# Data Management @ ESRF

## Rich metadata using Nexus, and linking to the EOSC

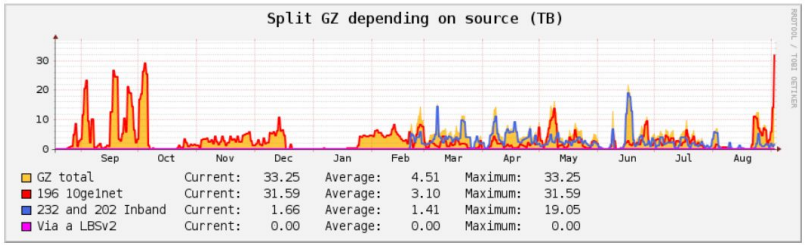
**Alex de Maria Antolinos**
**Software Engineer**
**Data Manager@Data Analysis Unit**
**Software Group**
**ESRF**

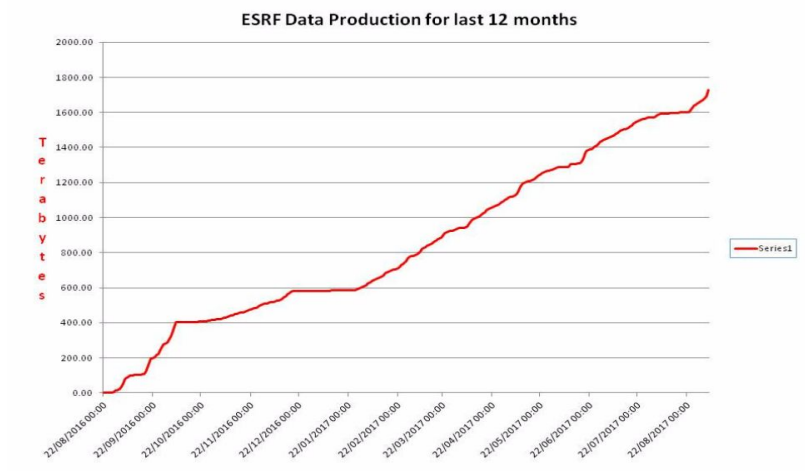
# Index

- **ESRF Data Policy**
  - Motivation
  - Data Policy General Principles
- **Implementation**
  - Overview
  - ICAT and Nexus
  - DOI
  - GDPR
  - Status
- **Panosc**

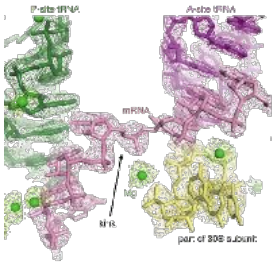
# Why a ESRF data Policy?

- Data is the **raw** material of science and is our **main product**








# Why an ESRF data Policy?

Data needs to be properly managed to allow:

- Verification
- Linking to publications
- Re-analysis
- New research
- Preservation of unique data sets

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

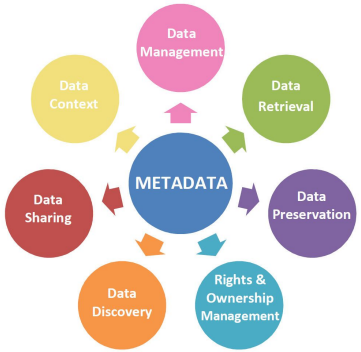
nature International weekly journal of science

**Data's shameful neglect**

*"Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly"*

*Nature* **461**, 145 (10 September 2009) | doi:10.1038/461145a

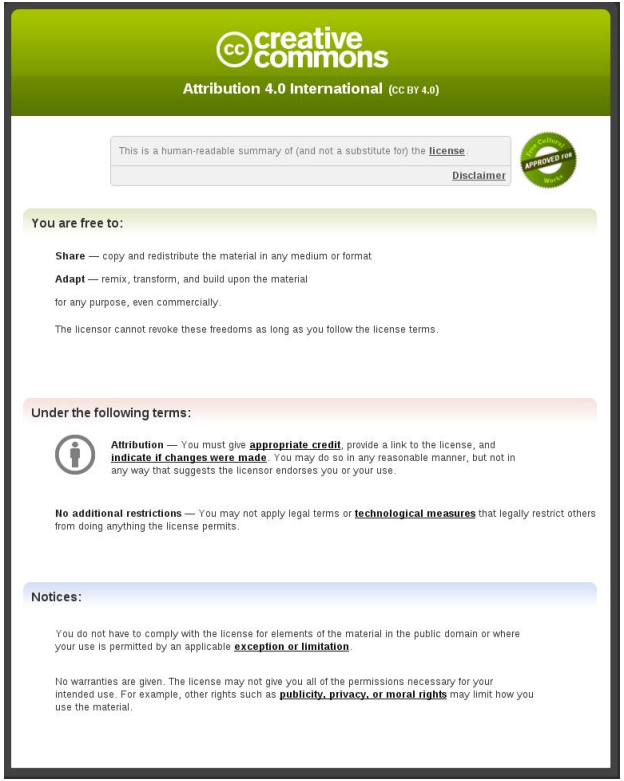
# Data Policy General Principles




- **Automatic capture** of data and metadata
- ESRF is the keeper (**custodian)** of the raw data and associated metadata
- Raw data and metadata will be selected, organized and look after in **well-defined formats** (curation)
- Raw data and metadata will be **READ-ONLY** for the duration of their life time
- **Proprietary research** (commercial) will be owned exclusively by the client who purchased the access and it is not covered by the data policy
- Restricted to the experimental team during the a **period of 3 years** (EMBARGO)
- Access to raw data and associated metadata is foreseen to be via a **searchable online catalogue (ICAT)**

