Making the most of data at the ESRF-EBS

by Andy Götz (ESRF)

CommDat workshop IUCR 2023, Melbourne

TALK OUTLINE

- Motivations
- Guiding principles
- Current status
- Ongoing developments
- Future plans
- Conclusion
- Acknowledgements

MOTIVATIONS

1. **Implement** the **ESRF Data Policy**
2. **Help Users** deal with **big** and **complex data**
3. **Support Raw and Processed data**
4. **Help Users publish + cite data** easily
5. **Give Users** the best **Data eXperience** possible
6. **Make data verifiable** and **re-usable**
7. **Provide a trustworthy data repository**
8. **Promote** and **adopt Open Science**

LOOKING BACK - DDDWG MEETING @ ROVINJ 2015

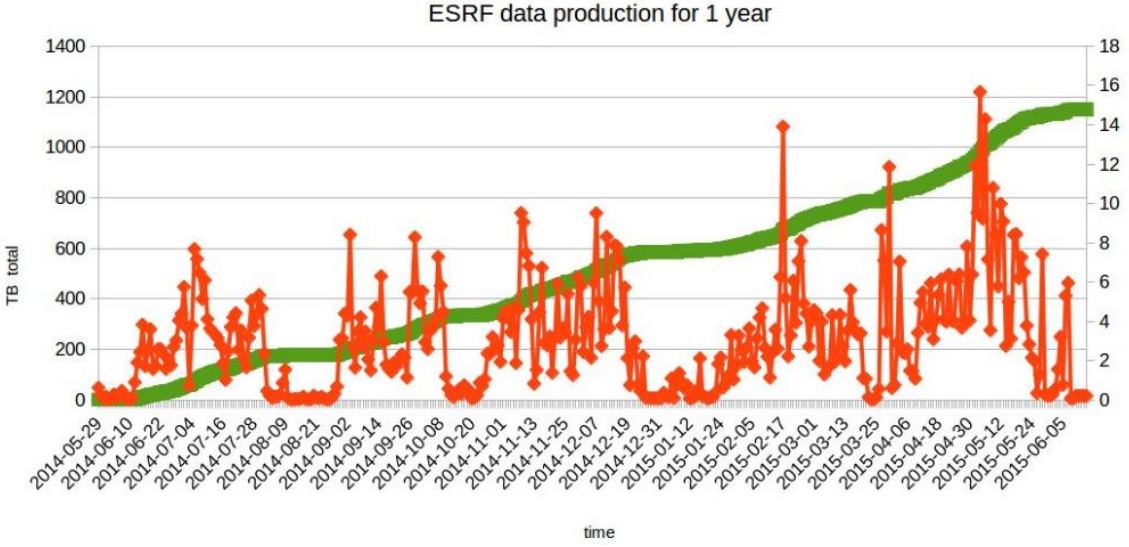
ESRF

- The European Synchrotron (Grenoble, France)
- 40+ Beamlines running 24/7
- Produced 1.2+ Petabytes of raw data in 2014-2015
- Metadata
  - Metadata well defined and managed for macromolecular crystallography (MX)
  - Non-unified approach for 35+ non-MX beamlines
- Upgrading source to a diffraction limited storage ring with 50+ more brilliance and coherence

LOOKING BACK - DDDWG MEETING @ ROVINJ 2015

ESRF

- The European Synchrotron (Grenoble, France)
- 40+ Beamlines running 24/7
- Produced 1.2+ Petabytes of raw data in 2014-2015
- Metadata
  - Metadata well defined and
    macromolecular crystallog
  - Non-unified approach for 3
- Upgrading source to a diffrac
  with 50+ more brilliance and



LOOKING BACK - DDDWG MEETING @ ROVINJ 2015

Goals for Metadata @ ESRF for ALL beamlines

1. Define metadata for all experimental techniques
2. Define data format(s) for automated data analysis
3. Annotate data with metadata and store in HDF5
4. Archive all metadata forever
5. Provide access to metadata

ICAT

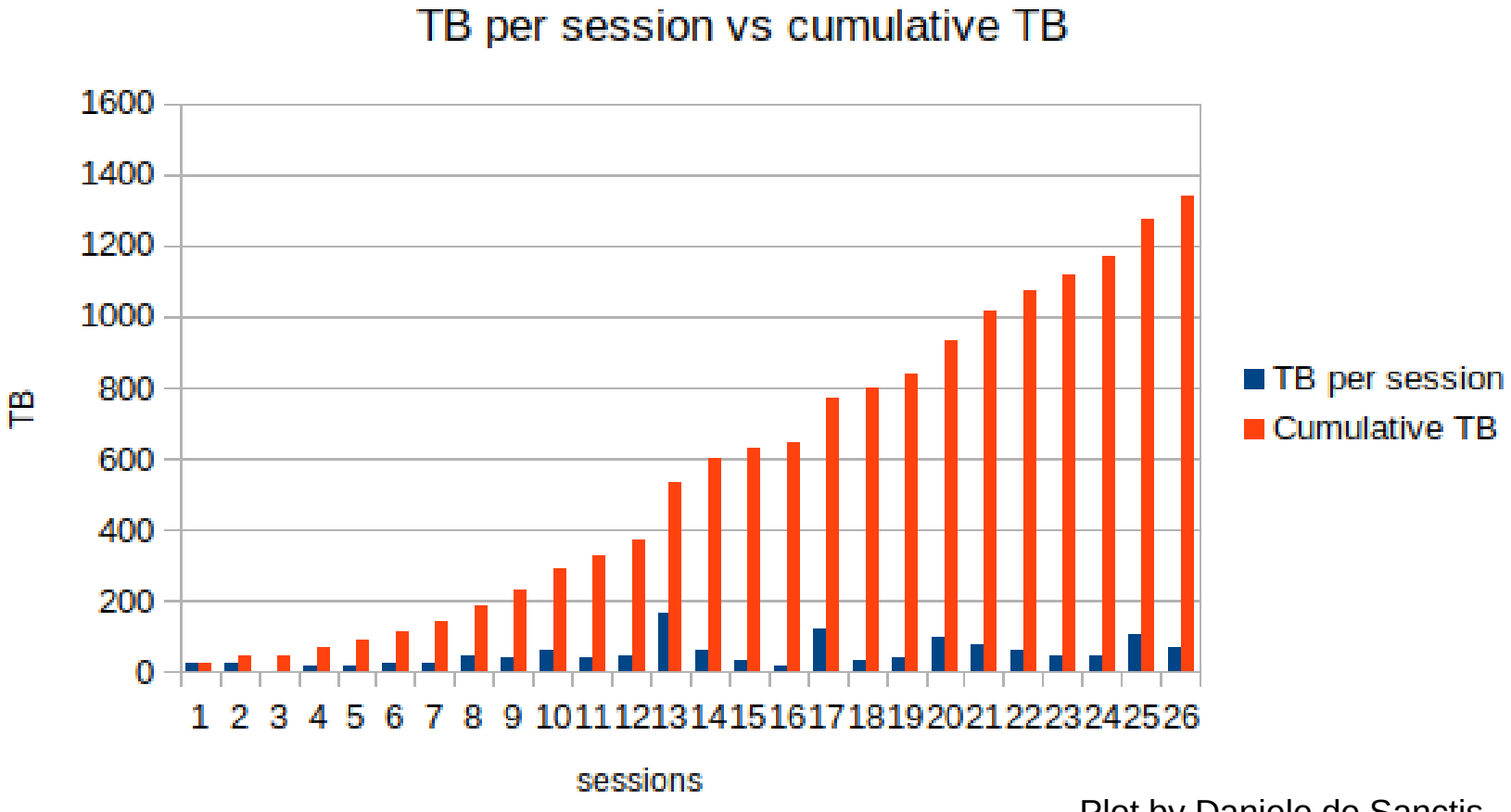
6. Implement DOI for data for provenance + publications
7. Provide users efficient download data service(s)
8. Archive (not-for-free) service to curate raw data
9. Implement the ESRF Data Policy (as soon as it has been defined by the management)

FAST FORWARD - 2023

1. **2015 - Council endorsed ESRF Data Policy**
    1. **Archiving** of metadata (forever) + raw data (10 years)
    2. **3 year embargo period** after which data from publicly funded beamtime are made open access under a CC-BY-4 licence
2. **2020 - ESRF-EBS 4th generation Storage Ring**
3. **2019-2022 - PaNOSC+ExPaNDS projects for FAIR data**
4. **2023 - Data Policy implemented on all beamlines**
    1. Automated ingestions of raw data + metadata ingested
    2. Two tape libraries with a potential capacity of 1 Exabyte each
    3. NeXus/HDF implemented on ~90% of beamlines
    4. ESRF-EBS currently producing 10+ Petabytes per year

