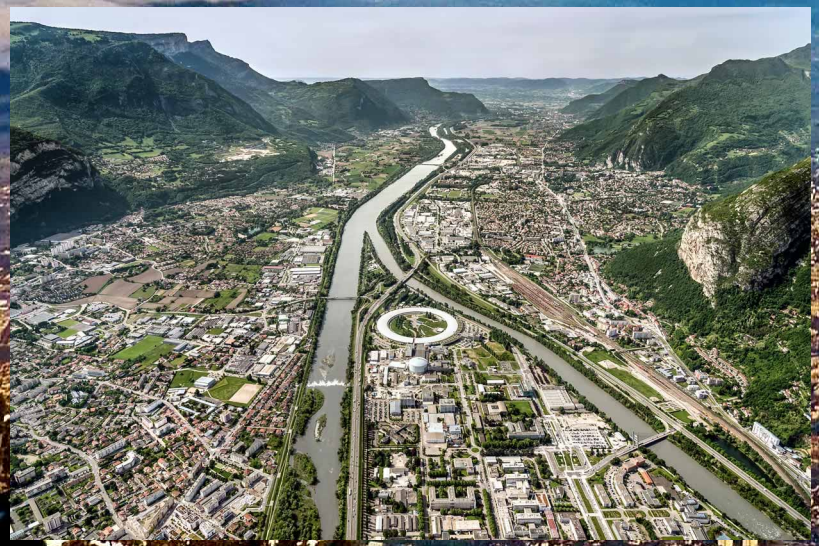
EXTENDING ICAT TO NEW SCIENTIFIC USE CASES

by **Andy Götz** (ESRF) on behalf of the
ICAT Collaboration

ICALEPCS
CAPE TOWN SOUTH AFRICA 2023

EXTENDING ICAT TO NEW SCIENTIFIC USE CASES

by **Andy Götz** (ESRF) on behalf of the ICAT Collaboration



ICALEPCS CAPE TOWN SOUTH AFRICA 2023

TALK OUTLINE

Scientific data lifecycle

ICAT metadata catalogue

Scientific use cases

Role of controls

Conclusions

ACCELERATING SCIENCE

ICALEPCS CAPE TOWN SOUTH AFRICA 2023

HOME COMMITTEE MEMBERS PROGRAMME REGISTRATION AUTHORS EVENTS EXHIBITORS AND SPONSORS KEY DATES TRAVEL CONTACT US

Welcome! Welkom! Wamkelekile!

It is with great pleasure that we welcome everyone to attend the 19th biennial ICALEPCS Conference. This year's conference holds a priceless moment in history for us (especially the organising committee) since it's the first time it's being hosted in Africa since its inception. For some it also presents a nostalgic feeling, being able to travel again after a period of uncertainty on resumption of collocated meetings. As such, this conference offers a fresh start and we get to meet again in person to discuss innovations and fashion collaborations from the various meetings.

In recent decades several instruments have come online to conduct different experiments to test and enable us to glean more insights into theories and improve our understanding of the universe at large. The MeerKAT dishes in South Africa make a star appearance, contributing towards this goal of ground-breaking science, with its highly sensitive receivers for radio wave observations. The need for equally matching control system software in scale and performance cannot be emphasised enough with the arrival of these powerful instruments. This sets the stage to introduce the theme for this year's conference as "Accelerating Control Systems Software for Ground-breaking Science". As we share our improvements and advances in our work with the community, we have a unique learning opportunity to be equipped to produce novel work and support the effort to churn out ground-breaking scientific work.

The South African Radio Astronomy Observatory (SARAO), managed by the National Research Foundation (NRF), is honoured to be hosting this edition of the ICALEPCS Conference. We wish to express our profound gratitude to our sponsors for empowering us to bring our ideas into reality.

"Accelerating Control Systems Software for Ground-breaking Science"

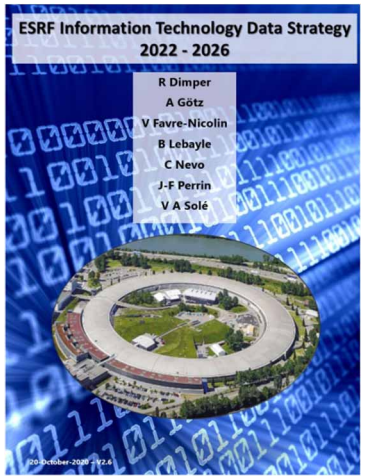
THE DATA LIFECYCLE

Data Processing, Analysis + Management

Controls

EBS DATA PRODUCTION FORECAST

YEAR	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Data production (PB)	2.8	3.4	7.9	4.7	10	20	30	40	50	60
Evolution (%)	-	*+21*	*+132*	*-41*	*+113*	*+100*	*+50*	*+33*	*+25*	*+20*



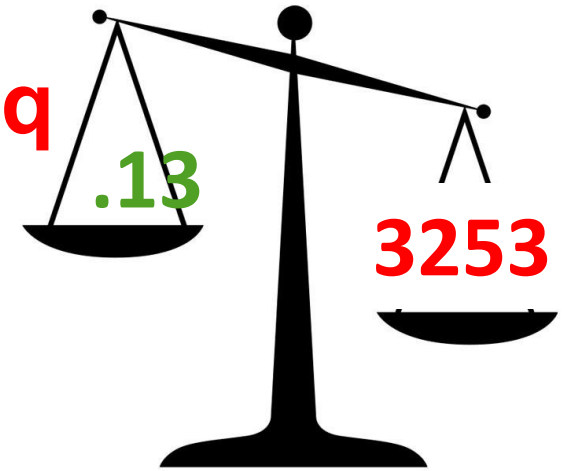
ESRF-EBS DETECTORS

Detector	Type	Max. data rate	# beamlines
	PCO EDGE	1 GB/s	10
	PCO DIMAX	8 GB/s	2
	**Eiger2**	**4 GB/s**	**8**
	PSI Eiger	2 GB/s	2
	Pilatus	2 GB/s	7
	Frelon	1 GB/s	8
	**Jungfrau**	**9 GB/s**	**1**
	Medipix	1 GB/s	6

