PAUL SCHERRER INSTITUT

PSI

OpenEM

WIR SCHAFFEN WISSEN – HEUTE FÜR MORGEN

Spencer Bliven :: Scientific Data Curation :: Paul Scherrer Institute

OpenEM Update

2024-03-05 AWI Department Meeting

Goals

- EM data should be FAIR and Open by default
- Standardized data management at all facilities
- Automatic metadata collection during acquisition
- Streamlined deposition in EMPIAR/EMDB/PDB
- Central data repository providing access to researchers & the public
  - Authenticated access during the embargo period
  - Open access after publication
  - Indexed by search engines or accessible by DOI

Open EM Data Network (OpenEM)

4 ETH Institutes ETH-RAT

5 Universities swissuniversities

University of Basel

Mohamed Chami, Timm Maier, *Yves Tittes*

UNIVERSITÄT BERN

David Kalbermatter, Benoît Zuber

UNIVERSITÉ DE GENÈVE

Andreas Boland, Orsolyz Barabas, Andy Howe, *Attila Nacsa*, **Robbie Loewith**

EPFL DCI Lausanne

Marco Cantoni, *Sofya Lakina*, Alexander Myasnikov, Alexandra Radenovic, **Henning Stahlberg**,

Unil UNIL | Université de Lausanne

Christel Genoud

PAUL SCHERRER INSTITUT PSI

Alun Ashton, *Spencer Bliven*, Gregor Cicchetti, Peter Hüsser, Michael Kallmeier-Glanz, Volodymyr Korkhov, Carlo Minotti, Elisabeth Müller, Gebhard Schertler

Empa Materials Science and Technology

Rolf Erni, *Despina Adamopoulou*

ETH zürich

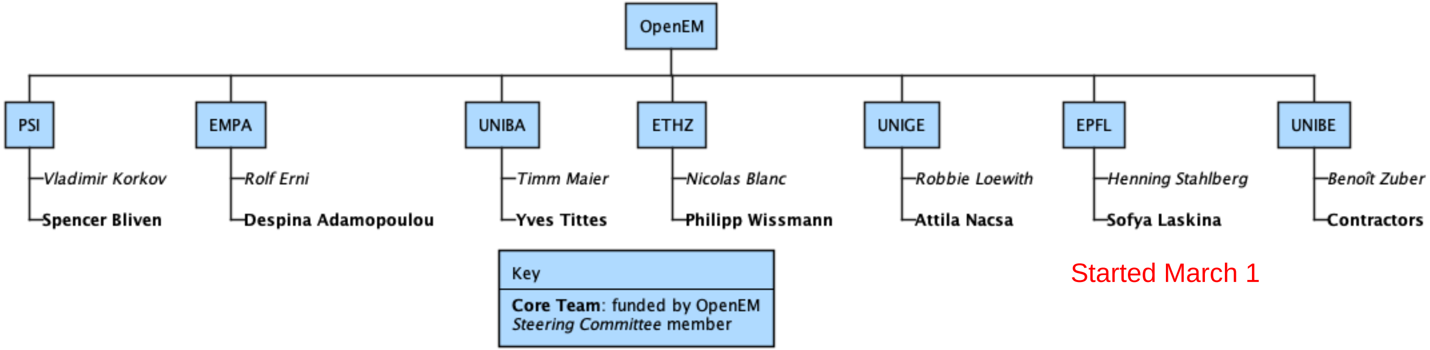
Matthew Baker, Nicolas Blanc, Daniel Böhringer, Christophe Briand, Christophe Copéret, Miroslav Peterek, Bilal Qureshi, Andrzej J. Rzepiela, *Philipp Wissmann*

Universität Zürich UZH

Open EM Data Network (OpenEM)

Open EM Data Network

- Two funding instruments
  - *ETH ORD.* PI: Henning Stahlberg. 1.5 MCHF
  - *Swissuniversities.* PI: Robbie Loewith. 0.92 MCHF
- 6.5 new positions
- Timeline: June 2023–Dec 2025



