DOI: 10.26314/GARR-Conf25-proceedings-14

Introducing the Elettra Scientific Data Lake: Concepts, Architecture and Select Applications

Roberto Pugliese, Matteo Billè, Marco De Simone, Iztok Gregori, Daniele Favretto, Francesco Guzzi, Aljosa Hafner, Fulvio Bille', and George Kourousias
IT Group, Elettra Sincrotrone Trieste, Italy

Abstract. The Elettra scientific Data Lake (EDL) represents a tailored adaptation of modern data lakehouse architecture for synchrotron facilities. By combining the flexibility of data lakes with the governance of data warehouses, EDL addresses the unique challenges of scientific data management including format heterogeneity, FAIR compliance, and real-time processing requirements. Built on heterogeneous on-site infrastructure spanning edge computing to HPC clusters, EDL supports custom web-based applications that transform raw experimental data into scientific insights while maintaining ISO27001 security standards

Keywords. data lake, scientific data, synchrotrons, data analysis, data management

1. Elettra Sincrotrone Trieste and Data Lakes

Elettra Sincrotrone Trieste is a multidisciplinary research infrastructure center operating two advanced light sources: the Elettra synchrotron, a third-generation electron storage ring (2/2.4 GeV) operational since 1993, and the FERMI free-electron laser. The facility serves 32 beamlines covering spectroscopy, diffraction, scattering, and imaging techniques, supporting researchers from over 50 countries. The upcoming Elettra 2.0 upgrade increases coherence by approximately 50 times, increasing X-ray brilliance and by more than two orders of magnitude [1]

Modern data lakes provide scalable repositories for storing vast amounts of raw data in native formats. Data lakehouses extend this concept by combining data lake flexibility with data warehouse performance and governance features. This hybrid architecture offers an optimal foundation for managing the complex, heterogeneous data streams generated by synchrotron experiments.

2. Scientific Data and the Elettra Data Lake (EDL)

Scientific data differs fundamentally from business data in its complexity and heterogeneity. Elettra experiments generate diverse formats including TIFF images, CIF crystallographic files, raw detector streams, CSV logs, and proprietary formats from specialized equipment. This diversity reflects the varied scientific techniques employed across beamlines.

The facility embraces open science through FAIR (Findable, Accessible, Interoperable, Reusable) principles. HDF5 serves as a primary container format for complex scientific

Conferenza GARR 2025 - Frontiere Digitali - Selected Papers
Bari 13-15 maggio 2025 - ISBN: 978-88-946629-4-8

data while maintaining crucial metadata. Digital Object Identifiers (DOIs) ensure persistent identification and citation of datasets, transforming experimental output into citable research products.

EDL adapts commercial data lakehouse concepts for scientific workflows by preserving native formats while building sophisticated metadata layers enabling cross-dataset discovery and analysis. The architecture supports streaming ingestion for real-time monitoring, automated quality assessment, and comprehensive provenance tracking linking raw data to processed results.

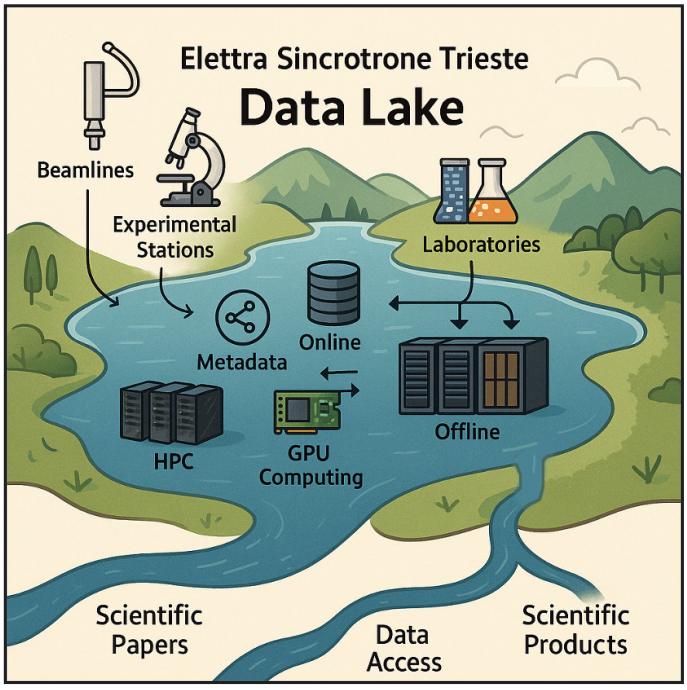


Fig. 1
Schematic representation of Elettra Data Lake

3. EDL Hardware Infrastructure

The Elettra Data Lake operates entirely on local infrastructure, ensuring data sovereignty and microsecond-level latencies critical for experimental workflows. This heterogeneous ecosystem spans from edge computing devices at beamlines handling multi-gigabyte-per-second data streams to centralized HPC resources.