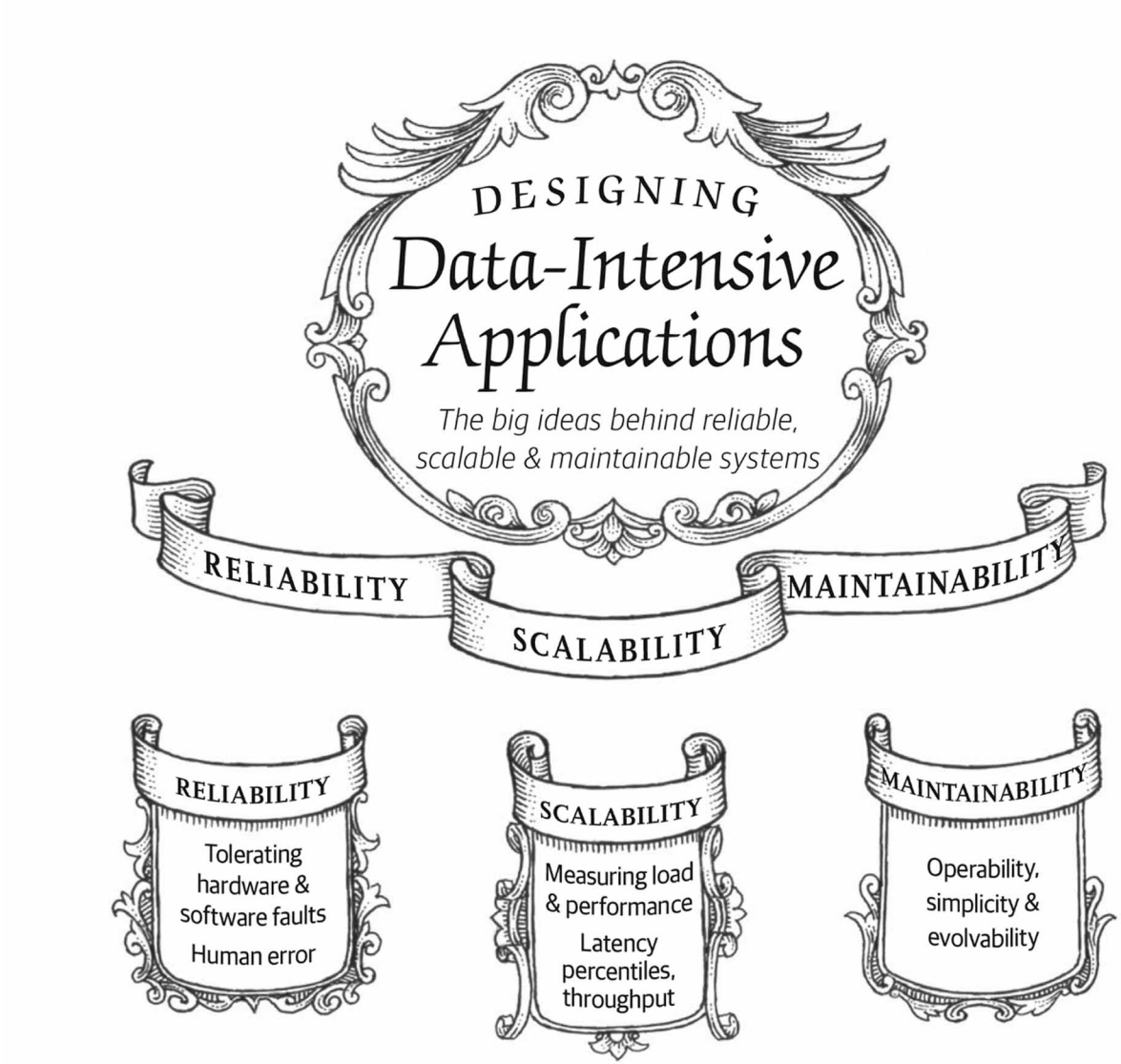
CARBON FOOTPRINT OF DATA

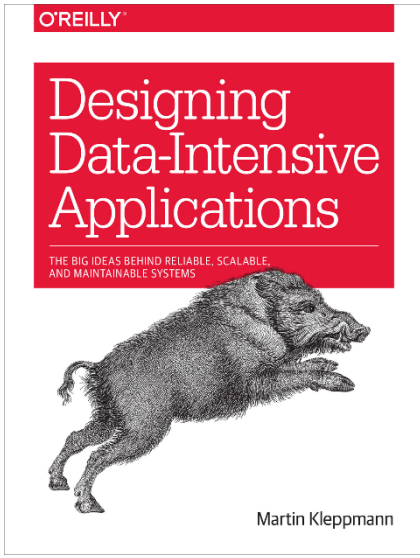
- **1 week experiment of 3 users** flying to ESRF and producing **200 GB** of data **= 3253 kilograms**
- **200 GB Data archived on tape** for **10 years** (full tape library)
  ~ 13 g * 10 yrs = 130 grams

ARCHIVING raw data for 10 years is 0.004 % of CO2eq of beam time to acquire the raw data!



THE DATA LIFECYCLE





WHAT IS ICAT?

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

Science and Technology Facilities Council

ICAT

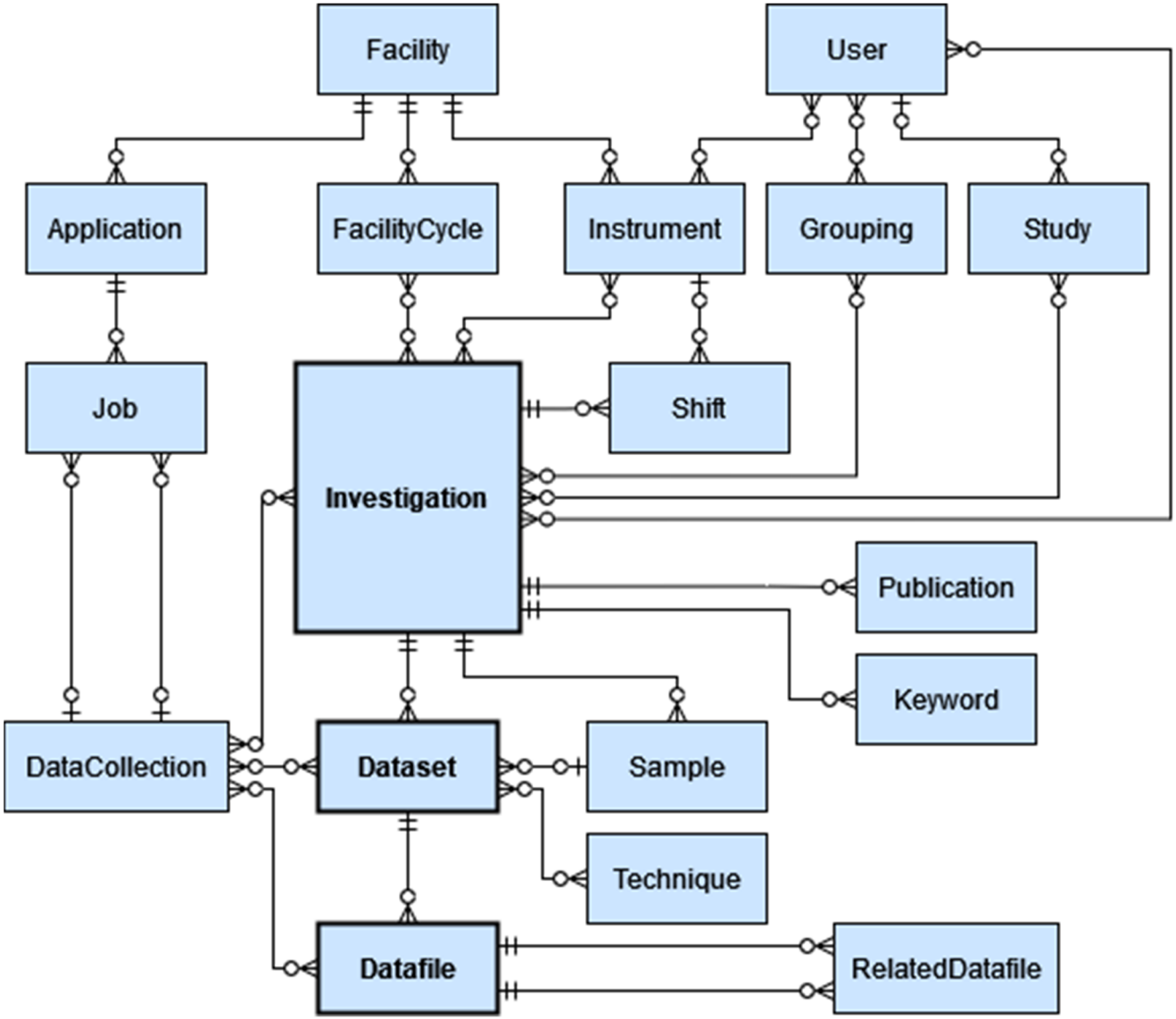
- ICAT supports research data management for **large-scale facilities and is** in production managing billions of datasets + files at ISIS, DLS, ESRF, HZB