# Data Policy

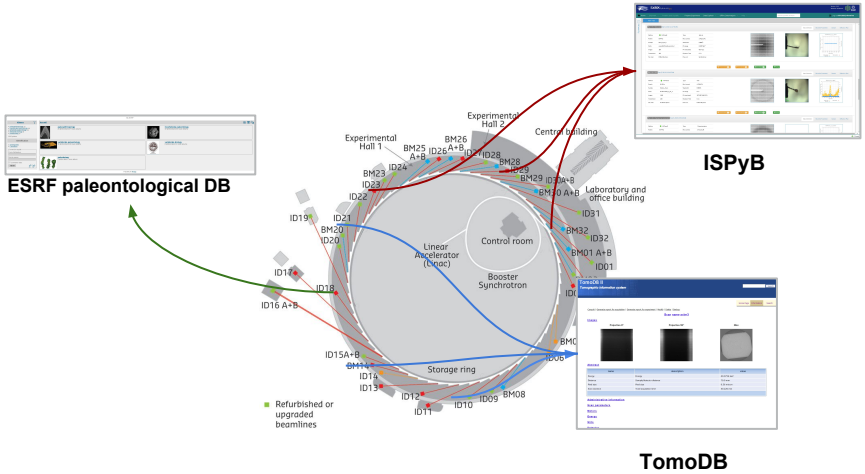
- After the embargo the data will be released under the license CC-By-4



# Implementation

# Implementation Overview

## Previous Situation

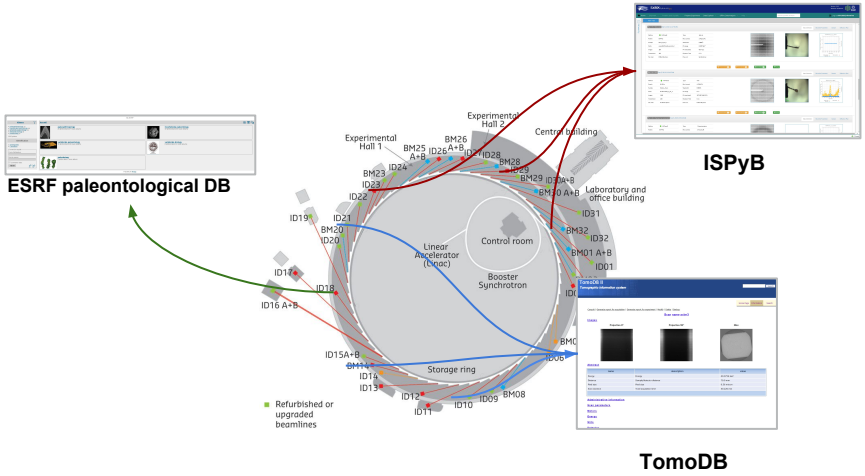



- **Raw Data**
  - Data are deleted from disk after 50 days
  - Full backups are kept for 2 years
  - No data management plan
  - No persistent identifiers

- **Metadata**
  - Not collected systematically
  - No online metadata catalogue for all beamlines
  - Experiment report is not public

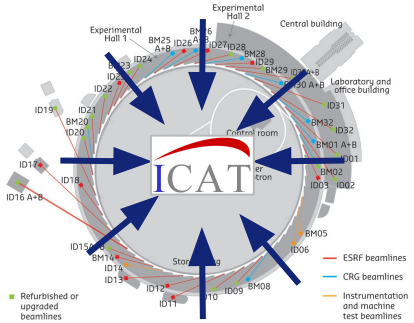
# Implementation Overview

## Current Situation




- Management of data was done by beamline or group
- No common strategy
- Different implementations of applications to manage data
- No archive
- Data was removed from disk after 50 days
- Data was backed up for 2 years

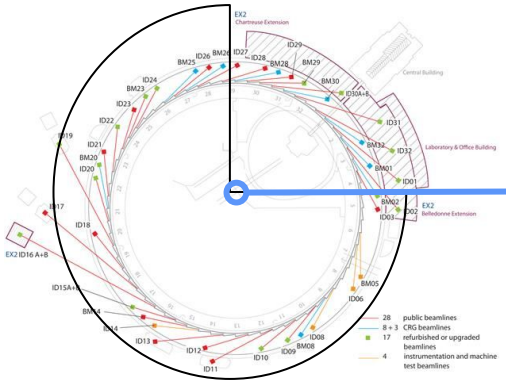
## Data Policy Implementation

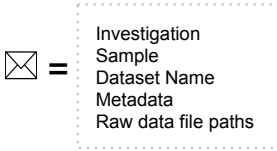



- A common implementation
- A common strategy for all beamlines
- Data and metadata storage is centralized
- Archive system
- Unique Persistent Identifier
- Same UI
- Based on a EU collaborative project

