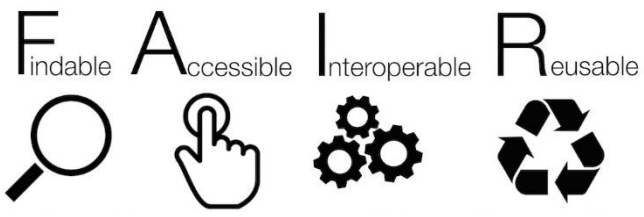
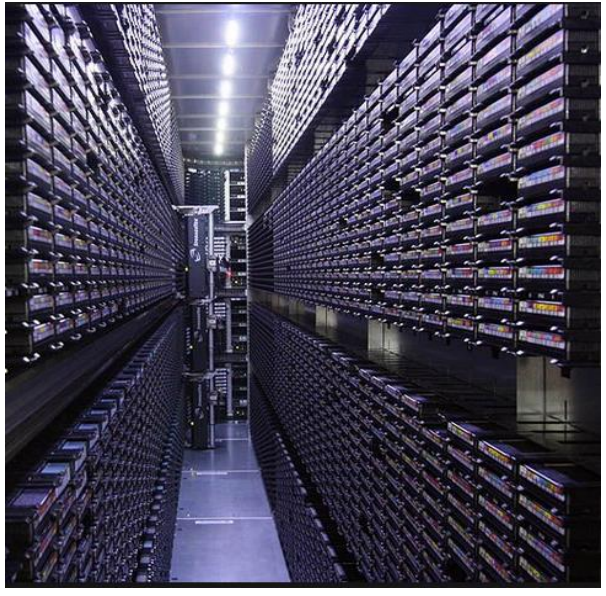
STATUS OF THE DATA POLICY SINCE 2015





The status of the ESRF Data Policy and its implementation since its approval by Council 2015

A.Götz

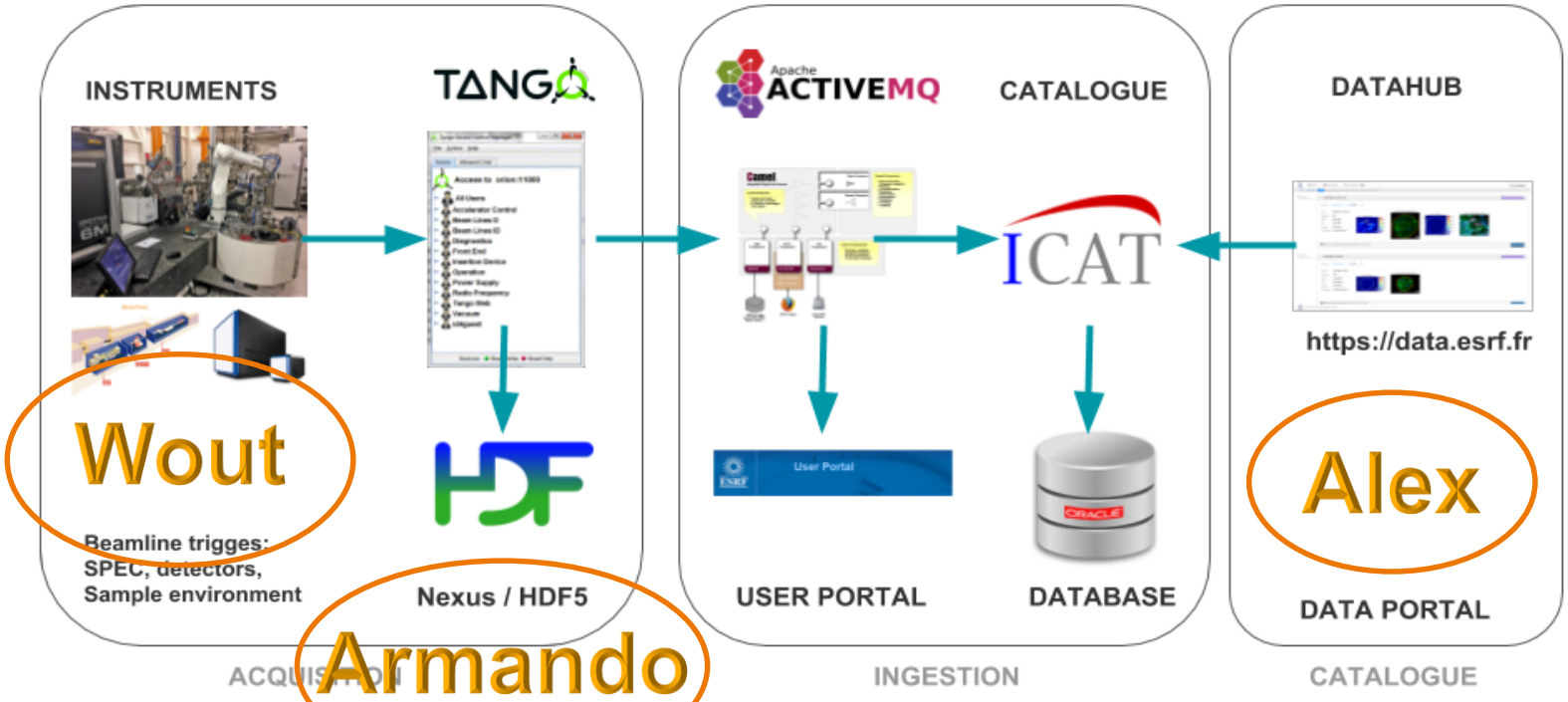
on behalf of all
implementers of the Data Policy

