# 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



- 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



TB / day



#### 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



- 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)
- 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 4<sup>th</sup> 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

TB per session vs cumulative TB



Plot by Daniele de Sanctis

# **PANOSC+EXPANDS IMPACT ON FAIR DATA AT ESRF**

- 1. First draft of **FAIR Data Policy** (thanks to <u>PaNOSC DP framework</u>)
- 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 <u>DS-Wizard</u>)
- 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+CERNVMFs 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 <u>PaNOSC search portal</u>)
- 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)

 $\rightarrow$  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)
  - Findable all raw data archived + have a DOI
  - Accessible data are licensed + downloadable
  - Interoperable standard metadata + data formats
  - Reusable 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





#### **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-tomany relations ...





#### 2023: 209 tables, 42 views

2003: 17 tables

# 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



Science and Technology Facilities Council



- 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
  - $\circ$  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

Making the most

Page 18

rotationAxis phiStart kappaStart omegaStart resolutionAtCorner undulatorGap1 undulatorGap2 undulatorGap3 beamSizeAtSampleX beamSizeAtSampleY centeringMethod actualCenteringPosition beamShape flux flux\_end workflowName workflowStatus crvstalSizeX Cr 2023 - MelbouncerystalSizeY

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

	Home All my sessions All my proposals					🖺 gaonach 🕶				
	MX2405	Sessions Shipr	nents Sample	changer						
Summary Acquisitions 21	Home Proposals	MX2405 MX 838	6 Collection							
Filter by: Scaling MR SAD Sample										
•         1         2         >         Show:         40         Total:         42										
01/07/2023 00:56:32 MXRPressE						Summary Beamline Parameters Acquisitions 4 Sample Autoprocessing 13 Workflow 4 Phasing 14				
Path: Adata/visitor/mx2405/id30a1/20230630/RAW_DATA/SMAD4MH2/SMAD4MH2-CD03	8134_B12-1									
Protein: SMAD4MH2			Best result@			MD phosing				
Sample: CD038134_B12-1			from grenades_paralk	elproc		MR Molecular Replacement from cell F4132				
Prefix: SMAD4MH2-CD038134_B12-1	F432	Compl.	Res.	Rmerge	I/Sigma					
Run: 1	Inner	100.0%	114.2 - 9.0	3.8	121.7					
# Images (Total): 1900 (2141)	Overall	100.0%	1142-25	14.6	31.4					
# mages (rota), 1600 (2141)	Overan			21.1	0111					
Transmission: 86.0982 %			Cubic system (F4 a=b=c	32)						
			197.9 Å			map.1 level = 0.2477 e/A*(4.56 rmsd)				
						PBD: refined.pdb <u>fullscreen</u> <u>unlead</u>				
	<u> </u>									
Automatic MR appears to have worked with the space group I4122,F23,F4132										
Dradafinad narameters: total notation = 260.0 degrees: Standard meth used : 900 v 200 i	m Dupamic aperture se	et to 50 um No strates	v results obtained fr	om EDNA chara	cterisation 360.0 de	was detecollection				
Predenied parameters, totan otation = 500.0 degrees, standard mesh used : 500 x 200 t	in. Dynamic aper tore se	a to so unitivo su ateg	y results obtained in	OIN EDITA Chara	cterisation. 500.0 de					
0 0 0 000 0 000	0000000	000000000	00000000	000 0	0					
autoPROC autoPROC_staraniso EDNA_proc grenades_CODGAS grenades_fastproc grenades_para	lelproc	Molecular_Replacemen	t_from_cell	XDSAPP	XIA2_DIALS					
01/07/2023 00:49:21 MXPressE						Summary Beamline Parameters Acquisitions 4 Sample Autoprocessing 13 Workflow 4 Phasing 6				
Path: Adata/visitor/mx2405/id30a1/20230630/RAW_DATA/SMAD4MH2/SMAD4MH2-CD03	8134_A12-1									
Protein: SMAD4MH2			Best result			MR phasing				
Sample: CD038134_A12-1			from EDNA_prov			MR Malecular.Replacement_from_cell F4132				
Prefix: SMAD4MH2-CD038134_A12-1	F 41 3 2	Compl.	Res.	Rmerge	I/Sigma					
Run: 1	Outer	100.0%	47.4-10.3	4.1	2.3					
# Images (Total): 1900 (2141)	Overall	100.0%	49.4-2.6	17.0	33.1					
# mages (rotal). 1000 (2141)	Great			201						
Transmission: 72.5 %			Cubic system (F41	32)		TK SA DI LA				