# Architecture

https://icat.esrf.fr



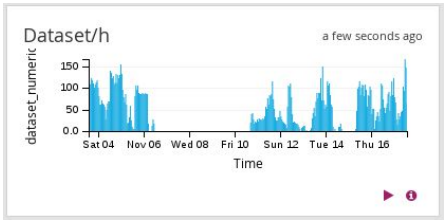


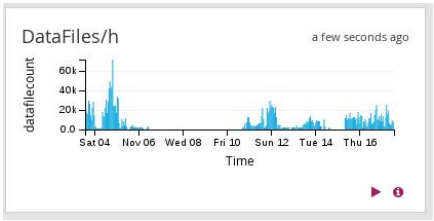











**PRODUCERS**

**CONSUMERS**

# ICAT and HDF5


SEND

SEND

SEND

PRODUCES

Metadata Manager
Tango Control Device

Ingester
Written in JAVA

ICAT 4.3
TOPCAT 1.1

WRITES

File containing
RAW DATA

File containing
RESULTS

HDF5 Master File
Nexus convention

# HDF5 + Nexus + ICAT

https://icat.esrf.fr


ICAT

# ICAT

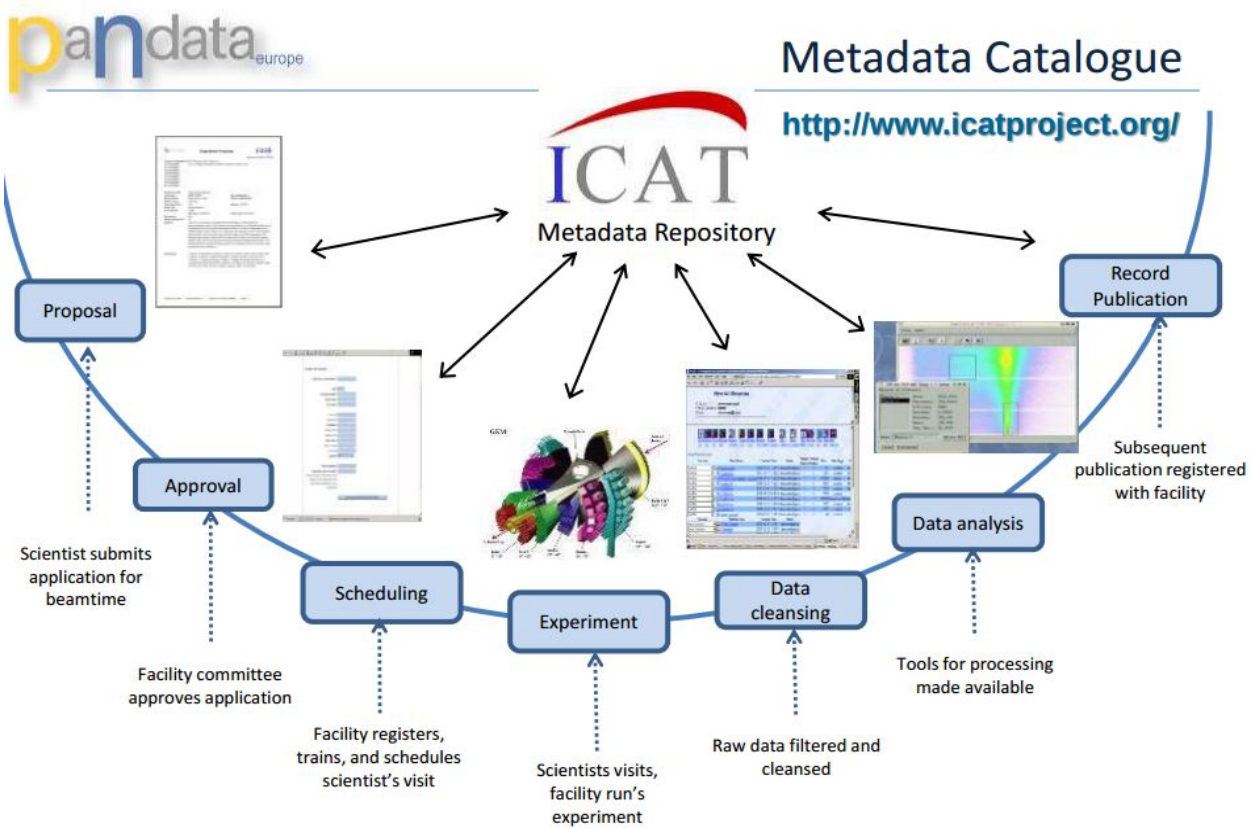
- ICAT was chosen for its **generic data model** which captures the scientific experiment
- ICAT not invented here (ESRF)
- **ICAT community** is collaborating to address all steps of data management






# ICAT

- ICAT is an open source **metadata management** system designed for **large facilities**



# NeXus Implementation@ESRF

**NXroot**

Top level. One per file.

**NXentry**

One group per measurement

**NXinstrument**

Describe the instrument.
Only one per NXentry

**measurement (@NXcollection)**

Flattened view of everything measured
Only one per NXentry

**sample (@NXsample)**

Define the physical state of the sample
during the scan

**NXdata**

The data to be plotted.
One NXdata group per plot

**user (@NXuser)**

Details of a user, i.e., name, affiliation, email address, *etc*

**NXsubentry**

Data or links to data for particular analysis

# HDF5 + Nexus + ICAT

https://icat.esrf.fr

- HDF5 as a mirror of ICAT on the local beamline file system
- Following the NEXUS convention

```
−<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'&#xA;&#x9;&#x9;" NAPItype="NX_CHAR">${scanType}</dataset_type>
    <folder_path ESRF_description="Scan starting date" ESRF_mandatory="Mandatory" NAPItype="NX_CHAR">${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"></group>
  +<group NX_class="NXsubentry" groupName="MX"></group>
  +<group NX_class="NXsubentry" groupName="PTYCHO"></group>
  +<group NX_class="NXsubentry" groupName="FLUO"></group>
  +<group NX_class="NXsubentry" groupName="TOMO"></group>
  +<group NX_class="NXsubentry" groupName="MRT"></group>
  +<group NX_class="NXsubentry" groupName="HOLO"></group>
  +<group NX_class="NXsubentry" groupName="WAXS"></group>
  +<group NX_class="NXsample" groupName="sample"></group>
  +<group NX_class="NXinstrument" groupName="instrument"></group>
  +<group NX_class="NXnote" groupName="notes"></group>
  </group>
```

