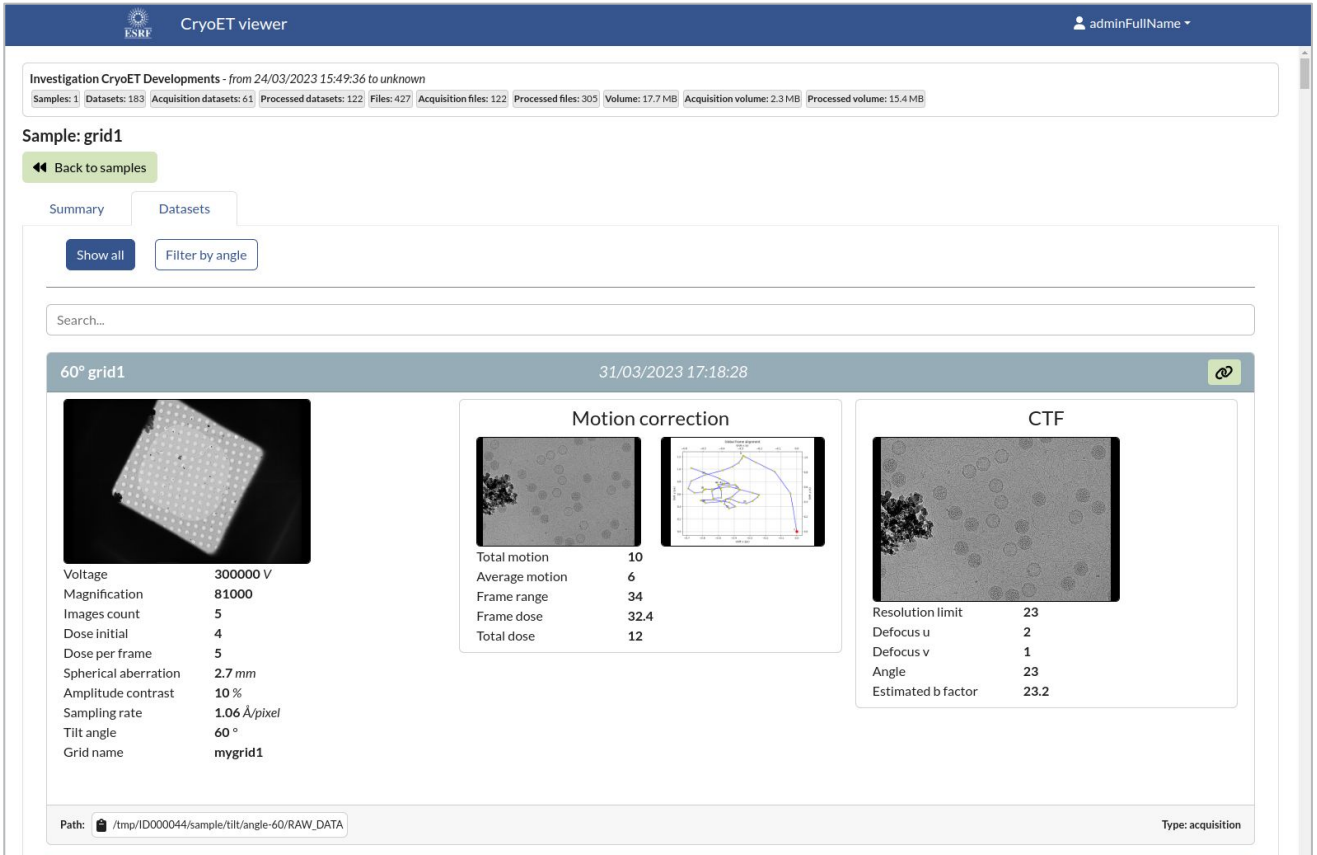
CryoET: metadata

- Dataset parameters are defined following **our** Nexus-like convention

```
<group NX_class="NXsubentry" groupName="EM">
   <protein_acronym ESRF_description="Protein acronym" NAPItype="NX_CHAR">${EM_protein_acronym}</protein_acronym>
   <voltage ESRF_description="Voltage" NAPItype="NX_CHAR">${EM_voltage}</voltage>
   <magnification ESRF_description="Magnification" NAPItype="NX_CHAR">${EM_magnification}</magnification>
   <images_count ESRF_description="Number of images in movie" NAPItype="NX_CHAR">${EM_images_count}</images_count>
   <position_x ESRF_description="Position X" NAPItype="NX_CHAR">${EM_position_x}</position_x>
   <position_y ESRF_description="Position Y" NAPItype="NX_CHAR">${EM_position_y}</position_y>
   <dose_initial ESRF_description="Dose initial" NAPItype="NX_CHAR">${EM_dose_initial}</dose_initial>
   <dose_per_frame ESRF_description="Dose per frame" NAPItype="NX_CHAR">${EM_dose_per_frame}</dose_per_frame>
   <spherical_aberration ESRF_description="Spherical aberration" NAPItype="NX_CHAR">${EM_spherical_aberration}</spherical_aberration>
   <amplitude_contrast ESRF_description="Amplitude contrast" NAPItype="NX_CHAR">${EM_amplitude_contrast}</amplitude_contrast>
   <sampling_rate ESRF_description="samplingRate" NAPItype="NX_CHAR">${EM_sampling_rate}</sampling_rate>
   <tilt_angle ESRF_description="tilt_angle" NAPItype="NX_CHAR">${EM_tilt_angle}</tilt_angle>