- **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, …

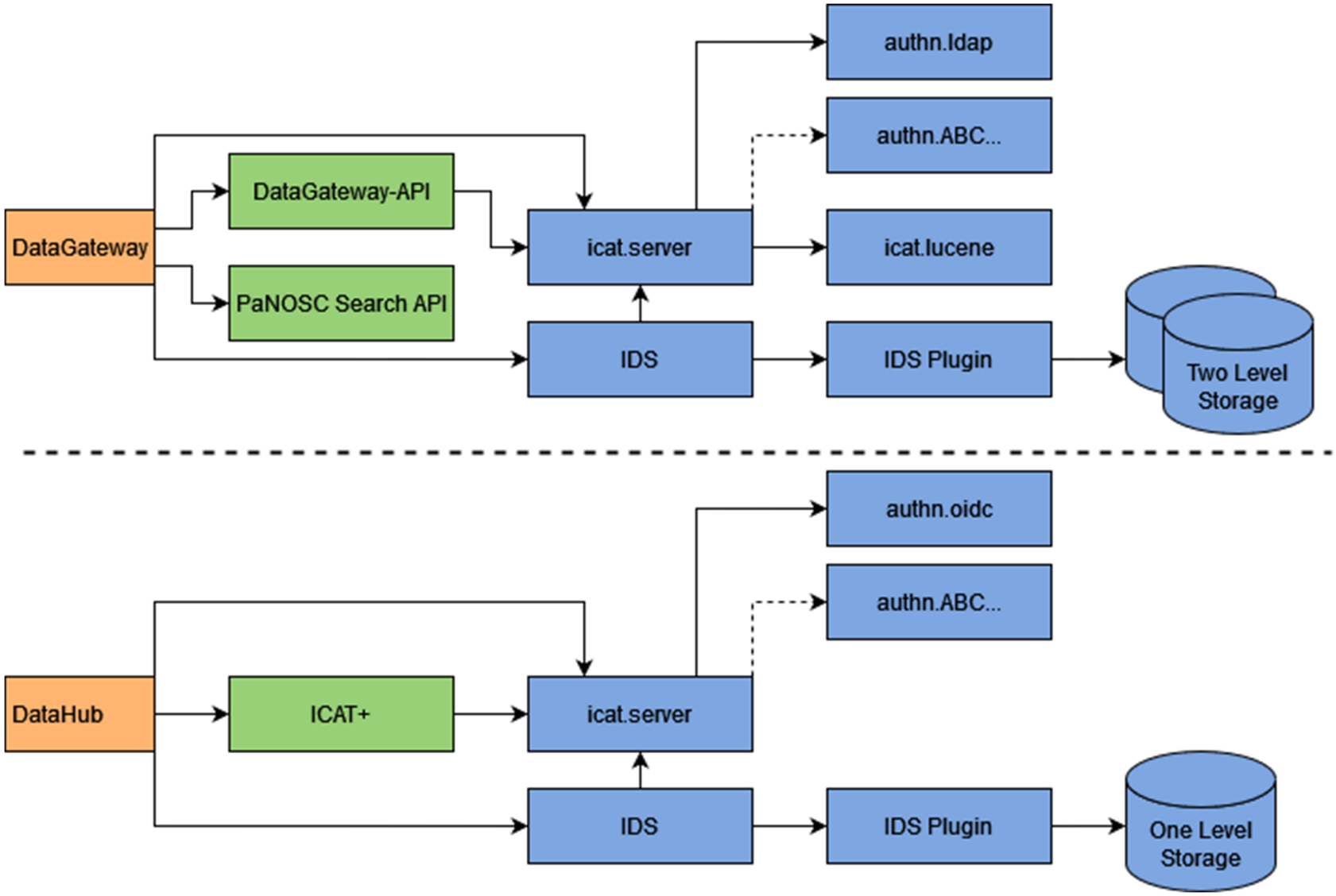
ICAT SITES

- **ISIS:** **1984**, 150 000 experiments, 20 000 000 files
- **DLS:** **2008**, 50 000 experiments, 5 000 000 000 files, **50 petabytes**
- **ESRF:** **2015**, 25 000 experiments, 2 000 000 datasets, 665 000 000 files, **12.4 petabytes**
- **HZB:** deployed in 2019
- **ALBA:** deployed in 2023
- **SESAME:** deployed in 2023
- **SIRIUS:** evaluating in 2023