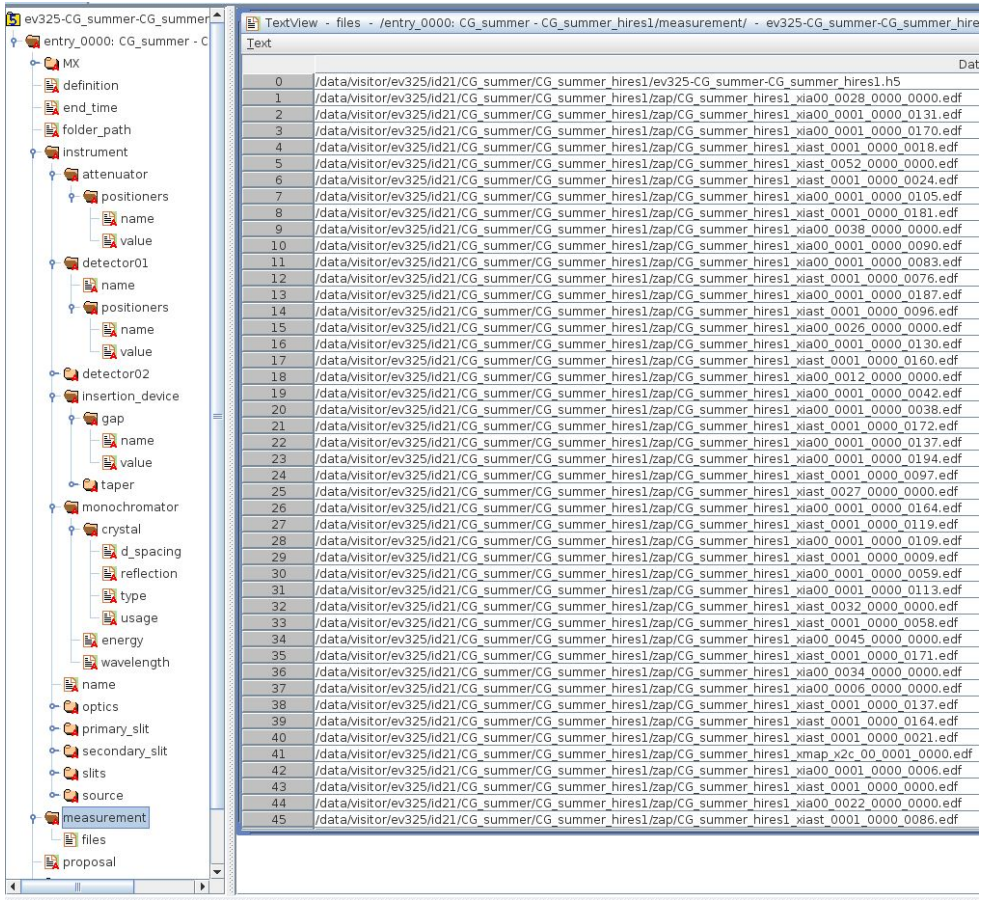
# HDF5 + Nexus + ICAT

https://icat.esrf.fr

```
−<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'&#xA;&#x9;&#x9;" NAPItype="NX_CHAR">${scanType}</dataset_type>
    <folder_path ESRF_description="Scan starting date" ESRF_mandatory="Mandatory" NAPItype="NX_CHAR">${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">${definition}</definition>
  +<group NX_class="NXsubentry" groupName="SAXS"></group>
  +<group NX_class="NXsubentry" groupName="MX"></group>
  +<group NX_class="NXsubentry" groupName="PTYCHO"></group>
  +<group NX_class="NXsubentry" groupName="FLUO"></group>
  +<group NX_class="NXsubentry" groupName="TOMO"></group>
  +<group NX_class="NXsubentry" groupName="MRT"></group>
  +<group NX_class="NXsubentry" groupName="HOLO"></group>
  +<group NX_class="NXsubentry" groupName="WAXS"></group>
  −<group NX_class="NXsample" groupName="sample">
      <name ESRF_description="Name of the sample" ESRF_mandatory="Mandatory" NAPItype="NX_CHAR">${Sample_name}</name>
      <description ESRF_description="Description of the sample" NAPItype="NX_CHAR">${Sample_description}</description>
      <chemical_formula ESRF_description="Chemical formula of the sample" NAPItype="NX_CHAR">${Sample_chemical_formula}</chemical_formula>
    −<group NX_class="NXpositioner" groupName="positioners">
        <name NAPItype="NX_CHAR">${SamplePositioners_name}</name>
        <value NAPItype="NX_CHAR">${SamplePositioners_value}</value>
      </group>
    −<group NX_class="NXenvironment" groupName="environment">
      −<group NX_class="NXsensor" groupName="sensors" ESRF_description="Parameters for controlling external conditions">
          <name NAPItype="NX_CHAR">${SampleEnvironmentSensors_name}</name>
          <value NAPItype="NX_CHAR">${SampleEnvironmentSensors_value}</value>
        </group>
      </group>
    </group>
  +<group NX_class="NXinstrument" groupName="instrument"></group>
  +<group NX_class="NXnote" groupName="notes"></group>
  </group>
```

