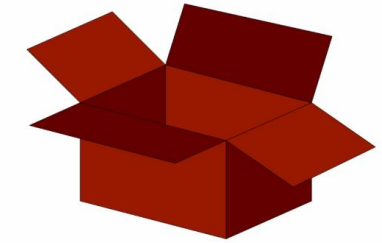
Scientific Data Policy of the European XFEL

Krzysztof Wrona
on behalf of the Data Policy and Data Reduction working group

European XFEL users' meeting 2024

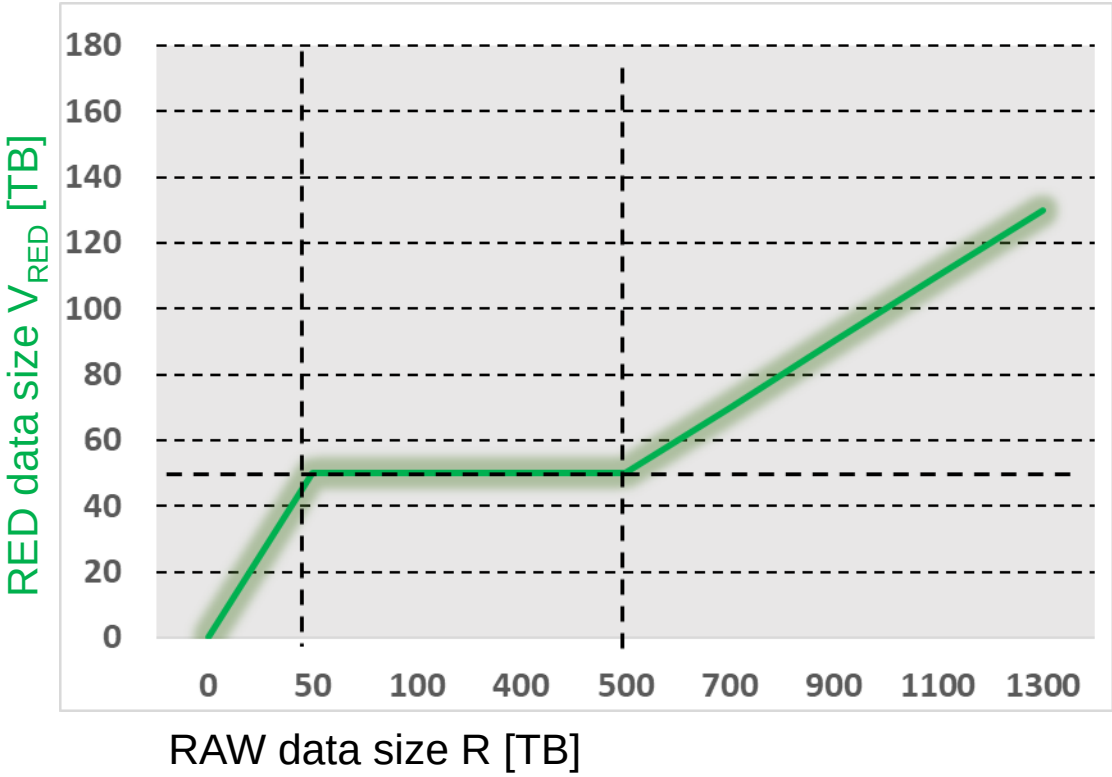
Satellite Workshop: Data Management, Analysis and Reduction at European XFEL
Friday, January 26

RED data concept



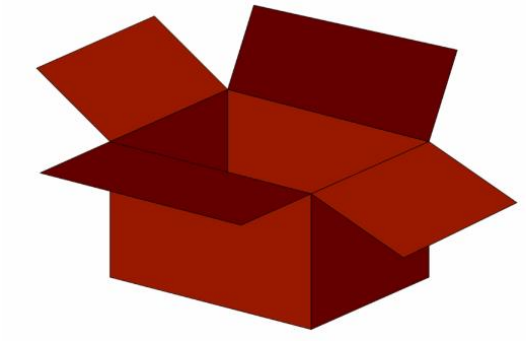
- The size of raw data **(R)** determines the retained volume.
- Limit specified in QoDS:

$$V_{RED} = \max(10\% \; R;\ \min(50TB;\ R))$$

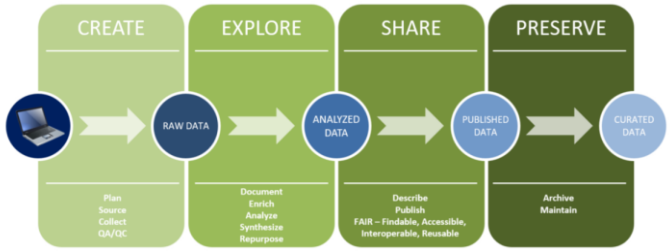


- If the size of raw data recorded for a proposal is:
  - **below 50TB**
    - ► you can retain data up to the size of raw data
  - **above 500TB**
    - ► you can retain 10%
  - **between 50TB and 500TB**
    - ► you can retain 50TB

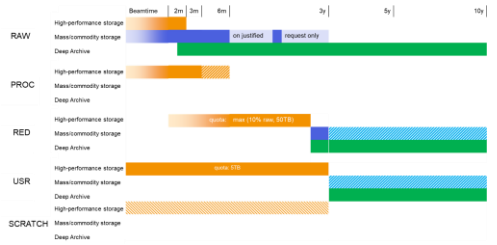
Highlights of the Revised Scientific Data Policy