DATA CHALLENGES OF THE ESRF-EBS – SSX@ID29

ESRF-EBS serial crystallography beamline (ID29) produced > 1.2 PB in first year



Plot by Daniele de Sanctis

PANOSC+EXPANDS IMPACT ON FAIR DATA AT ESRF

1. First draft of **FAIR Data Policy** (thanks to PaNOSC DP framework)
2. **Data Policy+data portal+e-logbook on 42/44 beamlines** (thanks to ESRF staff)
3. **Nexus/HDF5** on almost all beamlines (thanks to BLISS+NexusWriter+Lima)
4. **DMPs** implemented for all proposals (thanks to DS-Wizard)
5. **Data visualisation** in Jupyter + data portal + other web applications (thanks to H5Web)
6. **Jupyter-slurm** service used daily (thanks to jupyterhub-moss)
7. **VISA+CERNVMF**s deployed and being tested on beamlines (thanks to VISA)
8. **OASYS training** developed + implemented new algorithms (thanks to Vinyl)
9. **Search API** implemented and deployed (thanks to ICAT)
10. **Human Organ Atlas** deployed (thanks to PaNOSC search portal)
11. **Data transfer** solved for users (thanks to Globus + Aspera deployed)
12. **E-learning platform** ready to be adopted (thanks to e-learning.e-training.org)

→ **Open Science and FAIR data are becoming a reality @ ESRF**

GUIDING PRINCIPLES

- **Follow the FAIR guiding principles data for all techniques (not only for Structural Biology)**
  - **F**indable – all raw data **archived** + have a **DOI**
  - **A**ccessible – data are **licensed** + **downloadable**
  - **I**nteroperable – standard **metadata** + data **formats**
  - **R**eusable – electronic **logbook** + **sample** metadata
- **Follow the principles of Open Science**
  - Open Data, Open Source, Open Access, Open Infrastructure, …
- **Work closely with the scientific communities**
  - IUCr (CommDat, COMCIF), medical, materials science, …
- **Provide a sustainable open source software solution**

DATA MANAGEMENT STATUS @ ESRF

- **ISPyB + ICAT –** two working solutions at ESRF for structural biology (ISPyB) + all techniques (ICAT)

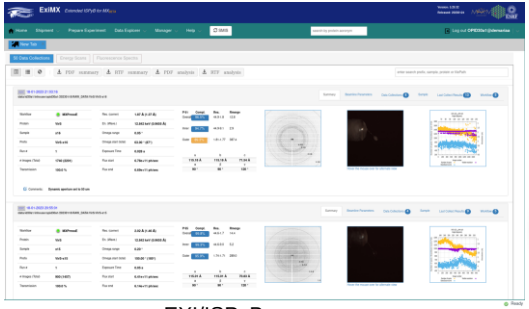
➢ **New development started in 2023 to merge the two solutions to provide a single platform for all techniques to provide FAIR data (raw and processed) to ESRF Users and Open Data Users**

➢ **ESRF-EBS - scientists strongly requesting automation of data processing for all techniques for users**

➢ **ESRF Data Repository CoreTrustSeal certified**

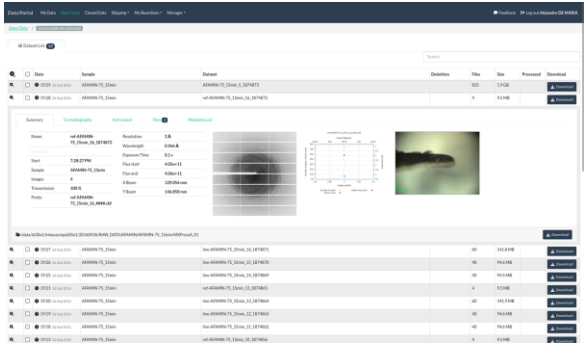
➢ **Collaborating with CommDat on Raw Data Journal**

TALE OF 2 DATABASES – 2002 TO 2022



EXI/ISPyB

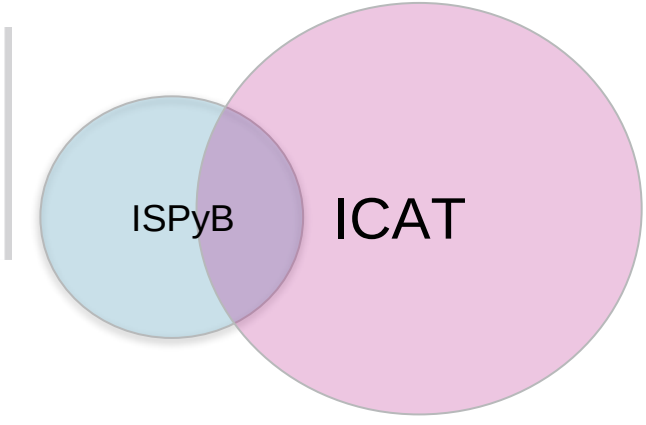
2002 - 2018



Data Portal/ICAT

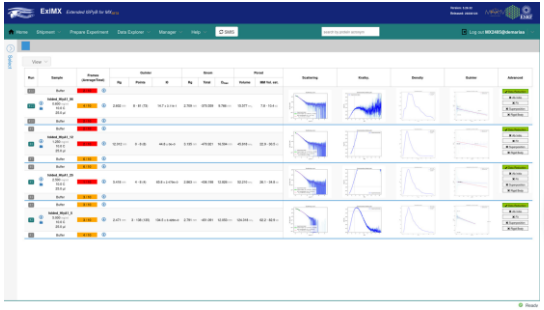
2020 - 2022

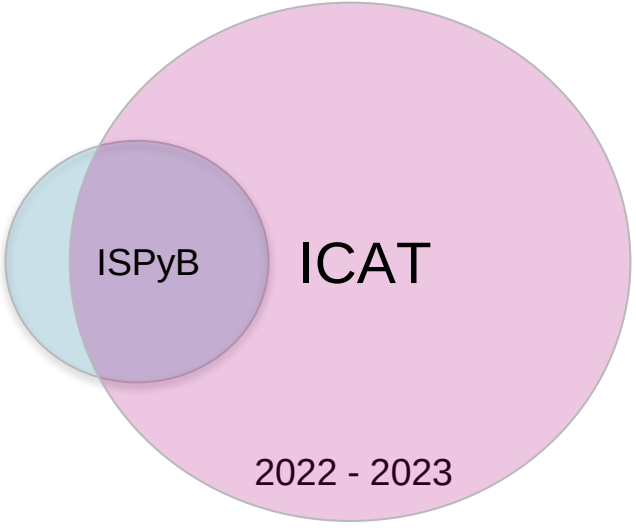
ISPyB's parcel label

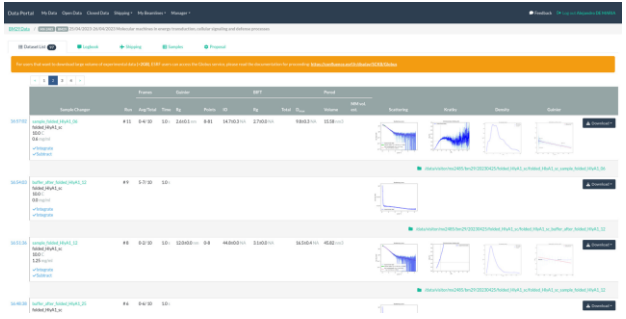


ICAT's parcel label

TALE OF 2 DATABASES – 2022 TO 2023



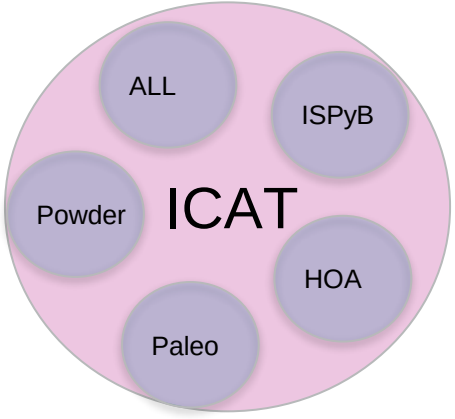




- Ongoing development to evaluate providing the ISPyB functionality with extended ICAT+
  - Join raw and processed data for all techniques
  - New solution must be better than ISPyB
  - Reprocessing of MX data a requirement