The on-site data center houses diverse computational resources including high-memory nodes for large-scale processing, GPU-accelerated systems for machine learning and reconstruction, and specialized hardware for domain-specific calculations. Storage employs a tiered architecture with NVMe for hot data, disk arrays for active datasets, and tape libraries for long-term preservation.

For offline data archiving, Elettra employs an IBM Spectrum Archive 4500 tape library equipped with 8 LTO-8 drives, providing a substantial 14 petabytes of uncompressed storage across 1200 LTO-8 tapes. This system leverages IBM LTFS alongside a custom, in-house developed software solution. Built on RESTful APIs with Python and utilizing Celery workers for job distribution, this software is fully Dockerized and features a scalable architecture designed for robust scientific data archiving. It ensures data integrity through double-copy storage and SHA512 checksums for verification. The custom archiving system is seamlessly integrated with the Virtual User Office (VUO) [2]. Raw scientific data is automatically saved in a dedicated tape pool in duplicate copies immediately after

production. Principal Investigators or beamline scientists can initiate the restoration of raw data copies from offline storage at any time. Furthermore, they have the autonomy to move entire investigations to offline storage, freeing up valuable space on their online storage. Both raw scientific data and full investigations are saved in double copies within dedicated tape pools.

The environment supports MPI for distributed processing across hundreds of cores and extensive GPU computing on both desktop workstations and server-grade accelerators. Sophisticated scheduling systems unify this heterogeneous ecosystem while ensuring experimental deadlines are met.

4. Custom EDL Applications for Scientific Data

The VUO web application serves as a comprehensive platform that collects and manages all information related to an experiment, from the initial request (proposal) through to the subsequent data collection. In addition to this core function, the application provides a unified login system for all internal company services and supports the implementation of the FAIR principles. Built on top of this ecosystem are dozens of specialized applications tailored to specific needs.

EDL's effectiveness manifests through custom web-based applications that transform raw data into insights. These tools provide intuitive interfaces while incorporating advanced user management, role-based access control, and audit trails aligned with Elettra's ISO27001 certification.

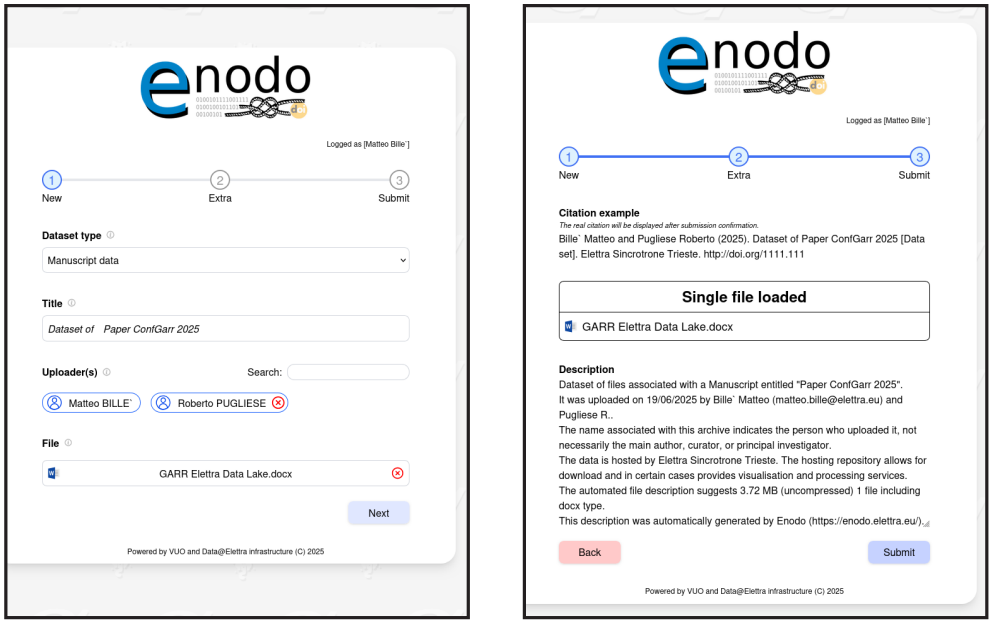


Fig. 2a - 2b

a) Enodo interface showing the complete set of input fields required to associate a dataset with a DOI.

b) pre-submission step, you where you can see the summary of your insertion

Some of these applications are:

- Enodo is a novel system inspired by Zenodo and WeTransfer. It allows users to upload manuscript datasets and associate them with a DOI (Digital Object Identifier), enabling citation in scientific publications. It ensures that datasets remain freely accessible and are persistently stored within the Elettra Data Lake, a FAIR and standardised repository. Publicly available on enodo.elettra.eu
- -XRFitVis provides an interactive environment for visualizing the results of XRF (X-ray Fluorescence) experiments. Built using web technologies, it allows researchers to access the tool both on-site and remotely. Publicly available on vuo.elettra.eu/go/xrfitvis
- -STP3 supports a dedicated micro-tomography beamline and operates on specialized hardware due to the computational demands of reconstruction. Its interface allows users to define optimal parameters and obtain full reconstructions of 100GB datasets in under 10 minutes. Used by the SYRMEP beamline.
- 4DHorizon is designed for visualizing ARPES (Angle-Resolved Photoemission Spectroscopy) data. It supports both 2D and 3D datasets and offers multiple adjustable parameters to modify the LUT and histogram, perform k-space transformations, extract the spectrum of the current slice, and enable smooth volume slicing for intuitive navigation through the volume. Used by the BaDElPh beamline.

