DataSTaMP

MAX IV Controls & IT Group

2022-06-0

“

The vision is to provide data generated at MAX IV a ‘permanent home’, and to increase the value of the data generated at MAX IV in terms of benefit to research in accordance with the FAIR-principles.

DataSTaMP

KAW has awarded MAX IV funding towards data storage and data management. Hence, MAX IV has designed a project called DataSTaMP - **Data STorage and Management Project** set out to improve and extend the data storage infrastructure, and to elevate the data management services offered to the user community at MAX IV.

- Total Budget: 75 MSEK
- Duration: Aug. 2019 - Dec. 2023
- Work effort: 8.7 FTE (in average)

Knut and Alice Wallenberg Foundation

Overview of Budget Plan

DataSTaMP Cost Component	2019	2020	2021	2022	2023	Total Project Costs
Tape storage	8 850	1 050	1 400	6 800	1 750	19 850 kSEK
Disc storage	4 000	7 500	4 000	0	0	15 500 kSEK
Cloud storage	360	720	720	720	1 080	3 600 kSEK
Servers	200	700	100	200	0	1 200 kSEK
Network	500	300	0	160	0	960 kSEK
Personnel functions	1 954	8 027	10 383	8 094	5 432	33 890 kSEK
TOTAL	**15 880**	**18 332**	**16 639**	**16 010**	**8 299**	**75 000 kSEK**

