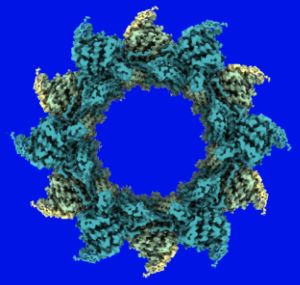
PSI Center for Scientific Computing, Theory and Data



OpenEM & SciCat

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

Scientific Data Lifecycle

PSI

PaNOSC, EOSC,
Google Dataset Search

Publications
DOI

SciCat
metadata catalogue for scientists

Archiving

Proposals (DUO)

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

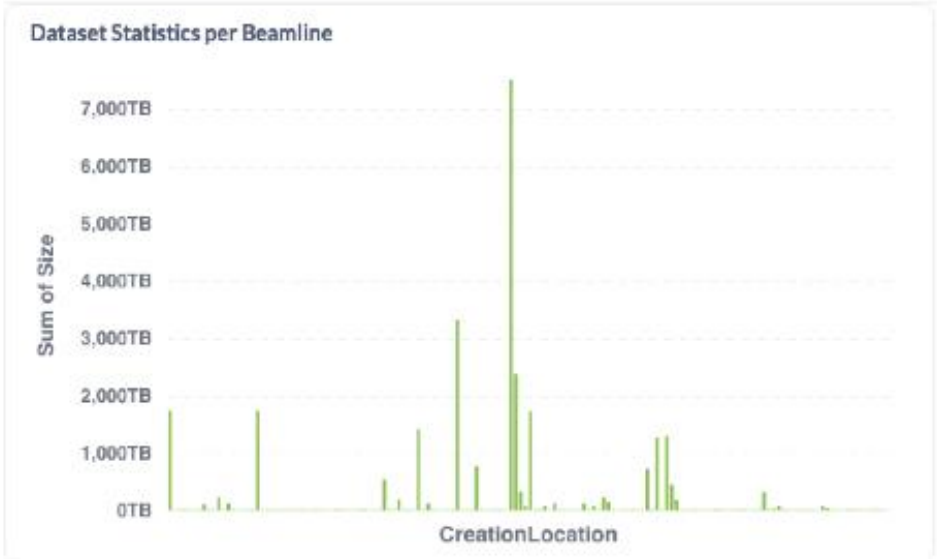
Acquisition Systems (BEC)

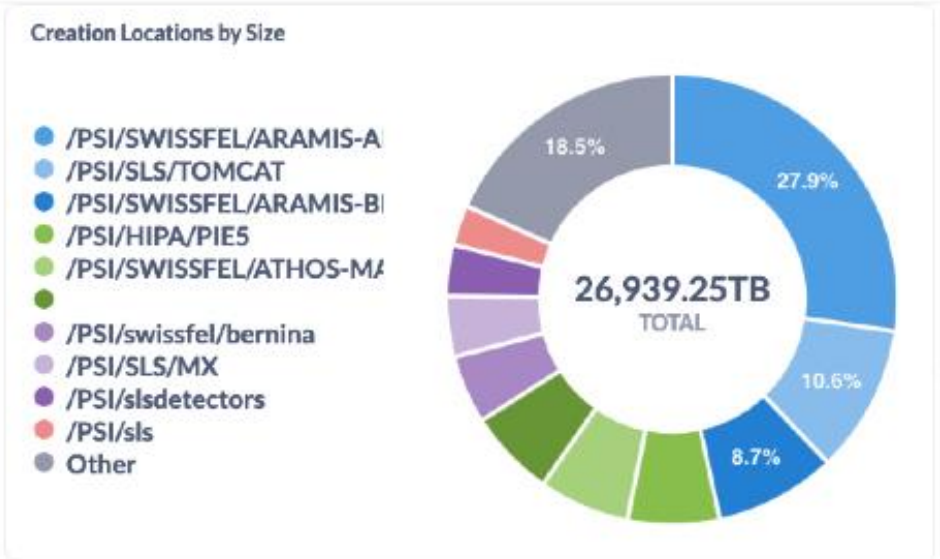
Workstations,
Clusters (ra/merlin)