OpenEM

PSI: *Vladimir Korkov*; **Spencer Bliven**

EMPA: *Rolf Erni*; **Despina Adamopoulou**

UNIBA: *Timm Maier*; **Yves Tittes**

ETHZ: *Nicolas Blanc*; **Philipp Wissmann**

UNIGE: *Robbie Loewith*; **Attila Nacsa**

EPFL: *Henning Stahlberg*; **Sofya Laskina**

UNIBE: *Benoît Zuber*; **Contractors**

Key

Core Team: funded by OpenEM

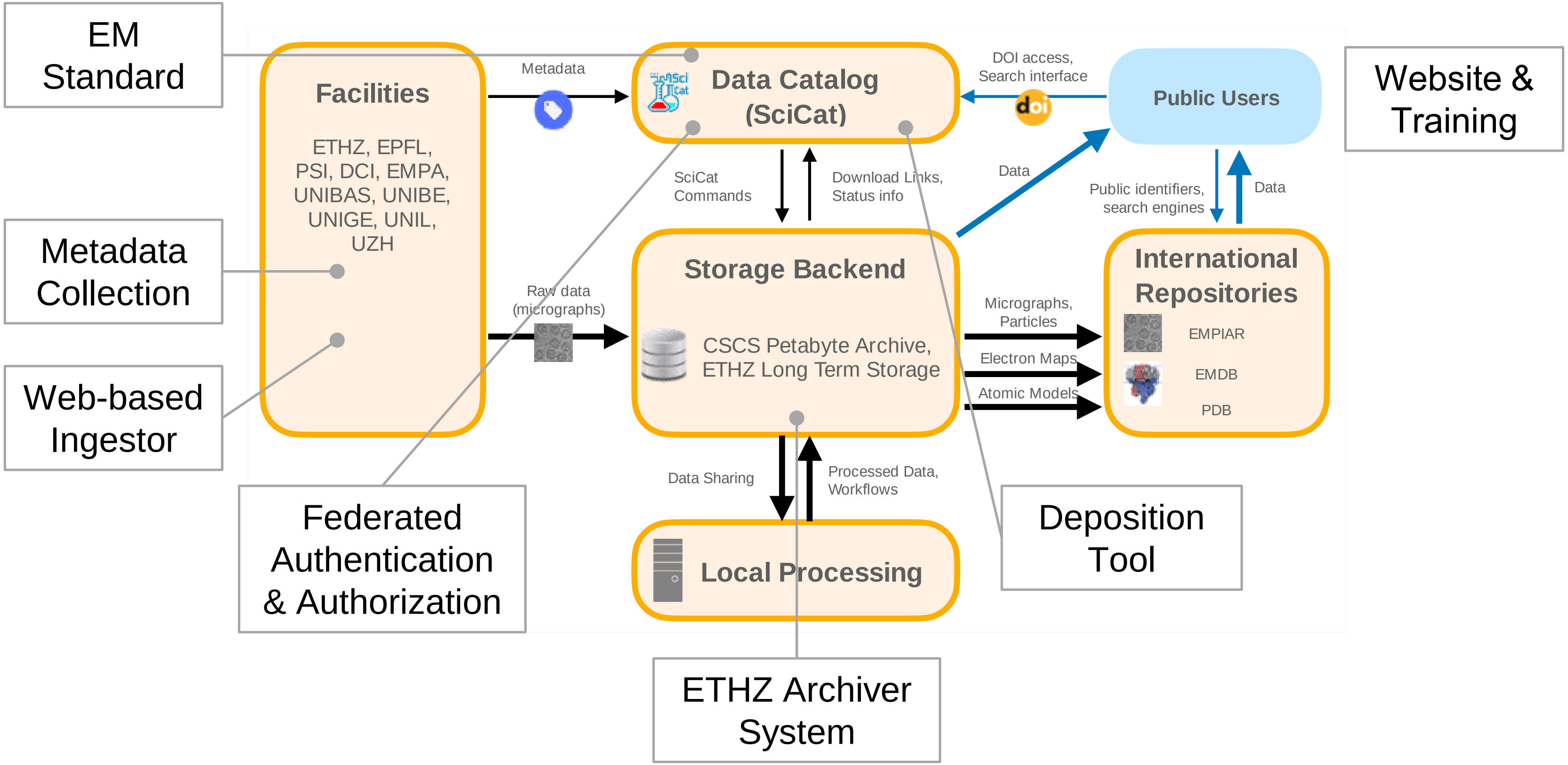
Steering Committee member

Started March 1

Project Management Tools

- Confluence for documents, meeting notes, and planning
- Github projects for tasks/Kanban
- Agile development
- Slack, sympa email lists
- https://swissopenem.github.io

