PSI Center for Scientific Computing, Theory and Data

OpenEM Open research data infrastructure for electron microscopy in Switzerland

Spencer Bliven[1*], Despina Adamopoulou[2], Sofya Laskina[3] , Carlo Minotti[1], Attila Nacsa[4], Yves Tittes[5], David Wiessner[6], Philipp Wissmann[7], Henning Stahlberg[3], Robbie Loewith[4]

[1]Paul Scherrer Institute (PSI), [2]Swiss Federal Laboratories for Materials Science and Technology (EMPA), [3]École Polytechnique Fédérale de Lausanne (EPFL) , [4]University of Geneva (UNIGE), [5]University of Basel (UNIBAS), [6]University of Bern (UNIBE), [7]Federal Institute of Technology Zurich (ETHZ). *spencer.bliven@psi.ch

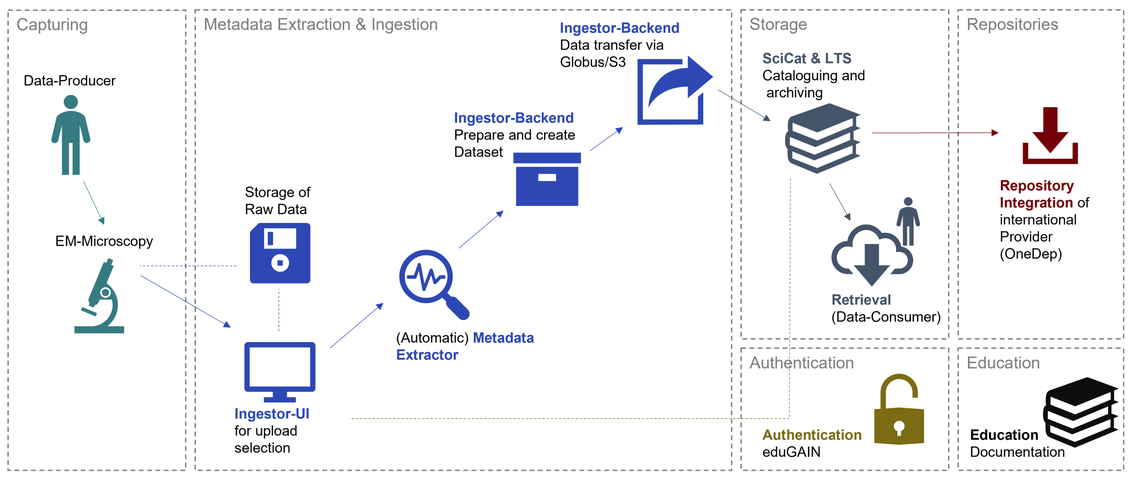
What is OpenEM?

The Open EM Data Network (OpenEM) is a consortium of Swiss electron microscopy (EM) facilities working together to improve Open Research Data (ORD) practices in the Swiss EM community.

OpenEM will target both researchers producing EM data and consumers of open data for additional science. Data producers benefit from more streamlined data collection, standardized facilities, easier deposition for publication, and adherence to data management policies. The wider availability of open EM data brings numerous benefits, including reproducing results, applying new techniques to old data, training AI & other new methods, and mining data for new insights.

https://swissopenem.github.io/

Architecture Overview



The OpenEM ecosystem is built on a microservice architecture. Users interact with the central SciCat Data Catalog at PSI, which provides the user interface, backend API, storage for metadata in MongoDB, and integration with external services. Authentication is supplied by the eduGAIN federation, allowing users from most academic institutes to access the service.