Alex, Maxime, Armando, Roberto, Christian, Emmanuel, Olof, ID16A, ID21, ID01, ID19, ID30A1, CM01, Management

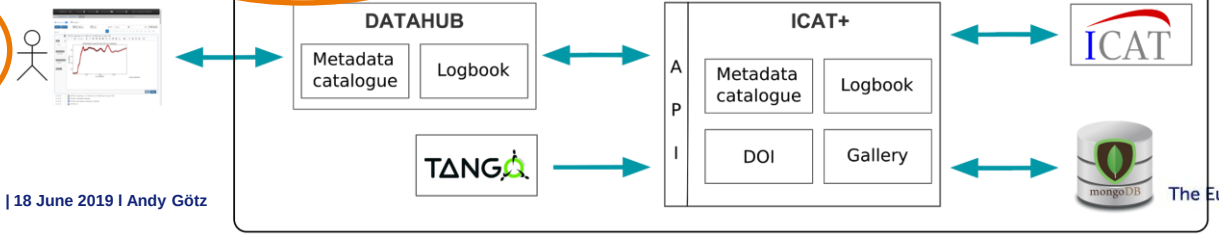
OUTLINE OF TALK

- Recall
- EU landscape
- Status and completion
- Open Data Now
- Data services
- Open issues
- Conclusion

TALKS ABOUT IMPLEMENTATION



Maxime



GOAL OF TALKS

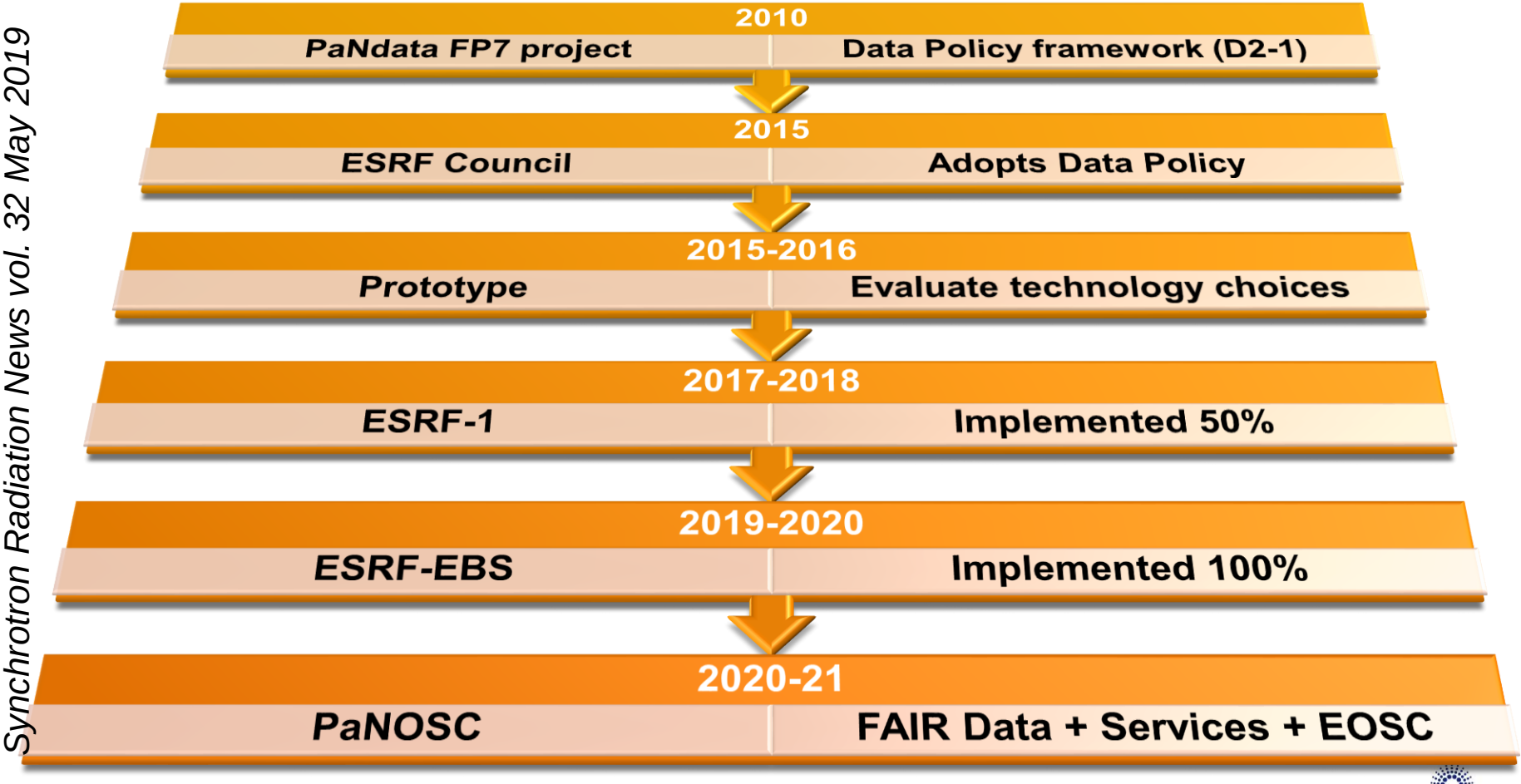
Explain how making **FAIR Data a Reality** (i.e. implementing the Data Policy) **helps Beamlines + Users**