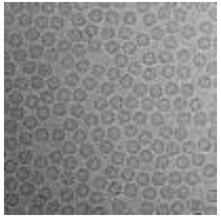
Architecture



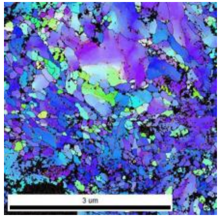
Open Standards Community for EM (OSCEM)

Diverse data types

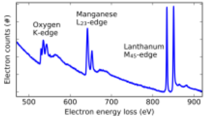
Raw Data



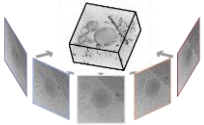
2D Micrographs
(EMPIAR-11016, Harder, EPFL)



Annotated Images
(Kunze and Sologubenko, ETHZ)



Spectrograms
Magnunor, Wikimedia



3D tomograms
teamtomo.org

More: ptychography, 4D STEM, …

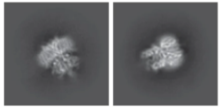
Sample Description

Instrument Settings

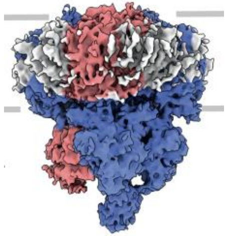
Workflows

Publications

Derived Data

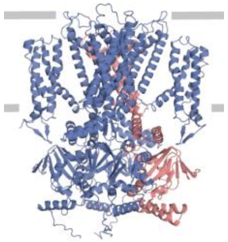


Particles & classes
(Barret, PSI)



Electron Maps (EMDB)

(EMD-12718, Barret, PSI)



Molecular Models (Protein Databank)

(7o4h, Barret, PSI)

More: Tomographic reconstructions, segmented models, …

PAUL SCHERRER INSTITUT PSI

Diverse Formats

Raw Data

Micrographs

- MRC
- TIFF
- Zarr/OME
- HDF5/EMD
- PRZ/numpy
- NXem

Metadata

- XML (diverse formats)
- SerialEM mdoc
- Image metadata (eg TIFF headers)

Manufacturers

- Thermo, Zeiss, Gatan, Jeol, ASI, …

Sample Description

- Proposal systems
- Labbooks

Instrument Settings

- SerialEMSettings.txt
- Configuration files

Workflows

- Project directories
- Workflow files (scipion, ccpem-pipeliner, etc)

Publications

- DOI
- DataCite records
- Git repos

Derived Data

CryoEM

- PDBx/mmCIF
- Map

Databases

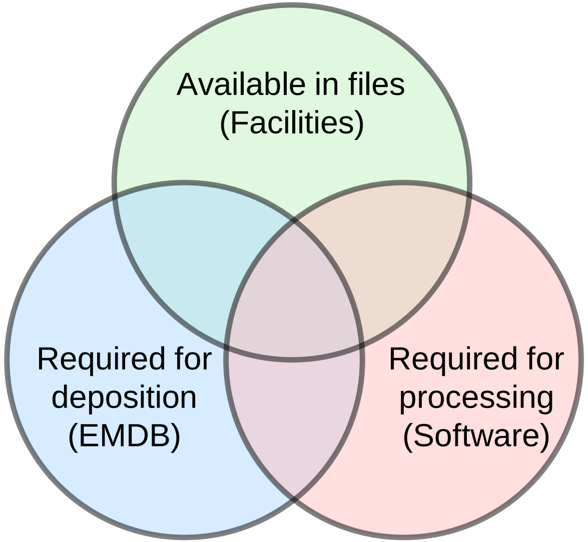
- EMPIAR/EMDB/PDB
- Zenodo/Institutional repos

Workshop February 22-23

- Open Standards Community for EM (OSCEM)
- Participants from facilities, software, and repositories
- https://indico.psi.ch/e/em-standards-2024
- Will draft an *ontology* and a *schema* for EM metadata
  - Encompasses both the minimal metadata for processing a dataset and the metadata required for depositing a dataset
  - Extensible to many techniques in EM (single particle, tomography, EELS
  - Derived from existing standards (CryoEM Ontology, PDBx, NXem)

