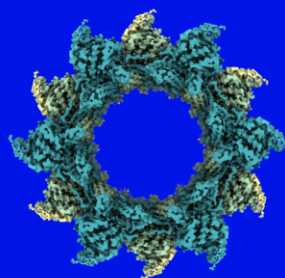


PSI Center for Scientific Computing,
Theory and Data



OpenEM & SciCat

Spencer Bliven :: 7903 Scientific Data Curation
AWI Department Meeting, 10 Dec 2024

Scientific Data Lifecycle

PaNOSC, EOSC,
Google Dataset Search

Publications
DOI



Archiving



Proposals (DUO)

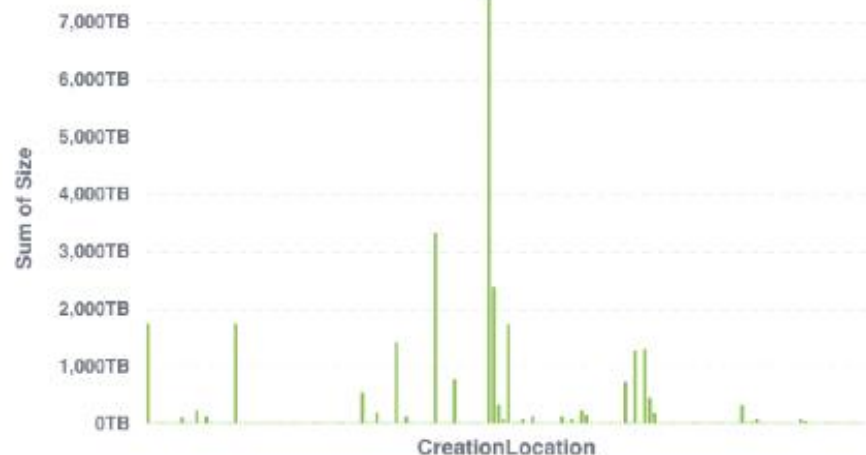
Beamlines (microscopes),
Simulations,
Electronic Lab Notebooks
(SciLog)

Acquisition Systems (BEC)

Workstations,
Clusters (ra/merlin)

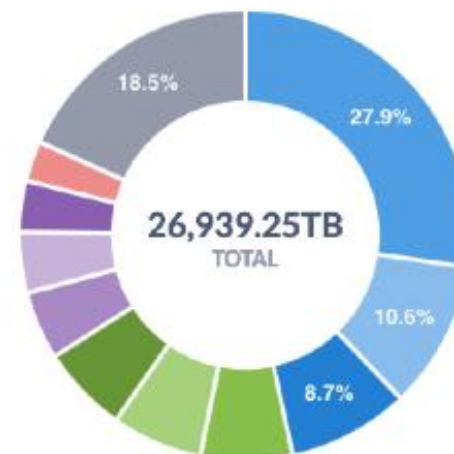
SciCat usage

Dataset Statistics per Beamline



Creation Locations by Size

- /PSI/SWISSFEL/ARAMIS-A
- /PSI/SLS/TOMCAT
- /PSI/SWISSFEL/ARAMIS-BI
- /PSI/HIPA/PIE5
- /PSI/SWISSFEL/ATHOS-MA
- /PSI/swissfel/bernina
- /PSI/SLS/MX
- /PSI/slsdetectors
- /PSI/sls
- Other