# HDF5 + Nexus + ICAT

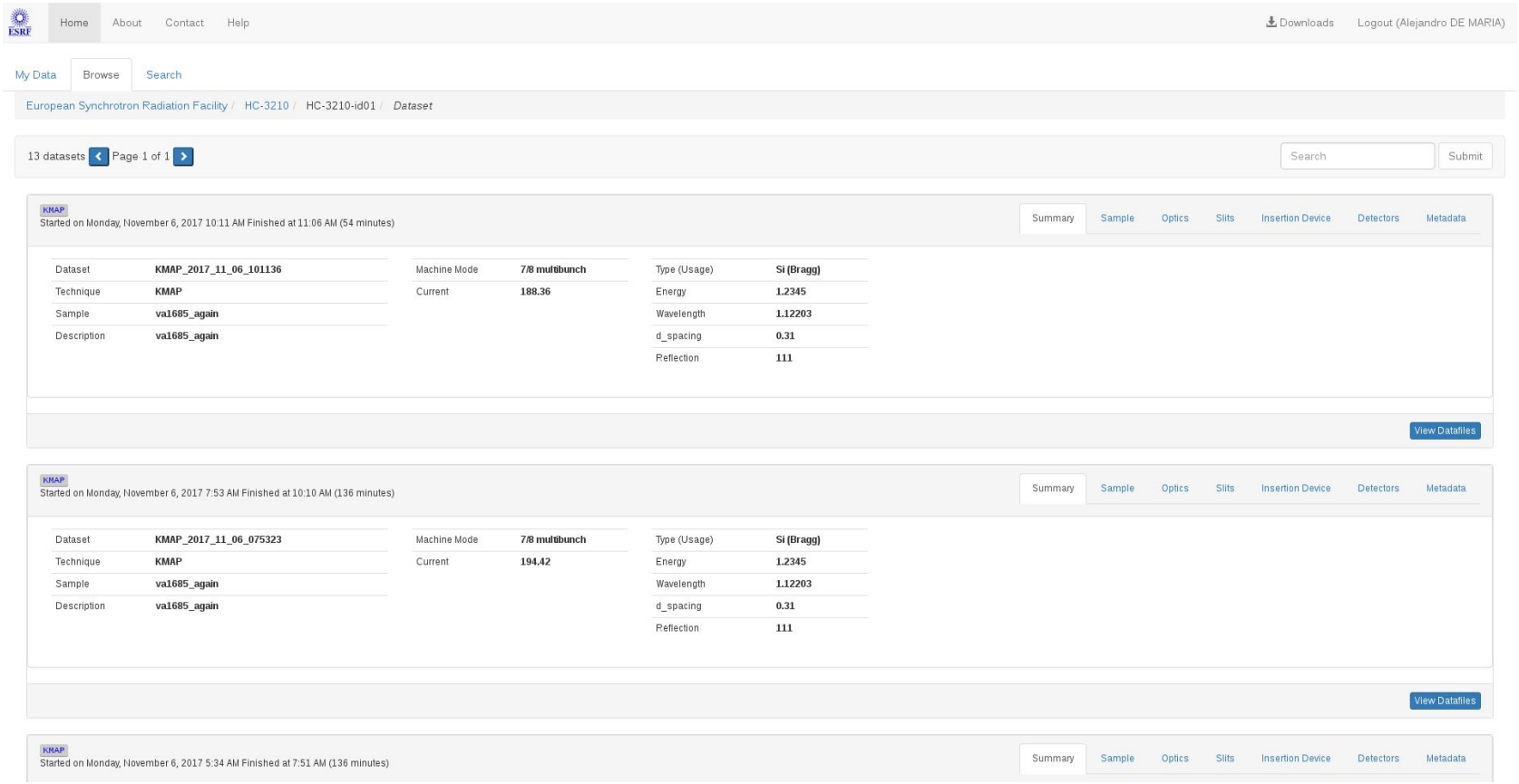
https://icat.esrf.fr



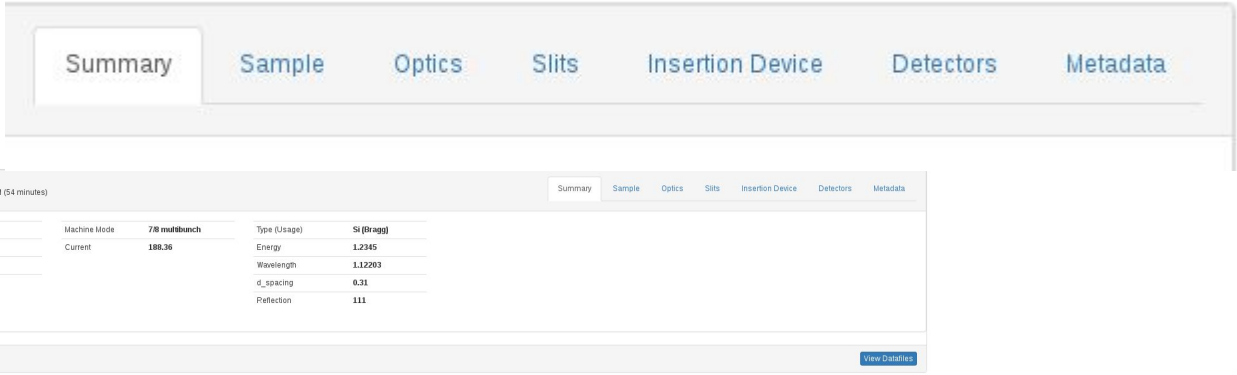
# TopCAT

ICAT Web User Interface

# TopCAT


Home
About
Contact
Help
Downloads
Logout (Alejandro DE MARIA)
My Data
Browse
Search
European Synchrotron Radiation Facility / HC-3210 / HC-3210-id01 / Dataset
13 datasets Page 1 of 1
Search
Submit
KMAP
Started on Monday, November 6, 2017 10:11 AM Finished at 11:06 AM (54 minutes)
Summary
Sample
Optics
Slits
Insertion Device
Detectors
Metadata
Dataset KMAP_2017_11_06_101136
Technique KMAP
Sample va1685_again
Description va1685_again
Machine Mode 7/8 multibunch
Current 188.36
Type (Usage) Si (Bragg)
Energy 1.2345
Wavelength 1.12203
d_spacing 0.31
Reflection 111
View Datafiles
KMAP
Started on Monday, November 6, 2017 7:53 AM Finished at 10:10 AM (136 minutes)
Summary
Sample
Optics
Slits
Insertion Device
Detectors
Metadata
Dataset KMAP_2017_11_06_075323
Technique KMAP
Sample va1685_again
Description va1685_again
Machine Mode 7/8 multibunch
Current 194.42
Type (Usage) Si (Bragg)
Energy 1.2345
Wavelength 1.12203
d_spacing 0.31
Reflection 111
View Datafiles
KMAP
Started on Monday, November 6, 2017 5:34 AM Finished at 7:51 AM (136 minutes)
Summary
Sample
Optics
Slits
Insertion Device
Detectors
Metadata

# TopCAT




KMAP

Started on Monday, November 6, 2017 10:11 AM Finished at 11:06 AM (54 minutes)

| | |
|---|---|
| Dataset | **KMAP_2017_11_06_101136** |
| Technique | **KMAP** |
| Sample | **va1685_again** |
| Description | **va1685_again** |

| | |
|---|---|
| Machine Mode | **7/8 multibunch** |
| Current | **188.36** |

| | |
|---|---|
| Type (Usage) | **Si (Bragg)** |
| Energy | **1.2345** |
| Wavelength | **1.12203** |
| d_spacing | **0.31** |
| Reflection | **111** |

View Datafiles

# TopCAT

ESRF | Home | About | Contact | Help | Downloads | Logout (Alejandro DE MARIA)

My Data | Browse | Search

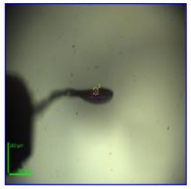
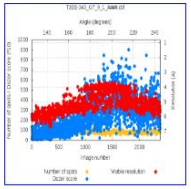
European Synchrotron Radiation Facility / [redacted] / *Dataset*

165 datasets < Page 1 of 3 >

Search | Submit

MX
Started on Sunday, November 19, 2017 5:04 AM Finished at 5:07 AM (3 minutes)

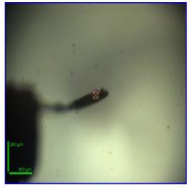
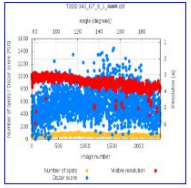
Summary | Positioners