ROLE OF ESRF

ESRF is a **Data Publisher**

In this role ESRF needs to publish high quality FAIR data

ESRF DATA POLICY TIMELINE



https://doi.org/10.1080/08940886.2019.1608119
Synchrotron Radiation News vol. 32 May 2019

RECALL OF ESRF DATA POLICY

1. Define, collect and keep rich **metadata forever**
2. Keep all **raw data** for at least **5 years** and aim for **10**
3. **High-level metadata** are made **public** during beam time
4. **Embargo period** of **3 years** during which proposal members have exclusive access to data
5. **Embargo period** can be **extended** on demand
6. **Data** are made **open access** after the embargo period
7. **HDF5** is the preferred **format** for metadata and data
8. Data Policy does **not** apply to **industrial beam time**

DATA POLICY @ PHOTON SOURCES

- Elettra • 2013
- ESRF • 2015
- MAXIV • 2015
- HZB • 2015
- HZDR • 2016
- PSI • 2016
- EuXFEL • 2017
- ALBA • 2017
- DESY • 2017
- SOLEIL • 2018
- Diamond • 2019

year in which Data Policy was adopted*

all institutes are still in the implementation phase

All institutes have adopted a Data Policy based on the PaNdata framework

*source: Calipsoplus NA1 survey

FAIR PRINCIPLES

Findable

Accessible

Interoperable

Reusable

EU LANDSCAPE

- European Open Science Cloud advocates FAIR data
- Many initiatives on how to make scientific data FAIR
- Journals are asking for FAIR data e.g. IUCr

IUCrJ
ISSN: 2052-2525
BIOLOGY | MEDICINE

Volume 6 | **Part 3** | May 2019 | Pages 341-343
https://doi.org/10.1107/S2052252519005918
OPEN ACCESS
Viewed by 716

Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography

John R. Helliwell,[a‡] Wladek Minor,[b§] Manfred S. Weiss,[c¶] Elspeth F. Garman,[d‡‡] Randy J. Read,[e‡‡] Janet Newman,[f§§] Mark J. van Raaij,[g§§] Janos Hajdu[h,i¶¶] and Edward N. Baker[j‡‡‡]

[a]School of Chemistry, The University of Manchester, Brunswick Street, Manchester M13 9PL, United Kingdom, [b]

IUCR FAIR DATA

IUCrJ
ISSN: 2052-2525
BIOLOGY | MEDICINE

Volume 6 | **Part 3** | May 2019 | Pages 341-343
https://doi.org/10.1107/S2052252519005918
OPEN ACCESS
Viewed by 716

Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography

John R. Helliwell,[a]‡ Wladek Minor,[b]§ Manfred S. Weiss,[c]¶ Elspeth F. Garman,[d]‡‡ Randy J. Read,[e]‡‡ Janet Newman,[f]§§ Mark J. van Raaij,[g]§§ Janos Hajdu[h,i]¶¶ and Edward N. Baker[j]‡‡‡

[a]School of Chemistry, The University of Manchester, Brunswick Street, Manchester M13 9PL, United Kingdom, [b]

IUCr Journals are now taking the lead by encouraging authors to provide a doi for their deposited original raw diffraction data when they submit an article describing a new structure or a new method tested on unpublished diffraction data. In the case of methods developed or tested with raw diffraction data, these data must be available to referees, and deposition of such data will eventually become compulsory. Permanent and prominent links will be provided from articles to the underpinning experimental data of each published research study.

We believe that these actions will maintain crystallography at the forefront of the effort for enhancing transparency and reproducibility of scientific results.

H2020 EOSC + FAIR PROJECTS

- **PaNOSC –** one of 5 cluster projects
- **ExPaNDS –** one of 5 national projects
- **EOSC –** European Open Science Cloud
  - EOSCPilot, **EOSC-hub, EOSC-Portal**
  - **FAIRsFAIR, RDA, GOFAIR, …**
- **FREYA –** Connected Open Identifiers
- **Calipsoplus –** JRA2 prototype DAAS
- **LEAPS –** data compression pilot project