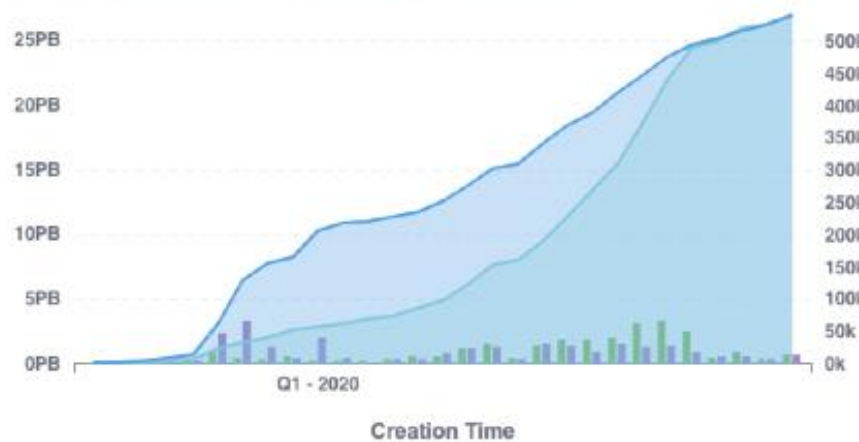


26.94PB

Total Archive Size

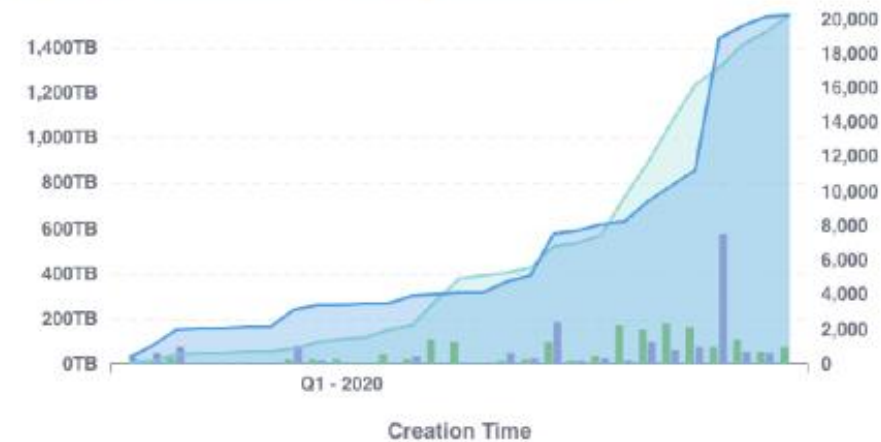
PSI Dataset Archive Statistics over Time

Size Cumulative Size Count Cumulative Count



PSI Dataset Retrieve Statistics over Time

Size Cumulative Size Count Cumulative Count



540.9k

Number of Datasets

88.37TB

Largest dataset

Version 3

- Loopback (javascript)
- Hard-coded job actions
- Code not following best practices
- Flexible scientific metadata
- Authentication & Authorization via PSI active directory
- Archiving to CSCS Petabyte Archive

Version 4

- NestJS (typescript)
- Configure jobs without code changes
- Better testing, CI/CD, and devops
- Optional **metadata schema** validation
- Authentication with eduGAIN **federated identities**
- Support **multiple storage systems** (adding ETHZ Long Term Storage)

What will change?



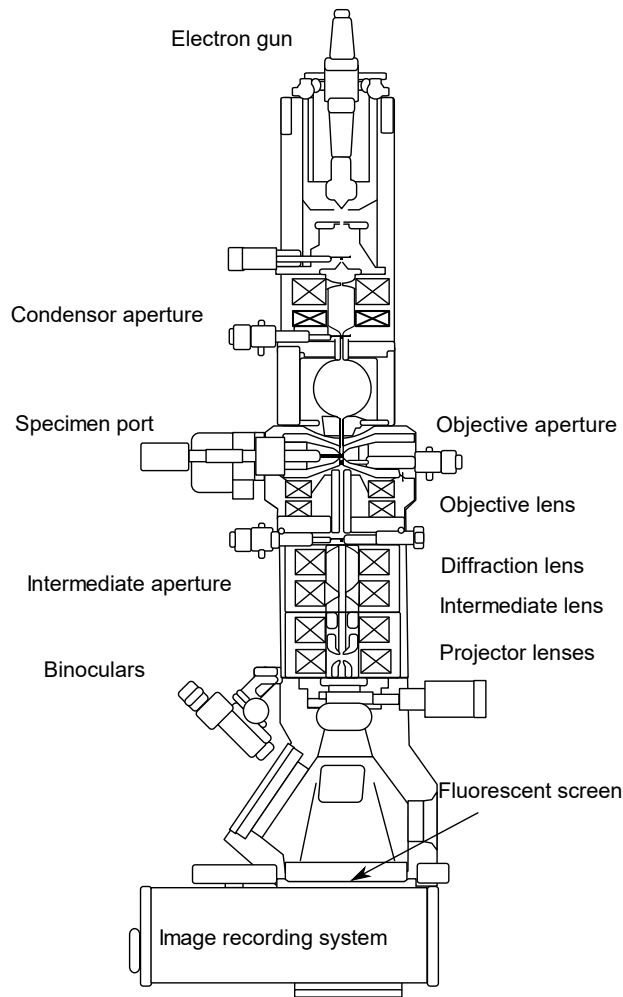
- Some changes to the REST API (renamed endpoints, added/removed properties)
- Single-binary CLI replaces datasetIngestor, datasetArchiver, datasetRetriever, ...
- Authentication with OpenID Connect, replacing token-based login for interactive use cases (but functional accounts will still be supported)
- Webpage for ingesting datasets interactively & monitoring data transfers
 - Support for the Qt-based `SciCat` GUI will be deprecated after this is feature complete
- Accept data from outside PSI (OpenEM electron microscopy facilities)
 - Requires installing an ingestor service to manage data transfer
- Support for Globus and S3 uploads from the internet

