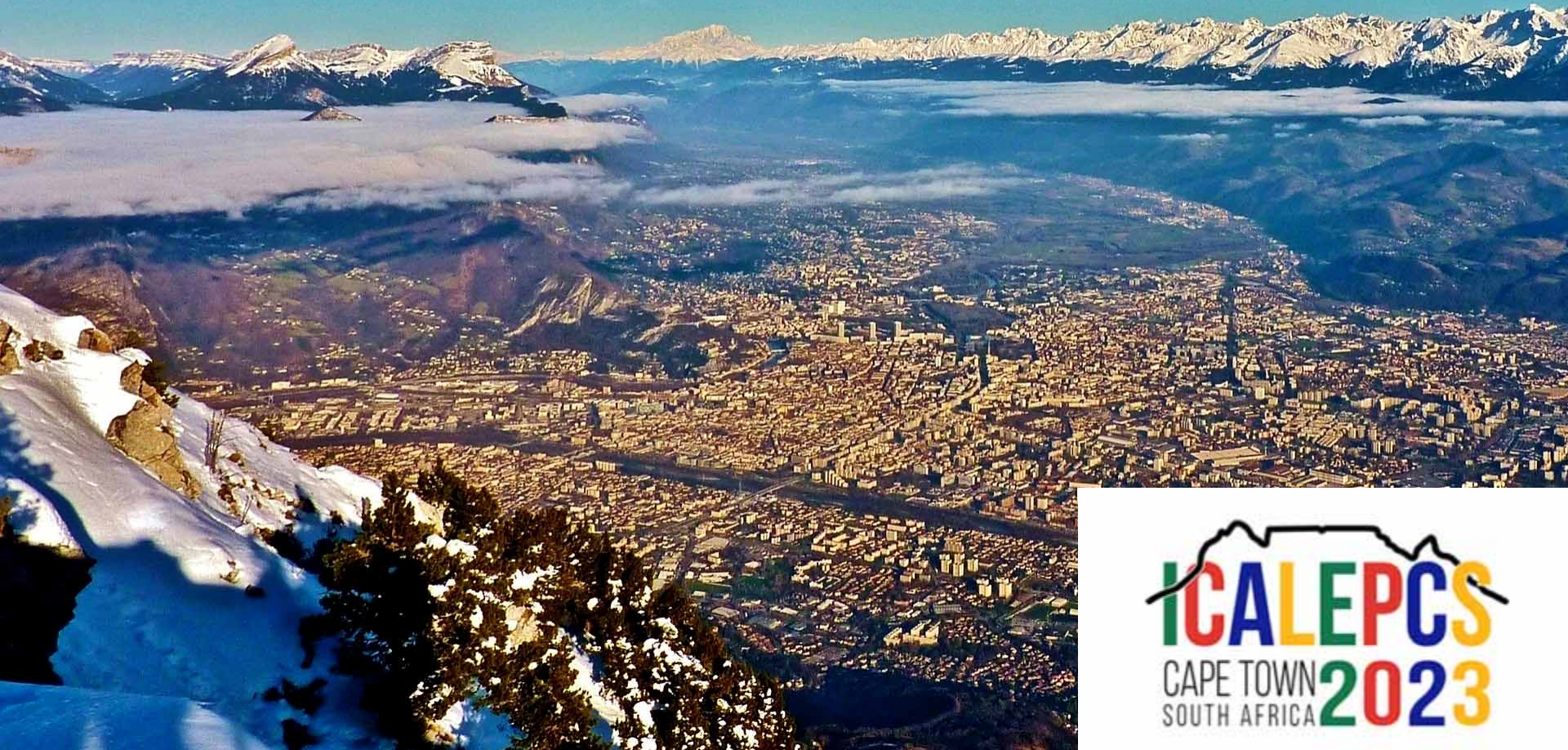


# EXTENDING ICAT TO NEW SCIENTIFIC USE CASES

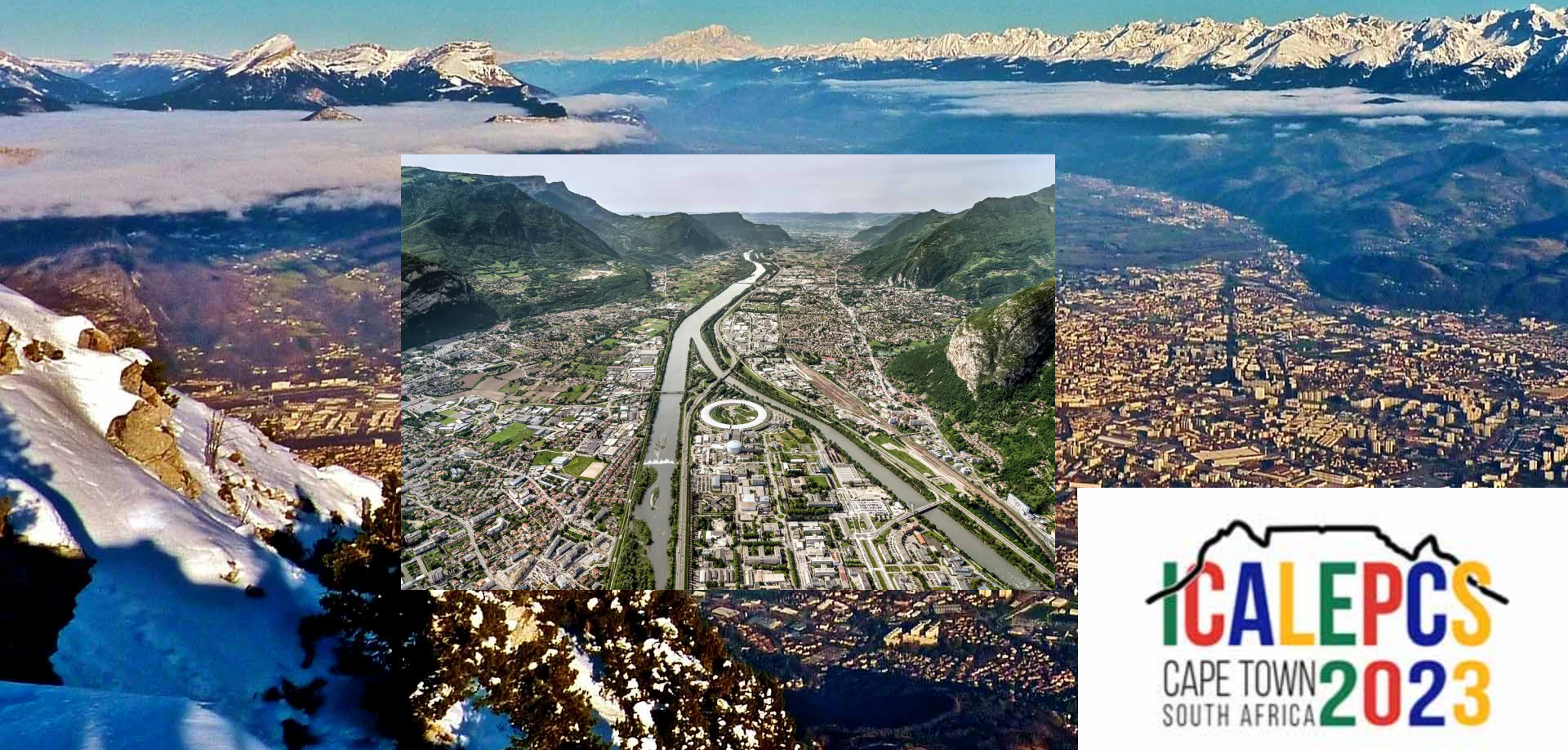
by **Andy Götz** (ESRF) on behalf of the  
ICAT Collaboration





# EXTENDING ICAT TO NEW SCIENTIFIC USE CASES

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



# TALK OUTLINE

**Scientific data lifecycle**

**ICAT metadata catalogue**

**Scientific use cases**

**Role of controls**

**Conclusions**





# ACCELERATING SCIENCE



[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**