➢ **Deadline - October 2023**

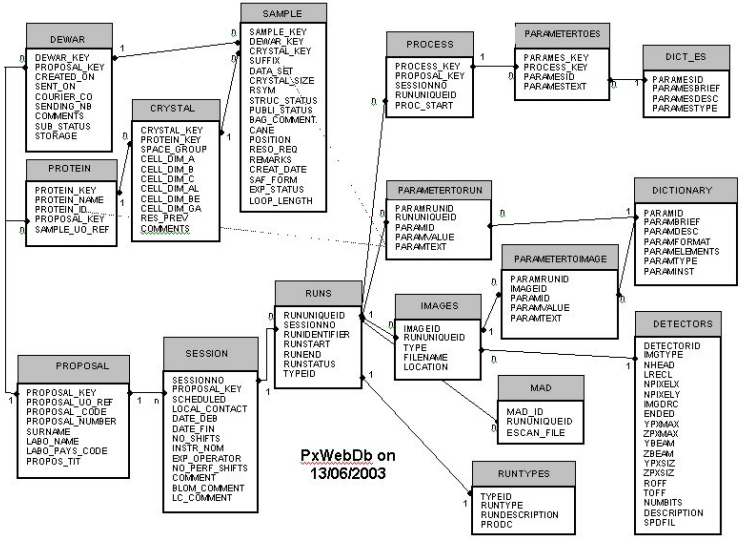
THEN THERE WAS ONE – 2024 …

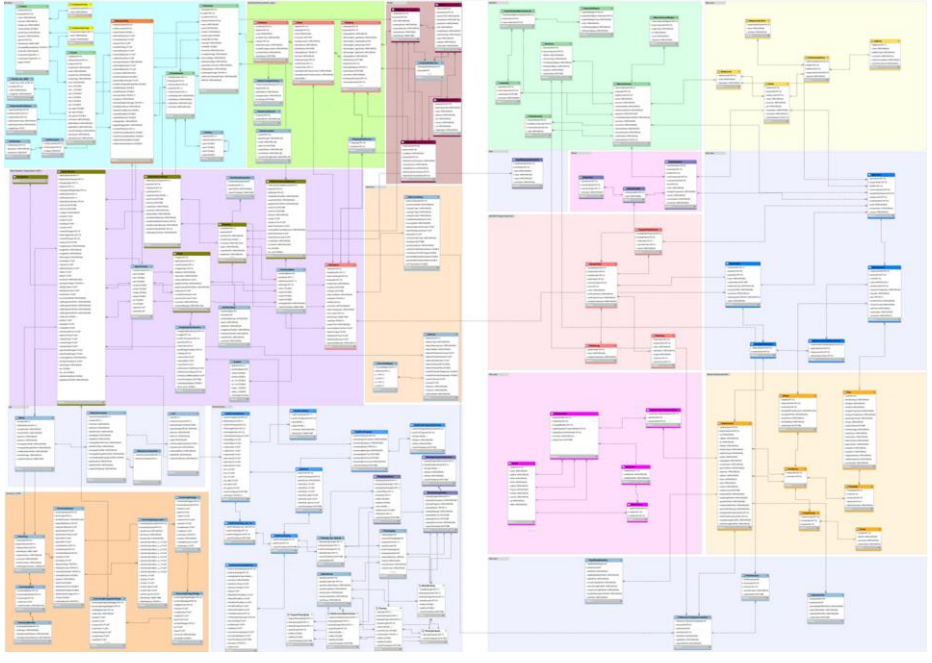


- Adopting a single solution based on a generic approach has advantages:
  - Facility level:
    - SB processed data linked to raw data
    - SB gets new improved user interfaces + reprocessing
    - Other BL profit from **SB capabilities**
    - Reduces maintenance and effort considerably
  - Beamline and MX's level
    - **Rapid** development
    - More **autonomy**
      - No need to new tables or columns
      - Metadata nomenclature is defined by scientist(s)/comunity
    - **Generic** backend
      - **Focus** on:
        - Doing **Science**
        - Define procedures and methods
    - **Modular architecture**
    - **Advanced user interfaces**
- Why **ICAT** and not **ISPyB** (or **SciCat** or **CKAN**) ?

From Schema-based in ISPyB …

- **After 20 years of adding tables and columns ISPyB has become unwieldy, many tables but few cases of many-to-many relations …**





2003: 17 tables

2023: 209 tables, 42 views

What is ICAT?

- **ICAT is a generic metadata catalogue developed and supported by STFC/UKIRT and ICAT collaboration**

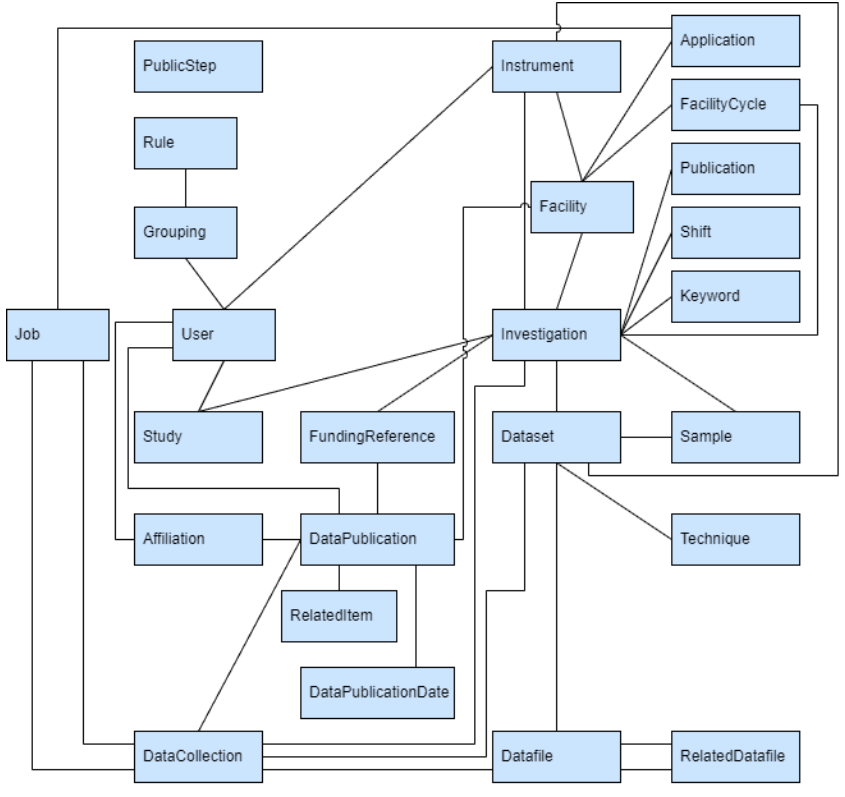
- ICAT supports research data management for **large-scale facilities and is** in production managing billions of datasets + files at ISIS, DLS, ESRF, HZB

- **ICAT is composed of a set of scalable components** :
  - ICAT Server:
    - Supports ORACLE and MariaDB databases
    - Rest/SOAP API
  - Authenticators: OpenID, SSO, DB, custom…
  - Fine-grained authorisation model based on roles
  - OAI-PMH metadata harvesting plugin
  - Search component based on Apache Lucene
  - python-icat: python client components
  - PANOSC/ExPands search API
  - Extensions e.g. e-logbook, bespoke portals, …

Schemas in ICAT …

- **A generic schema with domain-specific dictionaries offers a higher degree of flexibility for extending / adding new techniques;**
- **ICAT is a proven solution that just works out-of-the-box**

- **ICAT has around 40 tables** :
  - BLSessions
  - Users
  - Data collections
  - Data processing
  - Data publication
  - DOI
  - Techniques (PANET)
- **ESRF Extensions**:
  - E-logbook + notebook
  - Custom data portals
  - Sample Shipping + tracking



Main Entities

METADATA STORED AS PARAMETER LISTS

RawDataset

- experimentType
- startTime
- endTime
- actualSampleSlotInContainer
- actualContainerBarcode
- actualContainerSlotInSC
- dataCollectionNumber
- axisStart
- axisEnd
- axisRange
- overlap
- numberOfImages
- startImageNumber
- numberOfPasses
- exposureTime
- imageDirectory
- imagePrefix
- imageSuffix

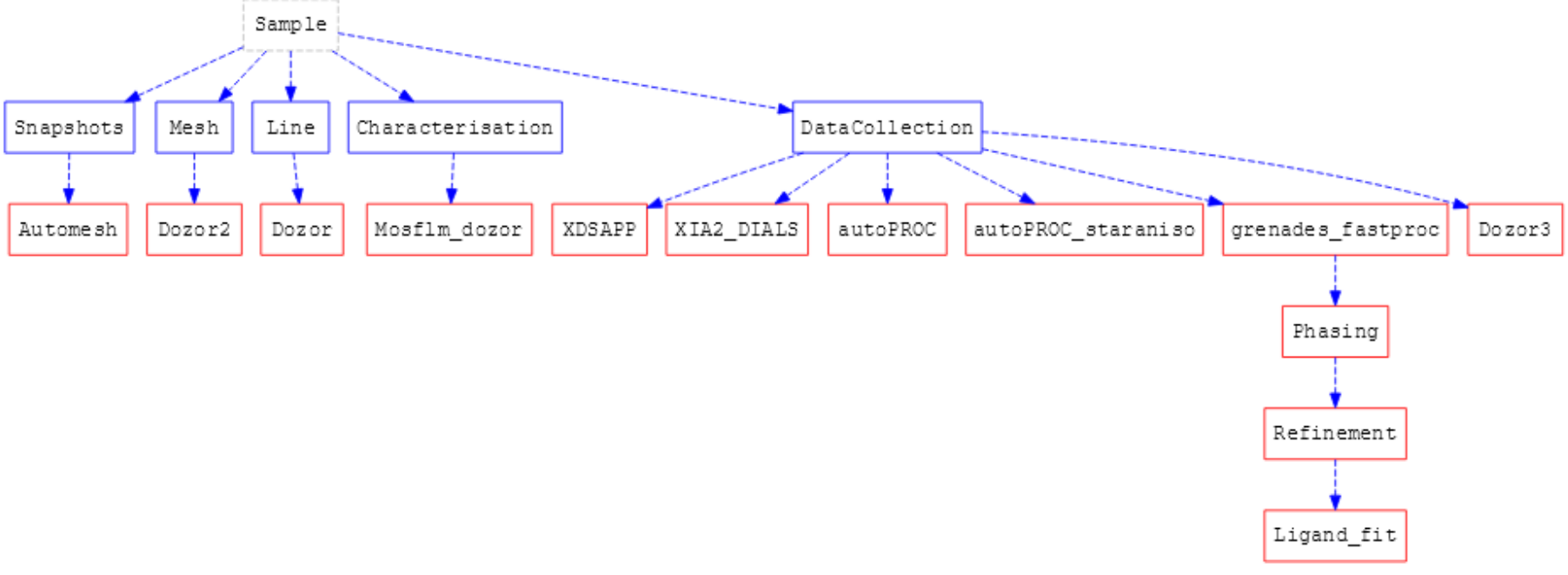
- rotationAxis
- phiStart
- kappaStart
- omegaStart
- resolutionAtCorner
- undulatorGap1
- undulatorGap2
- undulatorGap3
- beamSizeAtSampleX
- beamSizeAtSampleY
- centeringMethod
- actualCenteringPosition
- beamShape
- flux
- flux_end
- workflowName
- workflowStatus
- crystalSizeX
- crystalSizeY

- resolution
- detectorDistance
- xBeam
- yBeam
- xBeamPix
- yBeamPix
- slitGapVertical
- slitGapHorizontal
- transmission
- synchrotronMode

Similar to CIFs ?