PANOSC FACTSHEET

Call: Horizon 2020 InfraEOSC-04

Partners: ESRF, ILL, EuXFEL, ESS, CERIC-ERIC, ELI-DC, EGI

Description: cluster of ESFRI Photon and Neutron sources

Observers/non-funded: GÉANT, EUDAT, national RIs

Linked 3rd parties via EGI: DESY, STFC, CESNET

Status: Started 1/12/2018

Github: https://github.com/panosc-eu

Home page: https://panosc.eu

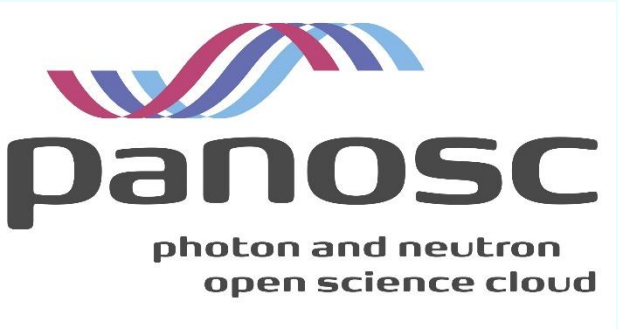
Twitter: @PaNOSC_eu #PaNOSC

Budget: 12 M€

Coordinator: ESRF

Started: 1/12/2018

Duration: 4 years



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

PANOSC OUTPUTS FOR ESRF

1. **Data policy –** define how to implement FAIR principles
2. **Data catalogue –** implement searching across data catalogues
3. **Data Management Plans –** develop tool for generating DMPs
4. **Metadata standard –** extend NeXus standard to new techniques
5. **Data analysis portal –** services for data reduction + analysis
6. **Jupyter notebooks –** enhance data visualization in Jupyter notebooks
7. **Software catalogue services –** pre-packaged software
8. **Simulation services –** OASYS and SIMEX photon simulation services
9. **AAI –** identity management with umbrellaID based on eduTEAMS
10. **Training –** training videos on data management, analysis, simulation
11. **ExPaNDS –** link up with all national photon sources
12. **EOSC –** link Photon and Neutron ESFRIs to EOSC