#### **Data visualization – MX autophasing**

Improving the users Data eXperience through visualization of processing



Page 22

#### Data Portal v2 – serial crystallography

#### Improving the users Data eXperience through efficient visualisation

	Home Calendar Proposals SS/	] <b>↓</b>	E gaor	nach 🔻 Logout
Stats Success Falled Processed Processing Error SSX-Chip SS  Total: 16 Summary Parameters 02/12/2022 16:44:51	Kjæt			
Protein: pnkfx	Sample: laagqq	Sample support: SSX-Chip	Experiment name: fvkhb	# Runs: 17
467.086 ind	exed (37.56%)	214,302 non-indexed hits (17.23%)	nmary	562.091 skipped images (45.2%)
		number statistics	cell a 30k median = 40.2777 m 20k 10k 40 40.2 40.4 40.6 10 10k cell alpha ct 10k 10k 10k 10k 10k 10k 10k 10k	cell b         cell c           median = 397,798         median = 130,2007           40         40           200         200           201         120           120         120           120         120           100         130.5
Run#		Run #1 summary (02,	/12/2022 16:44:51)	
#2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #11 #12 #13	Rideward 87,577 47,055 1 string 1 string	Permit 7_exec.4/ Angle (logres) 26 30 30 40 40 40 40 40 15 25 30 40 40 40 40 15 25 30 40 40 40 40 5 25 30 40 40 40 40 5 25 30 45 45 45 40 40 40 40 40 40 40 40 40 40 40 40 40	cell a         cccl           1500	ell b ter = 37.7998 ter = 132.2007 ter = 132.2007

02/12/2022 16:44:41

#### **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

ESRF	Data Portal	Data	Logistics 🕶	Instruments			Search investigation	Q	💄 Andy GÖTZ 👻
	home / IM-	<u>114 (26/0</u> 7	7/2023 on ID31)	/ datasets					
Investig	<b>gation</b> Experiment		<< < > >>	Page 1 of 2	Items 1-20 of 33	Show 20 🗸			Í
<u> </u>	Statistics		SRF_D	N_16					
	Datasets		0001	26/07/2023 13:28	:48				Summary
	Logistics		Dataset	0001	Distance -340.00				
			Start	26/07/2023	Energy 75.00		#ffoogan		
			End	13:28:48	Vibration 40.0 %	295 240 295			
				13:28:57		2200- 1 1 mar			
			Exp. Time	15		300	Andrew and the second s		
						integrab	asvg		
			🔒 /data/vi	sitor/im114/id31/20	)230726/RAW_DATA/ES	RF_DN_16/ESRF_DM	N_16_0001 🔥 🛃 Downlo	ad 🔻 🛛 Explore	