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset | [redacted]9 | Machine Mode | 7/8 multibunch | Flux | 3.93e+11 |  |  |
| Technique | OSC | Oscillation Overlap | 0.0 | Current | 168.81 | | |
| # Images | 2380 | Aperture | 50 um | Resolution | 4.00853173304 | | |
| Template | T3SS-343_G7_9_1_####.cbf | Wavelength | 0.966 | Transmission | 100 | | |
| Detector Distance | 501.375 | MX_beamShape | ellipse | | | | |
| Beam Size @Sample | 0.05, 0.05 | | | | | | |

View Datafiles

MX
Started on Sunday, November 19, 2017 4:50 AM Finished at 4:55 AM (5 minutes)

Summary | Positioners

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset | [redacted]4 | Machine Mode | 7/8 multibunch | Flux | 3.95e+11 |  |  |
| Technique | OSC | Oscillation Overlap | 0.0 | Current | 169.53 | | |
| # Images | 2400 | Aperture | 50 um | Resolution | 3.14539455615 | | |
| Template | T3SS-343_G7_8_1_####.cbf | Wavelength | 0.966 | Transmission | 100 | | |
| Detector Distance | 389.3375 | MX_beamShape | ellipse | | | | |
| Beam Size @Sample | 0.05, 0.05 | | | | | | |

View Datafiles

# DOI

10.15151/ESRF-ES-86410123

# MECHANISMS OF QUINOLINE-FAMILY DRUG ACTION AND DRUG RESISTANCE IN THE MALARIA PARASITE PLASMODIUM FALCIPARUM

**Authors**

Sergey KAPISHNIKOV
Jens ALS-NIELSEN

**Proposal**
LS-2704

**Publication year**
2018

**Beamline**
id16a

**Session date**
2018-02-16

**Category**
Life Sciences

**Publisher**
European Synchrotron Radiation Facility

**Abstract**
There is no abstract for this session.

**Experimental report**

Download

**Experimental data**

The data is under embargo until **2021** but could be released earlier. It is currently only available to download if you are a member of the proposal team.

Access data

**Citation**
Below is the recommended format for citing this work in a research publication.

*Kapishnikov S., Als-nielsen J. (2018). Mechanisms of quinoline-family drug action and drug resistance in the malaria parasite Plasmodium falciparum. European Synchrotron Radiation Facility (ESRF). doi:10.15151/ESRF-ES-86410123*

# GDPR

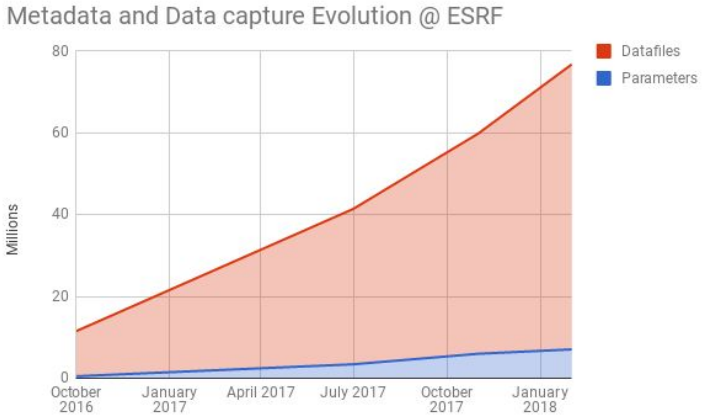
General Data Protection Regulation

# GENERAL DATA PROTECTION REGULATION (GDPR)

| Family GDPR Requirements | Definition | Question | Example |
|---|---|---|---|
| **Information** | Inform the persons concerned about the use of their personal data | Is there a procedure to prevent customers, prospects, employees from using their personal data? | Customers are not notified of the use of their data |
| **Consent** | Obtain the explicit consent of the persons concerned by the use of their personal data | How is consent obtained? | I accept to receive the newsletter: the box is not checked (choice of the customer) |
| **Purposes of the processing** | For any treatment the finality of the treatment must be explicit. | what is the purpose of the application)? | I do not use the data collected on the contact form of the site to send the newsletter |
| **Data minimization** | Use data strictly necessary for each treatment | What information do you need for proper treatment? | I make sure that I do not use all the customer data for example: name, first name, address, such as if I only need the first name |
| **Quality** | Ensure the good quality of personal data | How do you ensure that personal data are of the highest quality? Modalities, actors, frequency? | I update the phone number and e-mail address of the customer |
| **Right to access** | Allow people to access their personal data | Do you allow individuals to access their personal data? | The customer accesses his personal data on his dedicated customer area. |
| **Right to object** | To allow the person concerned to oppose at any time the processing of his / her personal data | Do you allow people to oppose processing on their persona data ? | Following the customer's request to stop receiving the newsletter, the customer no longer receives the newsletter |
| **Storage limitation** | Ensure that the personal data can only be retained for a period not exceeding that necessary for the purposes for which they are collected and processed | Are there procedures for managing the shelf-life of personal data ? | I ensure that the CV (paper and digital) of the non-hired candidates are deleted beyond 2 years (legal deadline) |
| **Right to erasure « Right to be forgotten «** | Allow to delete personal data from all applications | Is there a procedure to take into account the right to be forgotten? | Following the client's request to erase all his data, I make sure that the customer's data is deleted on all media. |
| **Data portability** | The right to retrieve part of their data in an open and machine-readable format. | Is there a procedure for requesting the recovery of personal data? | An employee wishes to obtain all of his personal data on an open format |

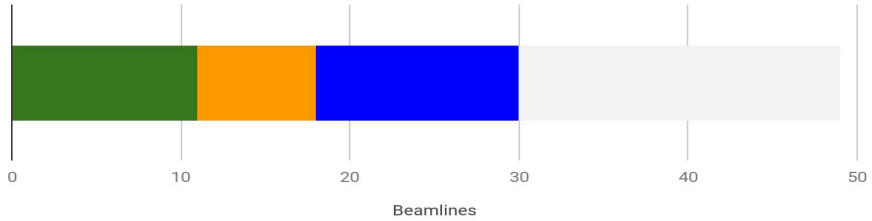
# Implementation of the ESRF Data Policy