New v4 REST endpoint for testing: <https://scicat.development.psi.ch/explorer>

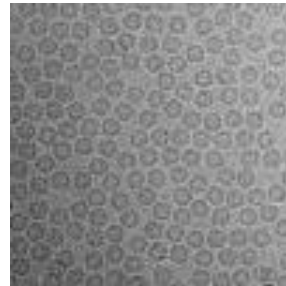


penEM

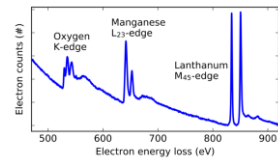
Electron Microscopy (EM)



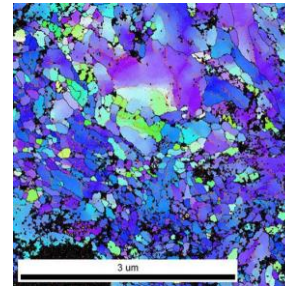
Raw Data



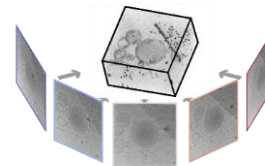
2D Micrographs
(EMPIAR-11016, Harder, EPFL)



Spectrograms
[Magnunor, Wikimedia](#)



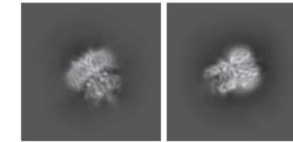
Annotated Images
(Kunze and Sologubenko, ETHZ)



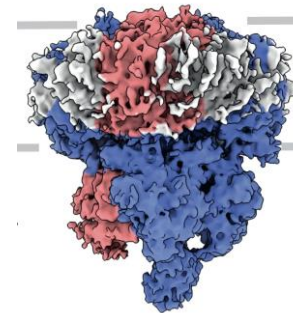
Tomogram tilt-series
teamtomo.org

More: ptychography, 4D STEM, VolumeEM, ...

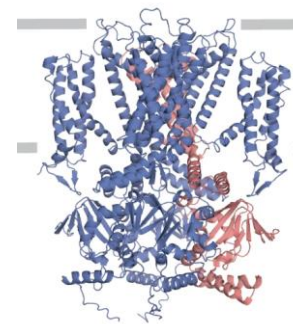
Derived Data



Class averages
(Barret, PSI)



3D reconstructed volumes
(EMD-12718, Barret, PSI)



Models
(7o4h, Barret, PSI)

More: Tomographic reconstructions, segmented models, ...

- Pressure from journals for open data
- Requirements from funders & institutes
- Easier data collection for microscope users
 - Standardized across Switzerland
 - Automatically log acquisition metadata
 - Rapid deposition in field-specific repositories
- Reuse
 - Verify/reproduce processing
 - Extracting additional results (especially from tomograms)
 - Method development, e.g. AI methods
 - Apply novel processing methods to older data



Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures

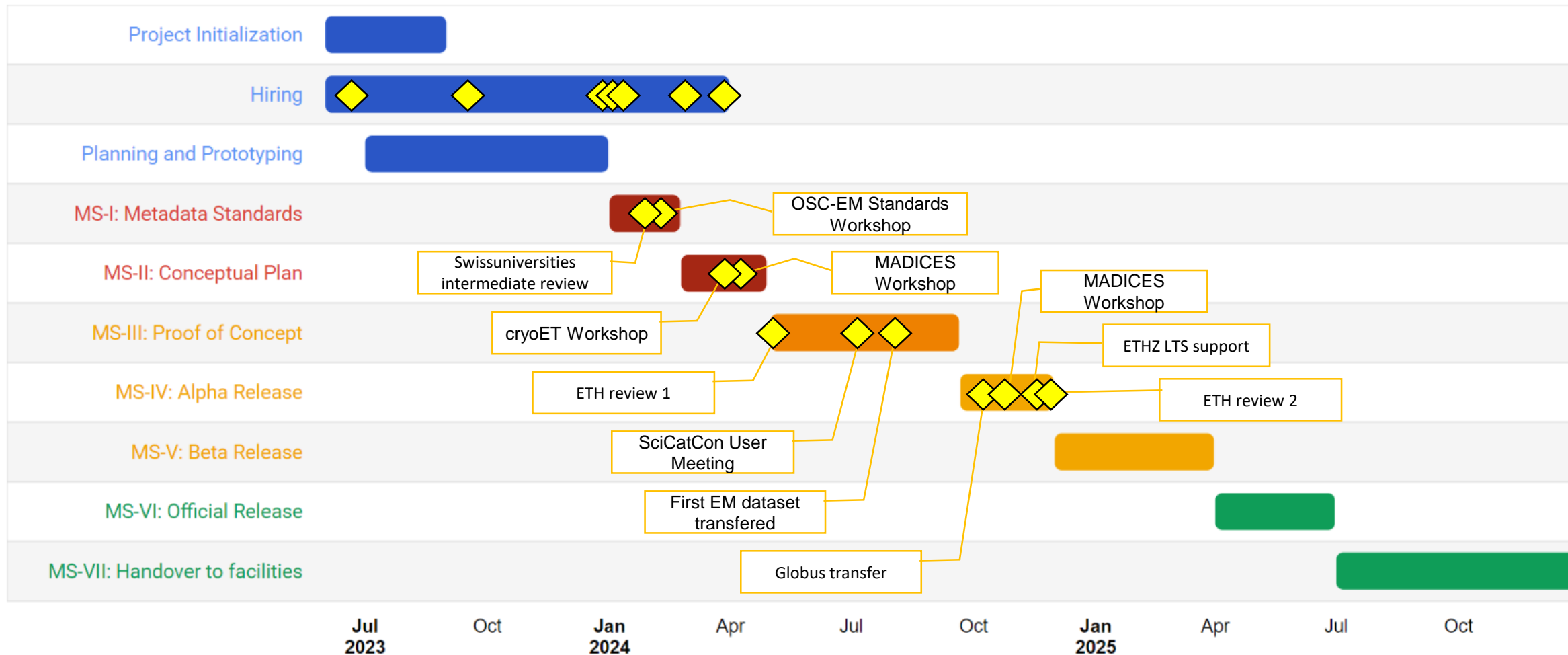
This year's prize celebrates computational tools that have transformed biology and have the potential to revolutionize drug discovery.