#### **Data processing with Ewoks – HTXRPD**



#### Data processing with Ewoks – 25 beamlines (so far)



"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

ata Portal My Data Open Data Closed Data Shipping • My Beamlines • Manager •		Data Parkal Wy Ode Geo Date Cossil Date Broader Broader Medizantina / Medizantina / Medizantina / Medizantina / Declarational Broader	
22/11/2022 28/11/2022 X-ray ideorption study of laser-shocked hernatile (Fe2O3) and	winth (%0	II Distanci ile 🗿 🗰 Lagbook 🔶 Disping 📾 Empire 🕢 Proposal	
😂 Dataset List 🕥 🛛 B Logbook 🔶 Shipping 🔛 Samples	Preposal	4100 Milliozada. Milliona Anto Marca C. Sect. Idead A Car	
+ Yerw D Take aphoto D Settings 🛦 FCit D Holp Q Search	365 found Date	March 30th 2023	
16.65.56 Al shield after the experiment	November 29th 2022	<ul> <li>PISSU BRIDDING FOR CONTRACT VIAge</li> <li>PISSU BRIDDI</li></ul>	
New shares to instant			
	November 28th 2022	PPC400-450MOF derived sample loaded in diversessnece reactor self. Base of the orthole with DN sizes (self-micros	
<ul> <li>04:02:28 shock - run 430.</li> </ul>		measure(DBF)(XA). MI connected ration	
Othersborn Parks 27+2203 8am/sap00 00,037     Saved in: /data/visitor/hc4919/id24/20221122/run_0430_shock_2022_11_28_06_0	0.11/nm.0430_shock_2022_11_28_06_00_11_nm_0430_shock_2022_11_28_06_00_11	IR quit into 3 parts	
Relative delays (setpoints)		exacto-interrupted and/restanted at 80:54 for N2 will	
Unive / Sec - Y.2.7is     Probe laser / SE: -12.5 ms (-3.3 ms / drive)		ences informating at 4 400 for entry in manys before CD=CD reacting (	
Drive energy:		nferenceshg05.irz~7.3	
<ul> <li>Request: 18.0 J</li> <li>On shot: 15.4 J.186% of request  (drive front-end energy: 14.6 J)</li> </ul>		sample ofig 1.1 - 4	
no. (410 mund		sample:56 mg	
140		PrCeOD nanowther 800	
10		PrCeO viges adds 800	
4.879		os situ musuments is ai posder padodis the function	
400 435		PtCeOD valvay-800: manual Sigma Aktitich 800-800	
		connect sample, did wit charge the name so it anotar/002.	
Average (we)		ungle num. 44.9 ag	
V1		experiment start at 212.06 notifizerable (2053)	
- BARRING CONTRACTOR		PMOF BITROD also	
-		sample name 202 mg/63 ng after reaction)	
A CONTRACTOR OF A CONTRACTOR OFTA CONTRACTOR O		experiment start at 1420 IPCeC0 403 MOE date W05	
Contraction of the second s		conference: 56.6 mg	
and the second se		after reaction 45.3 mg	
40 10 50 10 00 00 00 00 00 00		erong muntu we sent 02 toppher with H00	
¥2		PMOFILE CO-coldation same many, we dwy't alle each during articulary simulation denotes the manual (), dynatic	
		samplement. 13 mg alter markin	
		eperiment stated at 13.00	
		did a minister, experience and archive, internativelities many all the basicning of marcine, asses Testinghout in the and CO+O2 internativelial 1946 (and 1944) and 1944 (and 1944) and	der He after CD alered
		Missilanin	
		Mitaliania 02246.42 interact inc	h
		Holdson         Notacial         N.J.           All         All         All         All	N 23
		Hubbah 004,02 (0.14,04) (0.1 047 430 (0.14) 1 44 (0.14) 14 40 (0.14)	N 25 25 25 25
<ul> <li>Bit State S</li></ul>		Holman         Notation         Notation           Constant         40         90           All         40         90           J         40         90           J         40         90	N 25 25 25 25 25 25 25 25 25 25 25 25 25
<ul> <li>Statistic descent and the statistic descent and the stati</li></ul>		Holdent         Holdent         Ro           607         603         603           607         610         603           1         61         603           32         61         603           34         603         603           34         603         603	N 25 25 25 25 25 25

# Annotate experiments

- Chronological
- Automatically
- Manually
- Logbook types
  - Experiment
  - Beamline
  - Facility
- Electronic notebook

MS time (s) 360 - 630 647 - 898 954 - 1114 1133 - 1299 1333 - 1533

• Document style

#### Data Portal v2 – home page

Improving the users Data eXperience through efficient searching

ESRE Data Port	<b>tal</b> Data Logisti	ics - Instruments	Search investigation Q								
				T Public data is	Welcome to E. his application allows to find, access, in accessible to anyone. You need to be to See ESRF data on	SRF Data portal Ispect and download data acquired at gggd-in to visualize your data when it is plicy for more details	ESRF. sunder embargo.				
Searc Sea Or br			Q Search or er	nter address						search )	
Find, Secolar ( ) R Man											
Q a				0	*	0	0		H	-	
Acce		calendar.esrf	chat.esrf	data.esrf	confluence.esrf	esrf.cloud-hor	intranet.esrf	mail.proton	human-organ	_	



#### **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*



#### • 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

- <u>https://data.esrf.fr</u> and <u>https://human-organ-atlas.esrf.eu</u>
- ISPyB LIMS
  - Current version <u>https://github.com/ispyb</u>
  - New developments <u>https://gitlab.esrf.fr/icat/icat-plus</u>
- EWOKS workflows
  - Source <u>https://gitlab.esrf.fr/workflow/ewoks</u>
  - Documentation <u>https://ewoks.readthedocs.io/</u>
- ICAT metadata catalogue <a href="https://github.com/icatproject">https://github.com/icatproject</a>
- CoreTrustSeal <u>https://amt.coretrustseal.org/</u>