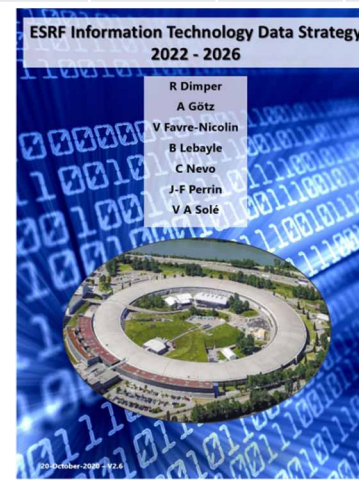
**Software for Ground-breaking Science”**

# THE DATA LIFECYCLE











# 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



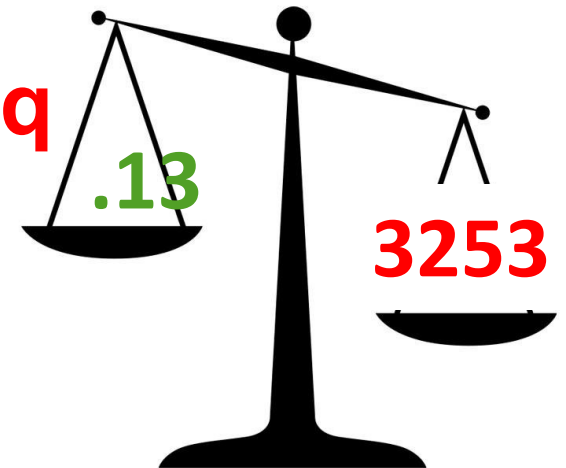
# ESRF-EBS DETECTORS

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

# CARBON FOOTPRINT OF DATA

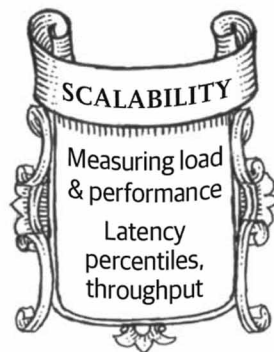
- 1 week experiment of 3 users flying to ESRF and producing **200 GB** of data = 3253 kilograms
- 200 GB Data archived on tape for 10 years (full tape library )  
~ 13 g \* 10 yrs = 130 grams