By [Ewen Callaway](#)



David Baker, Demis Hassabis and John Jumper (left to right) won the chemistry Nobel for developing computational tools that can predict and design protein structures. Credit: BBVA Foundation

Timeline



Updates



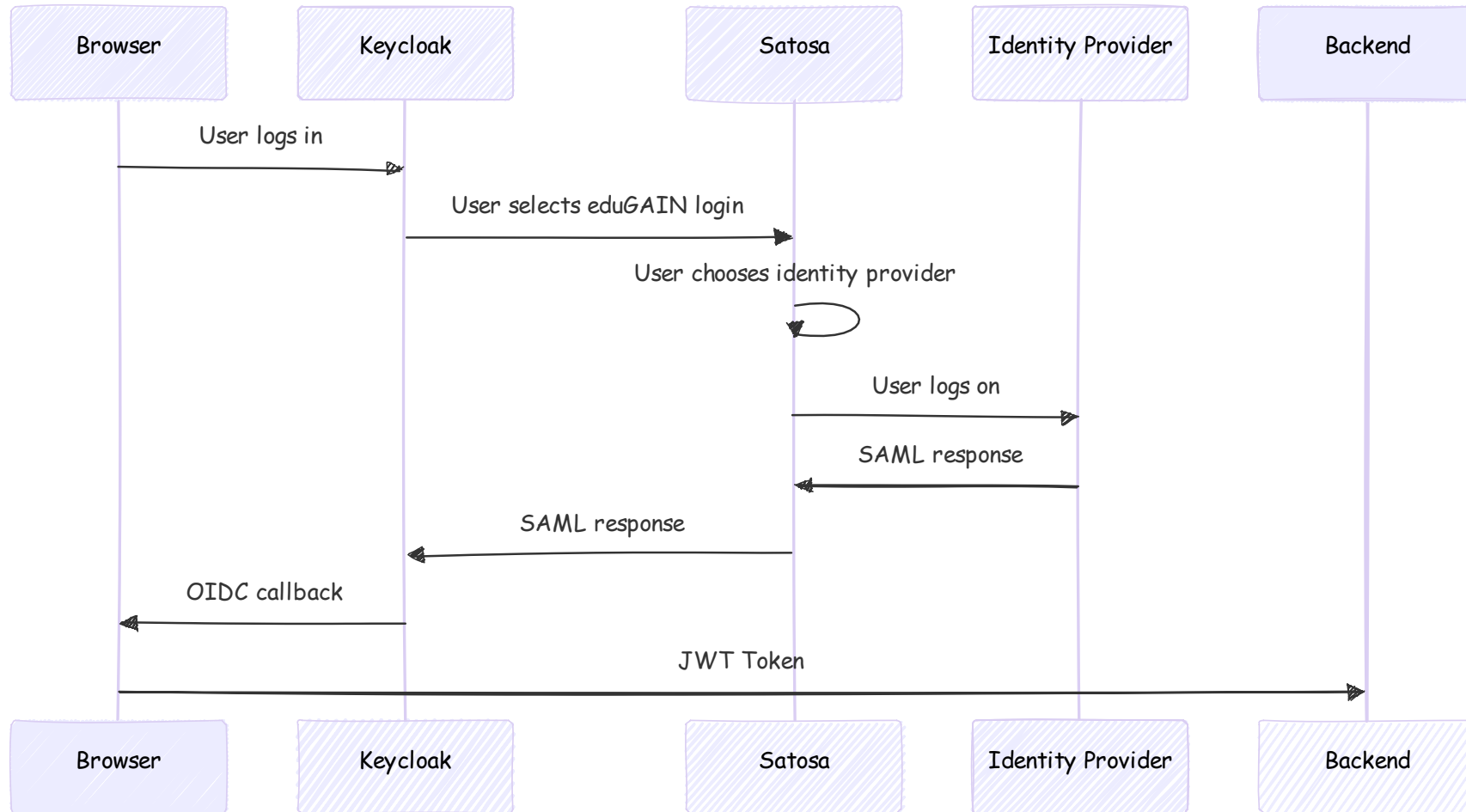
- Completed “Proof of Concept” Milestone in September to acquire & publish a high-resolution protein structure ([DOI:10.16907/a2ab7849-5de7-4e7f-8286-72ec73089ca8](https://doi.org/10.16907/a2ab7849-5de7-4e7f-8286-72ec73089ca8))
- Deployment of “alpha release” consisting of all services to Uni Basel expected in December
- Preparing for “beta release” at all facilities (including PSI) in late Q1-2025
 - Workshop including microscope operators scheduled for 13 Feb 2025 in Bern
- Submitted scientific & financial reports to ETH 20 Nov