   <grid_name ESRF_description="grid_name" NAPItype="NX_CHAR">${EM_grid_name}</grid_name>
   <group NX_class="NXcollection" groupName="motioncorrection">
      <total_motion ESRF_description="Total motion of the sample" NAPItype="NX_CHAR">${EMMotionCorrection_total_motion}</total_motion>
      <average_motion ESRF_description="Average motion" NAPItype="NX_CHAR">${EMMotionCorrection_average_motion}</average_motion>
      <frame_range ESRF_description="Motion frame range" NAPItype="NX_CHAR">${EMMotionCorrection_frame_range}</frame_range>
      <frame_dose ESRF_description="Dose/frame" NAPItype="NX_CHAR">${EMMotionCorrection_frame_dose}</frame_dose>
      <total_dose ESRF_description="Total dose" NAPItype="NX_CHAR">${EMMotionCorrection_total_dose}</total_dose>
   </group>
   <group NX_class="NXcollection" groupName="ctf">
      <resolution_limit ESRF_description="Limit of the resolution" NAPItype="NX_CHAR">${EMCTF_resolution_limit}</resolution_limit>
      <correlation ESRF_description="" NAPItype="NX_CHAR">${EMCTF_correlation}</correlation>
      <defocus_u ESRF_description="" NAPItype="NX_CHAR">${EMCTF_defocus_u}</defocus_u>
      <defocus_v ESRF_description="" NAPItype="NX_CHAR">${EMCTF_defocus_v}</defocus_v>
      <angle ESRF_description="" NAPItype="NX_CHAR">${EMCTF_angle}</angle>
      <estimated_b_factor ESRF_description="" NAPItype="NX_CHAR">${EMCTF_estimated_b_factor}</estimated_b_factor>
   </group>
</group>
```

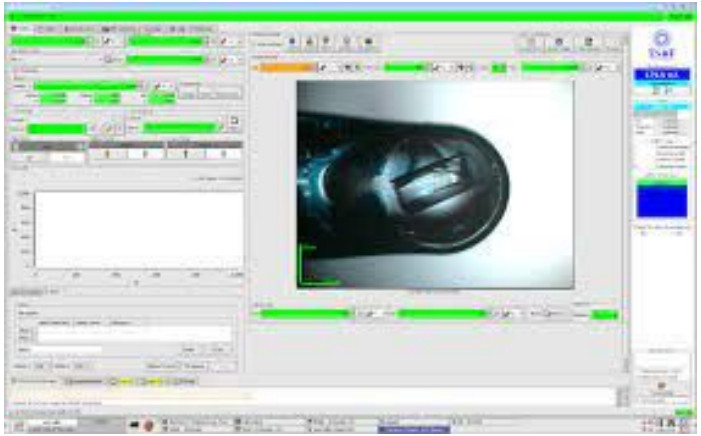
CryoET: UI

- Dedicated CryoET viewer (front-end talk later)