**ARCHIVING raw data for 10 years is 0.004 % of CO<sub>2</sub>eq of beam time to acquire the raw data!**





# THE DATA LIFECYCLE



O'REILLY

## Designing Data-Intensive Applications

THE BIG IDEAS BEHIND RELIABLE, SCALABLE,  
AND MAINTAINABLE SYSTEMS



Martin Kleppmann

# 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, ...

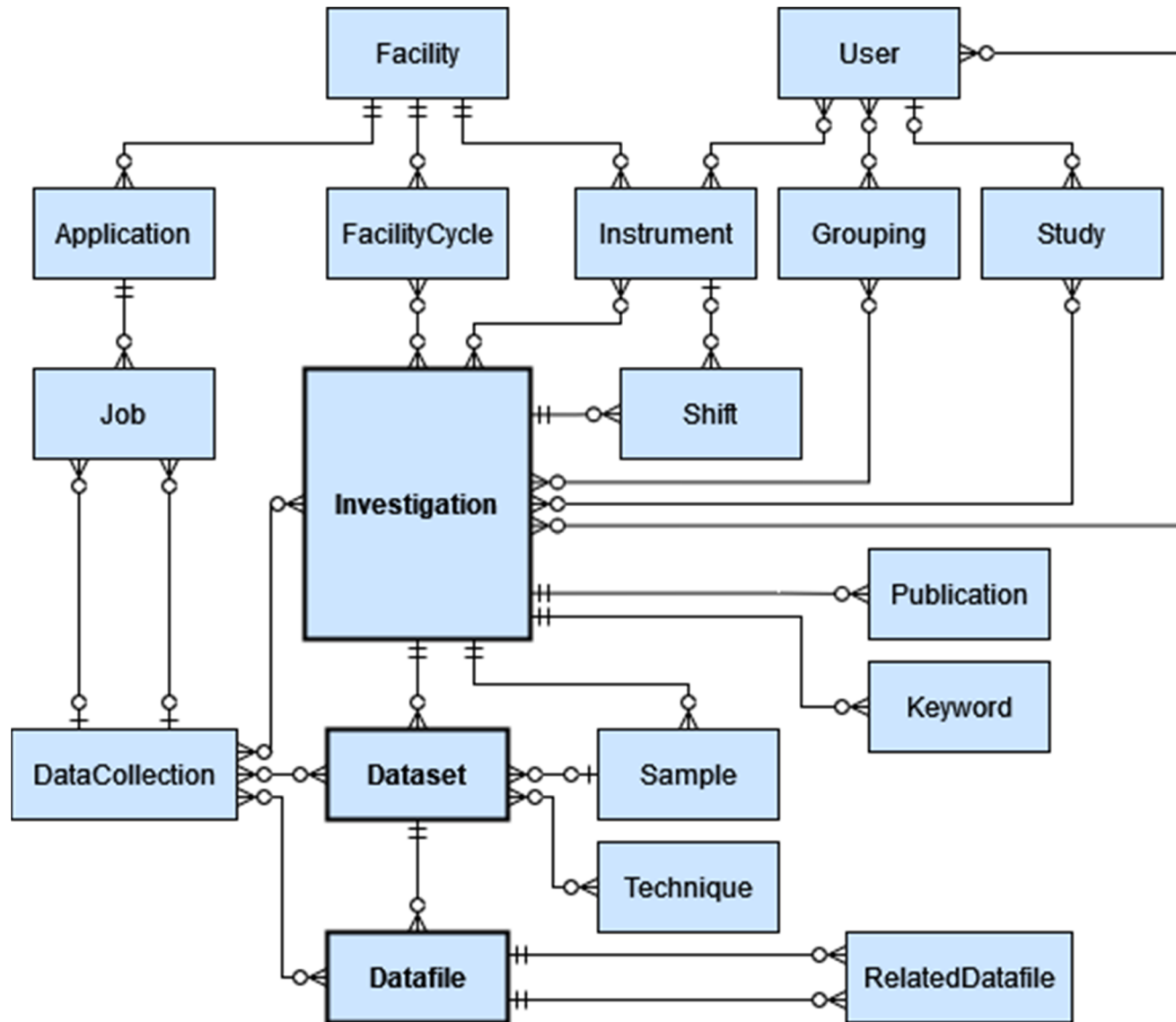


# ICAT SITES

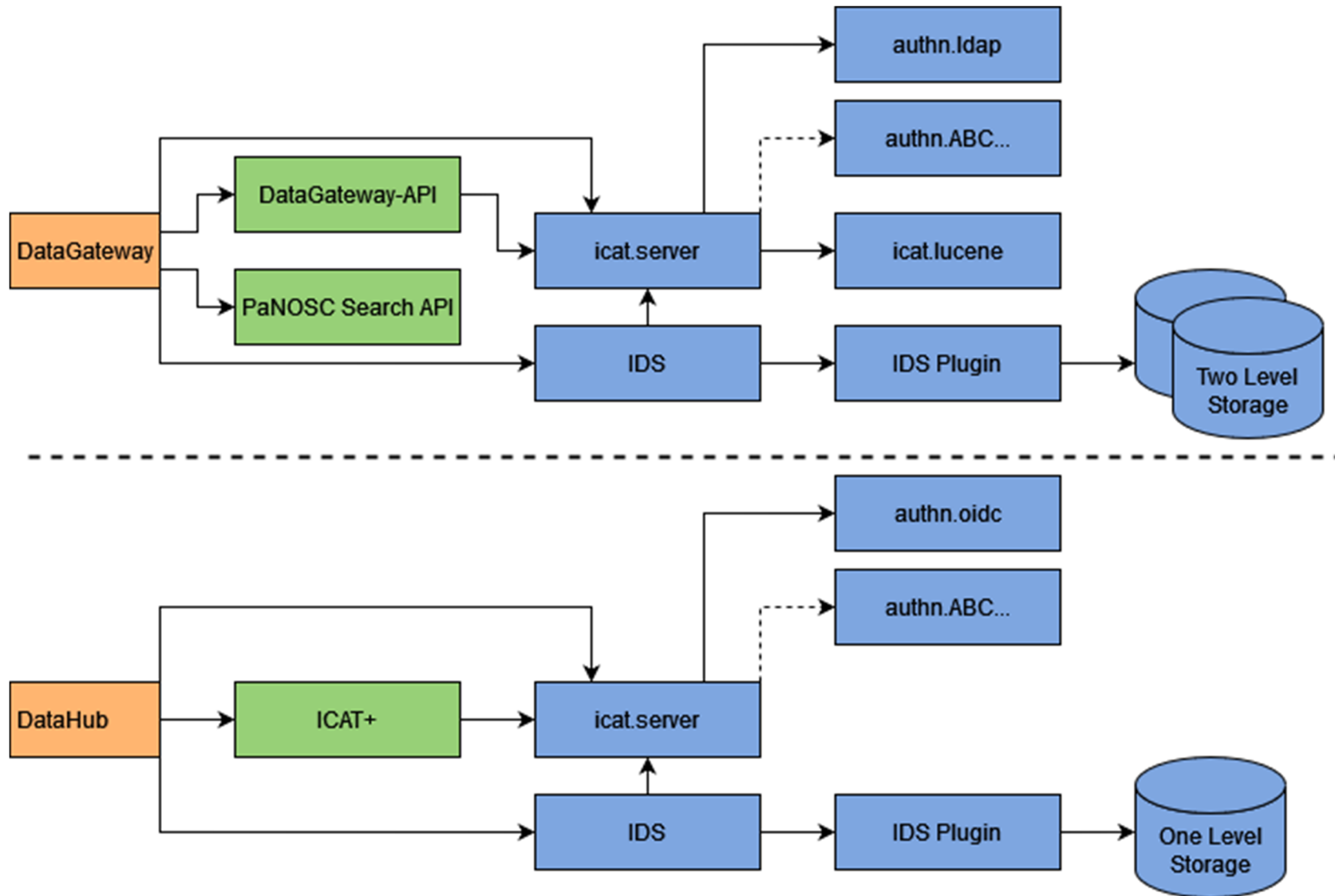
- **ISIS:** 1984, 150 000 experiments, 20 000 000 files
- **DLS:** 2008, 50 000 experiments, 5 000 000 000 files, **50 petabytes**
- **ESRF:** 2015, 25 000 experiments, 2 000 000 datasets, 665 000 000 files, **12.4 petabytes**
- **HZB:** deployed in 2019
- **ALBA:** deployed in 2023
- **SESAME:** deployed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