- **Metadata Collection**
  - Automatic capture of data and metadata
- **Data archiving**
  - Long term archiving in tape library **during 10 years**
- **Raw Data in HDF5**
  - HDF5 used as primary format for raw data
- **Open access of data**
  - **Persistent identifier** (DOI) associated to data from peer review proposals and open access data after an embargo period of 3 years




## Current Status

- Data policy already implemented on 11 beamlines, 7 in progress and 12 planned for 2018




Data Download Service:
https://icat.esrf.fr

Data Analysis Service:
COMING SOON

Keep update on status:
https://ww.esrf.fr/datapolicy

# Status(http://www.esrf.fr/datapolicy)

| | | | Data Policy Implementation | | | |
|---|---|---|---|---|---|---|
| Beamline | Status | Techniques | Metadata Collection* | Data archiving* | Raw Data in HDF5* | Open access to data* |
| ID01 | KMAP | | Implemented | In progress | | |
| BM01A | | | | | | |
| BM01B | | | | | | |
| ID02 | | | | | | |
| BM02 | | | | | | |
| ID03 | | | | | | |
| BM05 | Tomography | | In progress | In progress | | |
| ID06-LVP | | | | | | |
| ID06 | | | | | | |
| BM08 | | | | | | |
| ID09 | | | | | | |
| ID10 | | | | | | |
| ID11 | Tomography | | In progress | In progress | | |
| ID12 | | | | | | |
| ID13 | | | | | | |
| BM14 | | | | | | |
| BM15A | | | | | | |
| BM15B | | | | | | |
| ID16A | Fluo, Tomo | | Implemented | In progress | Implemented | |
| ID16B | Tomo | | In progress | In progress | | |
| ID17 | MRT, Tomography | | Implemented | In progress | | |
| ID18 | | | | | | |
| ID19 | Tomography | | In progress | In progress | | |
| ID20 | | | In progress | In progress | | |
| ID21 | Microscopy | | Implemented | In progress | | |
| ID22 | | | | | | |
| ID23-1 | MX | | Implemented | In progress | | |
| ID23-2 | MX | | Implemented | In progress | | |
| BM23 | | | | | | |
| ID24 | | | | | | |
| BM25A | | | | | | |
| BM25B | | | | | | |
| ID26 | | | | | | |
| BM26A | | | | | | |
| BM26B | | | | | | |
| ID27 | | | | | | |
| ID28 | | | | | | |
| BM28 | | | | | | |
| ID29 | MX | | Implemented | In progress | | |
| BM29 | BIOSAXS | | Implemented | In progress | | |
| ID30A-1 | MX | | Implemented | In progress | | |
| ID30A-3 | MX | | Implemented | In progress | Implemented | |
| ID30B | MX | | Implemented | In progress | | |
| BM30A | | | | | | |
| BM30B | | | | | | |
| ID31 | | | | | | |
| ID32 | | | In progress | In progress | | |
| BM32 | | | | | | |
| CryoEM | | | | | | |

***Techniques covered** - lists which techniques are currently concerned by the Data Policy

***Metadata collection** - status of metadata collection and storage in metadata catalogue for the listed techniques (orange = planned / in progress, green = implemented, grey = not planned / implemented yet)

***Data archiving** - status of long term archiving in tape library (green = data are being archived for 10 years in tape archive)

***Open access of data** - indicates if data is open access (green) or still under embargo (red), or no data archived (grey)

# PANOSC

**P**hoton **a**nd **N**eutron **O**pen **S**cience **C**loud proposal for

H2020 call **INFRAEOSC-04-2018**

***Connecting ESFRI infrastructures through Cluster projects***

# Conclusions

- PaNOSC is an opportunity for all partners to harmonise their data policies, provide data services and help user learn and adopt modern FAIR data management principles

- PaNOSC is an opportunity to integrate the PaNOSC partners with the European Open Science Cloud and e-infrastructure platforms i.e. EGI

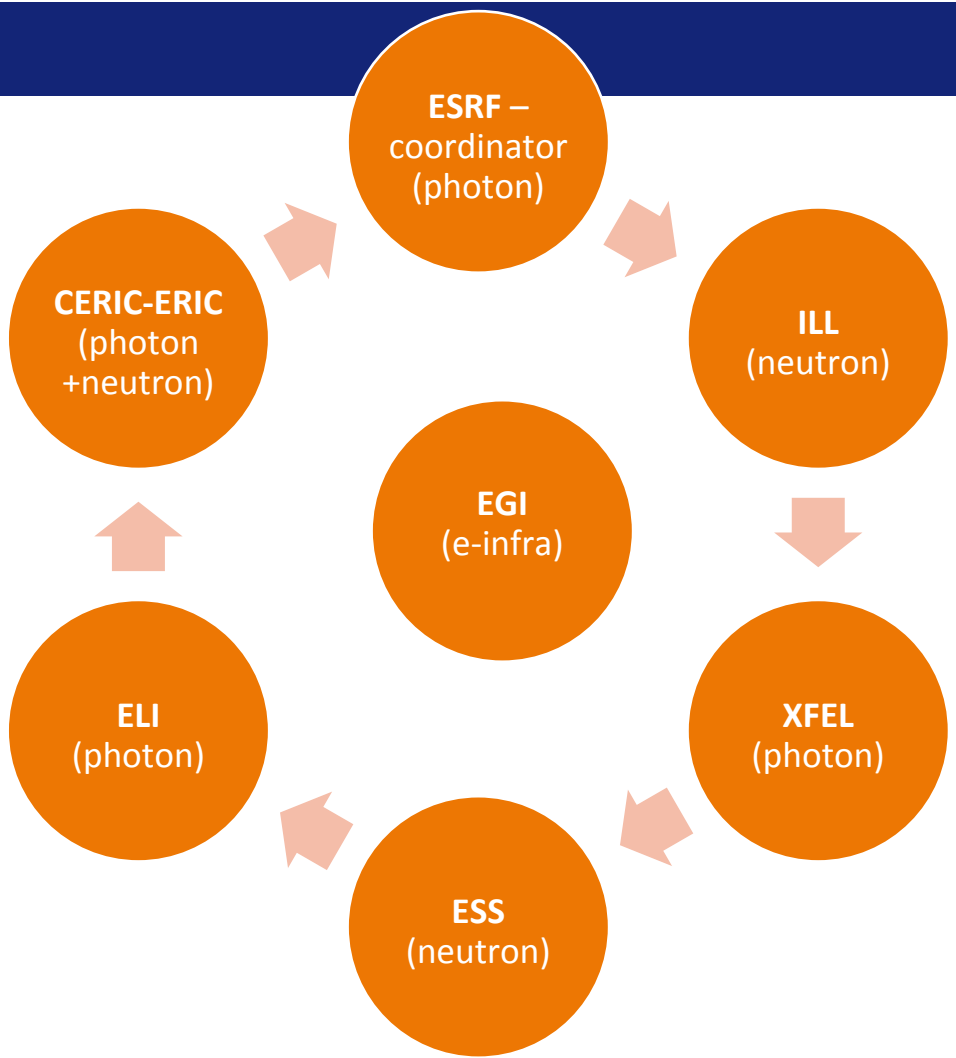
- PaNOSC addresses fundamental issues around data and providing data services