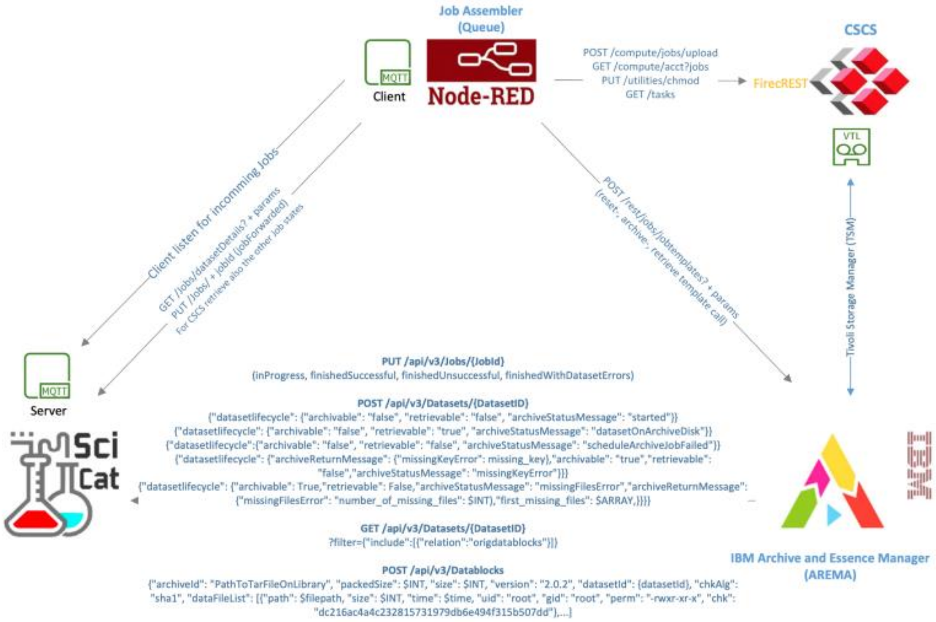
Current baseline: repository requirements (+)

Parameter	Type	EMDB required	SerialEM	EPU	In emd
Instrument Name	str	x	Yes	Yes	Yes
Illumination mode	str	x	No	Yes	Yes
Imaging mode	str	x	No	Yes	No
Electron source	str	x	No	Yes	No
Acceleration Voltage	int	x	Yes	Yes	Yes
C2 Aperture	int		No	Yes	Yes
CS	float		?	Yes	No
Nominal defocus (min/max)	float		Yes	Yes	■
calibrated defocus (min/max)	float		Yes	Yes	■
nominal magnification	float		Yes	Yes	No
calibrated magnification	float		No	No	No
speciman holder model	str		No	?	Yes
cooling holder cryogen	str		No	No	No
Temperature (min/max)	float		No	No	No
alignment procedure	str		No	No	No
software list	str		Yes	Yes	■
Detector / Camera	str	x	?	Yes	Yes
average dose per image	float	x	Yes	Yes	No
Energy filter	bool		Yes	Yes	No
energy filter slit width	float		Yes	Yes	No
detector pixels	int x int		Yes	Yes	No
Date of experiment	str		Yes	Yes	Yes
average exposure time	float		Yes	Yes	■
Tilt angle (min/max)	float		Yes	Yes	No
cryogen	str		No	No	No
Residual tilt	float		No	No	No
Details instrumet	str		No	No	No
specialist optics	str		Yes	Yes	■
spherical aberration corrector	str		No	No	No
chromatic aberration corrector	str		No	No	No
Microscopy settings	str	x	No	Yes	■
Detector mode	str		No	Yes	■
sampling interval	float		No	No	No
Movie frames per image	int		Yes	Yes	No
range of frames used	int		No	No	■
# of grids imaged	int		No	No	No
# of images	int		No	No	■
details for camera	str		No	No	No



Archiver System

PSI Archiver System



POST /compute/jobs/upload
GET /compute/acct?jobs
PUT /utilities/chmod
GET /tasks

Client listen for incomming Jobs

GET /Jobs/datasetDetails? + params
PUT /Jobs/ + JobId (jobForwarded)
For CSCS retrieve also the other Job states

POST /rest/jobs/jobtemplates? + params
(reset-, archive-, retrieve template call)

PUT /api/v3/Jobs/{JobId}
(inProgress, finishedSuccessful, finishedUnsuccessful, finishedWithDatasetErrors)