# ICAT CSMD SCHEMAS



# ICAT DEPLOYMENT



Browse

Facility Cycles

Experiments

My Data

Discover

Search

Download

## Data discovery and access for large-scale science facilities

### Browse, explore and visualise experimental data

Large scale facilities, such as synchrotrons, neutron and muon sources, lasers and accelerators, generate vast amounts of data that need to be managed in an efficient way, supporting data ingestion for long-term storage and archival, as well as data analysis and data publication workflows.

DataGateway provides a unified data discovery and data





# ISIS DATA GATEWAY

- Browse
- Facility Cycles
- Experiments
- My Data
- Discover
- Search



- Browse
- Facility Cycles
- Experiments
- My Data
- Discover
- Search
- Download

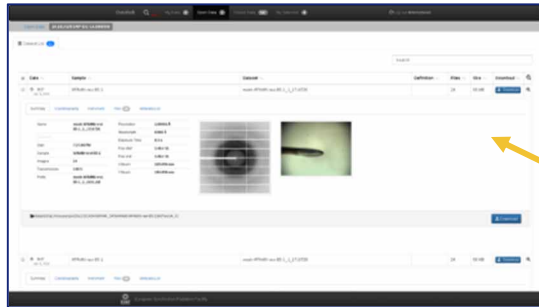
Types (3) ▾ Start date  End date  [Search](#)

For example "instrument calibration" or neutron AND scattering. See all [search options](#). 🔔 Only the top 300 results will be displayed for each type

[Display as cards](#) [Clear filters](#)

	Title <a href="#">Inclu</a>	Experiment <a href="#">Inclu</a>	RB Number <a href="#">Inclu</a>	DOI <a href="#">Inclu</a>	Size	Instrument <a href="#">Inclu</a>	Start Date From.. To...	End Date From.. To...
<input type="checkbox"/>	▼ SrF2 calibrati...	1 - SXD	32		11.63 MB	SXD	1989-05-08	2003-10-08
<input type="checkbox"/>	▼ Schultenite w...	2 - SXD	32		85.41 MB	SXD	1989-08-01	2003-10-08
<input type="checkbox"/>	▼ Holmium 25K...	3 - SXD	32		205.13 MB	SXD	1989-08-16	2003-10-08
<input type="checkbox"/>	▼ Rhenium Poly...	4 - SXD	32		14.8 MB	SXD	1989-08-19	2003-10-08
<input type="checkbox"/>	▼ ZrO2-9.4%Y2...	5 - SXD	32		42.17 MB	SXD	1989-09-05	2003-10-08