SciCat Statistics

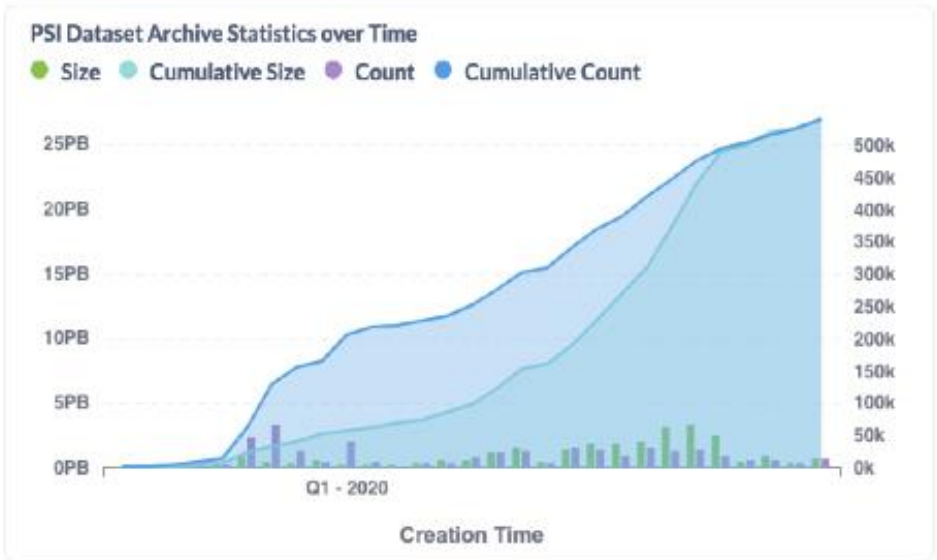
PSI

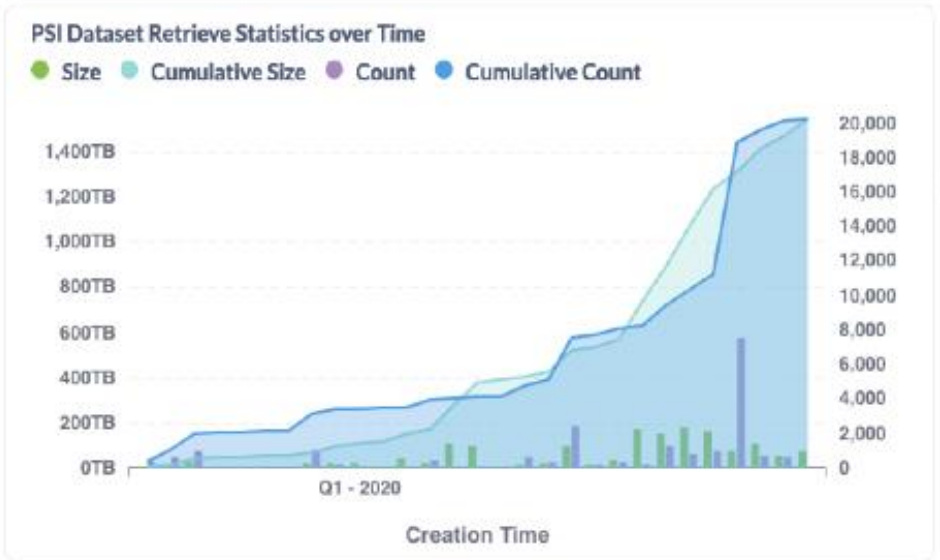
SciCat usage





26.94PB
Total Archive Size





540.9k
Number of Datasets

88.37TB
Largest dataset

http://metabase.psi.ch

SciCat Backend v4

PSI

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

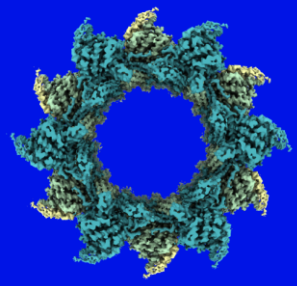
- NestJS (typescript)
- Configure jobs without code changes
- Better testing, CI/CD, and devops
- Optional **metadata schema** validation
- Authentication with eduGAIN **federated identities**
- Support **multiple storage systems** (adding ETHZ Long Term Storage)

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

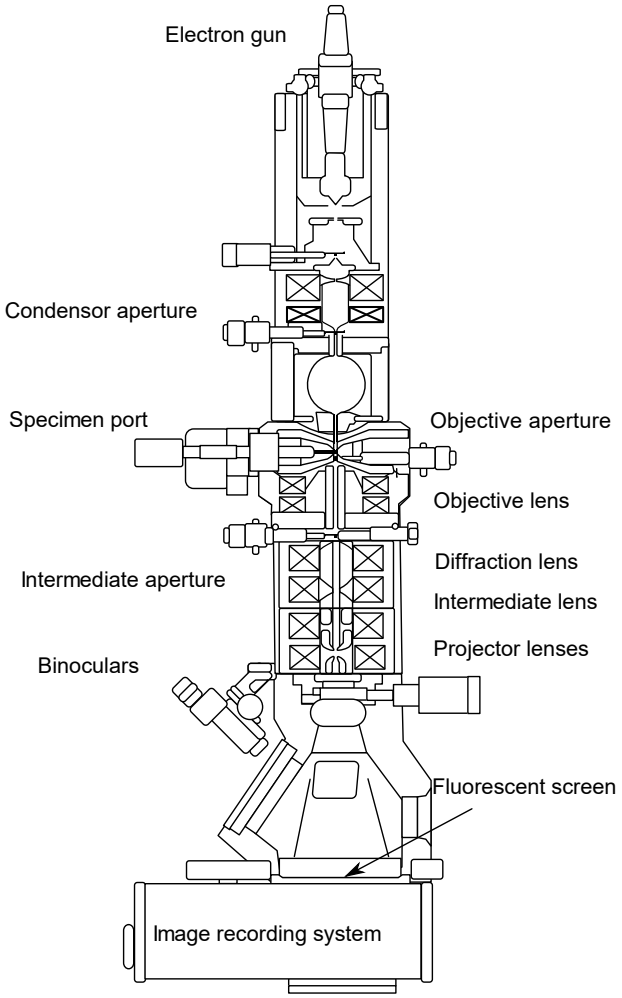
New v4 REST endpoint for testing: **https://scicat.development.psi.ch/explorer**

PSI

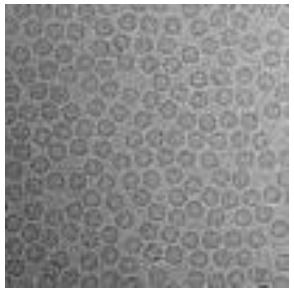


OpenEM

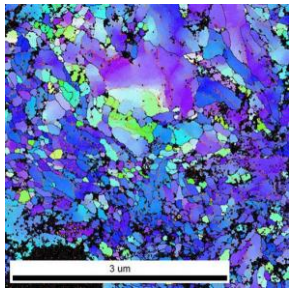
Electron Microscopy (EM)



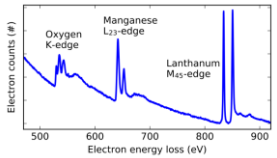
Raw Data



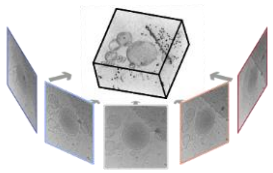
2D Micrographs
(EMPIAR-11016, Harder, EPFL)



Annotated Images
(Kunze and Sologubenko, ETHZ)