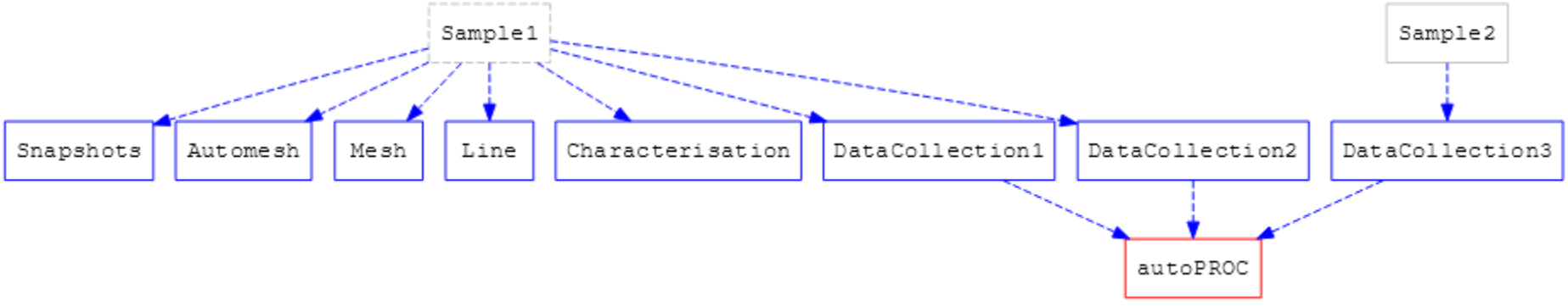
LINKING DATA PROCESSING + DATASETS

Workflow MxPress-E



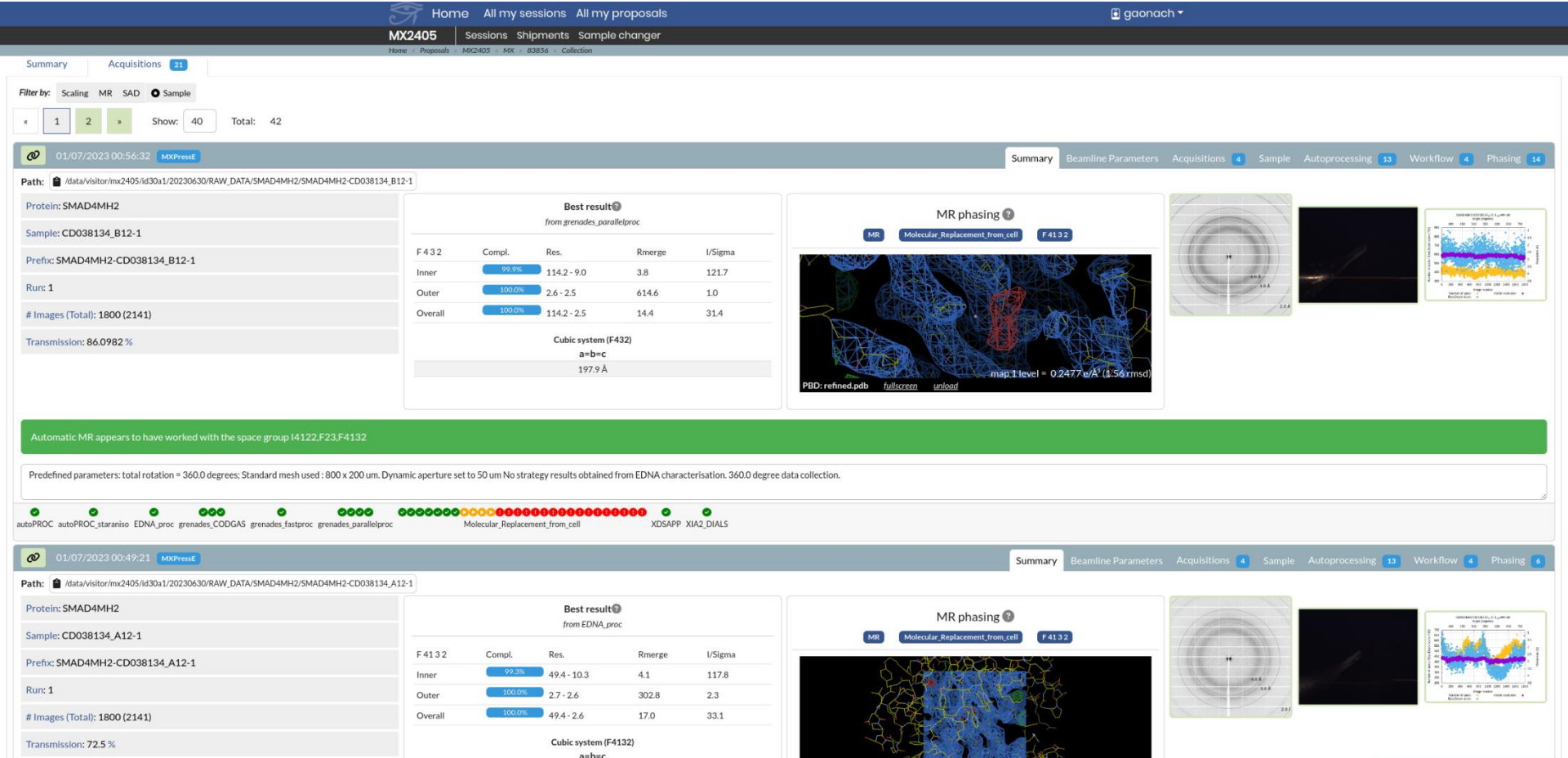
LINKING DATA PROCESSING + DATASETS

Merging



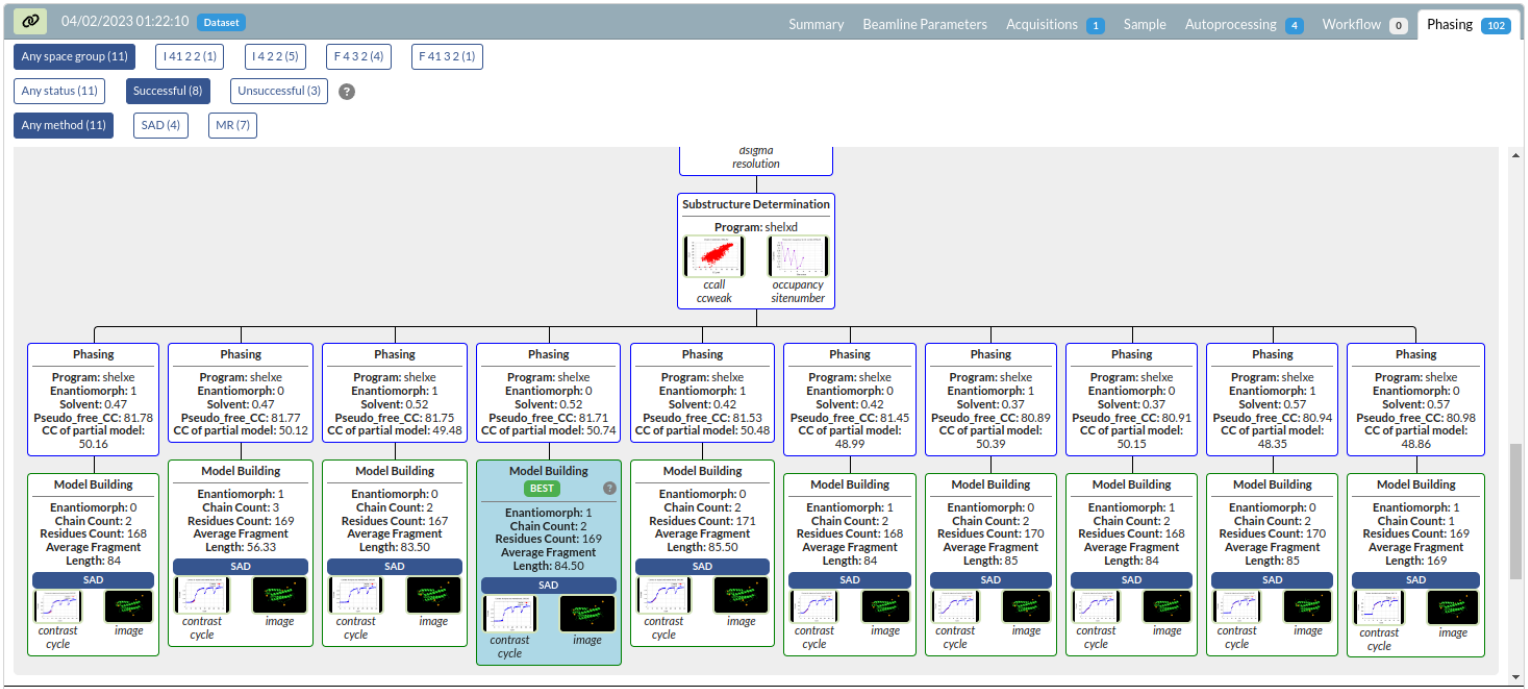
Data Portal v2 – MX auto processing

- **Improving the users Data eXperience through efficient presentation**



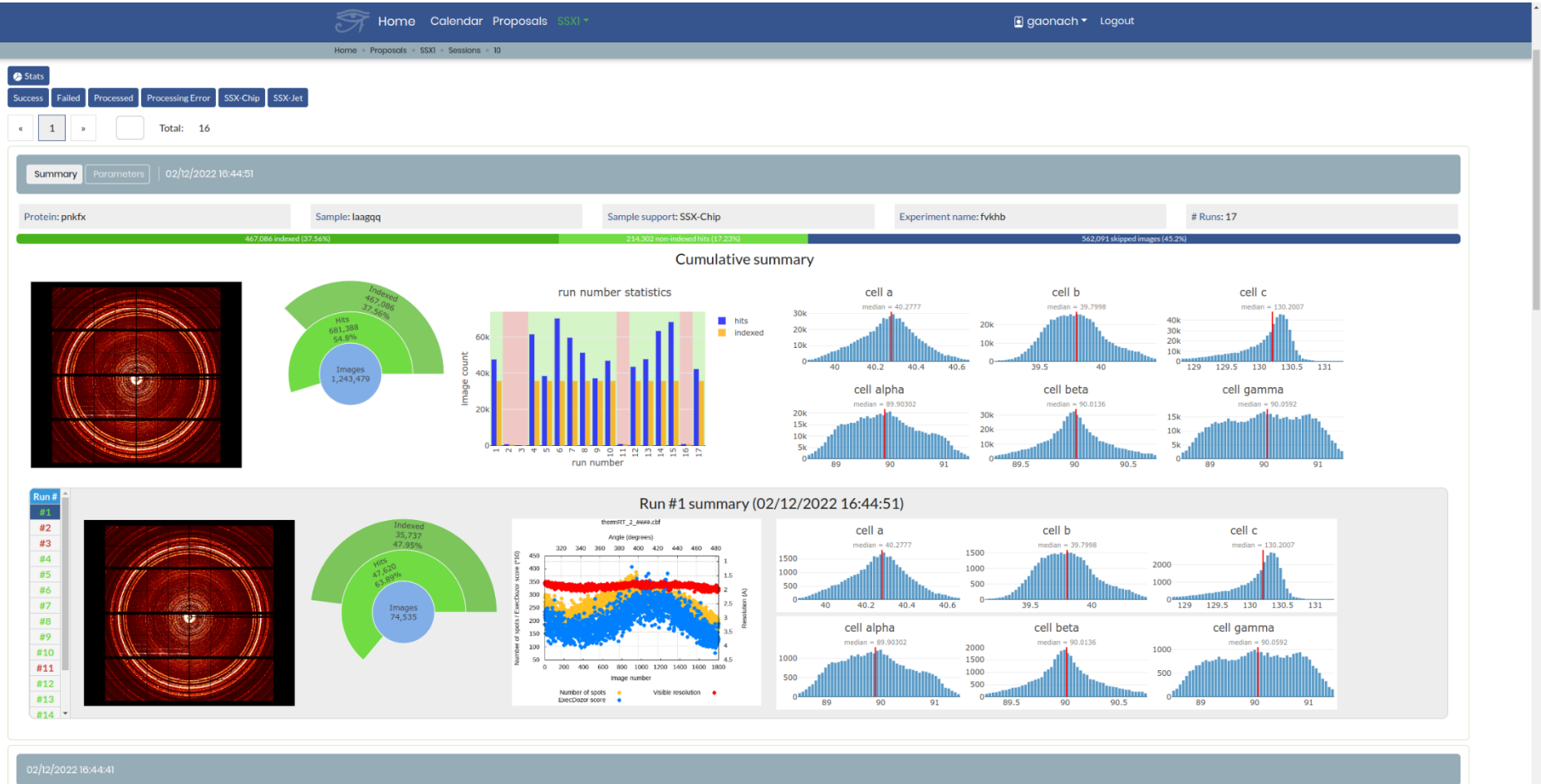
Data visualization – MX autophasing