- **PanOSC aims to make the platform for FAIR Data as a Service real !**

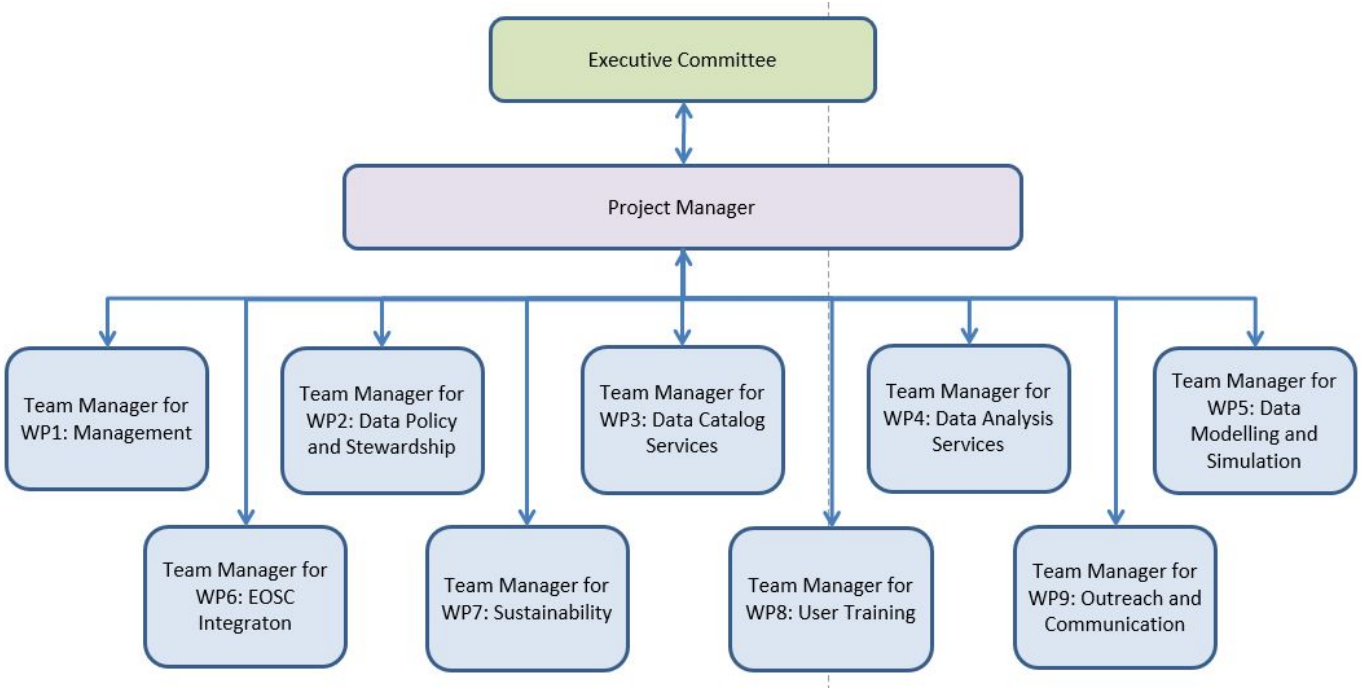
# Goals

| | |
|---|---|
| Data Stewardship | • Generalise the adoption of FAIR Open Data principles, federated data catalogs |
| EOSC | • Integrate in EOSC and OpenAire meta-catalog, AAI, data transfer + analysis services |
| Data Services | • Develop services for Data Analysis, Modelling and Simulation accessible locally and on EOSC |
| User Training | • Train users how to write DMPs, improve metadata, use DOIs for data, get credit for their data |

# Partners


ESRF – coordinator (photon)
ILL (neutron)
XFEL (photon)
ESS (neutron)
ELI (photon)
CERIC-ERIC (photon +neutron)
EGI (e-infra)

# Overview


Executive Committee
Project Manager
Team Manager for WP1: Management
Team Manager for WP2: Data Policy and Stewardship
Team Manager for WP3: Data Catalog Services
Team Manager for WP4: Data Analysis Services
Team Manager for WP5: Data Modelling and Simulation
Team Manager for WP6: EOSC Integraton
Team Manager for WP7: Sustainability
Team Manager for WP8: User Training
Team Manager for WP9: Outreach and Communication

# WP4 = Data Analysis Services

Goal: ***provide data analysis services for analyzing data available in data catalogs on the partner sites, EOSC platforms and/or public or commercial clouds***

T4.1 Harmonise technology, support, workflows between partners

T4.2 Prepare technological platforms for DAAS

T4.3 Harmonise access portals between partners and with EOSC

T4.4 Deploy Jupyter based cloud services

T4.5 Package applications for easy deployment

T4.6 Document use cases, including Jupyter notebooks

# THANKS