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

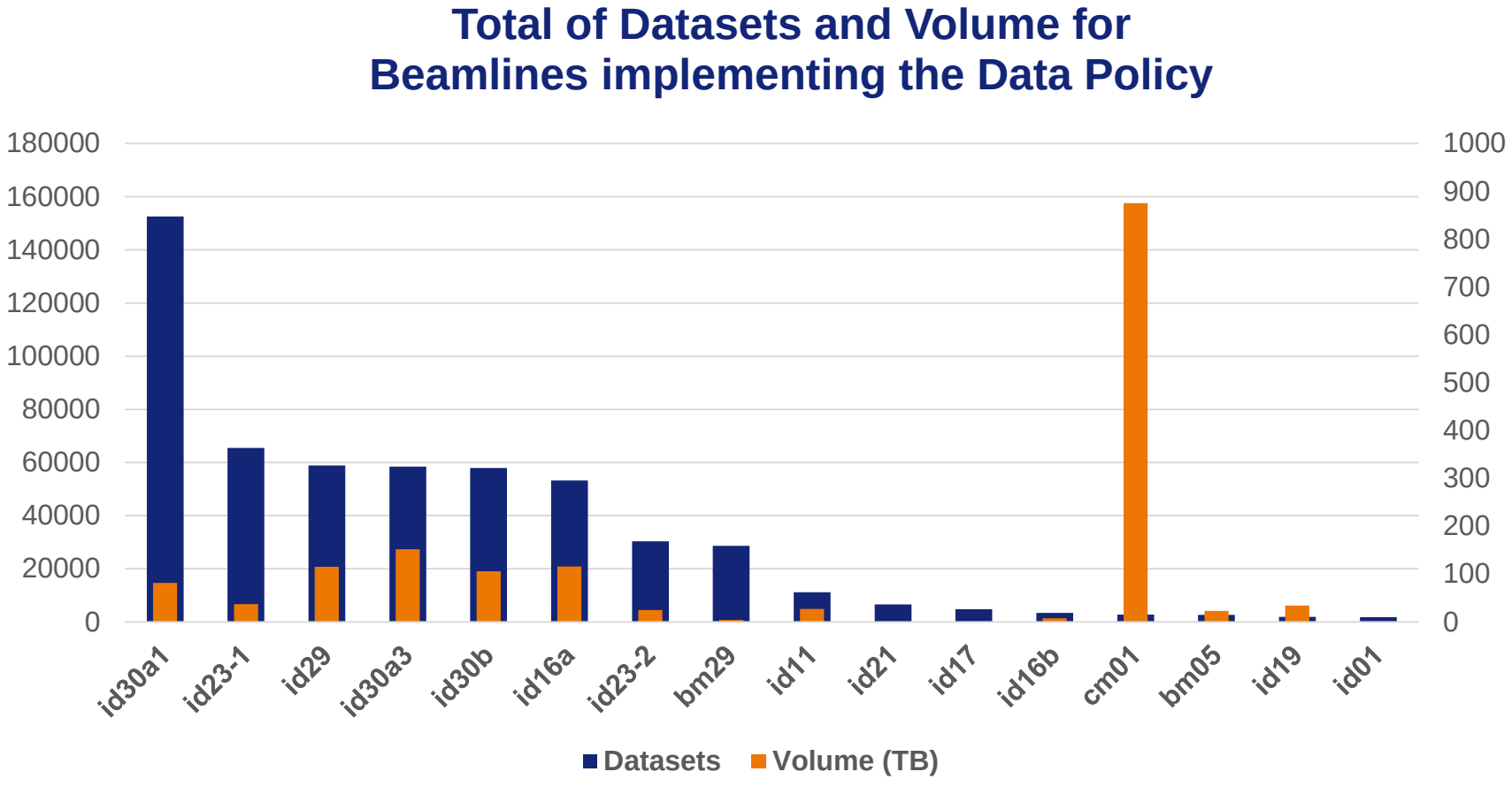
ESRF BEAMLINES STATUS @ STARTUP

IMPLEMENTED	BLISS – STARTUP	SPEC - STARTUP	ON HOLD
ID01 (kmap)	ID10	ID02 – 2019	ID03 - shutdown
BM05 (tomo)	ID11	ID06-LVP – 2019	BM18 – not built
ID11 (tomo)	ID13	ID06 – 2019	ID18 – not built
D16A (fluo, tomo)	ID15A	ID09 – 2020	ID24 - shutdown
ID16B (tomo)	ID15B	ID12 – 2020	BM01
ID17 (mrt, tomo)	ID19	ID20 – 2020	BM02
ID19 (tomo)	ID22	ID27 – 2020	BM08
ID21 (microscopy)	BM23	ID28 – 2020	BM16
ID23 (MX)	ID26		BM23
ID29 (MX)	ID29		BM25
BM29 (BIOSAXS)	BM29		BM26
ID30 (MX)	ID31		BM28
CM01 (SP)	ID32		BM30
			BM31
	https://www.esrf.eu/datapolicy		BM32

ESRF DATA POLICY STATISTICS

Total Samples	150 583
Total DOI	552
Total Sessions	1 886
Total Datafiles	157 966 298
Total Datasets	541 028
Total Parameters	18 096 849
Parameters/Dataset	33

ESRF DATA POLICY STATISTICS



SPEC + DATA POLICY

- **macros – metadata.mac excellent** set of macros developed by Roberto Homs
- **Beamlines –** used on 8 beamlines to implement Data Policy
- **How to Use –** see talks by Armando and Wout

BLISS + DATA POLICY

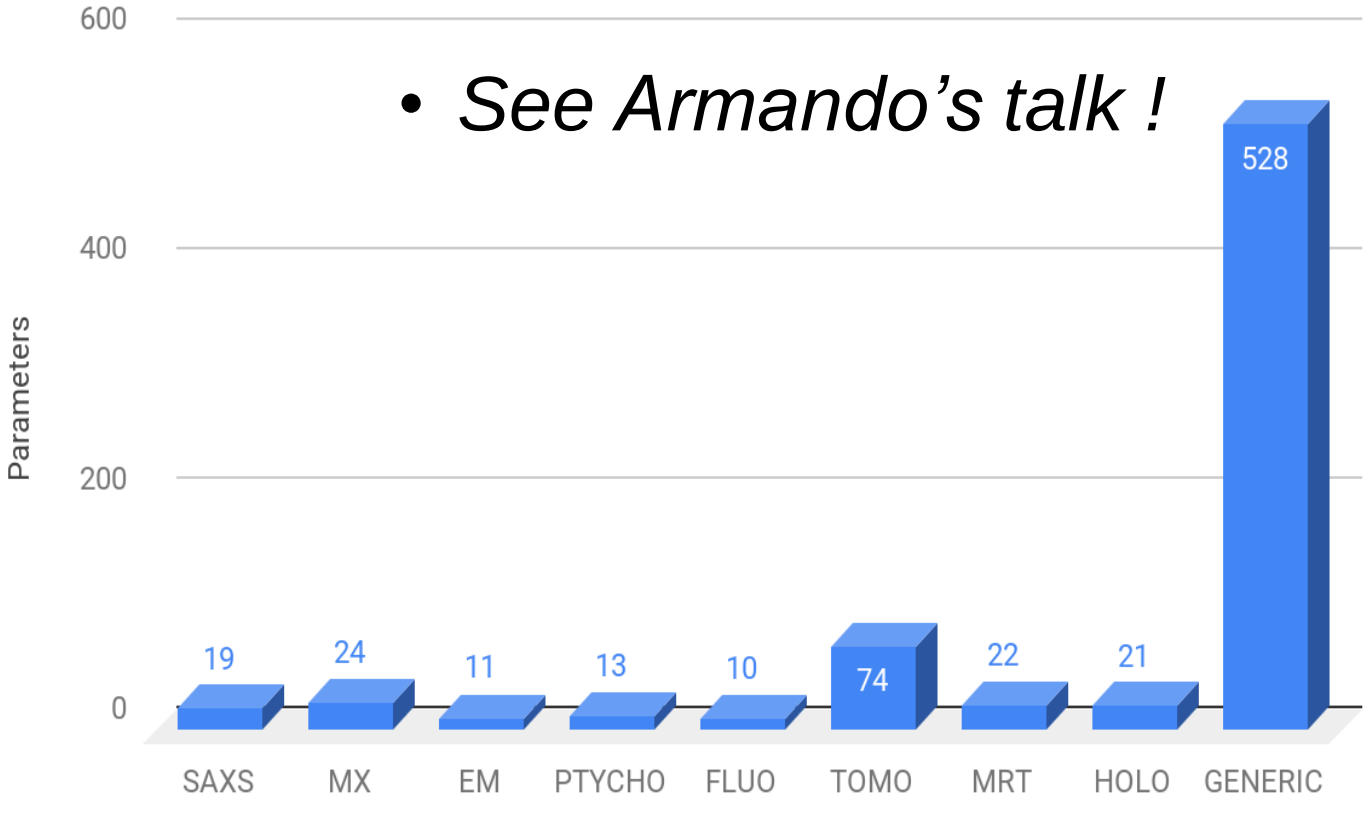
- **In progress –** move HDF5 file writer to external process
- **Implement triggers –** define and implement standard triggers for data policy i.e. newsample, newdataset, …
- **Planning -** Deploy on all BLISS beamlines in 2019 + 2020
- **Follow** issue - https://gitlab.esrf.fr/bliss/bliss/issues/592