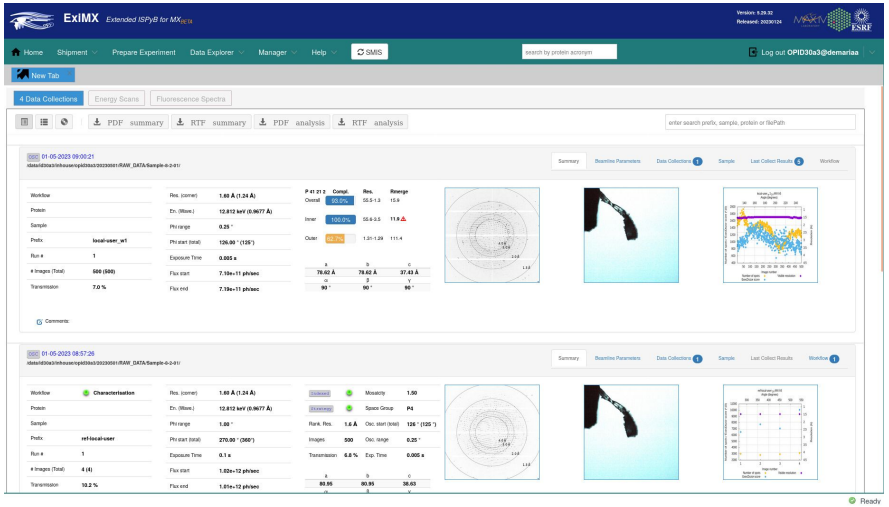


MX Use case

- Crystallography as state-of-the-art in automation at synchrotrons
  - ~ 20 years in developments
  - Fully automated analysis
  - Fully automated data acquisition = no users
  - Tailor made software:
    - MxCube (data acquisition)
    - ISPyB (LIMS)