The screenshot displays a web interface for a dataset titled "CryoEM of Apoferritin". The interface is organized into several sections:

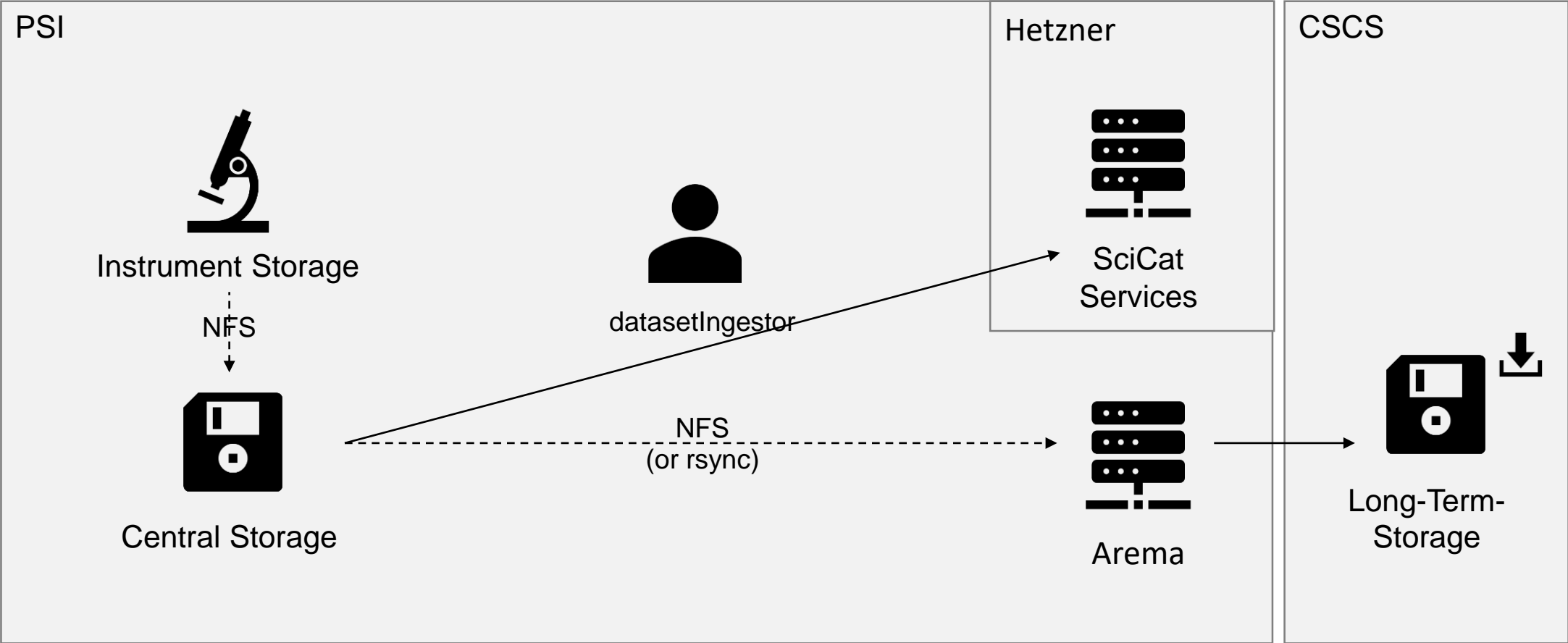
- General Information:** Name: CryoEM of Apoferritin; Description: cryoEM micrographs of Apoferritin at 0.415 Angstrom pixelsize; PID: 20.500.11935/e9958228-11b9-42ea-a099-813150c3ccea; Type: raw; Creation Time: 2024-08-30 10:51; Keywords: OpenEM, Single Particle Analysis.
- Creator Information:** Owner: Yves Tittes; Owner Group: a-35632; Access Groups: bioemtitang4.
- File Information:** Source Folder: /scicore/projects/scicore-p-malart-structbio/titye00/milestone1/data; Size: 2 TB; Data Format: TIFF image stacks.
- Related Documents:** Creation Location: /UNIBAS/BIOEM/TITANG4.
- Scientific Metadata:** A JSON-like structure containing acquisition parameters such as beamshift, beamtilt, calibrated_defocus, detector, dose_per_movie, energy_filter, exposure_time, gainref_flip_rotate, image_size, images_generated (21502), and imageshift.