MOVING TO HDF5 FOR RAW DATA

- **HDF5 used extensively for data analysis**
- **Lima** already **produces HDF5**
- **New Eiger detectors** will produce **HDF5 natively**
- **HDF5** reduces number of files, supports compression, metadata
- **Example - ID10** produced HDF5 during last run to reduce number of files a factor of hundreds to thousands (down from 75 million)
- **BLISS generates HDF5 natively**
- At **startup** more than **half the beamlines will produce HDF5**
- **HDF5 workshop** organized by **HDFGroup @ ESRF** in **September**
- **Shutdown is perfect time to adapt your programs to HDF5**
- **Proposal:** setup a **Users Group** for **HDF5 + Nexus** for **Beamlines**

NEXUS METADATA PARAMETERS

- *See Armando's talk !*



TIME TO OPEN YOUR DATA

* https://youtu.be/OM6XIICm_qo?t=644

FIRST CREATORS OF OPEN DATA

Matthew Bowler

+

Isai Kandiah

LATEST OPEN DATA FROM ESRF

MENU nature microbiology

Article | Published: 15 October 2018

Biogenesis and structure of a type VI secretion baseplate

Yassine Cherrak, Chiara Rapisarda, Riccardo Pellarin, Guillaume Bouvier, Benjamin Bardiaux, Fabrice Allain, Christian Malosse, Martial Rey, Julia Chamot-Rooke, Eric Cascales, Rémi Fronzes & Eric Durand

Nature Microbiology **3**, 1404–1416 (2018) | Download Citation

Data availability

The cryo-EM structures of the full complex TssKFG, TssK and TssFGE have been deposited in the Electron Microscopy Data Bank under ID codes EMD-0008, EMD-0010 and EMD-0009. The TssKFG, TssK and TssFGE models have been deposited in the PDB under ID codes PDB 6GIY, 6GJ3 and 6GJ1. Raw cryo-EM data are available on request.

should cite … doi:10.15151/ESRF-DC-186877747

CITING YOUR DATA IN PUBLICATIONS

- **The SHUTDOWN period is a perfect opportunity to write papers and publish your data!**

- **How to cite your data:**
  - *Create your DOIs before writing the publication so you can cite them*
  - *See Alex's talk !*

PUBLISHING VALUABLE DATASETS

SCIENTIFIC DATA

Data Descriptor | OPEN | Published: 11 December 2018

Time-resolved synchrotron X-ray micro-tomography datasets of drainage and imbibition in carbonate rocks

Kamaljit Singh, Hannah Menke, Matthew Andrew, Christoph Rau, Branko Bijeljic & Martin J. Blunt