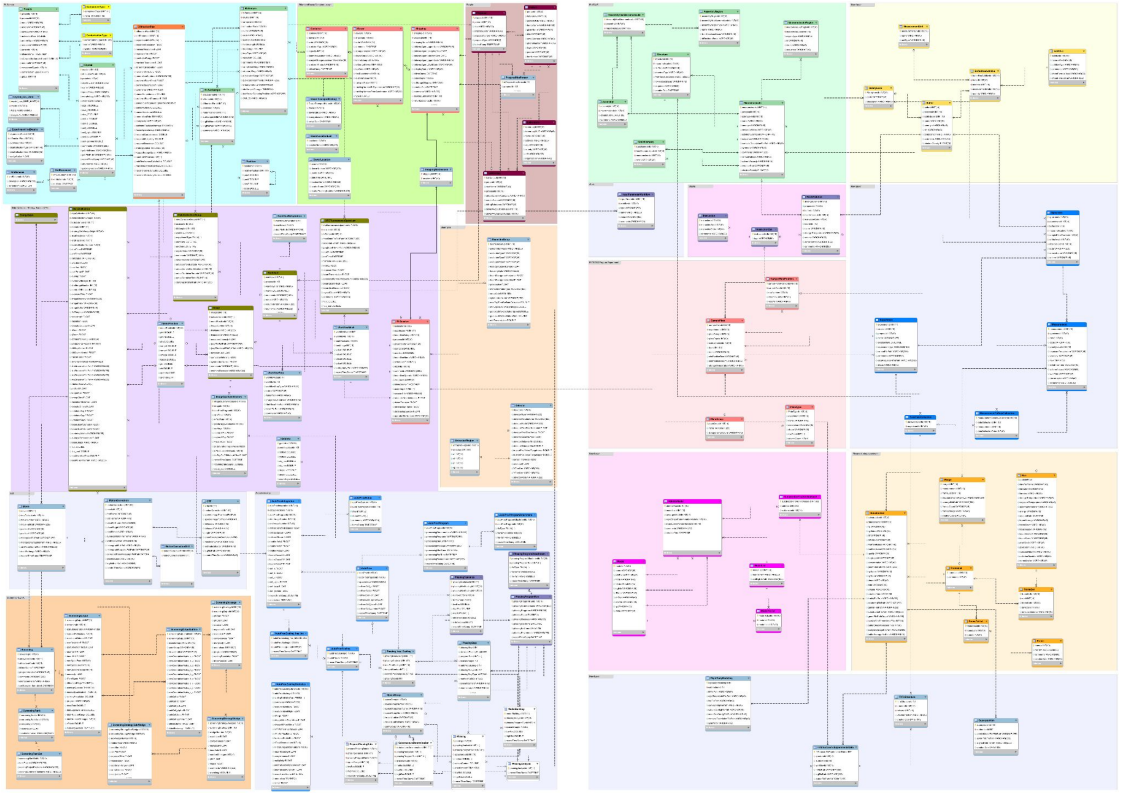
MxCube



ISPyB

MX Use case

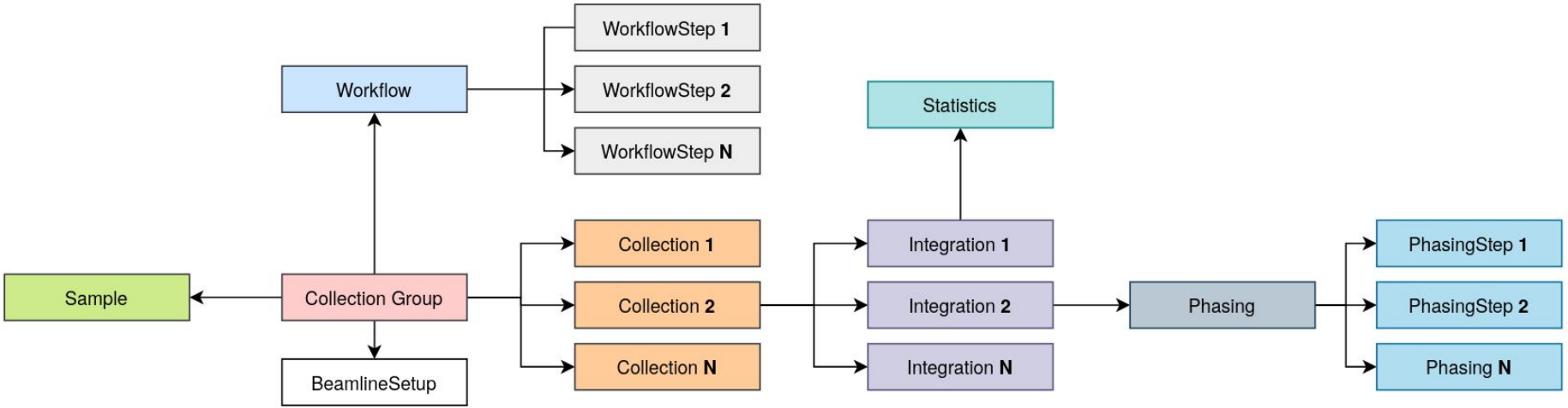
- From ISPyB to ICAT. WIP!!