A thumbnail image of a blue, circular protein structure is visible on the right side of the page, labeled "Apoferritin_thumbnail.png".

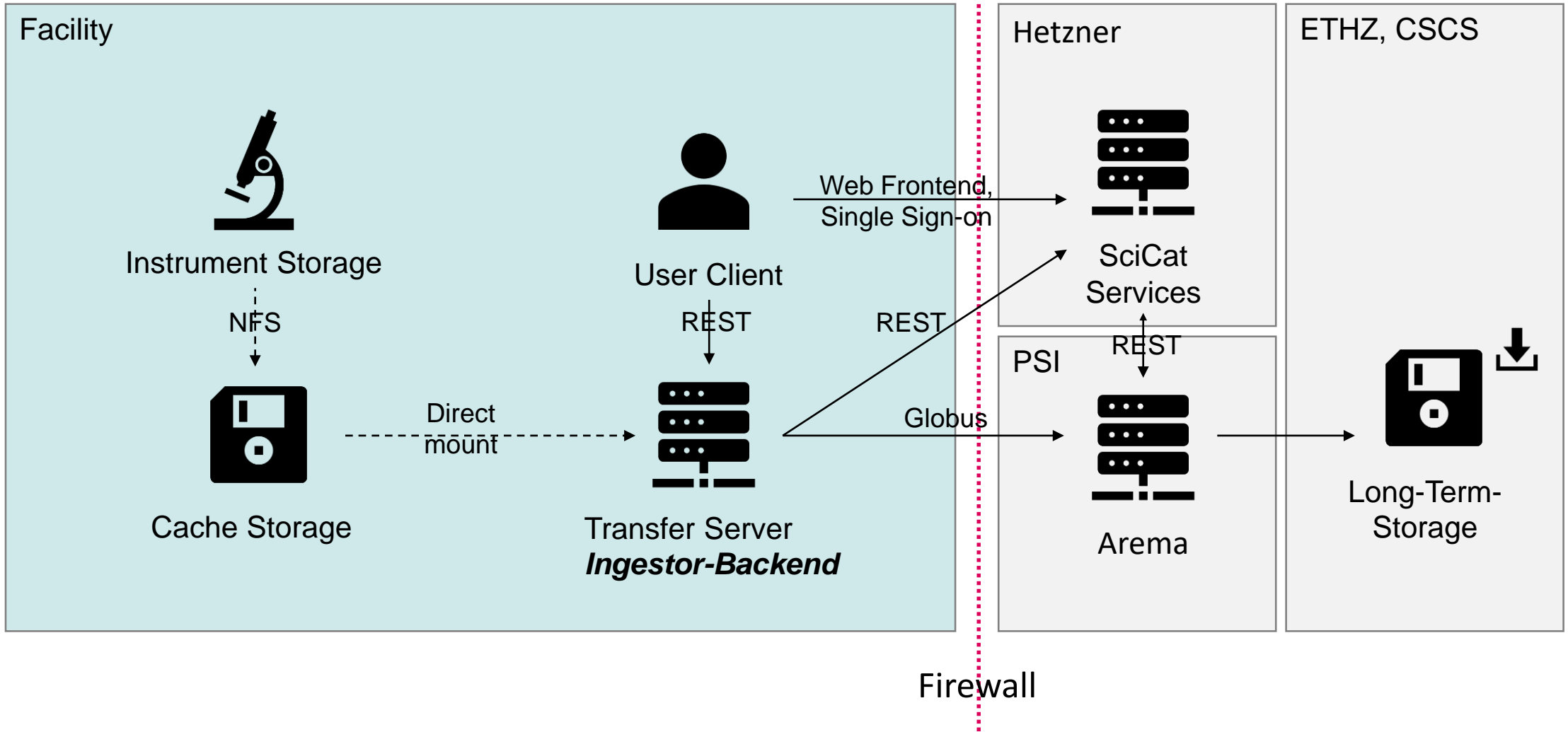
Authentication



Current Ingestor Architecture



Planned Architecture for external facilities



New Ingestor Frontend



Ingestor (OpenEM)