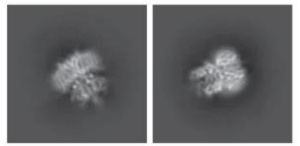
Spectrograms
Magnunor, Wikimedia



Tomogram tilt-series
teamtomo.org

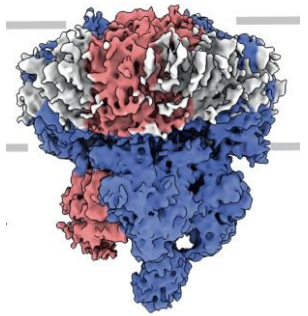
More: ptychography, 4D STEM, VolumeEM, …

Derived Data



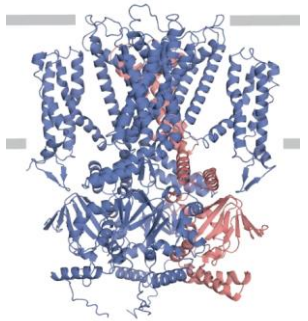
Class averages

(Barret, PSI)



3D reconstructed volumes

(EMD-12718, Barret, PSI)



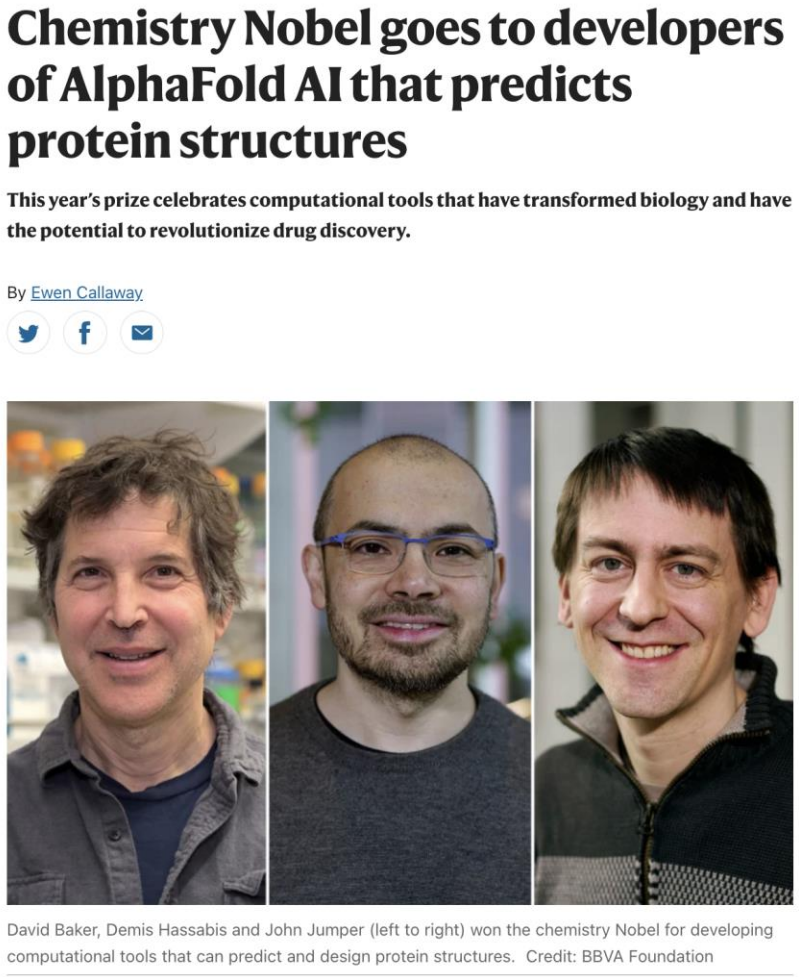
Models

(7o4h, Barret, PSI)

More: Tomographic reconstructions, segmented models, …

Motivation

- 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



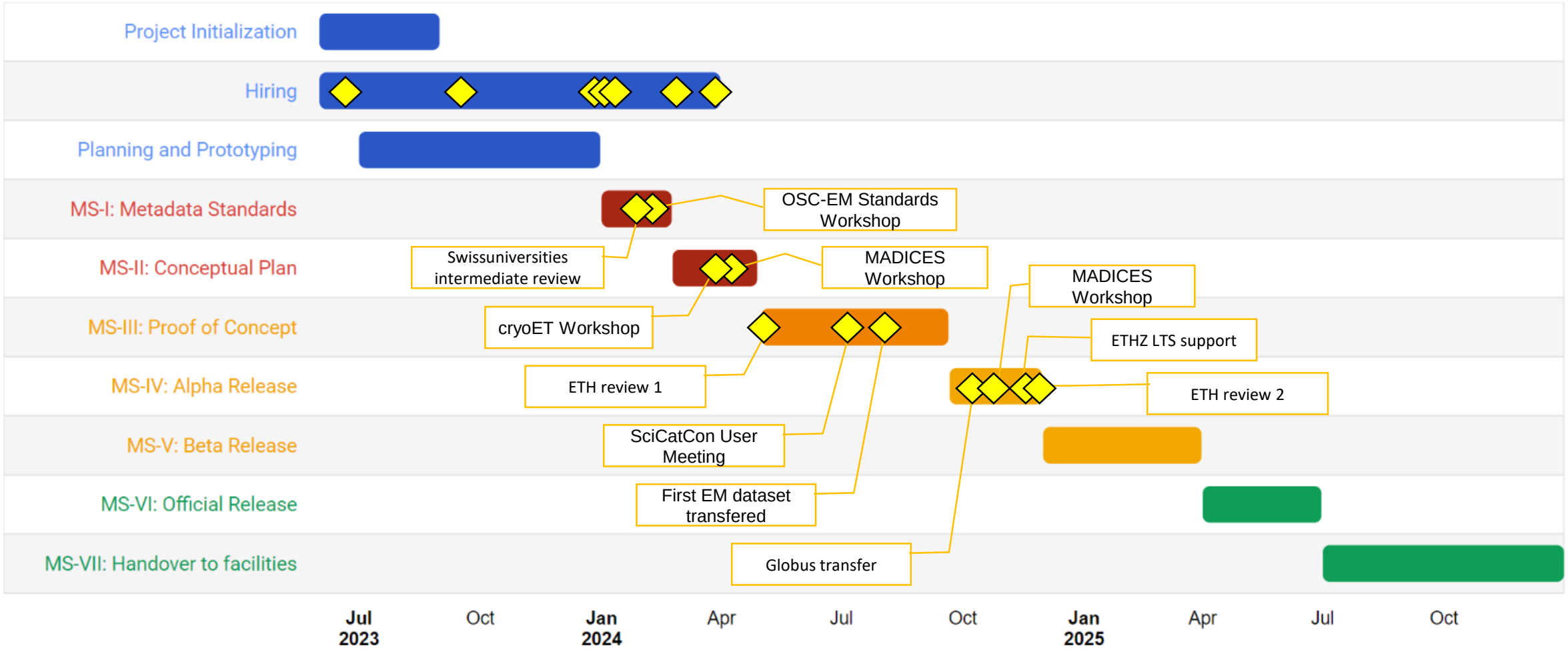
Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures

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

By Ewen Callaway

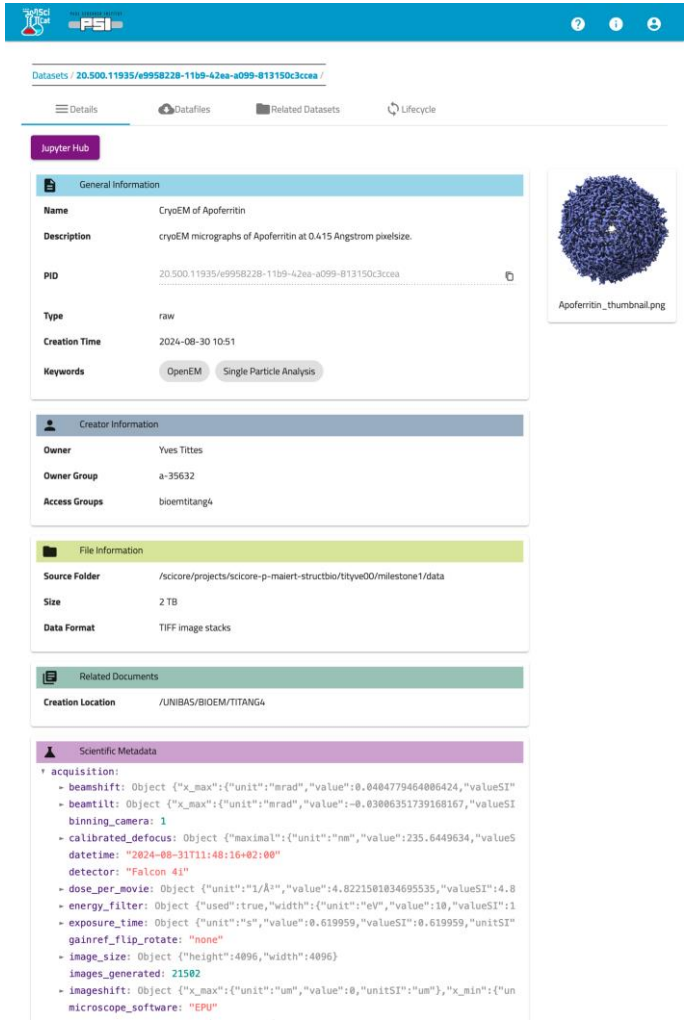
David Baker, Demis Hassabis and John Jumper (left to right) won the chemistry Nobel for developing computational tools that can predict and design protein structures. Credit: BBVA Foundation