- **Improving the users Data eXperience through visualization of processing**



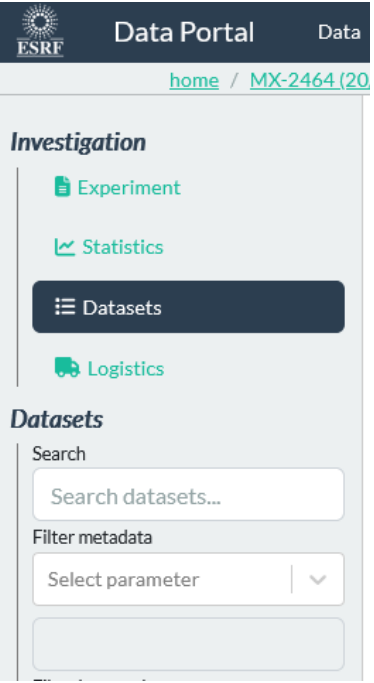
Data Portal v2 – serial crystallography

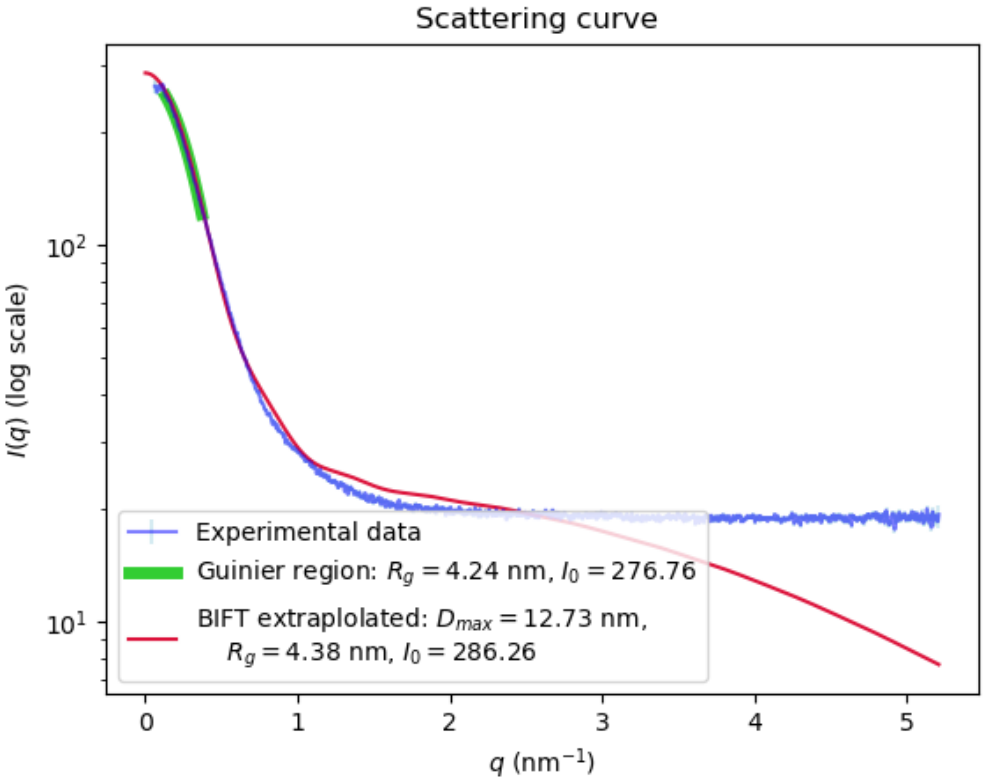
- **Improving the users Data eXperience through efficient visualisation**

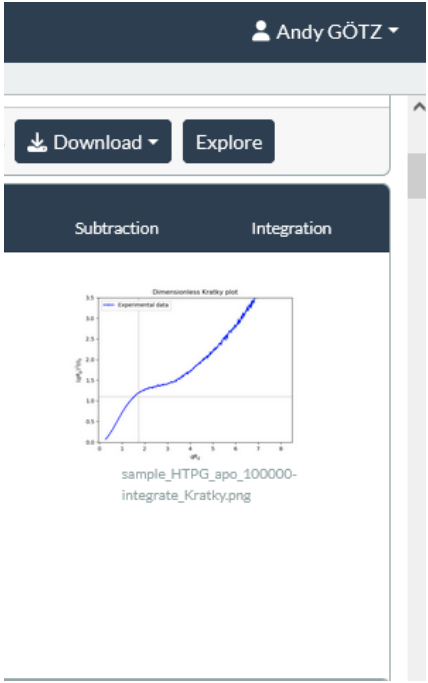


Data visualization - BioSAXS

- **Improving the users Data eXperience through visualization of processed results**