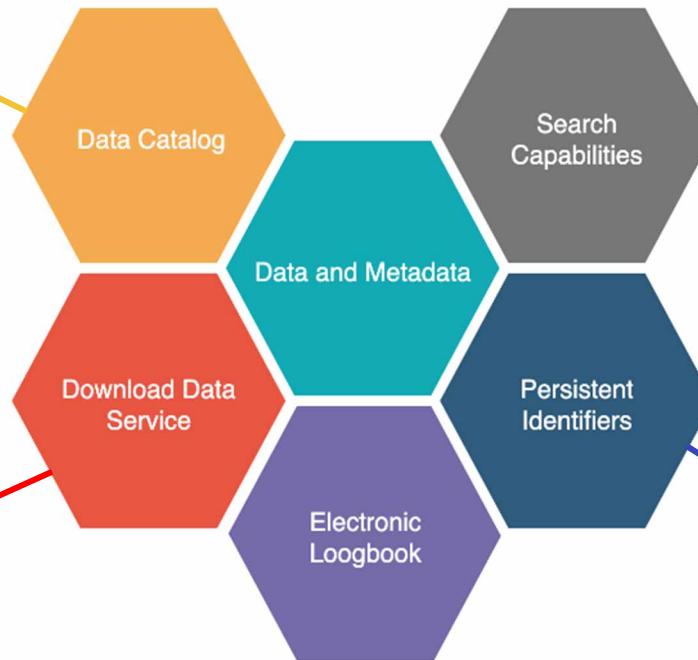
# ICAT+ EXTENDS ICAT CORE



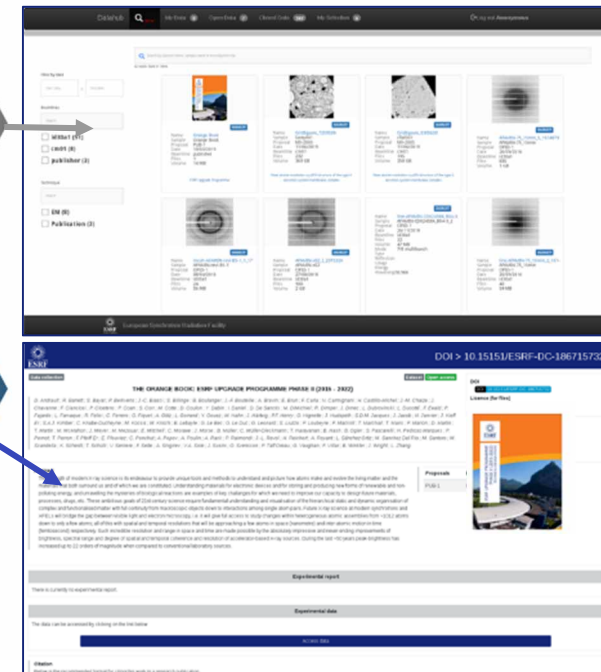
**Data Catalog**  
<https://data.esrf.fr>



**Data Service**  
Explore and Download data



## Search Service Search engine for Big Data



**Persistent Identifiers**  
Make your data findable and searchable

# ACCELERATING PD SCIENCE

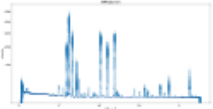
Logistics Instruments My selection 0 Search investigation... Andy GÖTZ

18 (13/09/2023 on ID31) / Datasets

<< < > >> Page 1 of 2 Items 1-20 of 32 Show 20

## ESRF\_CW\_16

0004 13/09/2023 17:36:46 Summary

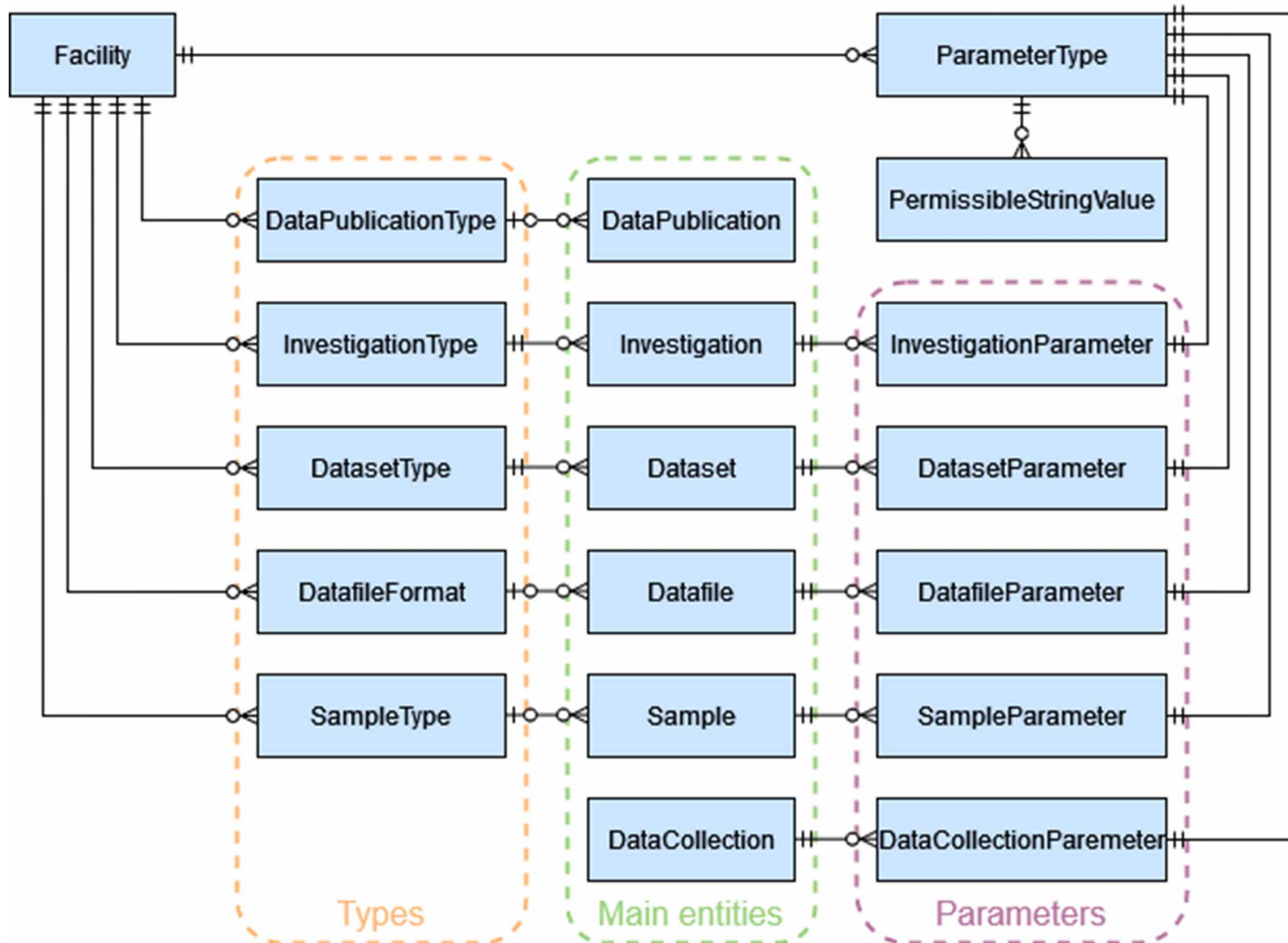
<b>Dataset</b>	<b>Distance</b>	
<b>Start</b>	<b>Energy</b>	
<b>End</b>	<b>Vibration</b>	
<b>Exp. Time</b>		