ICAT CSMD SCHEMAS

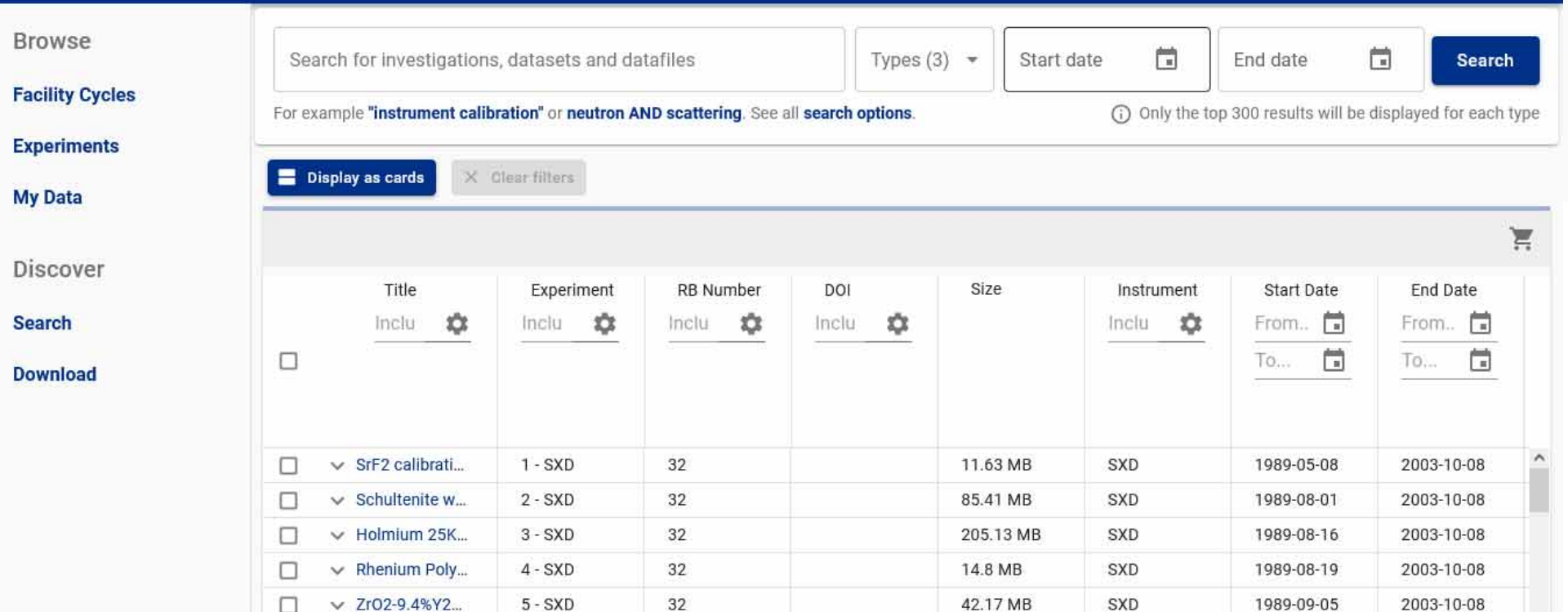


ICAT DEPLOYMENT

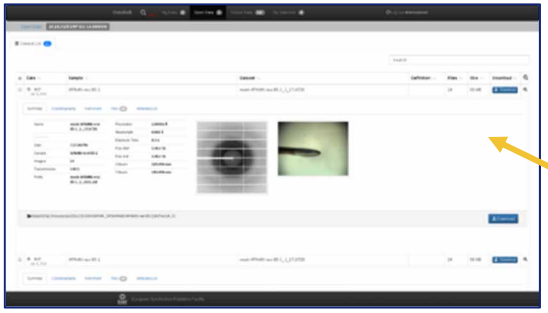


ISIS DATA GATEWAY

ISIS DATA GATEWAY

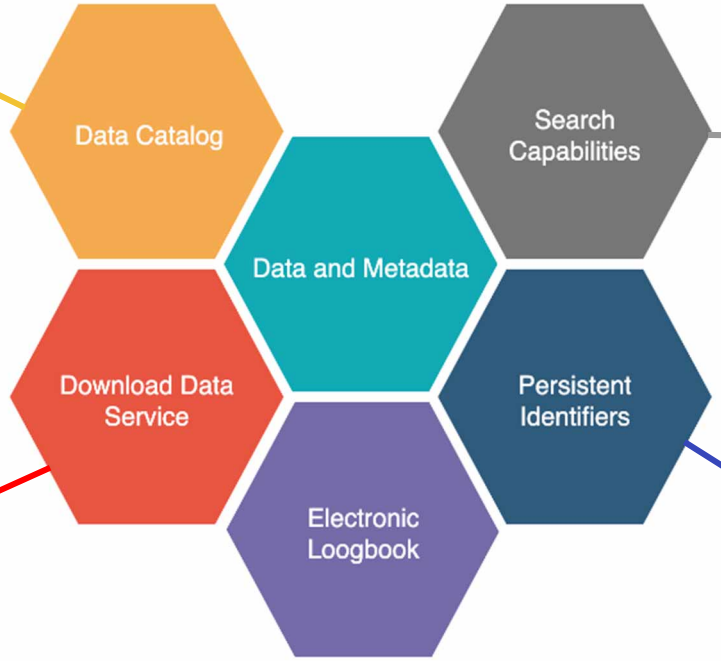


ICAT+ EXTENDS ICAT CORE

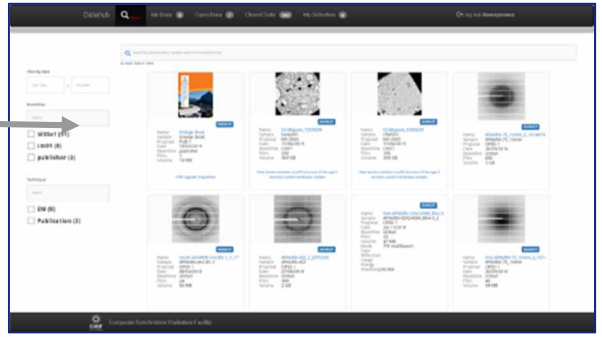


Data Catalog
https://data.esrf.fr

Data Service
Explore and Download data

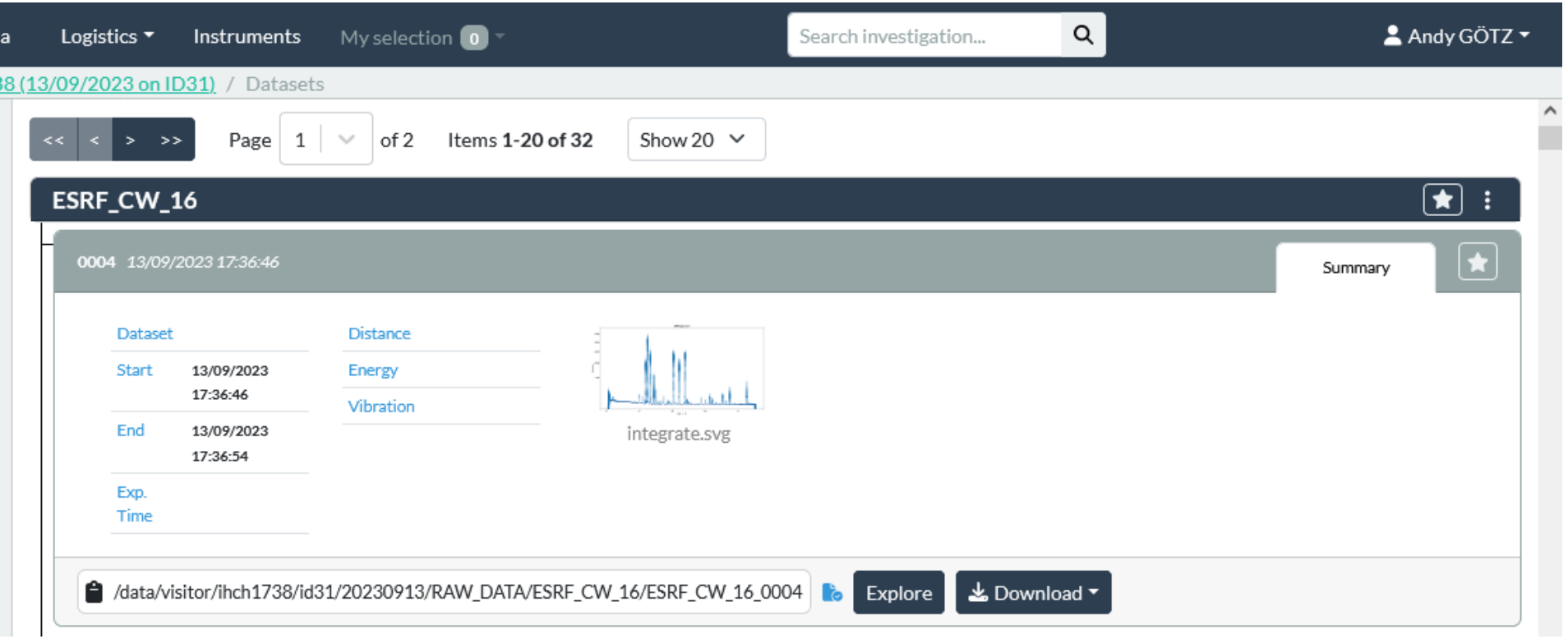


Search Service
Search engine for Big Data

