DSO Science Meeting on the LTDSF (April 30, 2026)

The meeting focused on the implementation and use cases of the Long-Term Data Storage Facility (LTDSF), particularly its S3-based storage system.

Participants explored several use cases, including archiving simulation results, building databases for AI workflows and reduced models, backing up data from CINECA, and supporting compliance with FAIR principles. Concerns were raised regarding data management policies, metadata requirements, and the risk of creating a "data swamp" without proper governance.

The group agreed on the need to establish a dedicated subgroup to develop policies and guidelines for LTDSF usage, addressing both technical and cultural aspects of data management. The importance of ensuring reproducibility, long-term data access, and clear procedures for data storage, access, quotas, and integration of persistent identifiers (PIDs) was emphasized.

The meeting concluded with agreement to form a smaller working group to define policies and to plan broader community training once initial guidelines are in place.

NEXT STEPS

- **Formation of a subgroup** (see below): Develop a policy document and guidelines for LTDSF usage, covering metadata, data management, and access. This should include:
    - Compliance with FAIR principles
    - PID minting
    - Storage quotas and lifecycle management
    - Procedures for approval, maintenance, and lifecycle of community databases and curated datasets
    - Requirements for metadata, immutability, and ability to be referenced (PIDs/DOIs)

Subgroup members:

- *Frank Jenko (DSO),*
- *Denis Kalupin (DSO),*
- *Rui Coelho (GUB),*
- *Pär Strand (DMP),*
- *Michal Owsiak (ACH),*
- *Maciej Brzezniak (LTDSF),*
- *Alessandro Pau (DATA),*
- *David Coster (DATA),*
- *Michele Marin (TSVV),*
- *Tobias Görler (TSVV),*
- *Daniel Told (TSVV).*

- **LTDSF team** (Maciej, Michal): Prepare and deliver a general training session on LTDSF usage, focusing on S3 functionality. Clearly communicate the trial status, quotas, and data retention limits (e.g. possible deletion after one year during the trial phase). ***Tentative deadline: May 15.***

- **Michal:** Encourage and support users in requesting trial access to LTDSF and submitting user stories via the dedicated tracker. (https://jira.eufus.psnc.pl/browse/ACH04SUPP-321)

Summary

Long-Term Data Storage Facility status

The discussion focused on the current status of LTDSF and potential data management solutions. David presented feedback from an initial training session conducted for a limited group of users testing connectivity across distributed locations. While the feedback was positive, concerns were raised about using LTDSF as a raw storage service, particularly regarding metadata enforcement and lifecycle management.

Maciej clarified that LTDSF is designed as a basic storage solution based on S3 architecture, not as a full data management platform. Pär highlighted the need to compare it with tools such as Rucio (https://rucio.cern.ch/) and Zenodo (https://zenodo.org/) and questioned whether it meets FAIR and long-term preservation requirements.

The group agreed to review user stories to better understand requirements and identify gaps between current capabilities and expected functionality.

Data Management and PID Implementation

The team discussed the Data Management Plan (DMP) and FAIR compliance, focusing on PID implementation. While LTDSF supports metadata assignment and data replication, it does not natively provide PID services or external data access mechanisms. It was noted that additional functionality, such as group sharing or external access, would require further policy decisions and implementation planning.

S3 Storage for Simulation Results

Two main use cases were discussed:

- Storing simulation results for cataloguing and reuse
- Archiving project data for long-term preservation

Concerns were raised about uncontrolled data uploads and insufficient metadata. It was estimated that approximately one petabyte of data per year could be stored across ~100 projects, reaching capacity within about eight years.

Discussion of Use-Cases for Data Storage

The group explored structured approaches to managing large datasets, including curated databases for simulations and AI workflows. A previously developed gyrokinetic database was highlighted as an example of underutilized resources, underlining the need for improved policies, visibility, and reuse strategies.

Several use cases were presented by David and discussed by the group, including:

- Maintaining simulation databases for AI model training
- Database maintenance and curation
- Backup and storage of HPC simulation results
- Data exchange platforms
- Preservation of data supporting publications
- Preservation of reference simulations (including I/O data, code versions, and computational environments)

The discussion emphasized the importance of shared platforms for collaborative AI applications, rather than isolated use cases. Key examples include curated multi-machine datasets with persistent identifiers and long-term simulation archives. Limitations of existing platforms such as Rucio and Zenodo (e.g., size constraints) further highlighted the need for a scalable and immutable storage solution.

The team also reviewed storage requirements across different time horizons, distinguishing between:

- Long-term storage (S3-based systems such as LTDSF)
- Temporary, high-performance storage (e.g. for data transfer and shared access)

LTDSF was identified as a suitable solution for storing data beyond HPC capacity, particularly for simulation results and validation datasets. The use of S3-compatible storage for AI workflows and machine learning pipelines was also noted. However, broader adoption is currently constrained by the lack of clear policies, specifications, and governance frameworks.

Finally, challenges related to the implementation of gyrokinetic IDS structures were discussed, including the significant coordination effort required and unclear cross-code use cases. The importance of preserving training datasets for reproducibility and long-term usability was emphasized, along with the need for stronger metadata standards.