Scientific Data **5**, Article number: 180265 (2018) | Download Citation

SCIENTIFIC DATA

Data Descriptor | OPEN | Published: 31 July 2018

High-throughput computational X-ray absorption spectroscopy

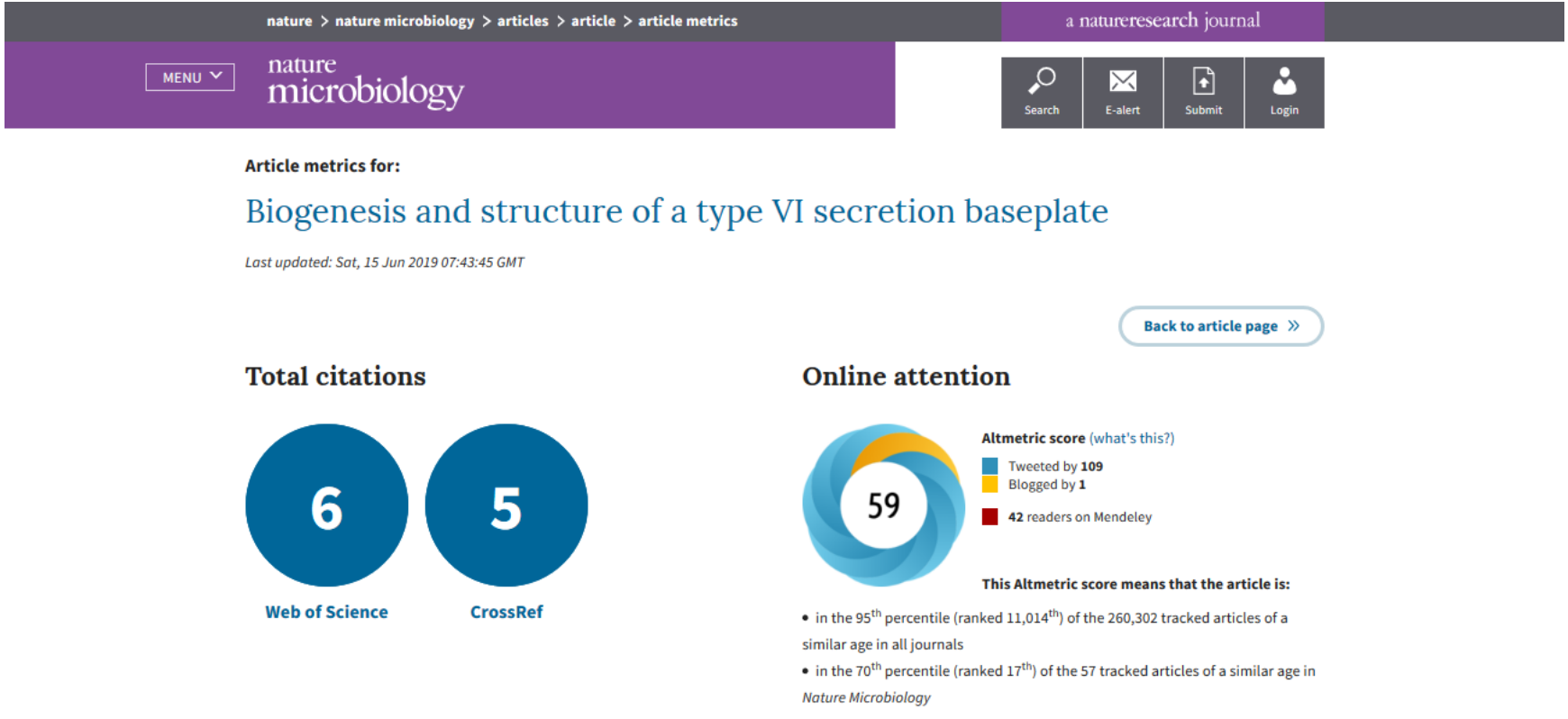
Kiran Mathew, Chen Zheng, Donald Winston, Chi Chen, Alan Dozier, John J. Rehr, Shyue Ping Ong & Kristin A. Persson

Scientific Data **5**, Article number: 180151 (2018) | Download Citation

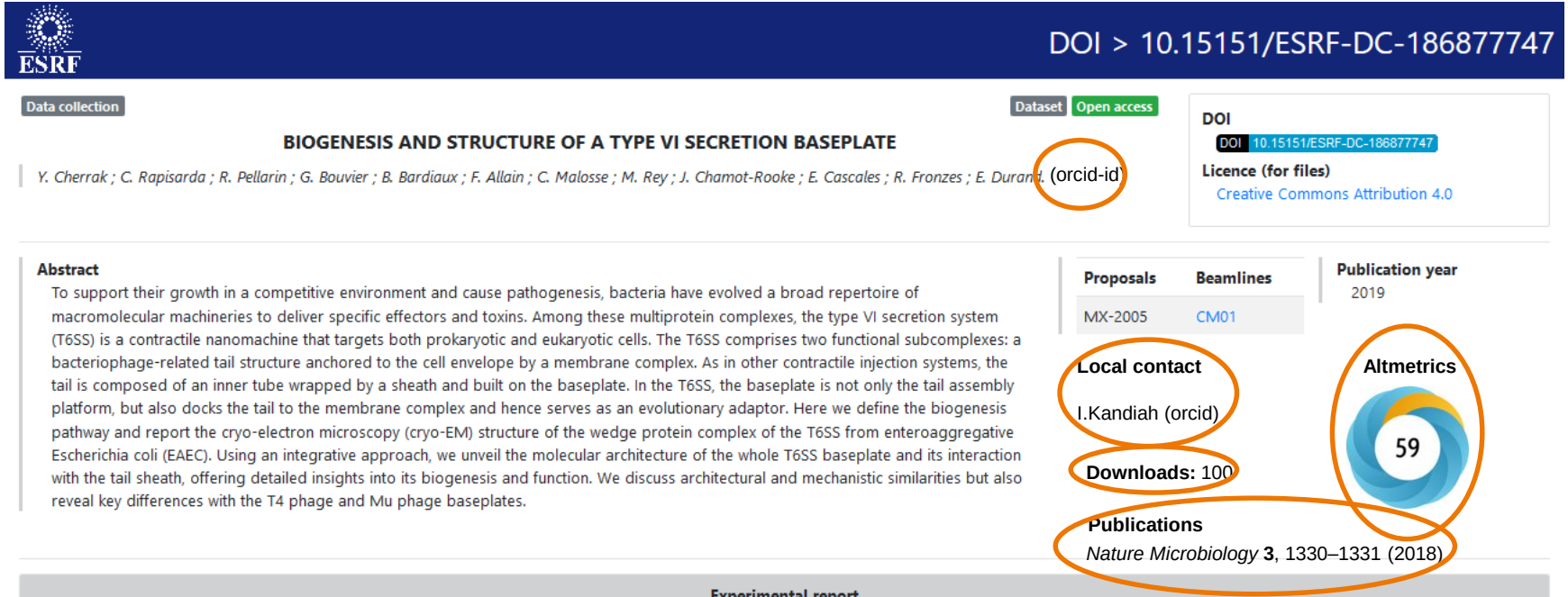
DOI'S – FUTURE DEVELOPMENTS

- **Abstract to be added to Proposal form**
- **Users to be informed (via email) of DOIs**
- **Add following to DOI's:**
  - Local Contact to DOI
  - ORCID ID to DOI
  - Links to publications (PUMA)
  - Number of downloads
  - Altmetrics for online metrics
- **DOI's for instruments i.e. beamlines**