Start 13/09/2023 17:36:46  
End 13/09/2023 17:36:54  
Exp. Time

[/data/visitor/ihch1738/id31/20230913/RAW\\_DATA/ESRF\\_CW\\_16/ESRF\\_CW\\_16\\_0004](/data/visitor/ihch1738/id31/20230913/RAW_DATA/ESRF_CW_16/ESRF_CW_16_0004) Explore Download



# 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](https://gitlab.esrf.fr/icat/hdf5-master-config/-/blob/master/hdf5_cfg.xml)

```
hdf5_cfg.xml 89.6 KB
Edit Web IDE
1 <?xml version="1.0" encoding="UTF-8"?>
2 <group NX_class="NXentry" groupName="{entry}">
3   <title ESRF_description="Name of the dataset" ESRF_mandatory="Mandatory" NAPitype="NX_CHAR">${scanName}</title>
4   <scanNumber ESRF_description="Scan number" ESRF_mandatory="Mandatory" NAPitype="NX_CHAR">${scanNumber}</scanNumber>
5   <proposal ESRF_description="Proposal code" ESRF_mandatory="Mandatory" NAPitype="NX_CHAR">${proposal}</proposal>
6   <dataset_type ESRF_description="Scan type can be 'step_by_step' or 'continuous'&#xA;&#x9;&#x9;" NAPitype="NX_CHAR">${scanType}</dataset
7   <folder_path ESRF_description="Scan starting date" ESRF_mandatory="Mandatory" NAPitype="NX_CHAR">${location}</folder_path>
8   <start_time ESRF_description="Scan starting date" ESRF_mandatory="Mandatory" NAPitype="NX_DATE_TIME">${startDate}</start_time>
9   <end_time ESRF_description="Scan ending date" record="final" ESRF_mandatory="Mandatory" NAPitype="NX_DATE_TIME">${endDate}</end_time>
10  <definition ESRF_description="Techniques used to collect this dataset" NAPitype="NX_CHAR">${definition}</definition>
11  <group NX_class="NXsubentry" groupName="SAXS">
12    <definition ESRF_description="Technique used to collect this dataset" NAPitype="NX_CHAR">${saxs_definition}</definition>
13    <version ESRF_description="Version" NAPitype="NX_CHAR">${saxs_definition.version}</version>
14    <directory record="final" ESRF_description="Data collection directory" NAPitype="NX_CHAR">${SAXS_directory}</directory>
15    <experimentType record="final" ESRF_description="Type of experiment" NAPitype="NX_CHAR">${SAXS_experimentType}</experimentType>
16    <runNumber record="final" ESRF_description="Run number" NAPitype="NX_CHAR">${SAXS_runNumber}</runNumber>
17    <prefix record="final" ESRF_description="" NAPitype="NX_CHAR">${SAXS_prefix}</prefix>
18    <maskFile record="final" ESRF_description="" NAPitype="NX_CHAR">${SAXS_maskFile}</maskFile>
19    <numberFrames record="final" ESRF_description="" NAPitype="NX_CHAR">${SAXS_numberFrames}</numberFrames>
20    <timePerFrame record="final" ESRF_description="" NAPitype="NX_CHAR">${SAXS_timePerFrame}</timePerFrame>
21    <concentration record="final" ESRF_description="" NAPitype="NX_CHAR">${SAXS_concentration}</concentration>
22    <comments record="final" ESRF_description="" NAPitype="NX_CHAR">${SAXS_comments}</comments>
23    <code record="final" ESRF_description="" NAPitype="NX_CHAR">${SAXS_code}</code>
```

# ACCELERATING MX SCIENCE

Logistics Instruments My selection 0 Search investigation... Andy GÖTZ

023 on ID23-1 / Datasets

**PRO-01-128\_BMS69-21** ★ ⋮

**Sample snapshot** **Diffraction** **automesh** **mesh** **line** **Quality indicator**

**Best auto processing** [Show all](#)

Friedel pairs unmerged  
Orthorhombic system (P2221)

	a	b	c			
	57.8 Å	134.4 Å	57.9 Å			
	Compl.	Res. low	Res. high	Rmerge	I/s(I)	cc1/2
inner	98.4%	134.4	8.8	65.7	6.1	0.9
outer	54.8%	2.4	2.3	108.9	1.2	0.5
overall	94.5%	134.4	2.3	50.0	3.6	0.9

Workflow MXPressF: X-centre + fbest + 180 degree dc on id23eh1 4 actions 01:24:54