Help

About



ingestor

Ingestor /

Control center



ID

New transfer



Fill out user-specific metadata



1 Select your ingestion method — 2 Fill out user-specific metadata — 3 Correct dataset-specific metadata — 4 Confirm inputs

Id*

is a required property

Title*

is a required property

Description

Status*

is a required property

Priority

First Name

Back

Next

New Ingestor Frontend



The screenshot shows the Ingestor frontend interface with several numbered callouts overlaid on it:

- 1. eduGAIN authentication**: Located at the top left, pointing to the user authentication area.
- 2. Select Data Source/Facility**: Located in the top left of the form, pointing to the 'Ingestion method' dropdown.
- 3. Browse remote data & select data format**: Located in the middle left of the form, pointing to the 'Id*' field.
- 4. Set authorization & license**: Located in the middle of the form, pointing to the 'Title*' field.
- 5. Proof automatically extracted metadata**: Located in the middle right of the form, pointing to the 'Status*' field.
- 6. Submit**: Located at the bottom right of the form, pointing to the submit button.
- 7. Monitor data transfer (globus)**: Located at the bottom right of the page, pointing to the 'Back' button.

The interface includes a top navigation bar with 'Ingestor (OpenEM)', 'Help', 'About', and a user profile icon. The main form contains fields for 'Id*', 'Title*', 'Description', 'Status*', 'Priority', and 'First Name', each with a 'is a required property' error message. A progress indicator at the top of the form shows steps: 1. Ingestion method, 2. Fill out user-specific metadata, 3. Correct dataset-specific metadata, and 4. Confirm inputs.



OSC-EM

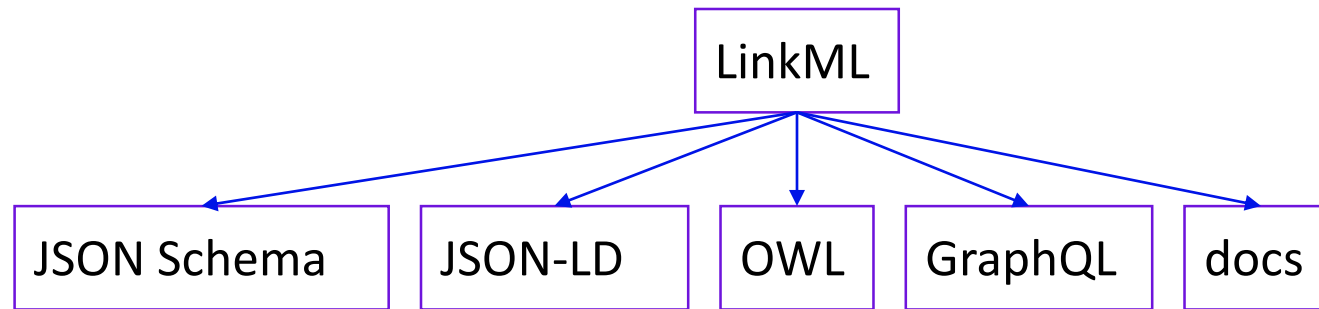
Open Science Community for Electron Microscopy (OSC-EM)



- Established in 2023 to bring together electron microscopy (EM) researchers, facilities, software developers, and data repositories to standardize EM metadata needed for data collection, processing, and deposition.
- Workshop 22-23 Feb 2024 with diverse participants
- Aims for interoperability with other ontologies and standards: [CryoEM ontology](#), [PDBx/mmCIF](#) dictionary, Helmholtz [EM Glossary](#), [NeXus-FAIRmat NXem format](#) and cryoET standards working group
- Active contributors include OpenEM facilities (swissopenem.github.io), the Instruct Image Processing Center ([I2PC](#)), and the EM Data Bank (www.ebi.ac.uk/emdb).
- Modular definition for different experimental methods and data processing stages (cryoEM, tomography, EELS, 3D reconstruction, etc).



- <https://github.com/osc-em>
- Schema in LinkML used to automatically generate JSON Schema, JSON-LD, OWL, GraphQL, etc, as well as documentation and a python SDK.