- Red(uced) data



- Data management plan



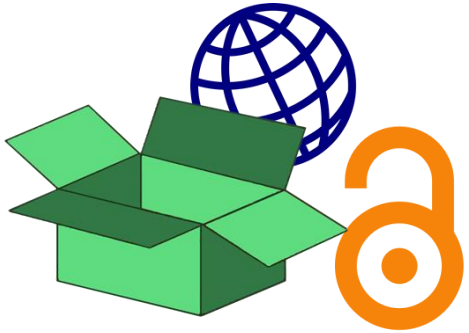
- Updated data retention periods

- PI's rights and responsibilities

- Supported data and metadata formats

- More on persistent identifiers (DOI, ORCID)

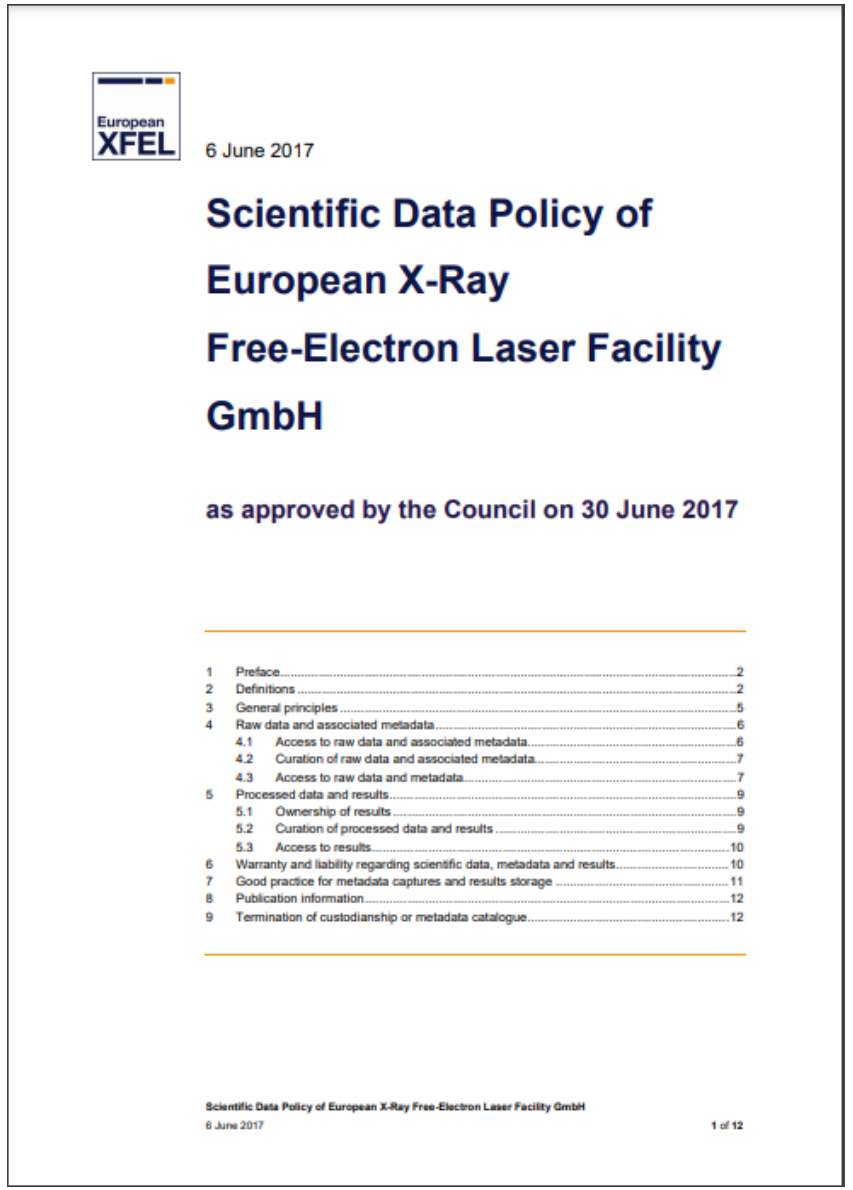
- Preserving user auxiliary data and metadata



- Open Access

Scientific Data Policy 2017 - 2024

- The current European XFEL Scientific Data Policy was approved by the European XFEL Council shortly before European XFEL transitioned from the Construction to the Operation mode.
- The policy is based on recommendations from the PaN-data European Strategic Working Group from 2011 following the majority of modifications from ILL and ESRF
- The policy defines the obligations and rights of the facility and its users with respect to the scientific data
- It allows a coherent approach to the data management services across different instruments and laboratories