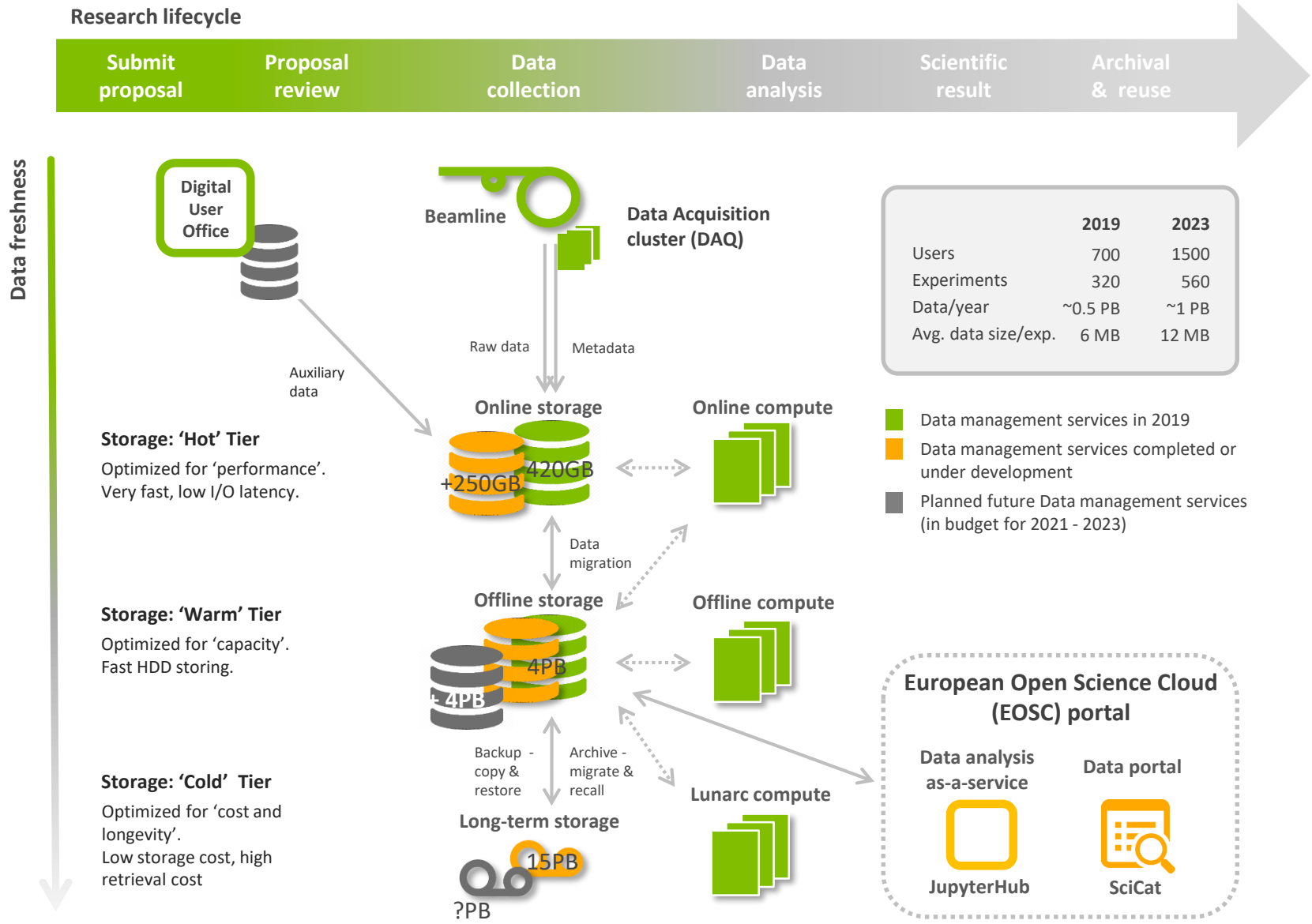
*54%

**46%

DataSTaMP have heavily invested in **Data Storage Infrastructure** in 2019 - 2020 [*]. The majority of the new and improved **Research Data Services** will be based on the new storage infrastructure and thus established during 2021 - 2023 [**].

Key Project Objectives

- **To provide a 'permanent' home for scientific data generated at MAX IV**
  - No need to move it if not necessary
  - Preservation period is according to budget & rate of data production
- **To provide a valuable data store (curated with meta-data)**
  - Data easy accessible, searchable, and reusable - in accordance with the FAIR principles
  - Could be open access later
- **To provide services for the long-term stored data**
  - Examples:
    - Collaborative tools for remote data analysis
    - Data browsing and visualisation
    - Full or partial data download
    - Automated migration of data between tape (cold) and disk (hot)
    - Prevent primary data loss or corruption by storing copy on a separate medium


Research lifecycle
Submit proposal
Proposal review
Data collection
Data analysis
Scientific result
Archival & reuse
Data freshness
Digital User Office
Beamline
Data Acquisition cluster (DAQ)
Raw data
Metadata
Auxiliary data
Online storage
+250GB
420GB
Online compute
Data migration
Offline storage
4PB
+ 4PB
Offline compute
Backup - copy & restore
Archive - migrate & recall
Long-term storage
15PB
?PB
Lunarc compute
Storage: 'Hot' Tier
Optimized for 'performance'.
Very fast, low I/O latency.
Storage: 'Warm' Tier
Optimized for 'capacity'.
Fast HDD storing.
Storage: 'Cold' Tier
Optimized for 'cost and longevity'.
Low storage cost, high retrieval cost
2019
2023
Users 700 1500
Experiments 320 560
Data/year ~0.5 PB ~1 PB
Avg. data size/exp. 6 MB 12 MB
Data management services in 2019
Data management services completed or under development
Planned future Data management services (in budget for 2021 - 2023)
European Open Science Cloud (EOSC) portal
Data analysis as-a-service
Data portal
JupyterHub
SciCat

Research Data Services provided to the **MAX IV user community**

Common Data Acquisition Service

Generic service for collecting experimental data which is capable of handling data flows from the next generation of high-speed detectors.

Long-term Storage Service

3-tiered data storage service with cold storage of inactive data on cost efficient tape storage and/or in the cloud.

Visualization Service

On-the-fly visualization and inspection of large data volumes generated by high-speed detectors at a beamline.

Remote Analytics Service

Collaborative platform including JupyterHub, to remotely perform interactive scientific analyses on collected data.

Research Data Service

Portal to search, browse, download data, metadata, publications and any other digital assets related to an experiment.

User Office Services

Advanced services to assist users in the various stages of the user workflow, incl. samples handling, mail-in of experiments and block sessions.

ExPaNDS <-> DataSTaMP

European Open Science Cloud (EOSC) Photon and Neutron Data Service

- Aims to provide coherent FAIR data services to the research community
- Standardize access to all scientific data collected at the photon and neutron facilities
- MAX IV involved in…
  - ExPaNDS-WP4: EOSC data analysis services for PaN national RIs
  - ExPaNDS-WP3: EOSC data catalogue services for PaN national RIs

Remote Analytics Service

Collaborative platform including JupyterHub, to remotely perform interactive scientific analyses on collected data.

Research Data Service

Portal to search, browse, download data, metadata, publications and any other digital assets related to an experiment.