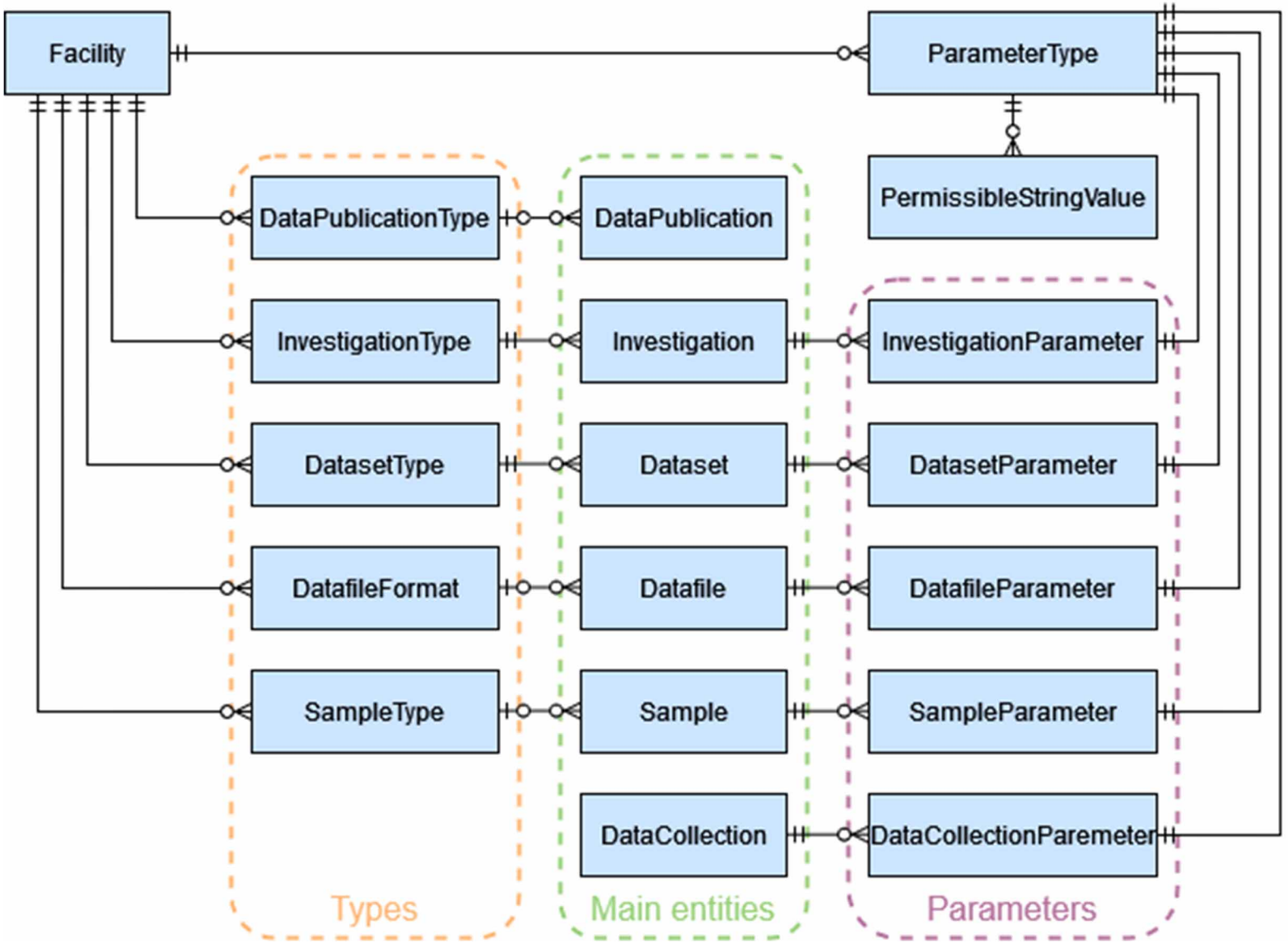


Persistent Identifiers
Make your data findable and searchable

ACCELERATING PD SCIENCE



ICAT PARAMETERS == DICTIONARIES



METADATA – NEXUS + ENHANCEMENTS

1. ESRF-EBS has adopted **Nexus** as standard vocabulary
2. Where **Nexus** is missing definitions we have defined new keywords following Nexus conventions
3. See **https://gitlab.esrf.fr/icat/hdf5-master-config/-/blob/master/hdf5_cfg.xml**