European XFEL

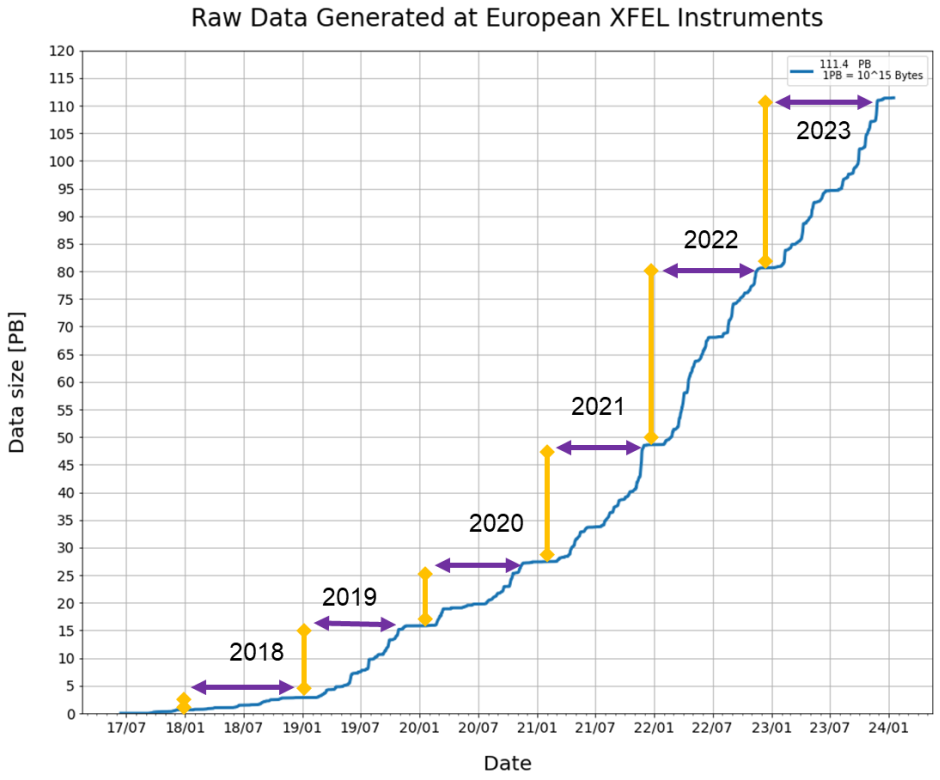
6 June 2017

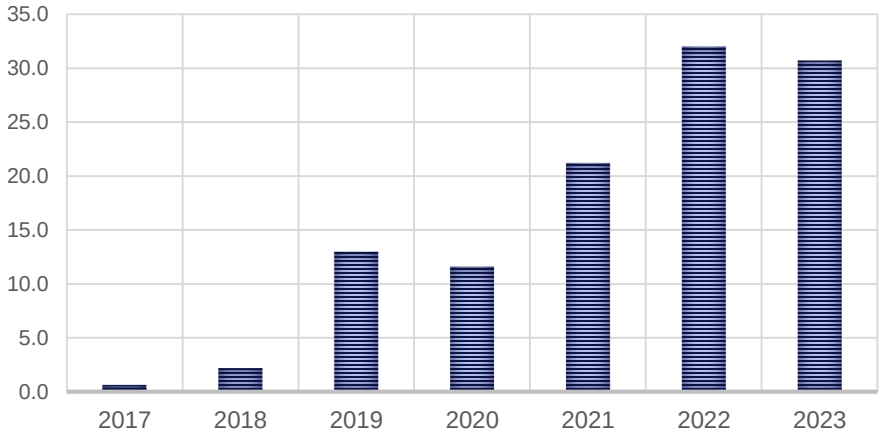
Scientific Data Policy of European X-Ray Free-Electron Laser Facility GmbH

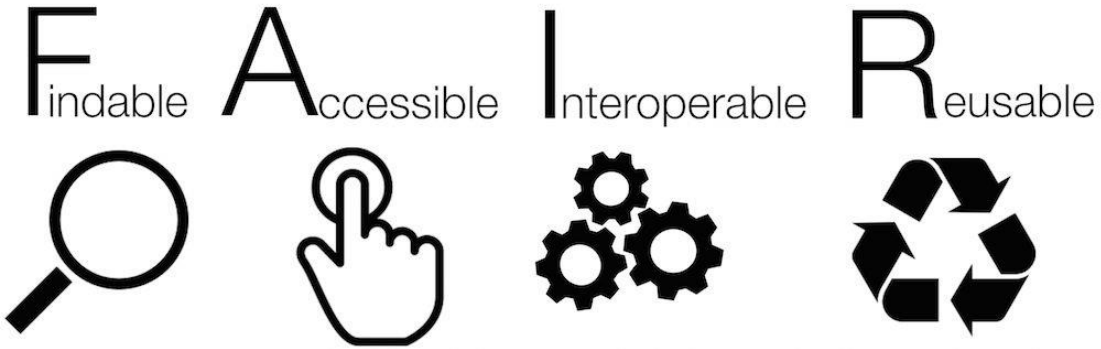
as approved by the Council on 30 June 2017

[in]Valuable Scientific Data

- We continue to generate massive amounts of scientific data
- The approach of storing all generated data long-term is becoming unsustainable
- We have an obligation to increase the value of the data by adhering to the FAIR principles







- In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Nature Scientific Data*.
- The ultimate goal of FAIR is to optimise the reuse of data
- The specific recommendations applicable for photon and neutron research infrastructures were codified in the PaNOSC project

- **Findable**: This is the first step for the reuse of the data, which is to find them.
- **Accessible**: Once the user has found the data, he must know how to access it
- **Interoperable**: Data needs to be integrated not only with other data, but also with applications or workflows for analysis, storage and processing.
- **Reusable**: other researchers can reuse all data.