**PRO-01-137\_BMS69-21** ★ ⋮

**Sample snapshot** **Diffraction** **automesh** **mesh** **line** **Quality indicator**

# ACCELERATING MX SCIENCE

Logistics Instruments My selection 0 Search investigation... Andy GÖTZ

023 on ID23-1 / Datasets

## PRO-01-128\_BMS69-21

Sample snapshot Diffraction automesh mesh line Quality indicator Best auto processing Show all

Friedel pairs unmerged  
Orthorhombic system (P2221)

	a	b	c
			57.9 Å
I/s(l)	6.1	0.9	
cc1/2	1.2	0.5	
	3.6	0.9	

01:24:54

Results of mesh scan

Diffraction signal

Mesh plot

mesh-PRO-01-128\_BMS69-21\_1\_2\_000102.h5



# ACCELERATING MX SCIENCE

ESRF Data Portal Data Logistics - Instruments My selection 0 - Search investigation... Q Andy GÖTZ -

Home / OA-17(07/10/2023 on ID23-1) / Datasets

Investigation

- Experiment
- Statistics
- Datasets 591**
- Jobs 1
- Logbook
- Prepare

Datasets

View as List Summary

Sample Select sample

MX

Show actions

Ranking shell

Selected statistics: Overall Rmerge

10 16 25

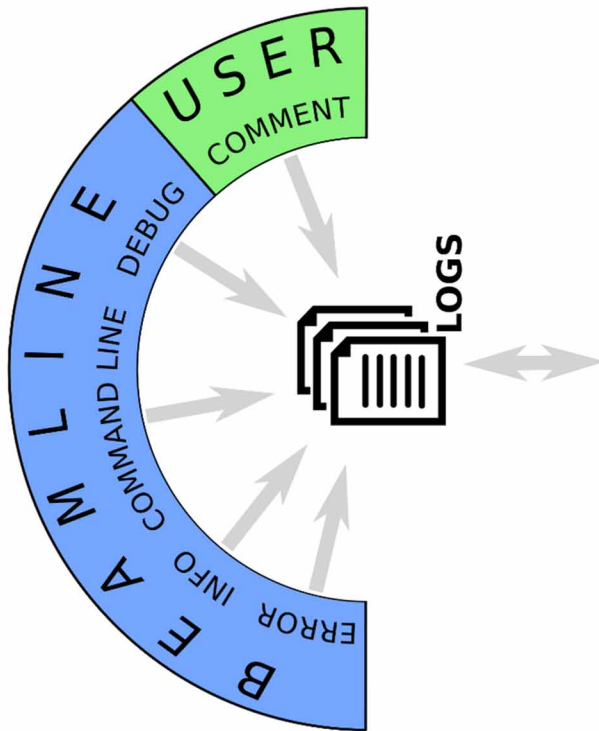
Legend:

- 1 not analysed
- 1 analysed
- 1 collected
- 1 autoprocessed

Puck: NA



# ELECTRONIC LOGBOOK



The screenshot shows a web interface for a logbook. At the top, there are navigation buttons: '+ New', 'Take a photo', 'View', and 'PDF'. Below these are search filters for 'Everywhere'. The main content is a list of log entries with timestamps and command-line text. A plot of 'offset normalized intensity, AU' vs 'energy, eV' is shown for one entry. On the right side, arrows point from specific log entries to labels: 'COMMENT', 'COMMAND LINE', 'INFO', and 'ERROR'. The text 'Web interface' is centered at the bottom of the screenshot area.

Timestamp	Log Entry	Annotation
08:03:00	OPTICS> # io/optics/figa def measure1 '_ccd_set_concat(1)'	
08:01:40	OPTICS> snap flux	→ COMMENT
08:01:25	OPTICS> dt	→ COMMAND LINE
03:46:39	OPTICS> New dataset: chof_root	→ INFO
01:20:21	OPTICS> zapxiimage thg 6.65 10.69 10 vbg 54.8 27.6 27 0 10 (zapug: #6, spec: #3)	
01:20:20	OPTICS> New dataset: cFeo42-_root2	
01:19:51	OPTICS> prdef Maps_554	
00:54:47	no new data collected	→ ERROR
00:51:57	OPTICS> dt	

# 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

Comments

Information

Errors

Command Lines

Machine

Order

Newest First

Oldest First

Date

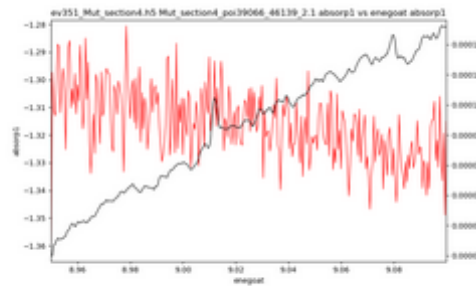
Select date

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

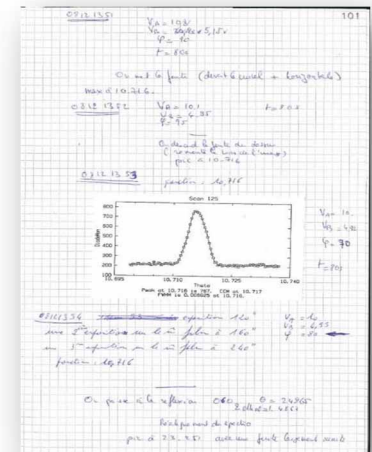
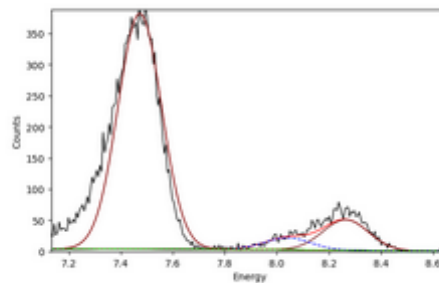
idet = 681803. ( 6.81803e+06 /s) p201 gain = 6

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



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

BM18  
BLC-14794

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

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

att1 chamber directly fine at full power.