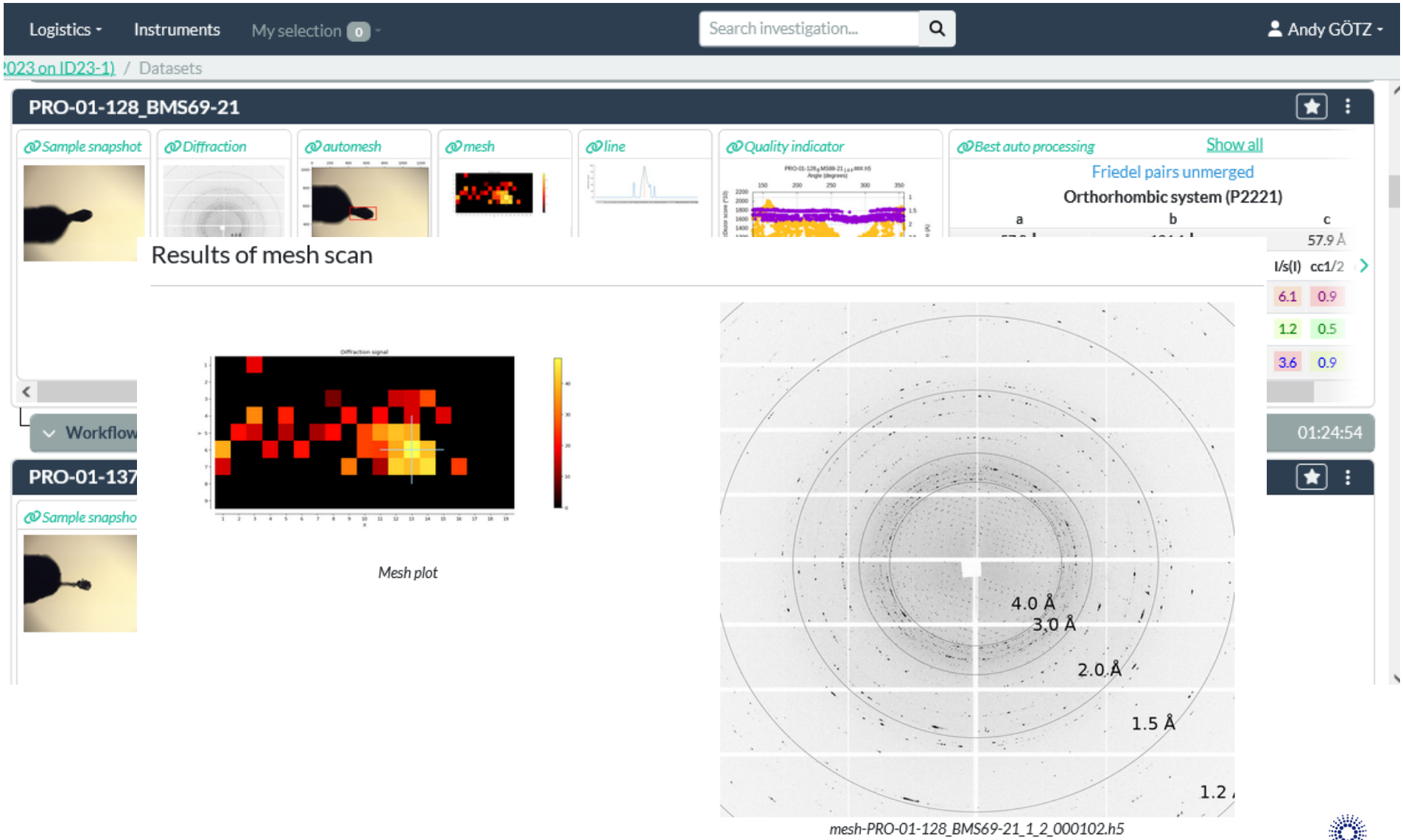
hdf5_cfg.xml 89.6 KB

Edit | Web IDE

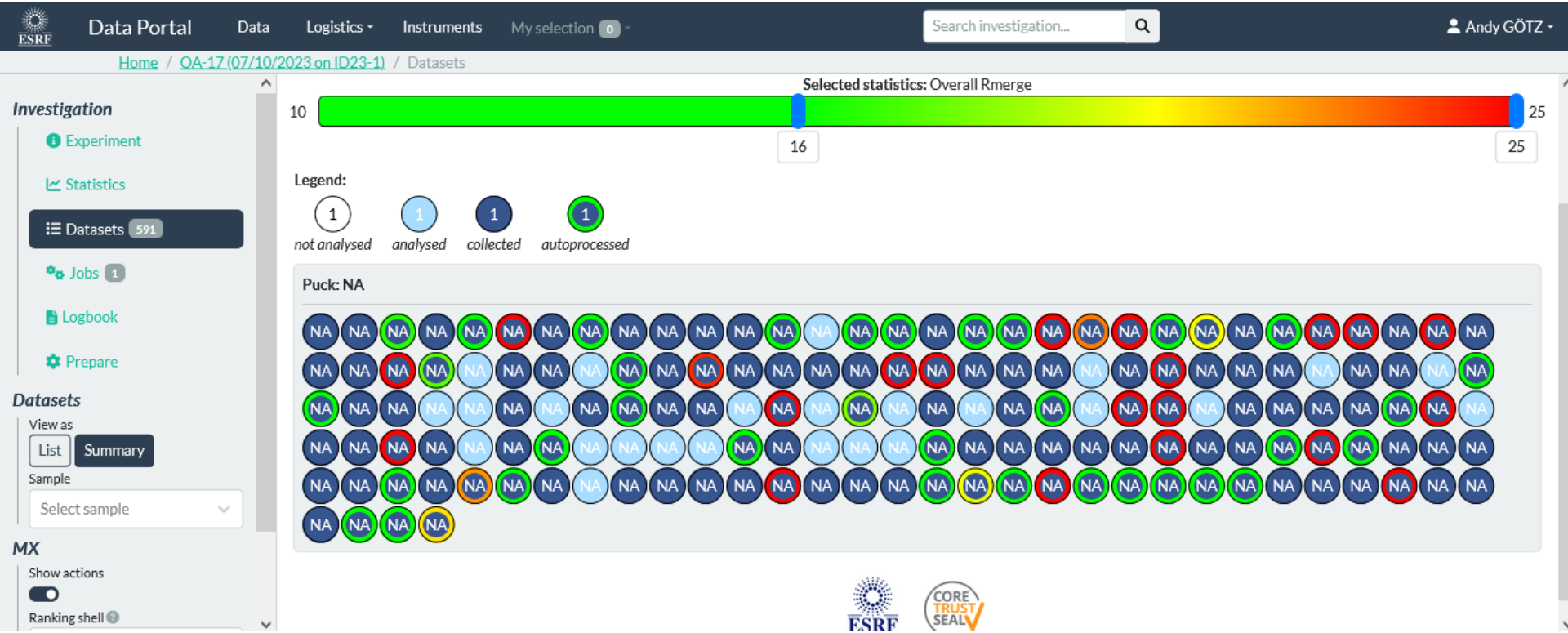
```
<?xml version="1.0" encoding="UTF-8"?>
<group NX_class="NXentry" groupName="${entry}">
   <title ESRF_description="Name of the dataset" ESRF_mandatory="Mandatory" NAPItype="NX_CHAR">${scanName}</title>
   <scanNumber ESRF_description="Scan number" ESRF_mandatory="Mandatory" NAPItype="NX_CHAR">${scanNumber}</scanNumber>
   <proposal ESRF_description="Proposal code" ESRF_mandatory="Mandatory" NAPItype="NX_CHAR">${proposal}</proposal>
   <dataset_type ESRF_description="Scan type can be 'step_by_step' or 'continuous'&amp;#xA;&amp;#x9;&amp;#x9;" NAPItype="NX_CHAR">${scanType}</dataset
   <folder_path ESRF_description="Scan starting date" ESRF_mandatory="Mandatory" NAPItype="NX_CHAR">${location}</folder_path>
   <start_time ESRF_description="Scan starting date" ESRF_mandatory="Mandatory" NAPItype="NX_DATE_TIME">${startDate}</start_time>
   <end_time ESRF_description="Scan ending date" record="final" ESRF_mandatory="Mandatory" NAPItype="NX_DATE_TIME">${endDate}</end_time>
   <definition ESRF_description="Techniques used to collect this dataset" NAPItype="NX_CHAR">${definition}</definition>
   <group NX_class="NXsubentry" groupName="SAXS">
      <definition ESRF_description="Technique used to collect this dataset" NAPItype="NX_CHAR">${saxs_definition}</definition>
      <version ESRF_description="Version" NAPItype="NX_CHAR">${saxs_definition.version}</version>
      <directory record="final" ESRF_description="Data collection directory" NAPItype="NX_CHAR">${SAXS_directory}</directory>
      <experimentType record="final" ESRF_description="Type of experiment" NAPItype="NX_CHAR">${SAXS_experimentType}</experimentType>
      <runNumber record="final" ESRF_description="Run number" NAPItype="NX_CHAR">${SAXS_runNumber}</runNumber>
      <prefix record="final" ESRF_description="" NAPItype="NX_CHAR">${SAXS_prefix}</prefix>
      <maskFile record="final" ESRF_description="" NAPItype="NX_CHAR">${SAXS_maskFile}</maskFile>
      <numberFrames record="final" ESRF_description="" NAPItype="NX_CHAR">${SAXS_numberFrames}</numberFrames>
      <timePerFrame record="final" ESRF_description="" NAPItype="NX_CHAR">${SAXS_timePerFrame}</timePerFrame>
      <concentration record="final" ESRF_description="" NAPItype="NX_CHAR">${SAXS_concentration}</concentration>
      <comments record="final" ESRF_description="" NAPItype="NX_CHAR">${SAXS_comments}</comments>
      <code record="final" ESRF_description="" NAPItype="NX_CHAR">${SAXS_code}</code>
```