ALTMETRICS


nature > nature microbiology > articles > article > article metrics
a natureresearch journal
MENU
nature microbiology
Search
E-alert
Submit
Login
Article metrics for:
Biogenesis and structure of a type VI secretion baseplate
Last updated: Sat, 15 Jun 2019 07:43:45 GMT
Back to article page »
Total citations
6
Web of Science
5
CrossRef
Online attention
59
Altmetric score (what's this?)
Tweeted by 109
Blogged by 1
42 readers on Mendeley
This Altmetric score means that the article is:
• in the 95th percentile (ranked 11,014th) of the 260,302 tracked articles of a similar age in all journals
• in the 70th percentile (ranked 17th) of the 57 tracked articles of a similar age in Nature Microbiology

FUTURE ESRF DOI …



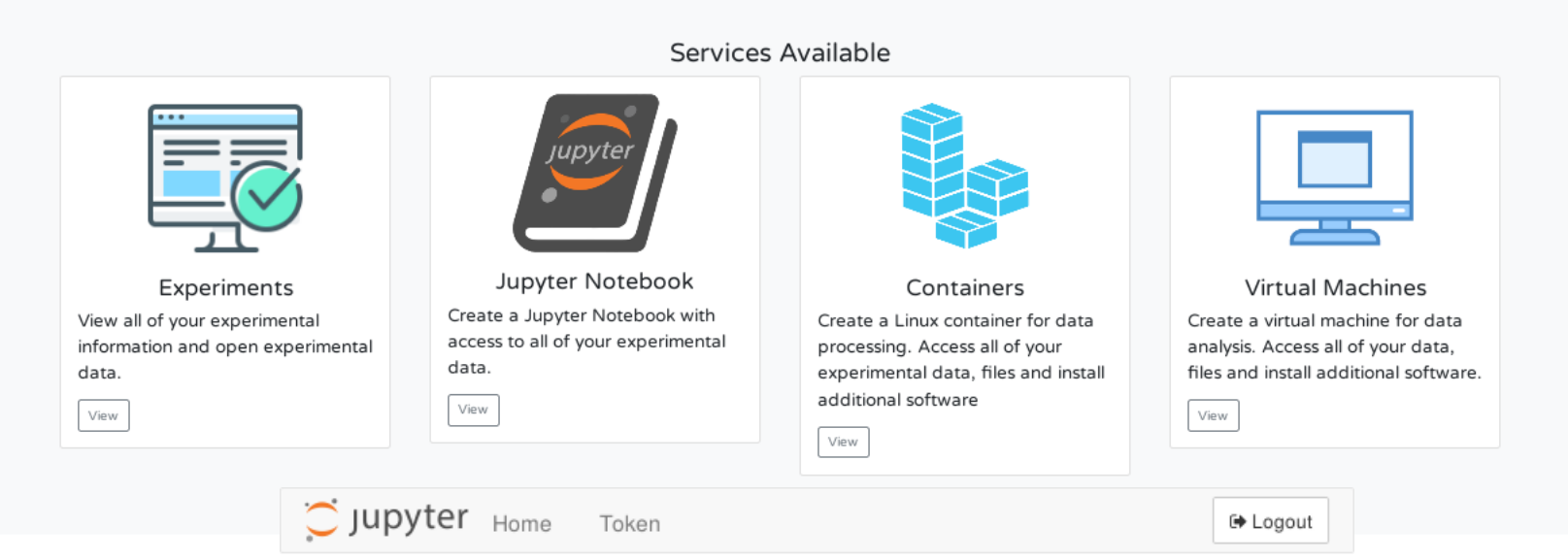
Proposed features

DATA SERVICES UNDER DEVELOPMENT

- CALIPSOplus – data portal prototype

- panosc photon and neutron open science cloud – will provide in production:
  - Jupyter notebook service (*jupyter.esrf.fr* soon)
  - Container service for running applications
  - Remote desktop service for VMs
  - Linking Data Portal to Data Services

CALIPSOPLUS DATA SERVICES



FUTURE DEVELOPMENTS

- **Web application for users to enter Proposal, Sample, + Dataset info**
- **Reliable download service (**globus**)**
- **Data dashboard per beamline**
- **Data reduction + compression**
- **Replace daily backup with archive**
- **Continue developing data portal**

OPEN ISSUES

- **How to update Data Policy**
  - Need to update DP for FAIR principles
- **Anonymous access**
  - Required for data mining + new algorithms
- **BAG proposals**
  - Not adapted to data policy / confidentiality
  - Need the list of “sub-authors” per sample
- **Process to extend embargo period**
- **What data analysis services to offer**

CONCLUSIONS

1. ESRF has the role of a data publisher
2. Data Policy on beamlines by end 2020
3. Users will profit from rich metadata + e-logbook, well structured data in HDF5/Nexus and long term archiving
4. DOIs will help to publish Open Data
5. New services for data download, reduction and analysis now possible