Process of Updating the Data Policy

04/2022
Concept document: SDP draft start based on PaNOSC recommendations

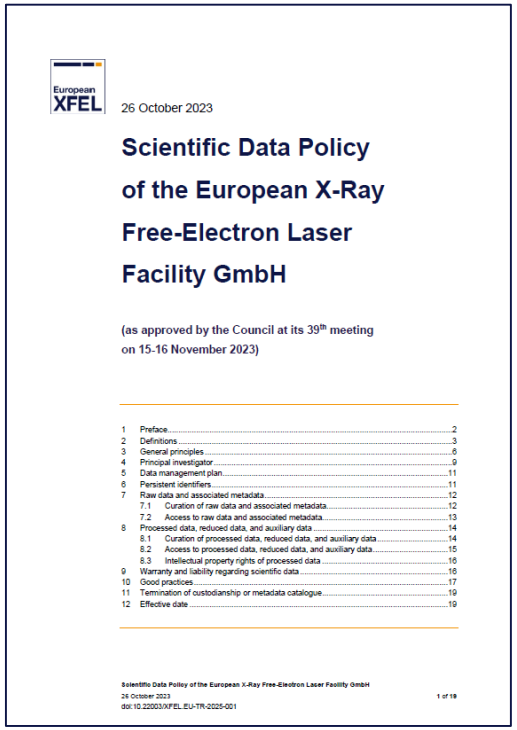
06/2022
Facility-wide working group creation

01/2023
Interim status presented at user meeting

03/2023
- Consultation with the EuXFEL Scientific Advisory Committee
- Legal consultations and adjustment started

06/2023
Consultations with selected users

11/2023
- SDP approved by the EuXFEL Council


European XFEL

26 October 2023

Scientific Data Policy of the European X-Ray Free-Electron Laser Facility GmbH

(as approved by the Council at its 39th meeting on 15-16 November 2023)

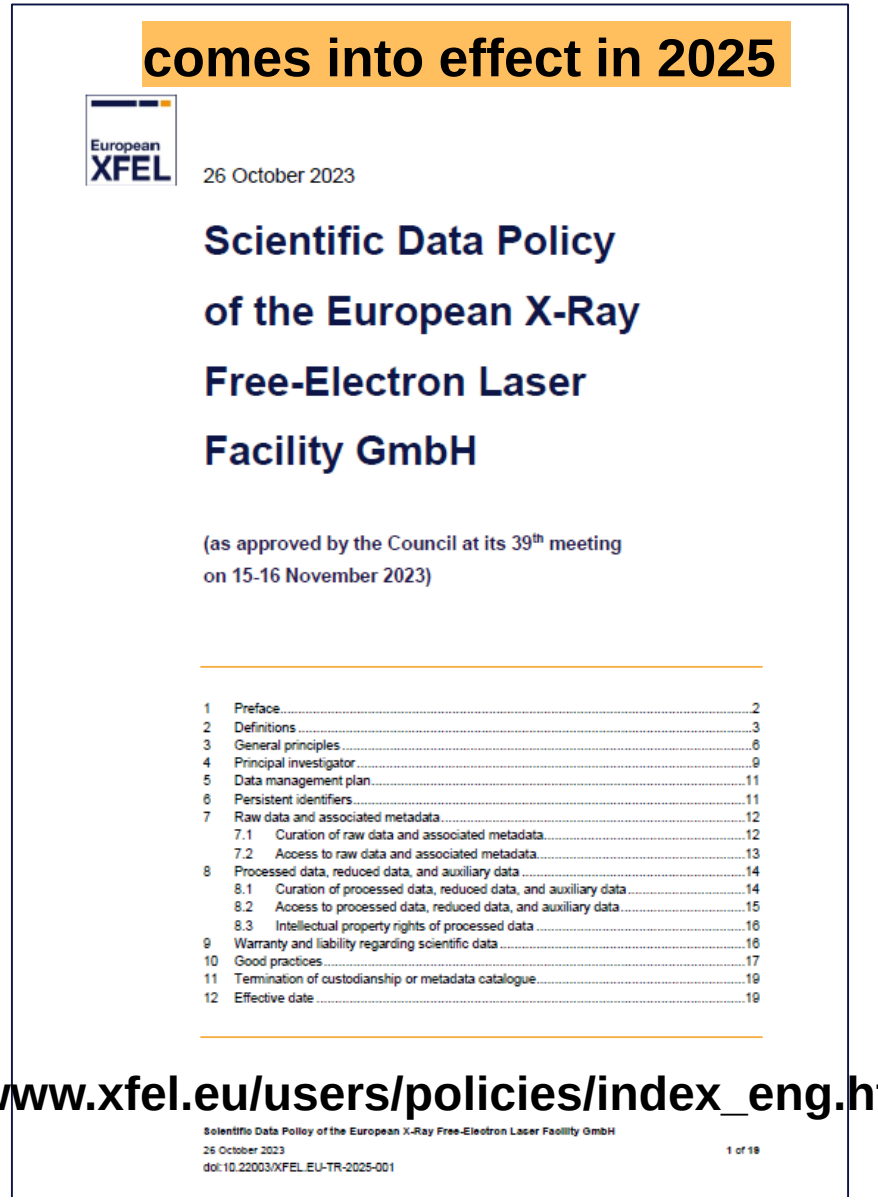
Scientific Data Policy of the European X-Ray Free-Electron Laser Facility GmbH
26 October 2023 1 of 19
doi:10.22003/XFEL.EU-TR-2025-001