ACCELERATING MX SCIENCE

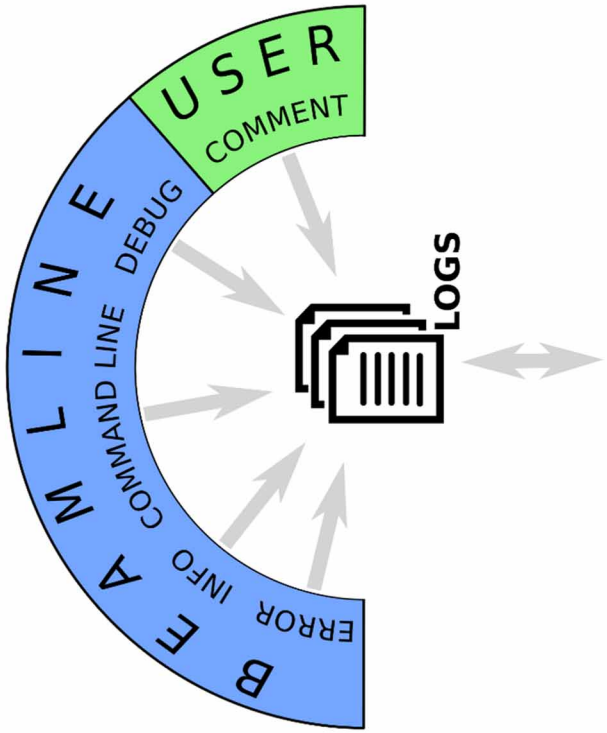
ACCELERATING MX SCIENCE

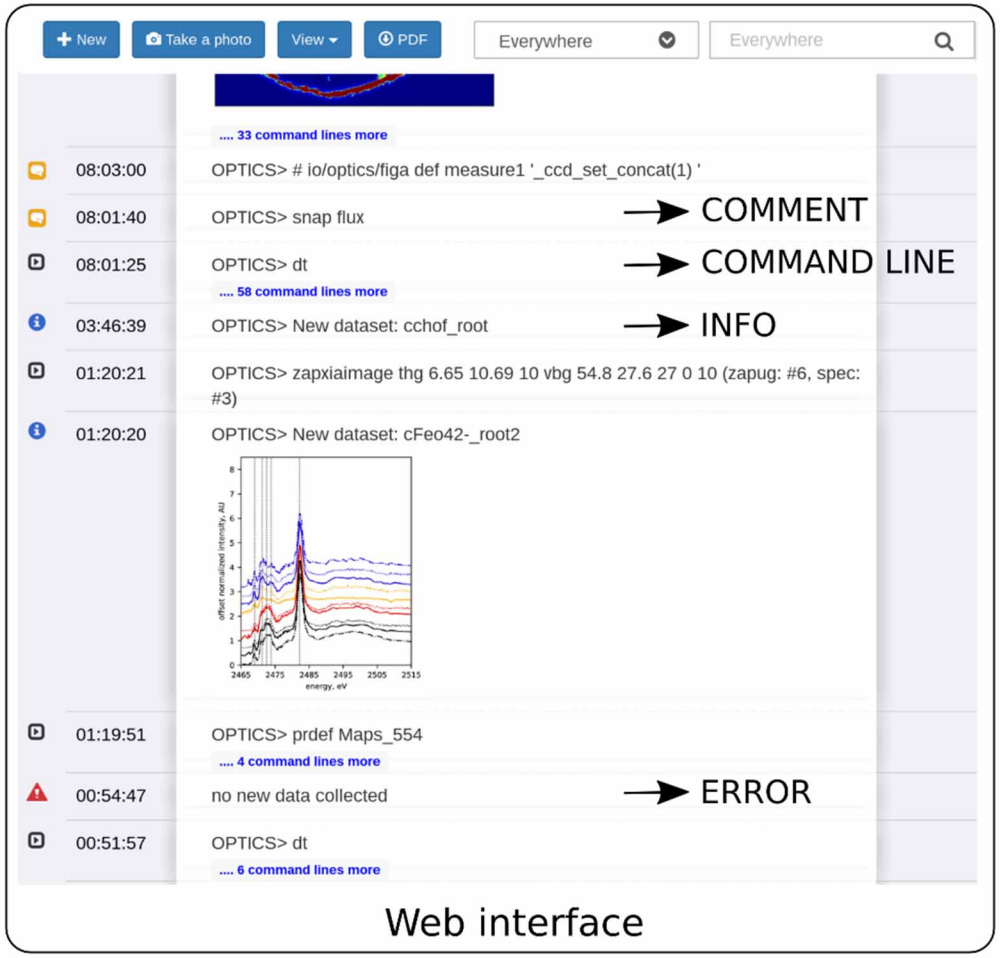


ACCELERATING MX SCIENCE



ELECTRONIC LOGBOOK





ELECTRONIC LOGBOOK

Data Portal | Data | Logistics | Instruments | My selection 0 | Search investigation...