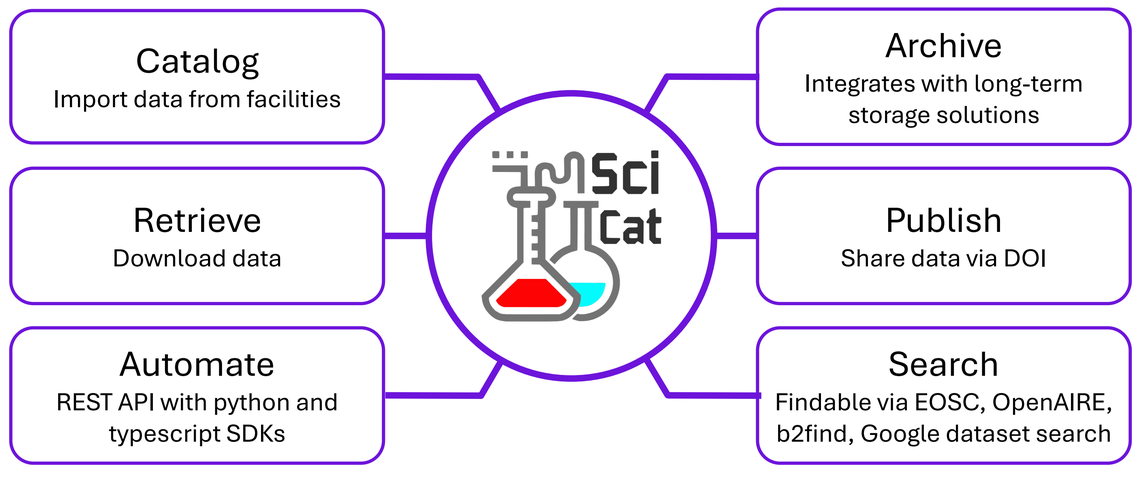
For adding data, an Ingestor service runs on the facilities to extract metadata in OSC-EM format and manage data transfer to the archiving system via S3 or Globus Transfer. SciCat's asynchronous job system then manages the storage of data on long-term storage (LTS) systems. Two tape-based storage systems are supported: the ETHZ LTS and the Petabyte Archive at the Swiss National Computing Centre (CSCS).

Users can manage datasets through SciCat, including publishing and assigning a DOI. Public data is indexed by search engines such as Google dataset search, European Open Science Cloud (EOSC), OpenAIRE, and b2find. Data can be easily retrieved from tape.

OpenEM also provides integration with the Worldwide Protein Data Bank (wwPDB) OneDep deposition system and the EM Public Image Archive (EMPIAR) for life science data.

Poster DOI: 10.5281/zenodo.13798727 This project was supported by the Open Research Data Program of the ETH Board and by a Swiss Open Research Data Grant (CHORD).

SciCat Data Catalog



SciCat is an open-source data catalog software used at a variety of synchrotrons and large facilities. Metadata from OpenEM partners is stored in the PSI SciCat Data Catalog alongside data from the Swiss Light Source, SwissFEL, and other PSI facilities.

Data Catalog: **https://discovery.psi.ch**
Code: **https://scicatproject.github.io/**

OSC-EM Metadata Schema

Well-structured metadata is a key to interoperable software. Processing EM datasets requires substantial amounts of metadata, which must be faithfully preserved from data collection, through data processing, and then published along with the data to enable reuse.

The Open Standards Community for EM (OSC-EM) was established to address the need for a common schema to preserve EM metadata across the full data lifecycle. A workshop in Feb 2024 included participants from EM facilities, software creators, and repositories. As a result, a schema has been developed for EM data. Schema terms are defined by existing ontologies where available: PDBx/mmCIF dictionary, CryoEM ontology, Helmholz EM Glossary, and NeXus-FAIRmat NXem format. The schema is defined in LinkML and converted to common schema formats to easy software integration (JSON Schema, JSON-LD, RDF, etc).

OpenEM provides conversion tools to import metadata from EM instruments to OSC-EM. Currently SerialEM and Thermo-Fisher EPU are implemented, with additional techniques planned. CryoEM and tomography metadata can also be exported to mmCIF format to ease deposition into the EM Data Bank (EMDB), Protein Data Bank (PDB) and EM Public Image Archive (EMPIAR).

https://github.com/osc-em/OSCEM_Schemas
Example: DOI 10.16907/a2ab7849-5de7-4e7f-8286-72ec73089ca8

Partners and Collaborators

OpenEM strives to maximize interoperability with other open services and databases. Many of the services are developed in tight collaboration with partners. Please contact us if you are interested!