Timeline

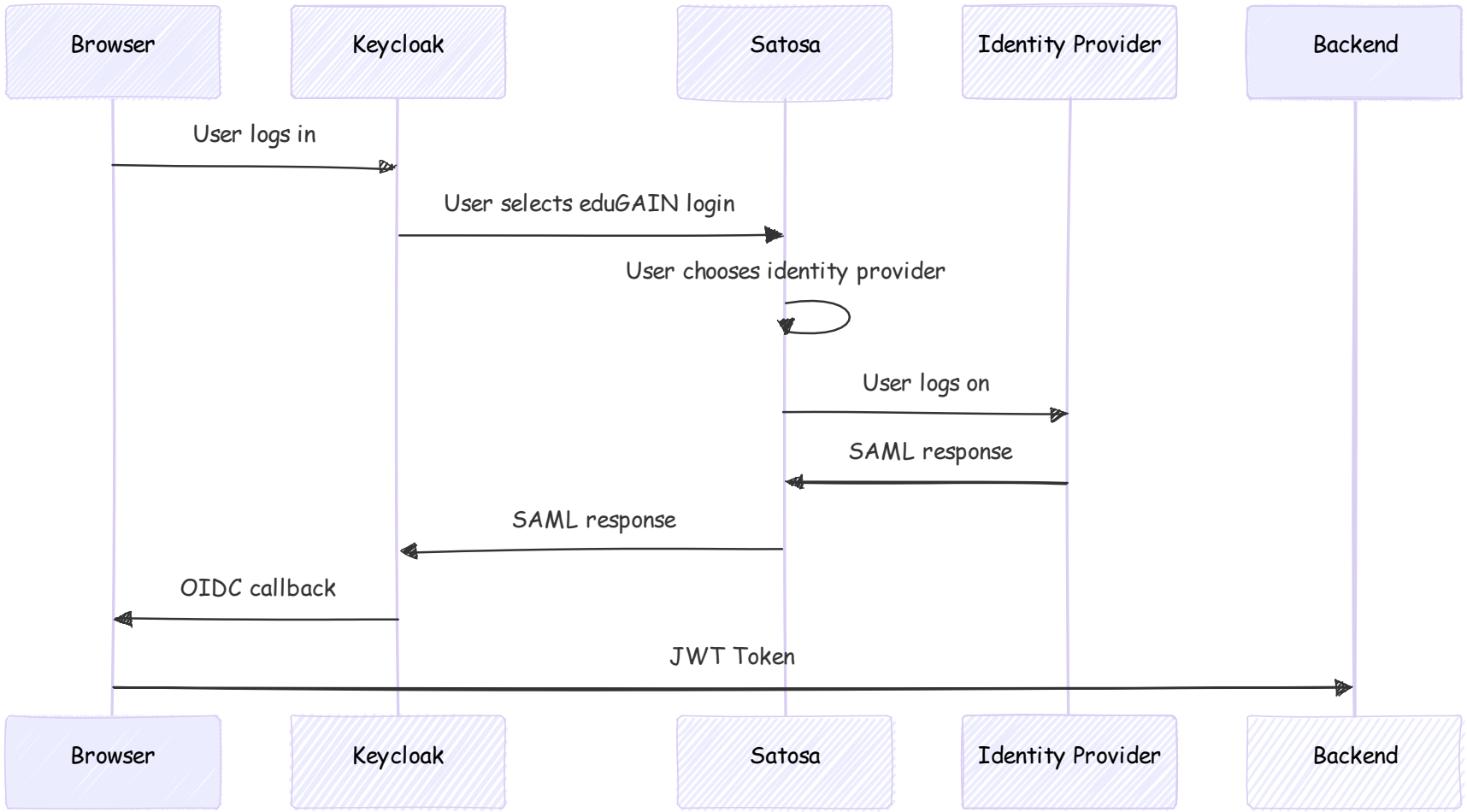

Project Initialization
Hiring
Planning and Prototyping
MS-I: Metadata Standards
MS-II: Conceptual Plan
MS-III: Proof of Concept
MS-IV: Alpha Release
MS-V: Beta Release
MS-VI: Official Release
MS-VII: Handover to facilities
OSC-EM Standards Workshop
Swissuniversities intermediate review
MADICES Workshop
cryoET Workshop
MADICES Workshop
ETHZ LTS support
ETH review 1
ETH review 2
SciCatCon User Meeting
First EM dataset transfered
Globus transfer
Jul 2023
Oct
Jan 2024
Apr
Jul
Oct
Jan 2025
Apr
Jul
Oct

Updates

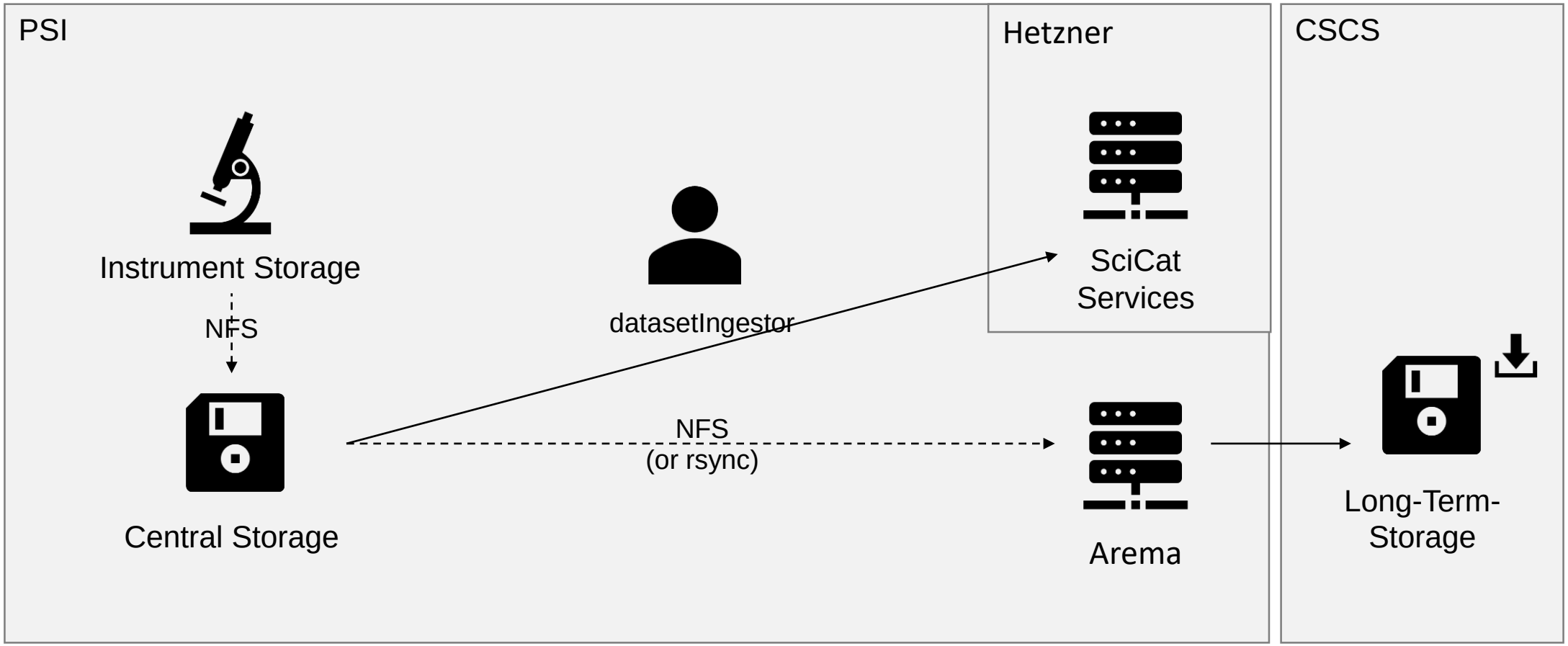
- Completed “Proof of Concept” Milestone in September to acquire & publish a high-resolution protein structure (DOI:10.16907/a2ab7849-5de7-4e7f-8286-72ec73089ca8)
- Deployment of “alpha release” consisting of all services to Uni Basel expected in December
- Preparing for “beta release” at all facilities (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