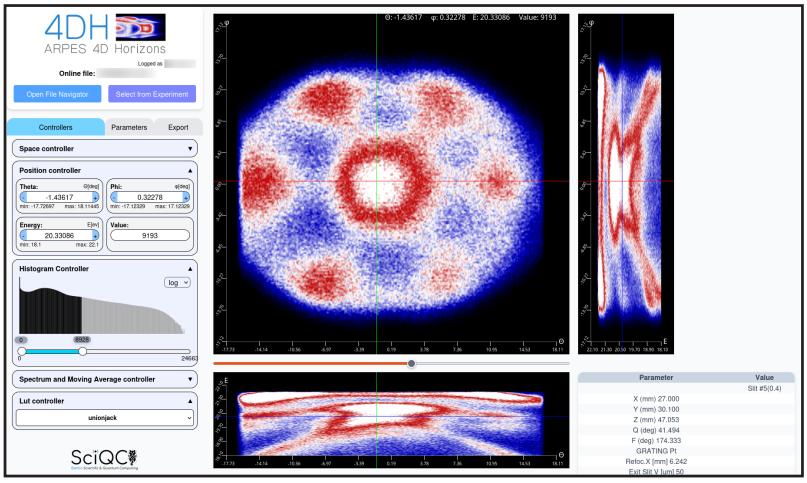


Fig. 3
Visualization of a 3D volume in the application. The left panel displays various visualization controls, while the right panel shows the rendered images along with key parameters and a navigation cursor for exploring the data

- Darkiver acts as a platform for file compression and decompression services. It offers a variety of conversion options, enabling users to upload files and quickly retrieve them in the desired output format. R&D in the context of PANOSC EOSC EU Node.
- eAI is a collective of experimental services and applications based on local LLMs. They meant to explore locally deployed services similar to ChatGPT but also Elettra specific applications for translation, summarization, scientific report generation and similar tasks. Available to Elettra personnel at BETA on eai.elettra.eu

Each application underwent co-development with scientific staff, ensuring interfaces match experimental workflows. Web-based architecture enables remote collaboration and real-time monitoring, proving invaluable for international research teams.

5. Conclusions and Future Perspectives

Synchrotron facilities generate data volumes that challenge traditional management approaches. EDL demonstrates that successful scientific data infrastructure requires deep integration with experimental workflows and flexibility to evolve with emerging methodologies. The heterogeneous hardware ecosystem provides the computational diversity necessary for the full spectrum of scientific analysis.

Custom applications showcase the importance of domain-specific tools in democratizing access to sophisticated analysis capabilities. As Elettra transitions to Elettra 2.0, the infrastructure must evolve correspondingly. Machine learning and Artificial Intelligence will play increasingly prominent roles in both analysis and experiment optimization. Through continued innovation, Elettra is establishing a model for transforming the data deluge into accelerated scientific discovery.

Acknowledgments

EDL requires competence and contributions from personnel beyond the list of authors. We acknowledge the contribution of the whole IT Group and of many beamline scientists of Elettra Sincrotrone Trieste.

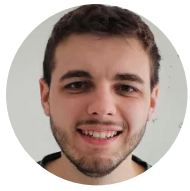
References

[1] https://www.elettra.eu/it/lightsources/elettra-2-0/elettra-2-0.html
[2] https://vuo.elettra.eu

Authors

Roberto Pugliese roberto.pugliese@elettra.eu

Roberto Pugliese, is the Deputy General Coordinator and IT Director at Elettra Sincrotrone Trieste. He holds a Ph.D. in Management, an MSc in Computer Science, an MBA, and PMP certification. Innovation Manager and Singularity University alumnus and ambassador, his research spans AI, robotics, telepresence, and data science, with publications in top journals like JSR and NIM. He supervises and directs the Elettra Data Lake.



Matteo Billè matteo.bille@elettra.eu

Matteo Billè is a scientific software engineer in the Scientific and Quantum Computing unit at Elettra Sincrotrone Trieste (quantum.elettra.eu). He is involved in data analysis projects with a focus on advanced visualization, in the development of the scientific data lake Data@Elettra, and in AI projects on local LLMs.

Scicomp Group sci.comp@elettra.eu