doi:10.22003/XFEL.EU-TR-2025-001

Updated Scientific Data Policy 2025+

- European XFEL Council **approved** the updated Data Policy in **November 2023**
- Updated Data Policy **will come into effect in 2025** for proposals which receive beamtime starting from 2026
- The implementation phase started recently, following the SDP approval

comes into effect in 2025



European XFEL

26 October 2023

Scientific Data Policy of the European X-Ray Free-Electron Laser Facility GmbH

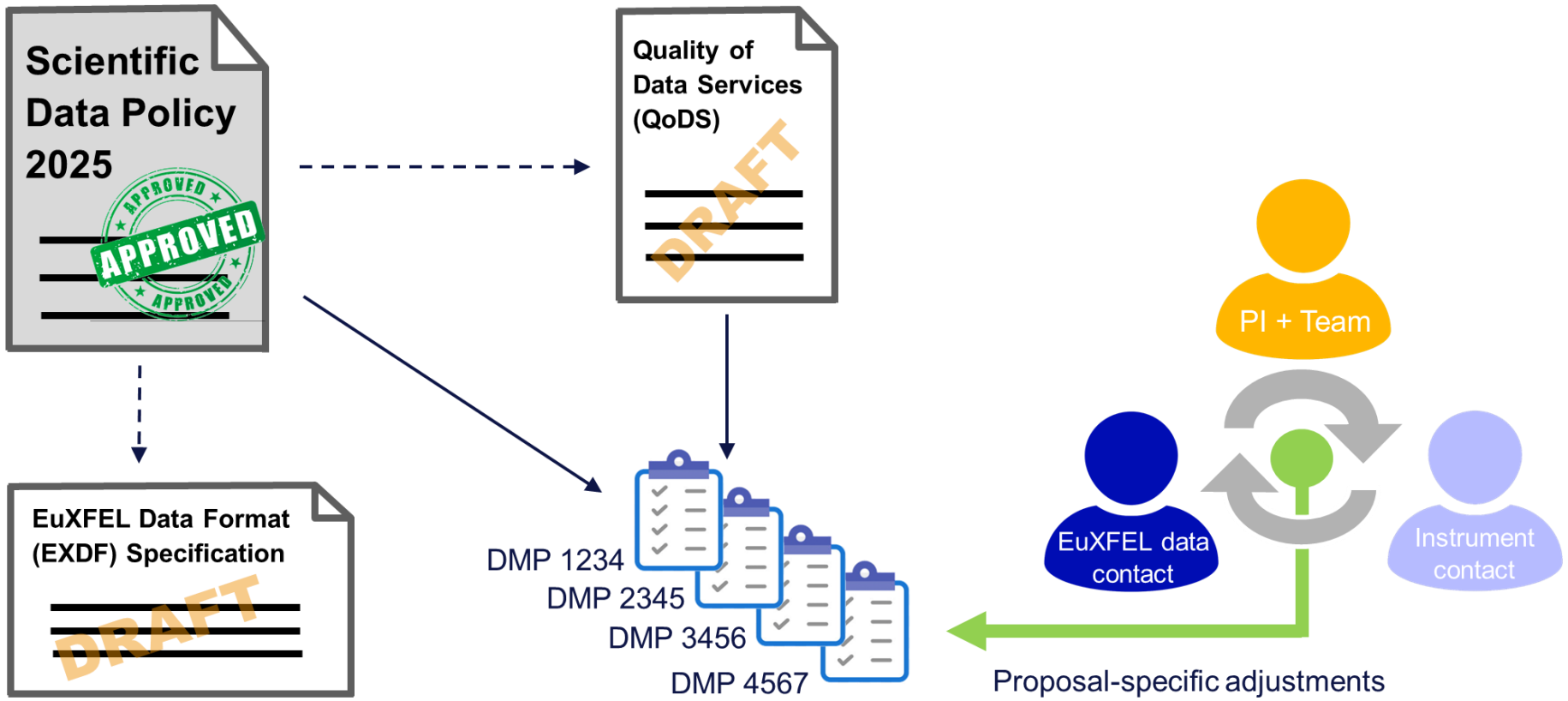
(as approved by the Council at its 39th meeting on 15-16 November 2023)

Scientific Data Policy of the European X-Ray Free-Electron Laser Facility GmbH
26 October 2023 1 of 19
doi:10.22003/XFEL.EU-TR-2025-001