- Import from SerialEM and Thermo Fischer EPU (more coming!)
- Export to mmCIF for deposition in EMDB/PDB OneDep
- Suitable for inclusion in SciCat `scientificMetadata` field

```
1  # Example OSC-EM dataset
2  ---
3  instrument:
4  |·microscope: Titan
5  |·illumination: FloodBeam
6  |·imaging: Brightfield
7  |·electron_source: FEG
8  |·acceleration_voltage: 300
9  |·c2_aperture: 70
10 |·cs: 2.7
11 acquisition:
12 |·holder: testitest
13 |·detector: Falcon 4i
14 |·detector_mode: counting
15 |·dose_per_movie: 0.5
16 |·date_time: "2024-01-01"
17 |·binning_camera: 2
18 |·pixel_size: 1.2
19 > grants: ...
24 > authors: ...
41 > sample: ...
75
```

Deposition in EMPIAR, EMDB, and PDB



- Export cryoEM data to EMPIAR, EMDB, and PDB
- Convert OSC-EM metadata to mmCIF
- Pre-fills many fields on the EMDB/PDB deposition website
- Directly transfer raw data to EMPIAR deposition

```
1 # Converter for JSON schema to mmCIF for PDB
2
3 This repository implements a file converter from OSC-EM JSON to PDBx/mmCIF.
4 The JSON schema is defined in [OSCEM](https://github.com/osc-em/OSCEM_Schemas/) and the PDBx/mmCIF format in ([PDBxL](https://mmcif.wpdb.org/dictionaries/ascii/mmcif_pdbx_v50.dic) Schema v50 ).
5
6
7 ## Running the converter and required inputs:
8 * converter executable
9 * with '--json' specify path to json file that contains metadata
10 * with '--dic' specify path to the PDBx/mmCIF dictionary file
11 * with '--conversions' specify path to [conversions table](https://github.com/openem/LS_Metadata_reader/). This table includes correspondance in names between OSC-EM and PDBx
12 * with '--level' specify the json element name that contains metadata entries. For SciCat that is usually "scientificMetadata"
13 * with '--append' specify if the metadata should be added to existing mmCIF to later deposit it in PDB
14 * with '--mmCIFfile' specify the path to existing mmCIF file. Throws an error if --append is false and --mmCIFfile is not specified
15 * with '--output' specify the file to write the newly created mmCIF with metadata entries
16
17 ## Checks against mmCIF
18 Converter explicitly parser through the PDBx definitions to extract as much data as possible. This allows for
19 * administrative categories sorted within mmCIF ( such as author information, grant, etc)
20 * em-related categories are sorted randomly, as there is no definitive sorting in PDB team as well
21 * file ends with information on atoms
22 * units in OSCEM definition are comared to PDBx ( converter for units will be implemented)
23 * numeric values are checked to be within a range allowed by PDBs
24 * values are checked to be within a list of attributes allowed by PDBx. This is additionally enhanced to match via regular expressions or ceratin logic. |
```

OpenEM Websites

- Public project website: <https://swissopenem.github.io>
- ETH ORD Portal: <https://open-research-data-portal.ch/projects/open-em-data-network/>
- Github: <https://github.com/SwissOpenEM>

OSC-EM Standard

- <https://github.com/osc-em>

SciCat Data Catalog

- Data repository: <https://discovery.psi.ch>
- Published datasets: <https://doi.psi.ch>
- SciCat website: <https://scicatproject.github.io>
- v4 test deployment: <https://scicat.development.psi.ch/explorer>

Acknowledgements



- Scientific Data Curation Group
 - Leo, Carlo
 - Welcome Omkar & Frédéric!
- Datacenter & DB Services
 - Peter Huesser, Michael Kallmeier-Glanz, Bernard Bumbak
- SciCat Contributors (Max Novelli *et. al*)
- OpenEM (**PI**; *Core Team*)
 - EPFL: Marco Cantoni, *Sofya Lakina*, Alexander Myasnikov, Alexandra Radenovic, **Henning Stahlberg**
 - PSI: Alun Ashton, Gregor Cicchetti, Peter Hüsser, Michael Kallmeier-Glanz, Volodymyr Korkhov, Carlo Minotti, Elisabeth Müller, Gebhard Schertler
- EMPA: Rolf Erni, *Despina Adamopoulou*
- ETHZ: Matthew Baker, Nicolas Blanc, Daniel Böhringer, Christophe Briand, Christophe Copéret, Miroslav Peterek, Bilal Qureshi, Andrzej J. Rzepiela, *Philipp Wissmann*
- UNIGE: Andreas Boland, Orsolyz Barabas, Andrew Howe, *Attila Nacsa*, **Robbie Loewith**
- UNIBAS: Mohamed Chami, Timm Maier, *Yves Tittes*
- UNIBE: David Kalbermatter, Benoît Zuber, *David Wiessner*