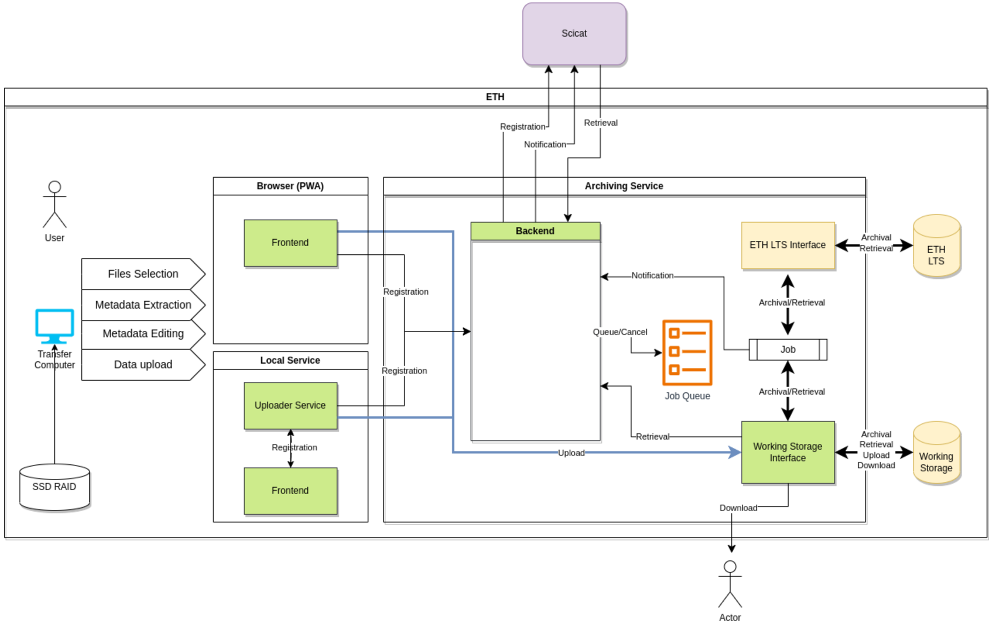
POST /api/v3/Datasets/{DatasetID}
{"datasetlifecycle": {"archivable": "false", "retrievable": "false", "archiveStatusMessage": "started"}}
{"datasetlifecycle": {"archivable": "false", "retrievable": "true", "archiveStatusMessage": "datasetOnArchiveDisk"}}
{"datasetlifecycle":{"archivable": "false", "retrievable": "false", "archiveStatusMessage": "scheduleArchiveJobFailed"}}
{"datasetlifecycle": {"archiveReturnMessage": {"missingKeyError": missing_key},"archivable": "true","retrievable": "false","archiveStatusMessage": "missingKeyError"}}}
{"datasetlifecycle": {"archivable": True,"retrievable": False,"archiveStatusMessage": "missingFilesError","archiveReturnMessage": {"missingFilesError": "number_of_missing_files": $INT),"first_missing_files": $ARRAY,}}}}

GET /api/v3/Datasets/{DatasetID}
?filter={"include":[{"relation":"origdatablocks"}]}

POST /api/v3/Datablocks
{"archiveId": "PathToTarFileOnLibrary", "packedSize": $INT, "size": $INT, "version": "2.0.2", "datasetId": {datasetid}, "chkAlg": "sha1", "dataFileList": [{"path": $filepath, "size": $INT, "time": $time, "uid": "root", "gid": "root", "perm": "-rwxr-xr-x", "chk": "dc216ac4a4c232815731979db6e494f315b507dd"},...]

- RabbitMQ for job notifications from SciCat
- Node-RED dispatch system
- Arema scheduler
- Tivoli Storage Manage for tape interface
- FirecREST for CSCS operations
- SciCat receives status updates via REST API

ETHZ Archiver System



- Initial data transfer to S3 staging location
  - Resumable uploads (tus)
- Archive/retrieve from *ETH Long-term storage* via posix interface
- Interact with Scicat Job API via REST
- Deployment with Cellery/Kubernetes
- Monitoring: grafana, prometheus

SciCat changes

Needed SciCat Features

- Backend undergoing a major update to v4.0.0 (typescript, NestJs, better configuration, …)
- Federated authorization via OIDC
- User management system
  - Independent of PSI-AD
  - Fine-grained authorization
  - More detailed accounting for billing
- New data transfer mechanism
  - Accessible from outside PSI
  - Bandwidth for 3-4 PB/year
  - Considering Globus or S3
- Web-based upload for decentral data

Thanks!

- Data Curation group (Leo, Carlo, Ali)
- OpenEM team
- SciCat team
- Archiver team (Pedro, Michael, Bernard)
- EM Facility staff