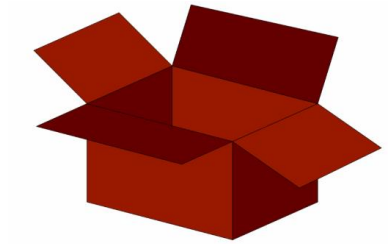
https://www.xfel.eu/users/policies/index_eng.html

doi:10.22003/XFEL.EU-TR-2025-001

Scientific Data Policy – big picture



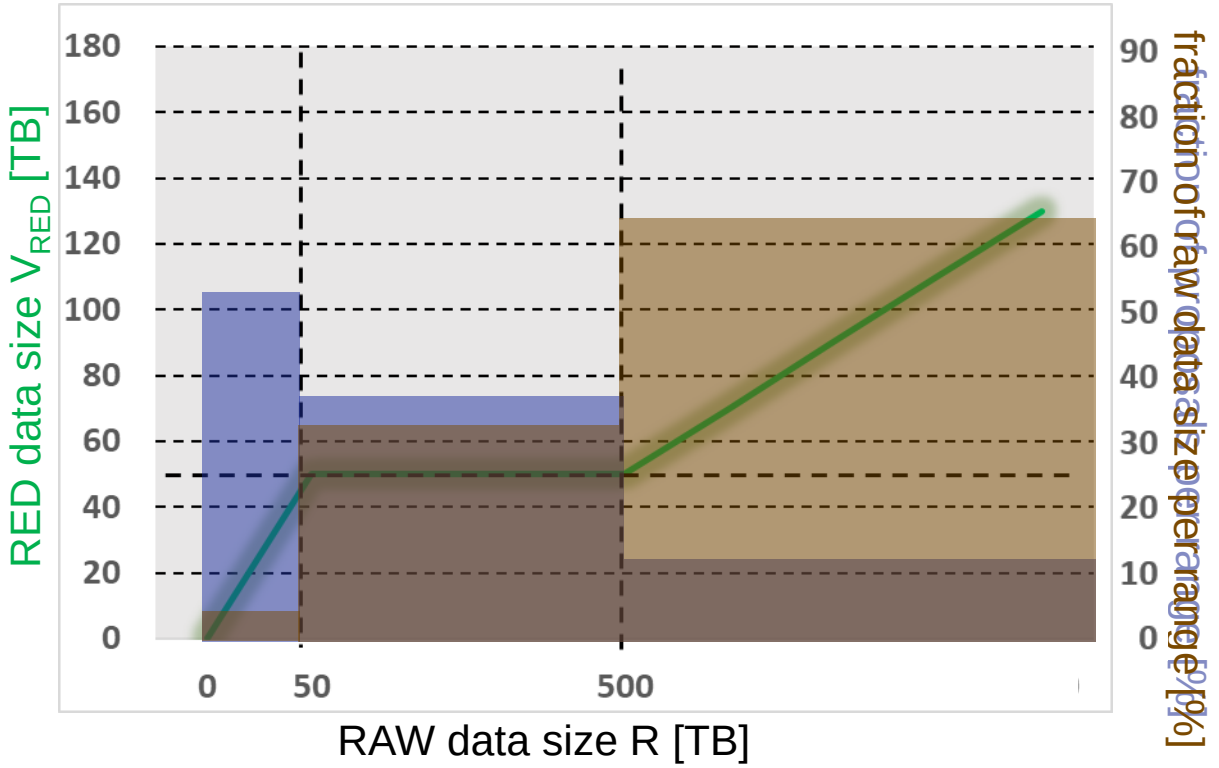
RED data concept



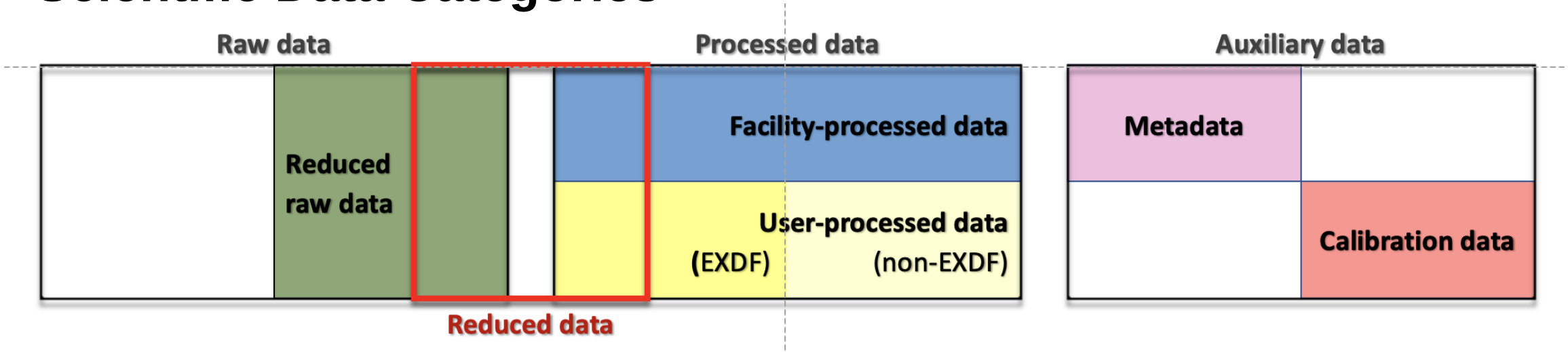
- The size of raw data **(R)** determines the retained volume.
- Limit specified in QoDS:

  $V_{RED} = \max(10\% \; R; \min(50TB; R))$

  (The parameters in the formula may evolve in the future.)
- If the size of raw data recorded for a proposal is:
  - **below 50TB**
    - ► you can retain up to the size of raw data
  - **above 500TB**
    - ► you can retain 10%
  - **between 50TB and 500TB**
    - ► you can retain 50TB



