Quality of Data Services

This document defines the levels of quality of data services and retention regulations for different categories of scientific data.

1. Data categories

Figure 1 shows scientific data categories and their hierarchical relation as defined in the Scientific Data Policy in section 2. Scientific data comprises of raw, processed and auxiliary data.

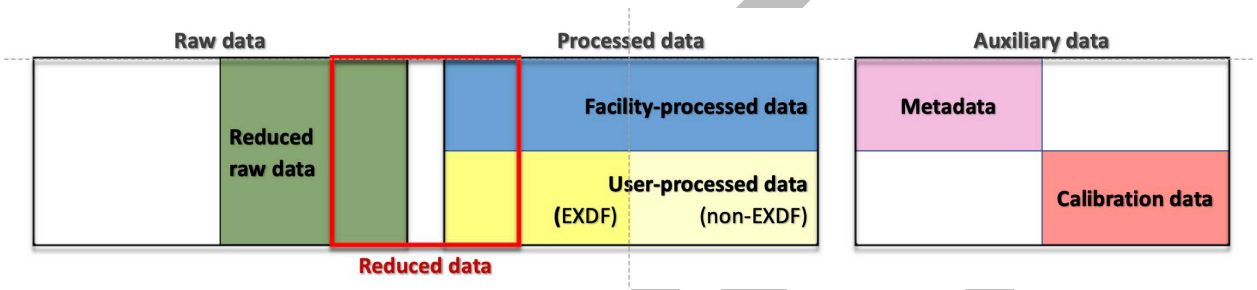

Figure1. Categories of scientific data as defined in the Scientific Data Policy

2. Data Storage classes

Data storage classes are part of the implementation concept and facilitate grouping the data categories according to the retention rules and applied levels of quality of data services. The following data storage classes are introduced here:

RAW: contains raw data as per definition 2.15 or, if applicable, reduced raw data as per definition 2.16 of the Scientific Data Policy.

PROC: contains facility-processed data as per definition 2.18 of the Scientific Data Policy.

RED: contains reduced data as per definition 2.20 of the Scientific Data Policy.

USR: contains auxiliary data as per definition 2.21 of the Scientific Data Policy, and user-processed data as per definition 2.19 of the Scientific Data Policy.

SCRATCH: contains temporary processed data as per definition 2.17 of the Scientific Data Policy.

OPEN: contains an openly accessible, immutable selection of RED and USR.

CAL: contains calibration data as per definition 2.22 of the Scientific Data Policy.

3. Quality of data services

Scientific data are stored on and accessible from storage systems which offer different levels of quality of service. In the European XFEL environment, the following storage systems are available:

High-performance storage: implemented with the focus of high performance and redundancy, tailored for large scale data processing and analysis.

Mass storage: implemented with the focus on high capacity and scalability, balancing read/write performance and latency against cost per unit storage.

Deep Archive: implemented with the focus on long-term data preservation, prioritising storage density and safety at the expense of access latency.

4. Data Retention

Figure 2 shows the retention periods for the defined storage classes and taking into account quality of data services.