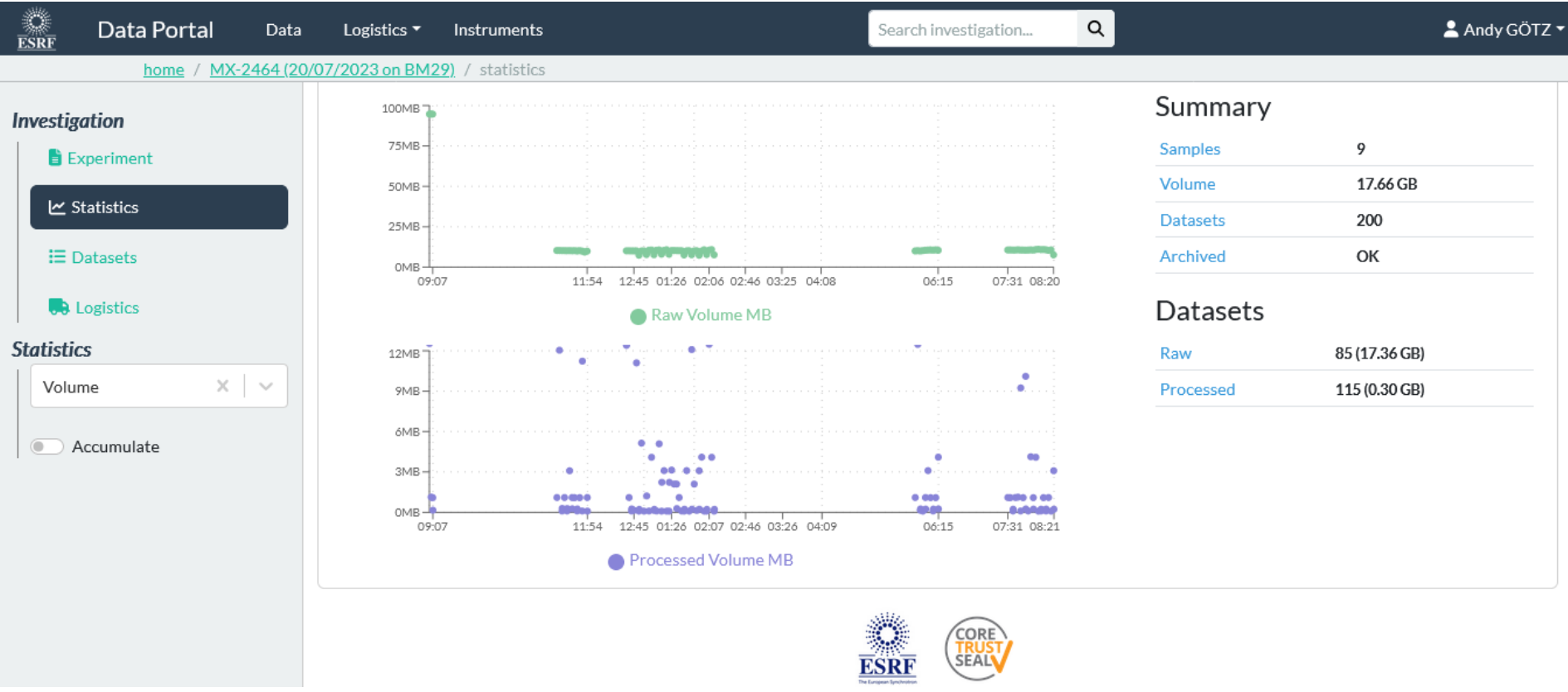






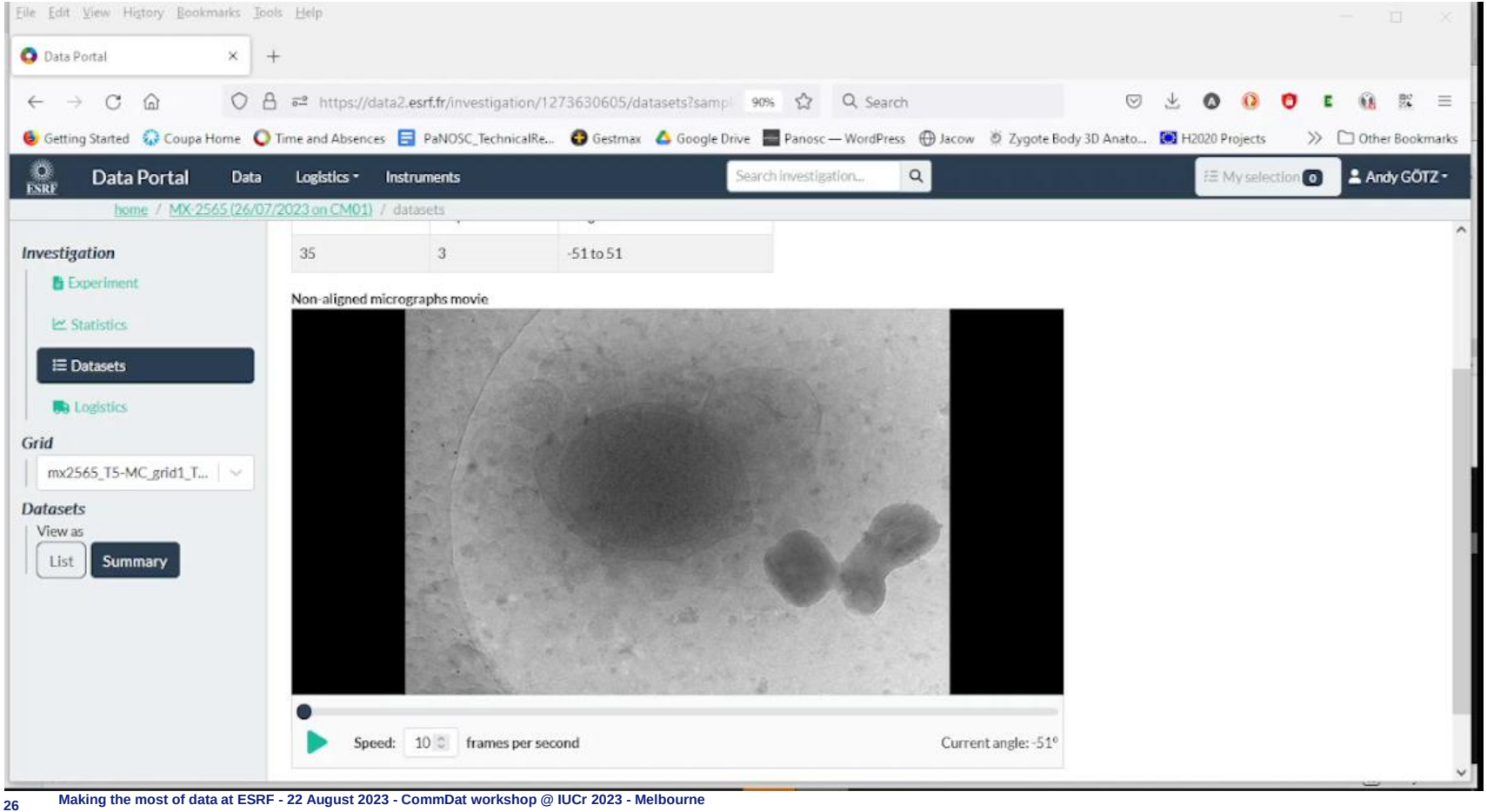
Data visualization - BioSAXS

- **Improving the users Data eXperience through visualisation**



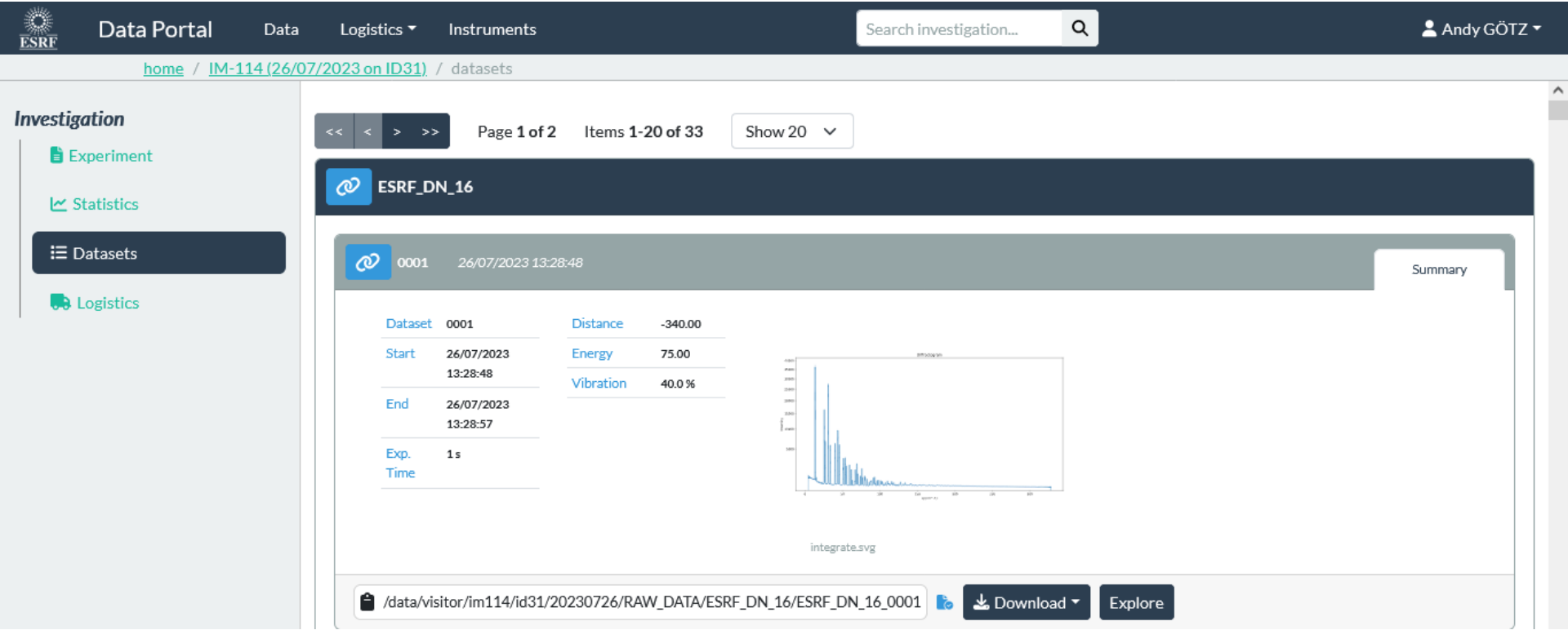
Data visualization - CryoET

- **Improving the users Data eXperience through visualisation**



Data visualization – HTXRPD

- **Improving the users Data eXperience through visualization of processed results**