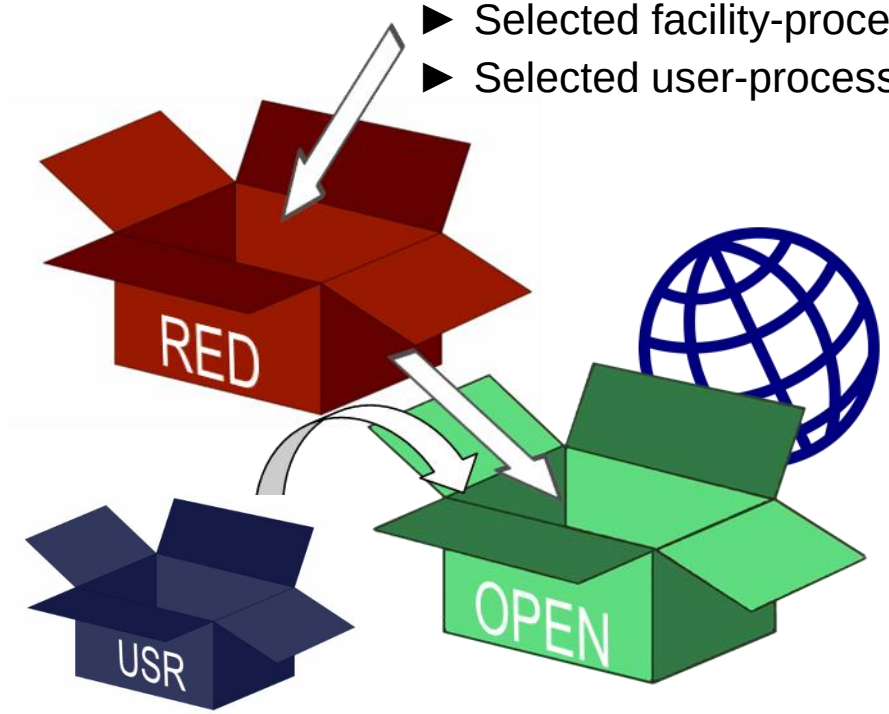
Scientific Data Categories



- Reduced data content (examples):
  - Selected runs of high-quality raw data, portions otherwise not modified
  - All runs of raw data, portions comprised of event-selected detector frames
  - Only facility-processed (EXDF) data, likely selected and/or transformed
  - A mixture of selected raw data and facility-processed data or user-processed data in EXDF format
- Auxiliary data (examples)
  - Detector geometry
  - Sample images or metadata
  - Logbook records
  - Processing scripts
  - Information about used software

OPEN data – what, when, how?

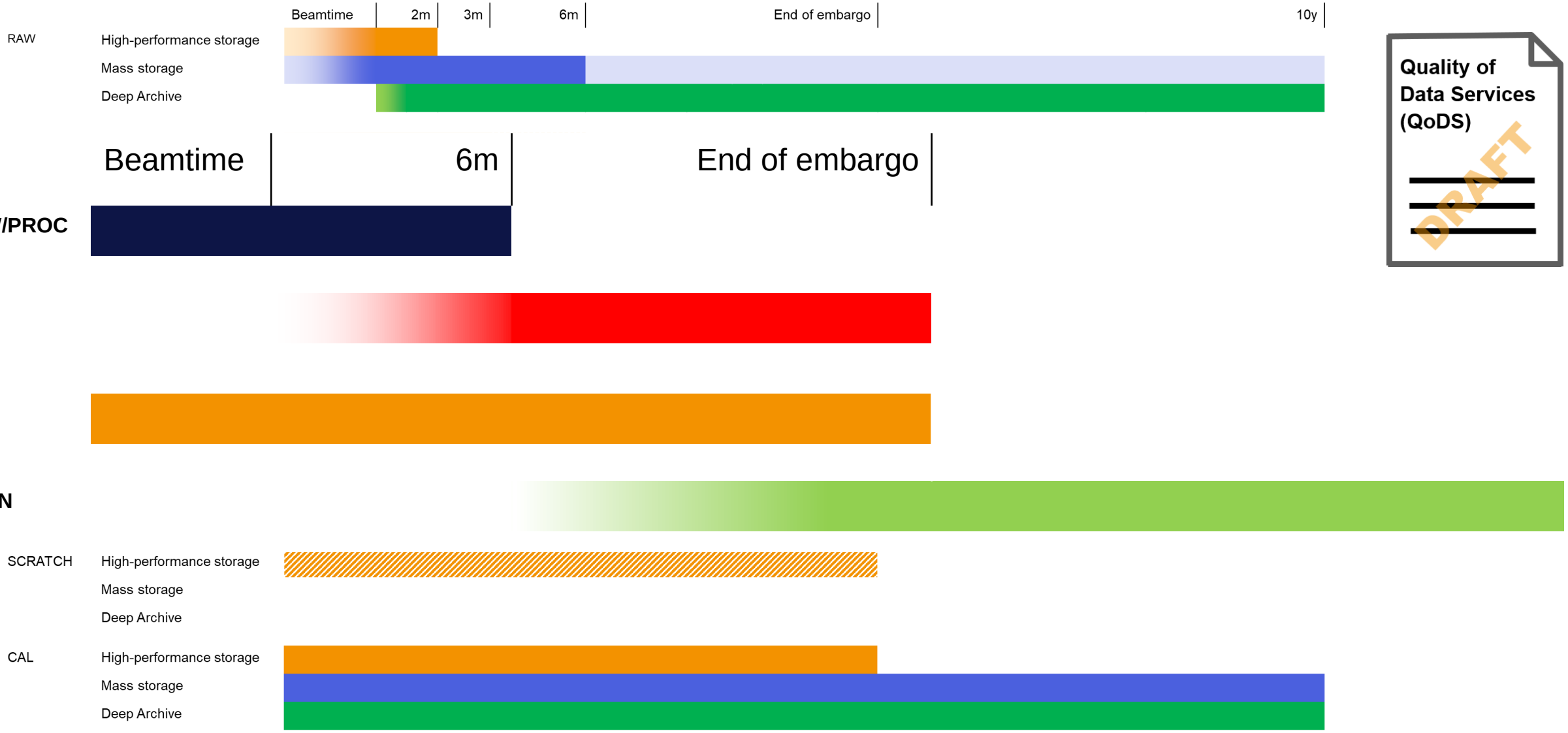
- Selected raw data
- Selected facility-processed data
- Selected user-processed data in the European XFEL-supported format