Home / EV-351 (04/10/2022 on ID21) / Logbook

Investigation

- Experiment
- Statistics
- Datasets 65
- Logbook
- Prepare

Logbook

Search

Search

Type

- [x] Comments
- [ ] Information
- [ ] Errors
- [ ] Command Lines
- [ ] Machine

Order

- (•) Newest First
- ( ) Oldest First

Date

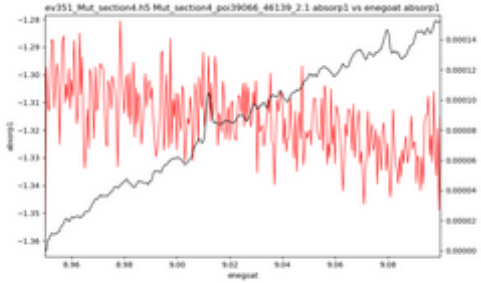
Select date

iodet = 200847. (2.00847e+06 /s) p201 gain = 7

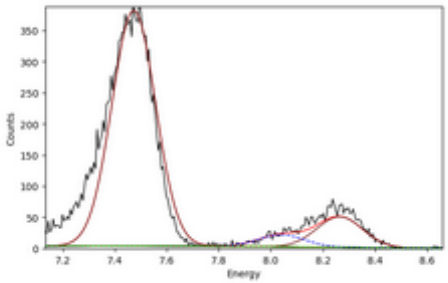
idet = 681803. (6.81803e+06 /s) p201 gain = 6

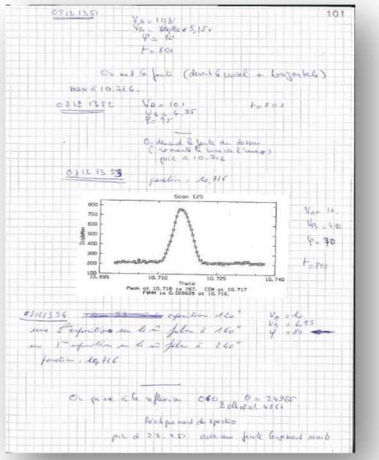
XANES scan on hole to check if it is clean for XANES. NO holder

fsopen(); sct(1, p201, fx2,calc_fx2); fsclose()



fsopen(); sct(1, p201, fx2,calc_fx2); fsclose()





ACCELERATING SCIENCE

of 11 — + Automatic Zoom

ESRF

test of new optic, commissioning of new filters, characterization of inline monochromator

BM18
BLC-14794

[15:49:29] [2023-09-01] 2 days of conditionning of filters in OH1 with beam.

The different axis are moving back and forth in the beam.

att1 chamber directly fine at full power.

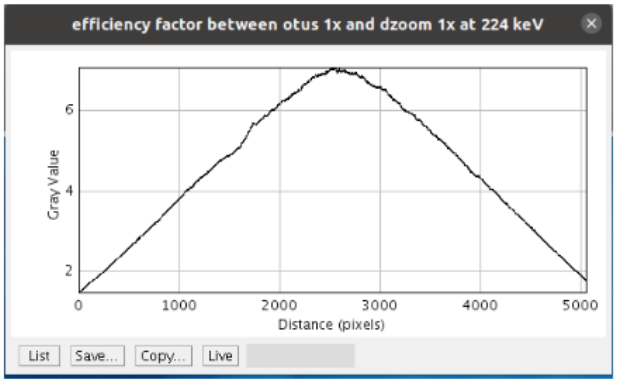
Att2, basically fine, but with sudden pressure peaks that close the valves. It tens to decrease in frequency.

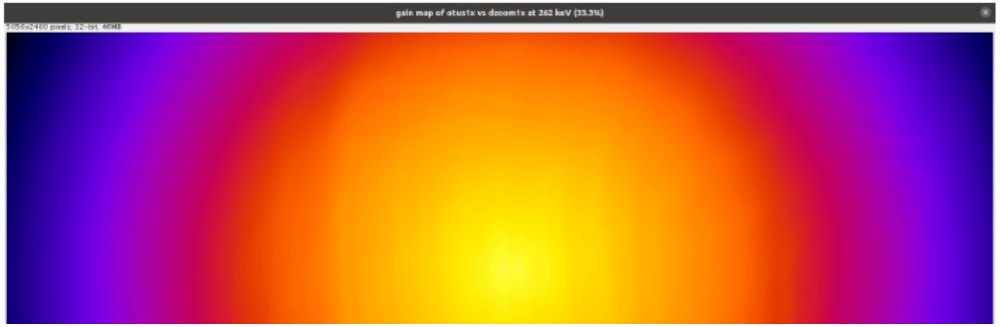
Att3, much more degassing, espeically on t
with continuous movement of the axis. Anyt
days and 2 nights, it is not really decreasing
used in monochromatic only, it should be fir

[11:20:43] [2023-09-01] test of quinary slits afte
direction. The slice beam seems very precis

At next shutdown, the limit switch should be

— + 100%





FAIR DATA POLICY – MEANS

- **Findable means implementing**
  - **Digital Object Identifiers** for datasets (DOIs), **long term data archiving**, **metadata** standard (**Nexus/HDF5**), metadata **catalogue** (**ICAT**), **search API** (OAI-PMH/**PaNOSC**), data portal (**ICAT+**)

- **Accessible means implementing**
  - **Data Policy** (esrf.eu/datapolicy), data made available under a **licence** (CC), **download** protocol (e.g. http, rsync or globus)

- **Interoperable means implementing**
  - Standard **metadata** vocabulary (**Nexus**), standard data **format** (**HDF5**), use community standards (Gold Standard)

- **Guidelines available from - PaNOSC, ExPaNDS, FAIRsFAIR, RDA, EOSC**

FAIR DATA READY FOR REUSE

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

CONCLUSION

- **ICAT** is a **stable, scalable** and **maintainable metadata catalogue solution** for **very large data sources**

- **ICAT** has **multiple frontends** and **extensions** which allows it to adapt to easily to **complex scientific use cases** with **dictionaries of parameters e.g. MX, CryoEM, powder, spectroscopy, tomography …**

- **Scientists** need **processed data** and **advanced web interfaces** to find, explore and exploit them

- **FAIR DATA** needs **good metadata** from the **control system** and **good sample metadata + e-logbooks**

ACKNOWLEDGEMENTS

- **The ICAT collaboration**
- **ALL maintainers of ICAT at STFC/CSD**
- **ALL colleagues at ESRF** with special thanks to the following people for their work developing ICAT:

Alejandro de Maria (data manager + web)
Marjolaine Bodin (data manager + web)
Mael Gaonach (frontend developer)
Max Nanao (scientist)
Romain Talon (scientist)
Didier Nurizzo (scientist)