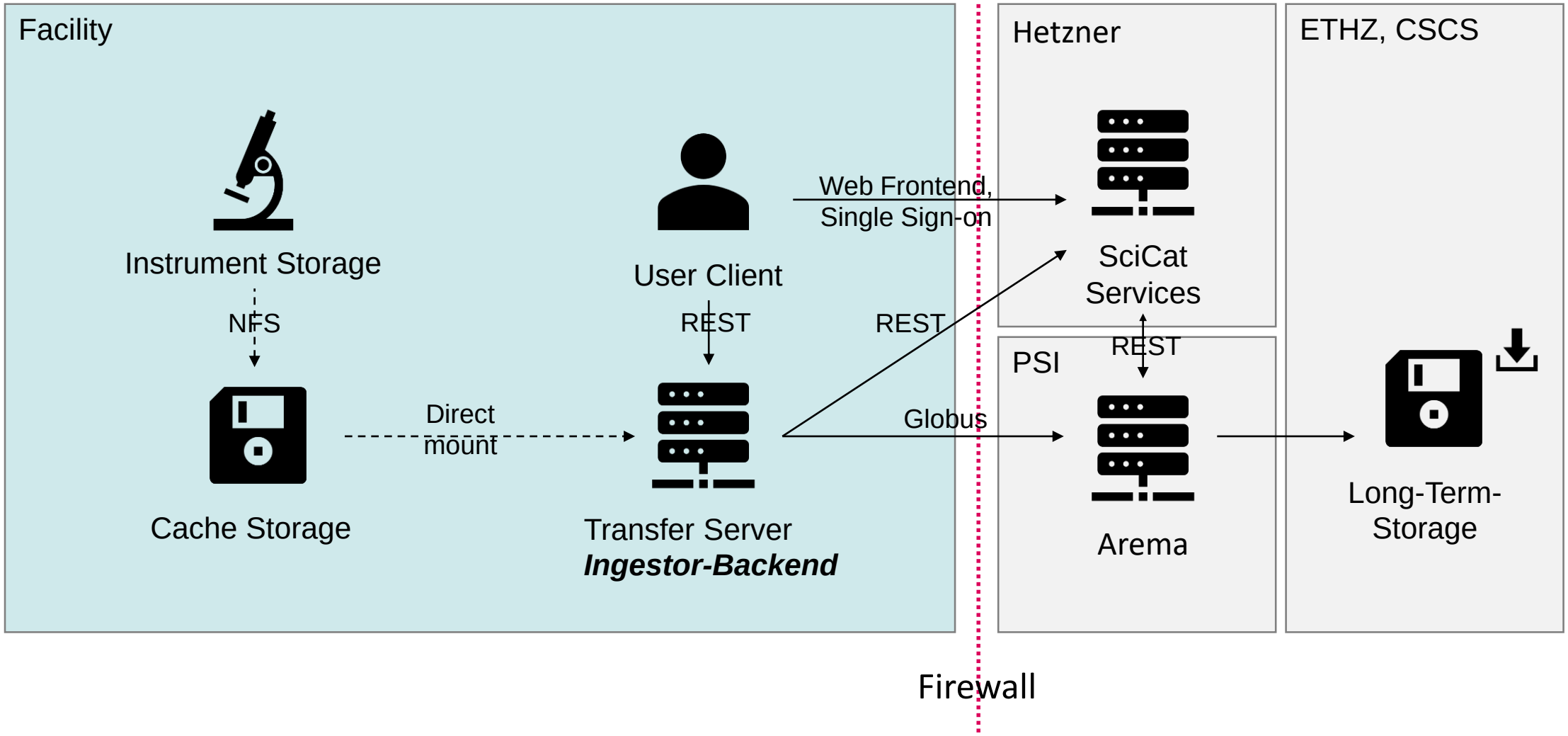
Authentication



Current Ingestor Architecture

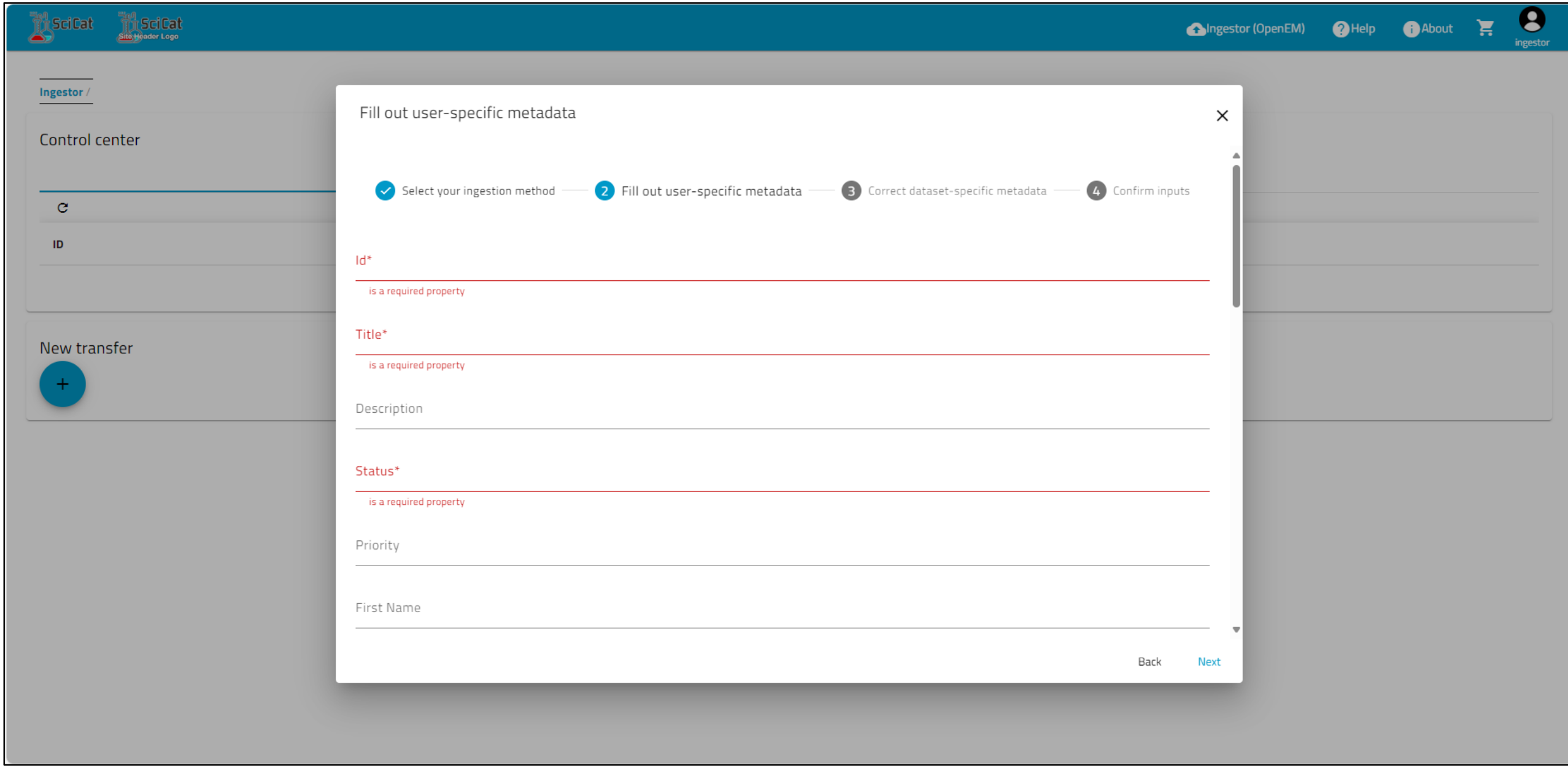


Planned Architecture for external facilities



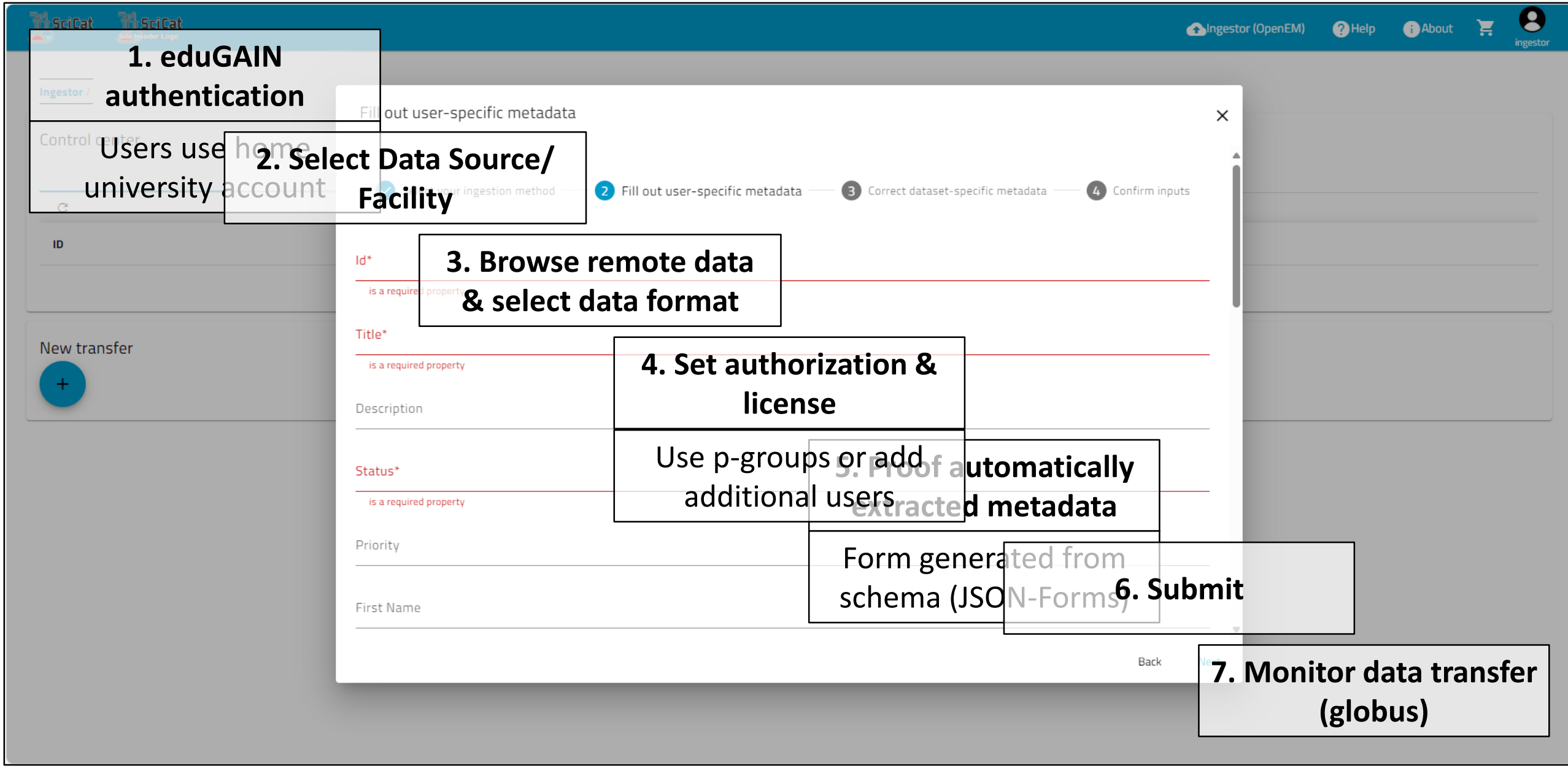
New Ingestor Frontend

PSI



New Ingestor Frontend

PSI

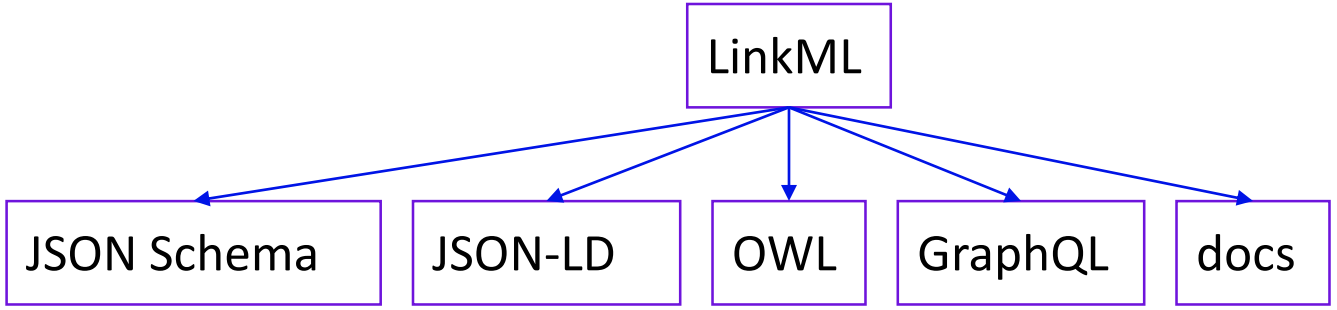


Open Science Community for Electron Microscopy (OSC-EM)