Data processing with Ewoks – HTXRPD

Acquisition

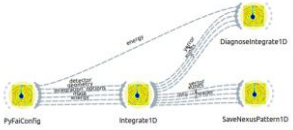
In-memory data

NeXus Writer

Device Server (Lima)

RAW_DATA

Ewoks Workflow

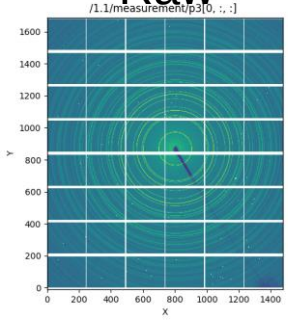


blissoda

triggering logic
parametrization
feedback
handling

Raw

Raw



blissdata

data access

Bliss 2.0

Ewoks

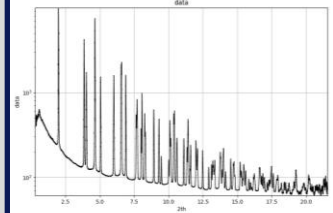
Computation (Slurm or BL computer)

Ewoks Workflow

Data provenance
Reprocess

Results

Processed



Feedback
Visualization
flint, daiquiri, mxcube

PROCESSED_DATA

Data portal

Data processing with Ewoks – 25 beamlines (so far)

Tomography

BM05, BM18, ID11, ID16B, ID17, ID19

30 tasks

SAXS/WAXS

BM02, ID09, ID11, ID31

18 tasks

Spectroscopy

BM23, ID24

15 tasks

Fluorescence

ID16A, ID16B, ID21

11 tasks

Dark-field Microscopy

ID06, ID11

16 tasks

Imaging

ID16b, ID21

4 tasks

MX Beamline Automation

ID23-1, ID23-2, ID30A-1, ID30A-3, ID30B

292 tasks

BioSAXS

BM29

Under construction

Special Diffraction

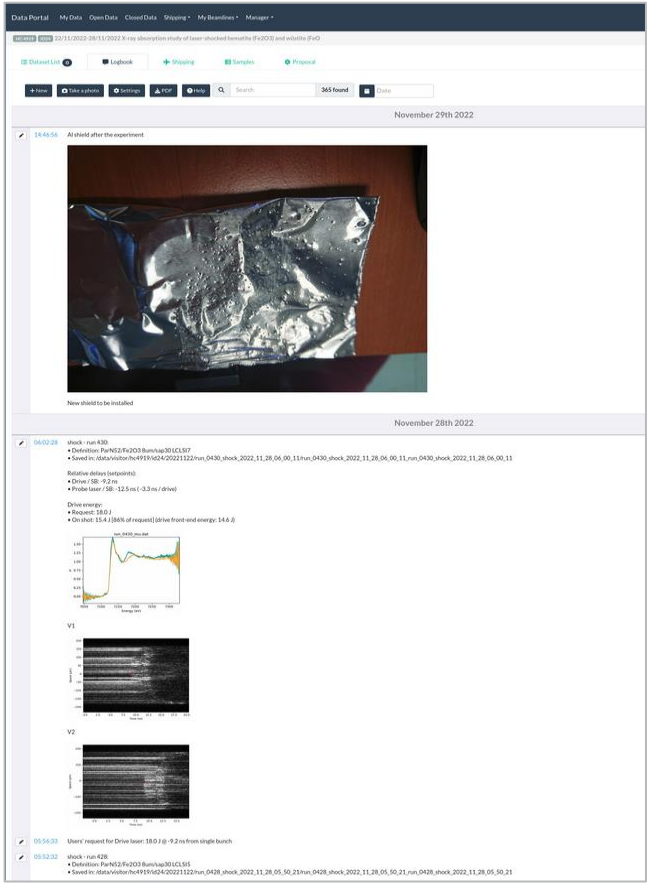
ID11, ID22

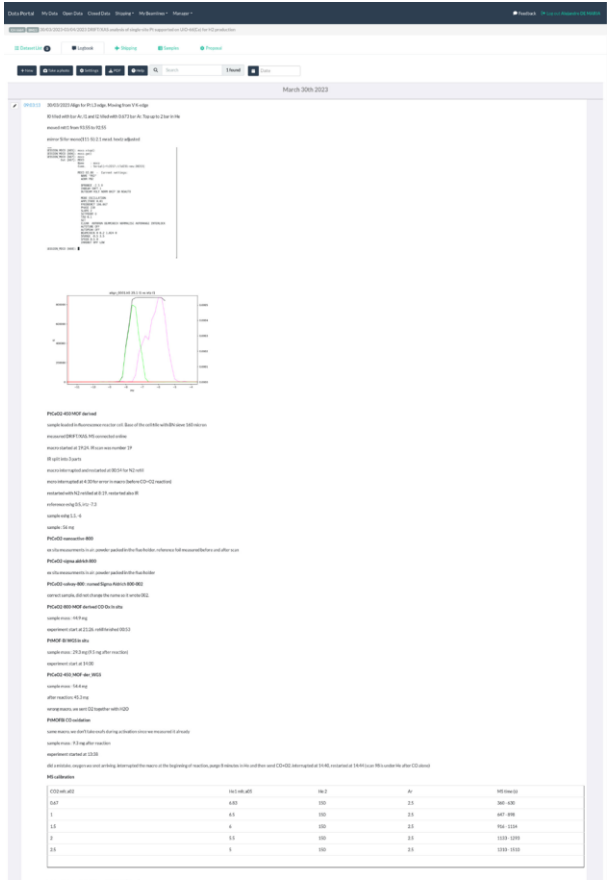
5 tasks

"*HDF5 is the unsung hero that makes all of this possible*" – Wout de Nolf

Electronic logbook + notebook

- **Logbooks and notebooks make experiments and data FAIRer**

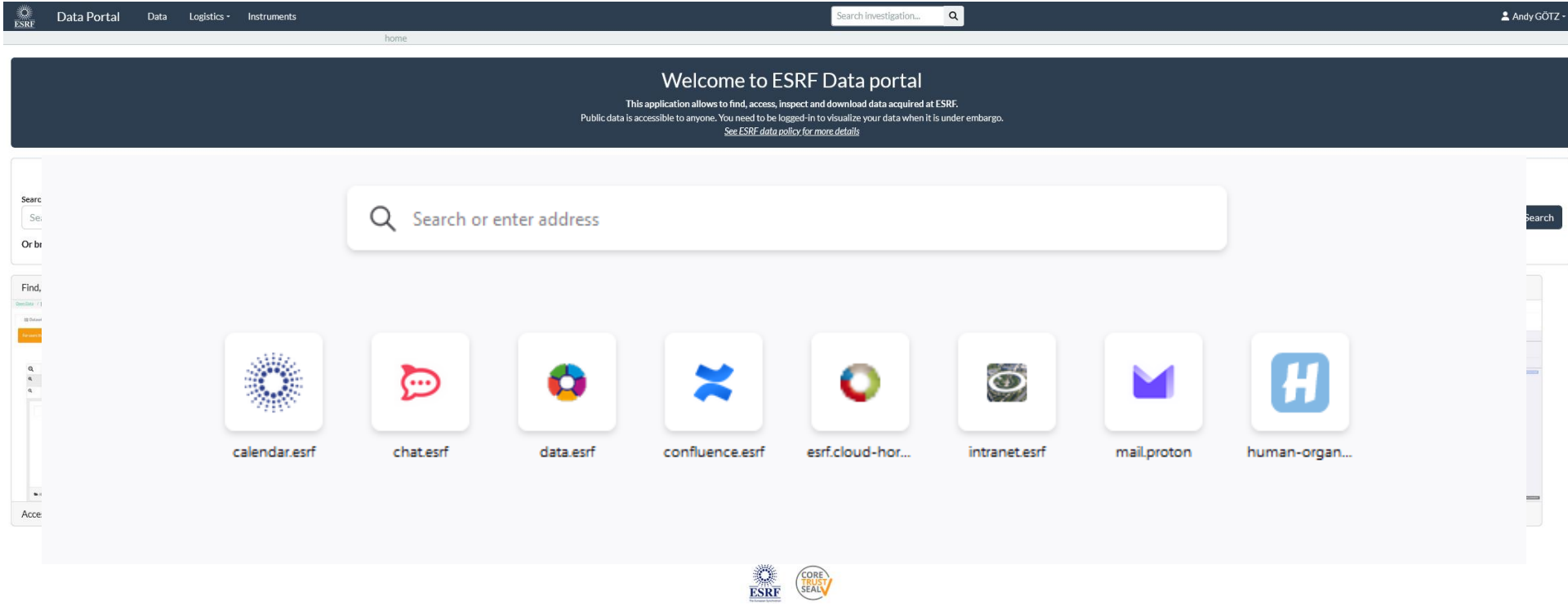




- **Annotate experiments**
  - Chronological
  - Automatically
  - Manually
- **Logbook types**
  - Experiment
  - Beamline
  - Facility
- **Electronic notebook**
  - Document style