User Office Services

Advanced services to assist users in the various stages of the user workflow, incl. samples handling, mail-in of experiments and block sessions.


MAX IV
HZDR
HELMHOLTZ ZENTRUM
DRESDEN ROSSENDORF
PAUL SCHERRER INSTITUT
PSI
HZB
Helmholtz
Zentrum Berlin
SOLEIL
SYNCHROTRON
Partners
ExPaNDS
egi
UKRI
ALBA
Elettra Sincrotrone Trieste
DESY.
diamond

Project **scope** of work

WP1: Data management of experiments

The main goal of work package is to create seamless access to, and easy management of data sets from the variety of data catalogues related to user experiments.

WP2: Experimental data & metadata

Goal is to provide a acquisition software that handles high data rates, support various of detectors, and automatically record metadata.

WP3: Data evaluation

Aims at making collected data useful and reusable by providing tools and platforms for the user community post-visit, to share, process and analyse raw data remotely.

WP4: Data storage

The objective of the work package is to establish a state-of-the-art storage infrastructure where research data produced at MAX IV, can be securely stored and preserved for at least 10 years.

Long-term storage service

Backup: *Prevent primary data loss or corruption by storing a copy of the scientific data on a separate medium.*
Archival: *Preserve scientific data taken at MAX IV beamlines, raw and processed, for 10+ years.*

IBM TS4500 tape storage system

- Max storage capacity: 15PB (100PB)
- Tape drives: 6 x LTO-8 (max 12)
- Tapes: 36 I/O slots, 1320 tapes (max 1784)
- Data transfer rate: 2,1 GB/s
- System designed for HA

POD for data staging

- Cabinets: 5 x 19'' server racks
  - 10 x PDUs with 24 x c13, 4x c19 outlets.
- UPS: 60 kW. 4 modules a 20kW. (15 min full load.)
- Cooling: 2 x LCP CW, 30kW/LPC
- Integrated fire extinguish. NOVEC 1320.
- 2 x CMC + water leak detection
- Monitoring alarm: 4 temp sensors + 1 humidity sensor

Visualization and Research Data Service

Sci'Zoo' ecosystem supports the scientific metadata recording and preservation.

Beamline experimental station
Sardana to supervise, control and acquire data from instrumentation

Sardana

SciDog

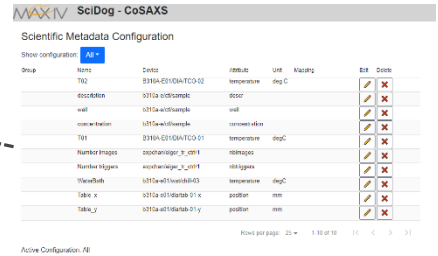
SciFish

Kafka

SciCat

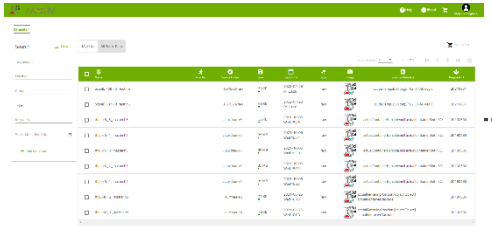
Elogy

SciToad

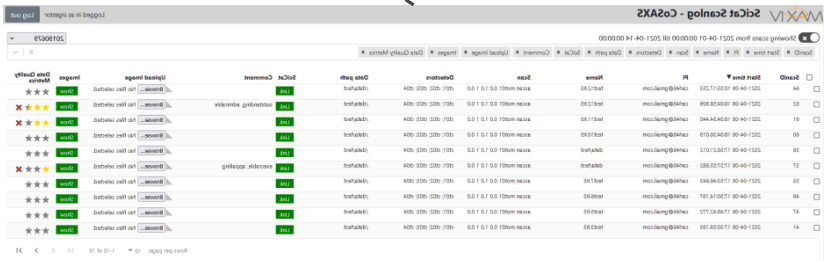


- Web UI to configure what metadata to record
- Variables fetched from control system (TangoDB)

- Electronic logbook
- Experimental notes



- Web UI to show overview of datasets
- Researcher can access, download and mint DOIs



- Web UI to show overview of scans
- Feature to comment, upload images and rate scans

Open Data Portal · PaNOSC

Datasets available for search in EOSC portal when DOI minted by researcher.


MAX IV