- 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, Helmholz 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).

OSC-EM Schema

- 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.

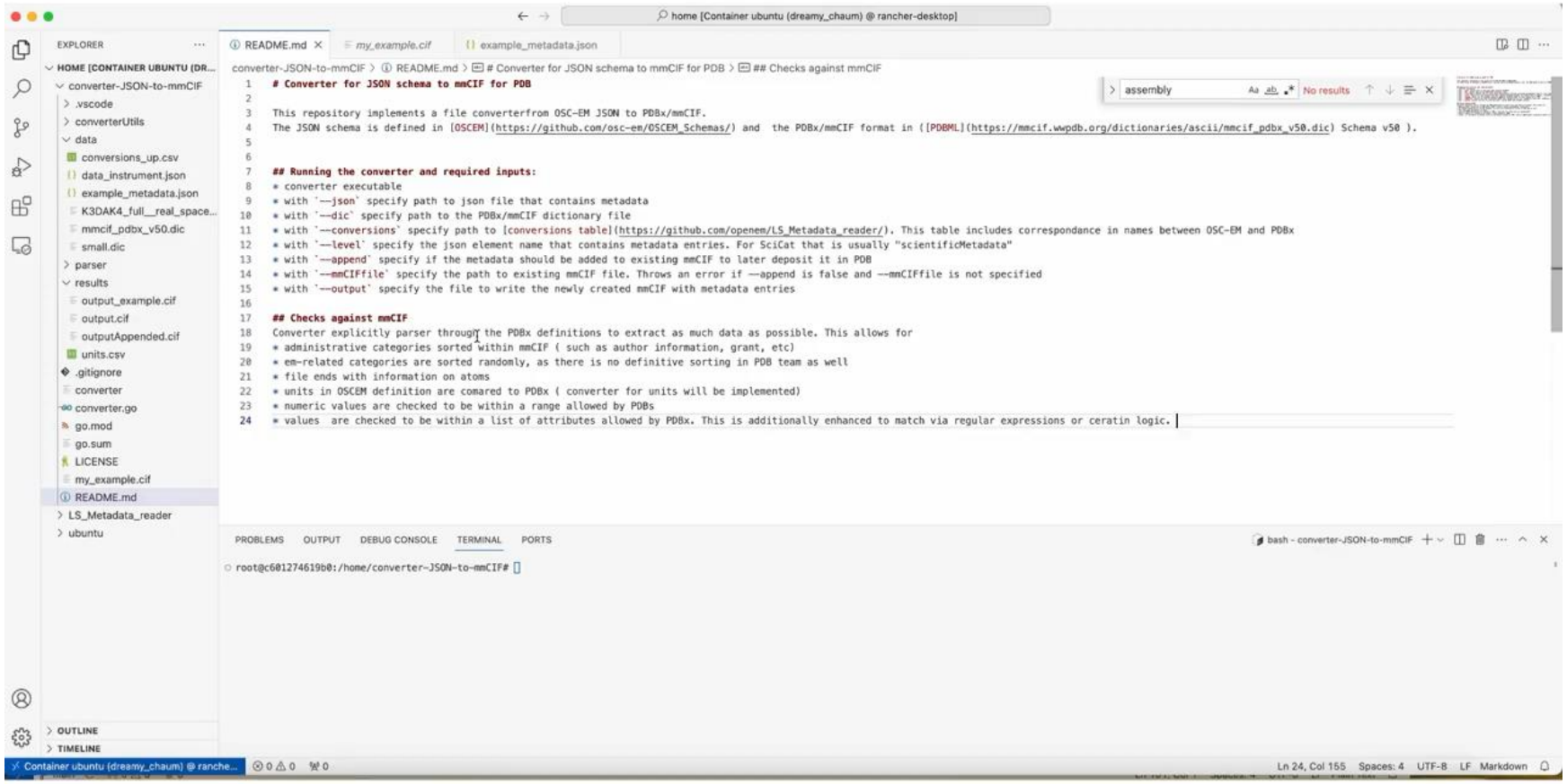


- Import from SerialEM and Thermo Fischer EPU (more coming!)
- Export to mmCIF for deposition in EMDB/PDB OneDep
- Suitable for inclusion in SciCat `scientificMetadata` field

```yaml
# Example OSC-EM dataset
---
instrument:
  microscope: Titan
  illumination: FloodBeam
  imaging: Brightfield
  electron_source: FEG
  acceleration_voltage: 300
  c2_aperture: 70
  cs: 2.7
acquisition:
  holder: testitest
  detector: Falcon 4i
  detector_mode: counting
  dose_per_movie: 0.5
  date_time: "2024-01-01"
  binning_camera: 2
  pixel_size: 1.2
grants: …
authors: …
sample: …
```

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



Resources

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*