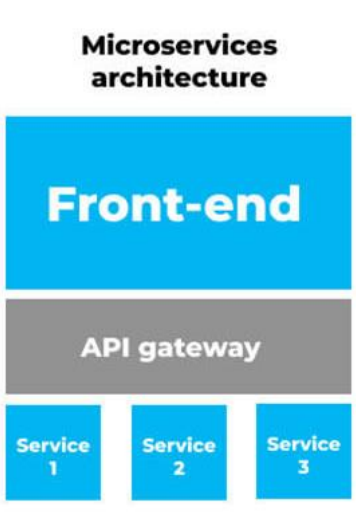
Data Portal v2 – home page

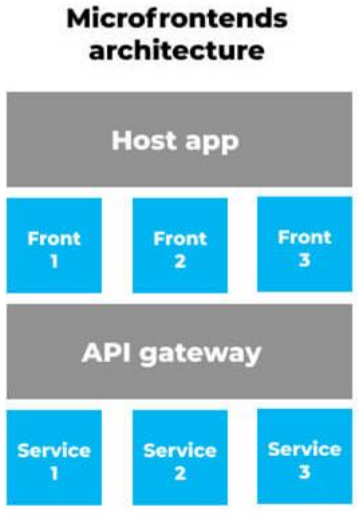
- **Improving the users Data eXperience through efficient searching**

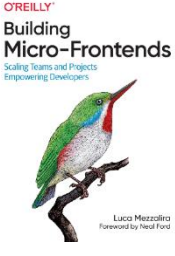


Architecture based on microservices + micro frontends

- **Web interfaces** are **key** to providing a **good Data eXperience**
- **Micro-frontends** adopted to **developing scalable flexible** applications

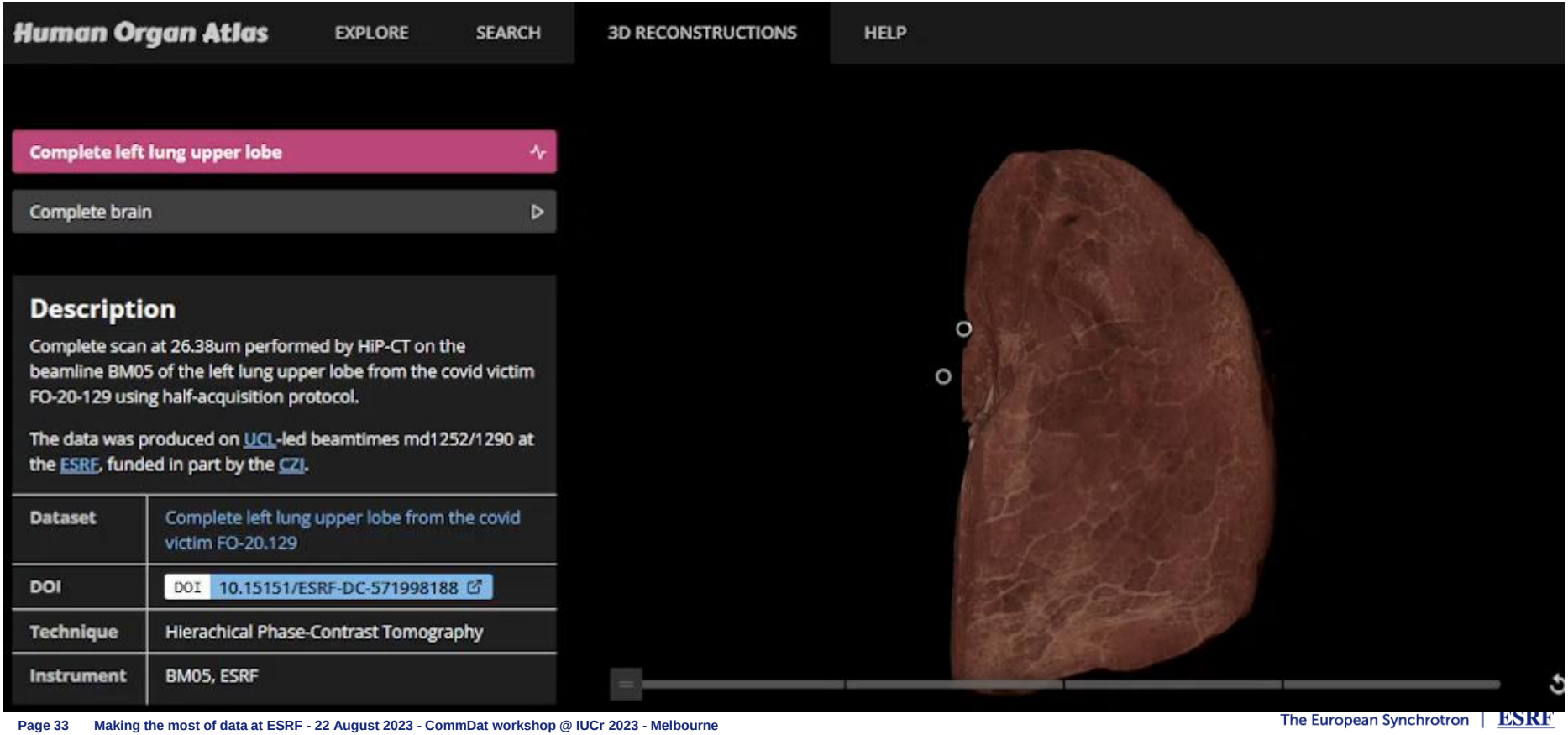






DATA COLLECTIONS OF HIGH QUALITY DATA

Human Organ Atlas - https://human-organ-atlas.esrf.eu



FUTURE PLANS

- Finish evaluation (in 2023) of ISPyB in ICAT for MX + other techniques for raw and processed data
- Implement data reprocessing for ESRF MX Users
- Continue developing workflows for all techniques
- Ensure ESRF MX raw data metadata comply with IUCr Raw Data Journal requirements
- Release new Collections of Open Data
  - Human Organ Atlas – adding hundreds of datasets
  - Paleontology – 200 TB of 20 years of processed data
  - Materials Science – power electronics (AMP), batteries

GOOD, BAD AND UGLY

- **GOOD**
  - Providing Users with a single place to find and deal with data
  - Processed data improves efficiency of beamtime
  - Electronic logbooks help documenting experiments
- **BAD**
  - Getting users to cite data DOIs
  - Defining & Collecting sample information
  - Tracking data reuse and non-use
- **UGLY**
  - Standardising metadata for data (re)processing
  - Minimising “dark” data i.e. unusable data

CONCLUSION

- **ESRF** is developing a **common platform** based on **ICAT** for **raw** and **processed** data for all techniques
- **Workflows** automating **data processing** for many techniques (**HDF5** is unsung **hero** of data automation)
- **All Users** get access to **processed** data
- **IUCr CommDat** is helping facilities to adopt **FAIR data** practices through ***Raw Data Letters*****,** can we do more?
- The question is not if raw data long term curation is feasible anymore but ***how to give all Users the best Data eXperience to interpret and reuse the data***

ACKNOWLEDGEMENTS

- **ESRF Core Team:**

  Alejandro de Maria, Marjolaine Bodin – ICAT+ ISPyB
  Mael Gaonach, Axel Bocciarelli, Loic Huder, Giannis Koumotsos – web interfaces
  Wout de Nolf, Olof Svensson, Loic Huder, Henri Payno – workflows
  Marcus Oscarsson, Antonia Beteva – MxCuBE developers
  Max Nanao, Romain Talon, Didier Nurizzo – MX scientists
  Daniele de Sanctis – SSX beamline scientist
  Isai Kandiah – EM beamline scientist

- **ESRF-EBS beamline scientists**
- **ISPyB international collaboration**
- **EU projects – PaNOSC, ExPaNDS + STREAMLINE**

SOME USEFUL LINKS

- **ESRF data portal**
  - https://data.esrf.fr and https://human-organ-atlas.esrf.eu
- **ISPyB LIMS**
  - Current version - https://github.com/ispyb
  - New developments - https://gitlab.esrf.fr/icat/icat-plus
- **EWOKS workflows**
  - Source – https://gitlab.esrf.fr/workflow/ewoks
  - Documentation – https://ewoks.readthedocs.io/
- **ICAT metadata catalogue -** https://github.com/icatproject
- **CoreTrustSeal -** https://amt.coretrustseal.org/