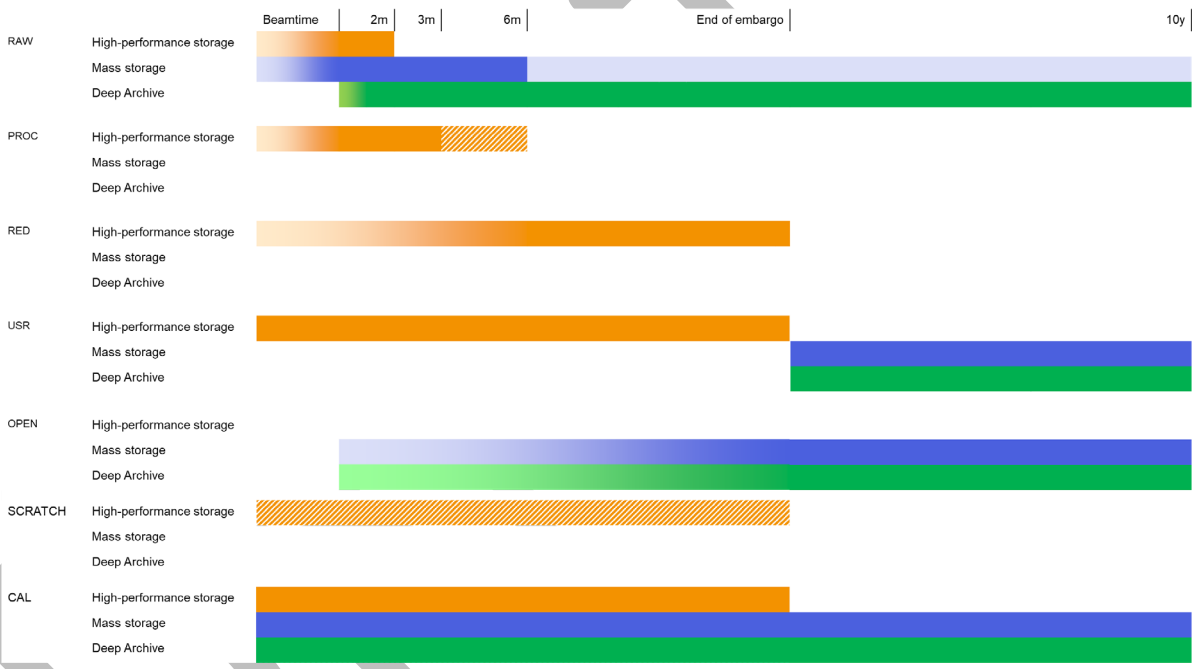

Figure 2. Retention periods for the defined storage classes and storage systems

In the following the retention regulations are described in detail:

RAW data storage class

Initially, the raw data is stored on high-performance storage for two months and in parallel, copied to mass storage and deep archive. RAW data can be removed from high-performance storage by European XFEL GmbH earlier than two months if storage space is required for data taking from subsequent experiments. RAW data are removed from mass storage after six months. Beyond this point, the raw data will only be restored from deep archive to mass storage (a) during the embargo period upon a formal request from the PI or (b) after the embargo period upon a formal request from any party interested in use of the data. Such requests must be initially assessed by the instrument lead, and should they consider the request justified, they must then make a formal request to the

European XFEL Scientific Data Manager. Final approval will be provided by the European XFEL Executive Data Manager.

Only a single copy of RAW data will be preserved in deep archive for long-term storage.

PROC data storage class

PROC data is initially stored on high-performance storage for three months. The storage period can be extended once for another three months. For operational reasons, the storage period can be shortened by the European XFEL GmbH if the reproducibility from raw data is guaranteed.

RED data storage class

RED data is primarily dedicated for user analysis during the embargo period. The content of the RED storage class must be finalized within the first six months after the beamtime and subsequently becomes read-only. On request of the PI, the entire or partial content of the RED storage class can be made open before the default embargo period, e.g. in the context of a journal publication. At the time of opening a dataset it becomes immutable.

The maximum size of the RED data is defined with the following formula:

max(10% of RAW data size; min(50TB;RAW data size))

RED data is kept on high-performance storage until the end of the embargo period. After the embargo period, this data is entirely accessible under the rules defined in the Scientific Data Policy and as implemented in terms of the OPEN data storage class (see below).

USR data storage class

USR data is kept on high-performance storage until the end of the embargo period and then it is moved to mass storage and deep archive. Selected USR data can be made open access under the rules defined in the Scientific Data Policy and as implemented in terms of the OPEN data storage class. After the embargo period, the remaining USR data are accessible upon request from the PI. Such requests must be initially assessed by the instrument lead and should they consider the request justified, they must then make a formal request to the European XFEL Scientific Data Manager. Final approval will be provided by the European XFEL Executive Data Manager.

There is a fixed quota of 5TB for USR data. In order to ensure data safety, filesystem snapshots and backups are provided during the embargo period.

OPEN data storage class

The purpose of the OPEN data storage class is to facilitate open access in conjunction with registered DOIs and curated immutable data sets. Data of this storage class derives from the RED storage class, representing a complete copy upon expiry of the embargo period for a given proposal. The PI is strongly encouraged to enrich RED data with relevant auxiliary data and other scientific data stored in USR in order to comply with the FAIR principles.

Optionally, selected subsets of data in RED may be copied to OPEN before the end of the embargo period on request of the PI, e.g. with the purpose of publishing results derived from the RED selection.

OPEN data is stored long-term on mass storage and in deep archive.

SCRATCH data storage class:

Scratch disk space is provided for temporary storage of user data for the duration of the embargo period. The lifetime of data in SCRATCH is not guaranteed, and the scratch space will be regularly cleaned.

CAL data storage class

Initially, the data is stored on high-performance storage and copied to mass storage and deep archive. CAL data is openly accessible from mass storage for general re-use.