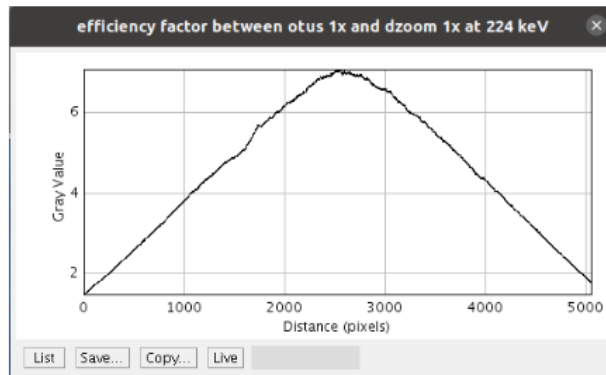
Att2, basically fine, but with sudden pressure peaks that close the valves. It tends to decrease in frequency.

Att3, much more degassing, especially on tl with continuous movement of the axis. Any 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 after 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](http://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)**

- **Reusable** means implementing

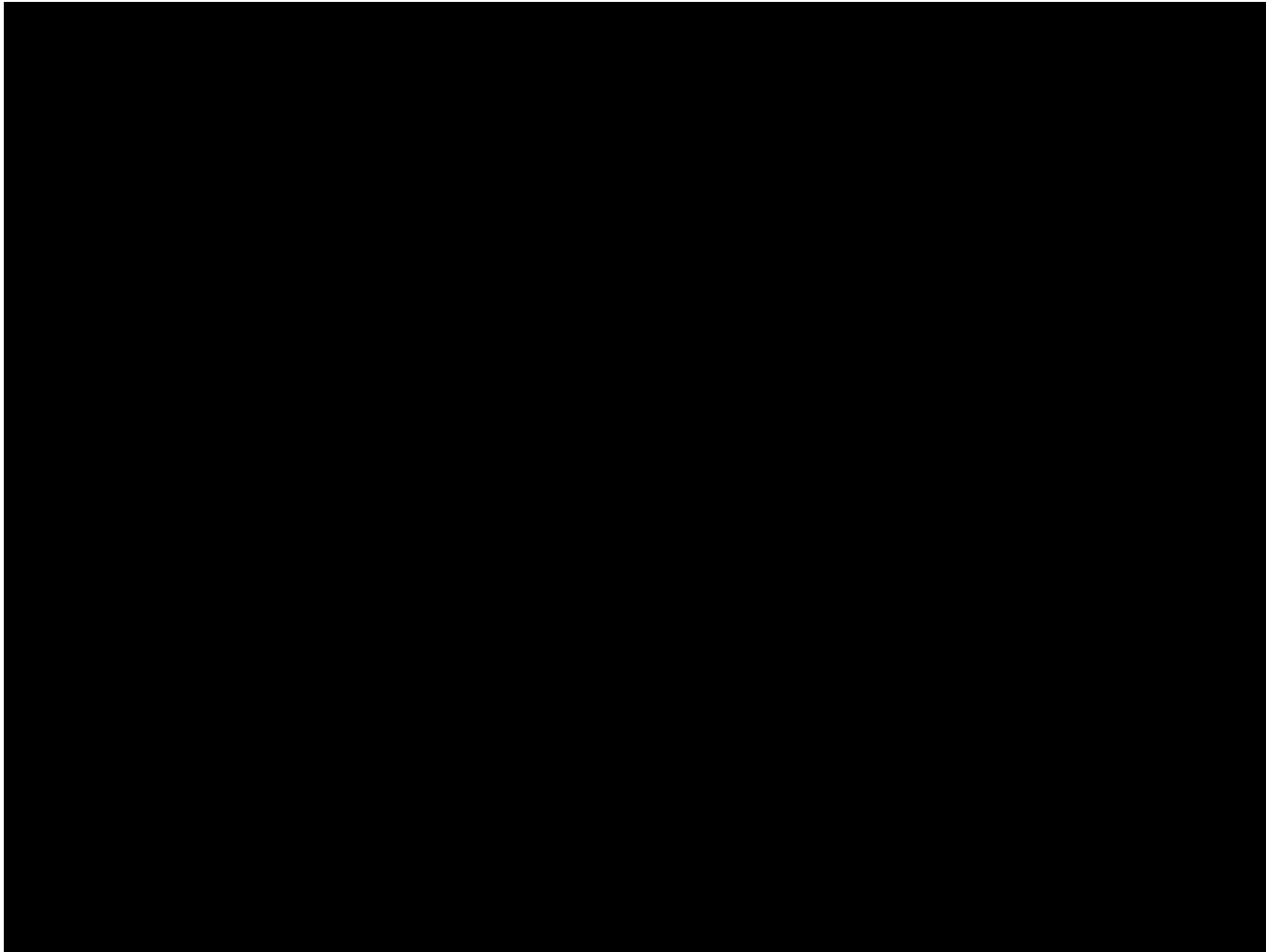
- **High quality metadata**, **electronic logbook (ICAT+)**, **open data licence (CC)**

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



# FAIR DATA READY FOR REUSE

Human Organ Atlas - <https://human-organ-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)**