ISPyB Database

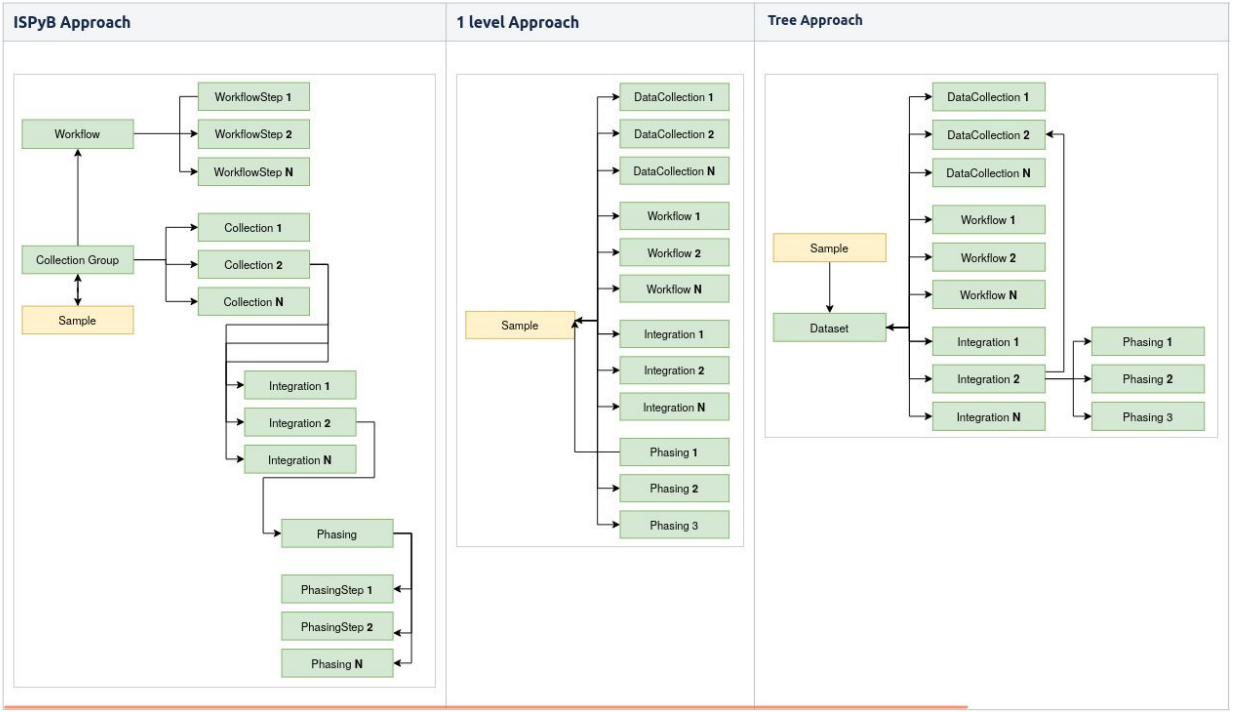
MX Use case

- Identify entities



MX Use case

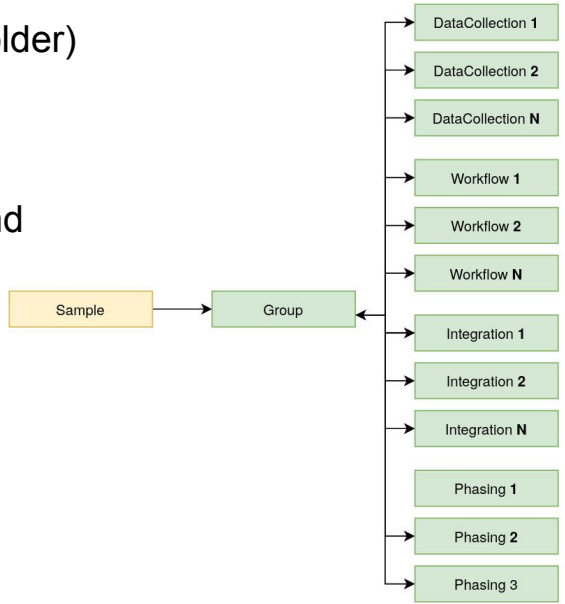
- From ISPyB to ICAT



ISPyB Database

MX Use case

- Proof of concept
  - ISPyB2ICAT procedure
    - Export ISPyB data to datasets (1 dataset = folder)
  - Ingest the data with soft links
  - Modify the current ISPyB UI to use ICAT as backend



ISPyB Database

Implications

- Handling processed data means:
  - Database:
    - Increased number of datasets (x2? x10? x100?)
    - Increased number of datafiles
    - Increased volume of data to archive
  - User interfaces:
    - Dedicated UI (frontend talk)
      - More developments = more people involved/teams
      - Scalability and maintainability
  - Data policy: some open question that the data policy makers need to answer:
    - Will processed data be open?
    - Will processed data be archived?
    - Will processed data be preserved? How long?
    - Who can upload data? When? where?

Summary

Datasets	1776288
Beamlines	49
Total Volume	10.4 PB
Total Number of files	603624263

Statistics 27/04/23

Conclusions

- The data acquisition and analysis standardization has allowed to reach higher levels of automation

- Handling processed data is a must when:
  - ++ number of samples
  - ++ fully automated pipelines

- Assessment of the quality of the data in real time:
  - Increase productivity and makes an efficient use of the beamtime

- ICAT seems to be the best option but need to be tested
  - We have presented an approach that is simple and powerful but has drawbacks like consistency
  - Performance
  - Scalability
  - Maintenance

Acknowledgement

- Marjolaine Bodin
- Mael Gaonach
- Andy Goetz
- Wout de Nolf
- Olof Svensson
- Axel Bocciarelli
- Loic Huder
- Data Automation Unit