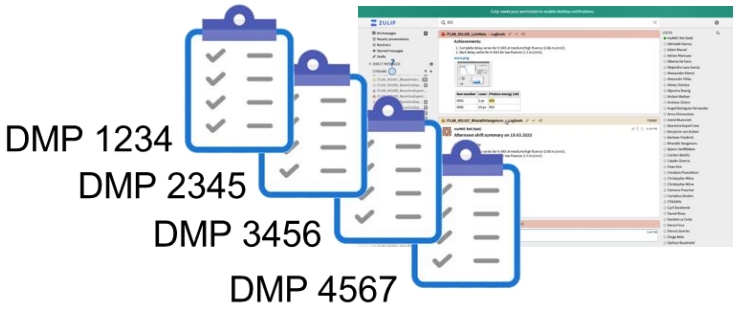
- Selected auxiliary data
- Selected user-processed data in any format

- You can define specific data sets, e.g. by including all the scientific data necessary to reproduce the results of the corresponding journal publication or those related to a certain scientific question.
- The data can be opened at any time during the embargo period
- DOI will be generated for each defined dataset
- At the end of the embargo period, RED data becomes automatically OPEN

New Data Retention Scheme



Scientific Data Policy Implementation – major tasks



Data Management Plan

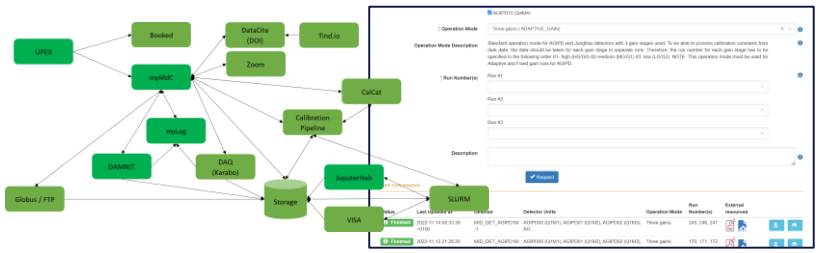
Data reduction

Selection

- By experiment setup: runs
- Temporal: trains, pulse pattern
- By event: “hit” frames
- Spatial: ROIs, modules

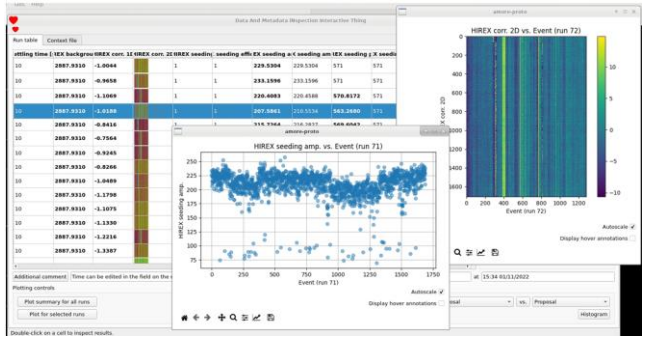
Transformation

- Temporal averaging (e.g. trains)
- Spatial integration (e.g. azimuthal)
- Conversion to photon counts
- Compression (lossless / lossy)



Data management services

Metadata harvesting



Validation of reduced data and monitoring of data quality

Summary and outlook

- The updated SDP aims to **increase your experience, ensure successful beamtimes and data analysis** up to the publication phase, making data more FAIR throughout
- The updated SDP will **come into effect from January 2025** and will be applicable to beamtimes starting from 2026
- Data reduction tools and services are being implemented; data reduction at the instruments has started
- DMPs will go into a pilot phase in early 2024, to be incrementally expanded and streamlined by experience
- Let us know if you want to contribute (feedback, discussion, pilot) to the implementation of the SDP: **data-policy@xfel.eu,** we will distribute further information using the **computing@xfel.eu** mailing list