

PSI Center for Scientific Computing,
Theory and Data

The Open Data Network for Electron Microscopy (OpenEM)

SciCatCon 2024



EPFL



PSI



Empa

Materials Science and Technology

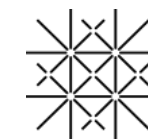
ETH zürich



**DUBOCHET
CENTER
FOR IMAGING**

Unil

UNIL | Université de Lausanne



**University
of Basel**



u^b

**UNIVERSITÄT
BERN**

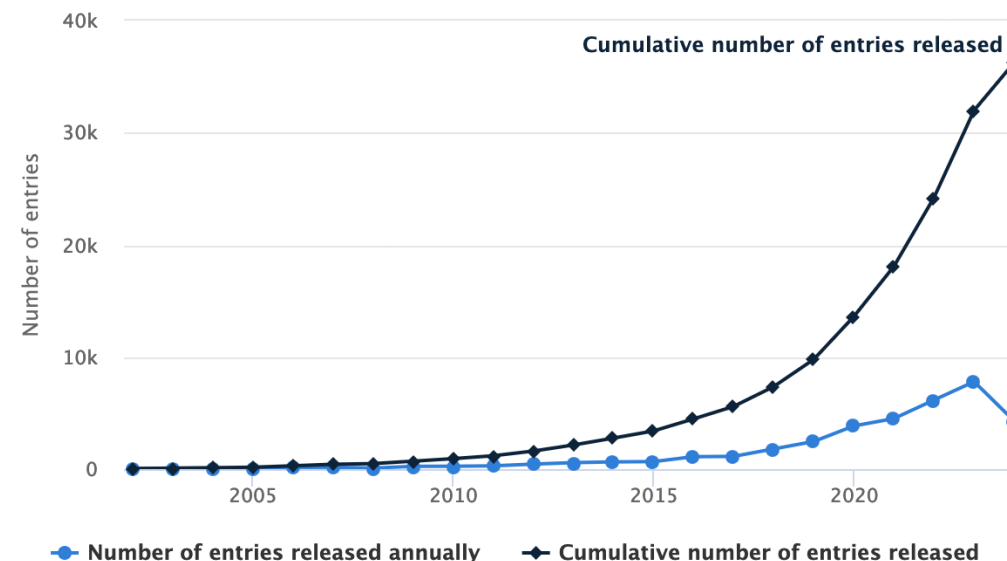
Spencer Bliven
3 July 2024

Swiss Electron Microscopy Facilities in OpenEM

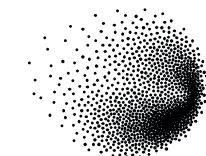
OpenEM Facilities

Facilities	8
Microscopes	50
Yearly microscope users	500
Data Produced	6.4 PB/year

EMDB entries released per year and cumulatively



4 ETH Institutes



PSI

EPFL



Empa

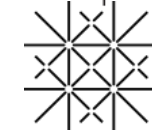
Materials Science and Technology

ETH zürich

4 Universities

Unil

UNIL | Université de Lausanne



University of Basel

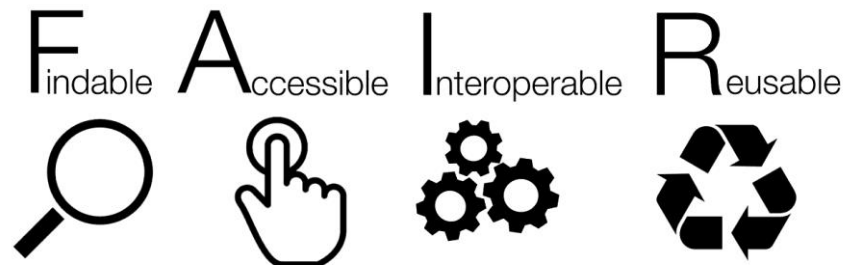


u^b

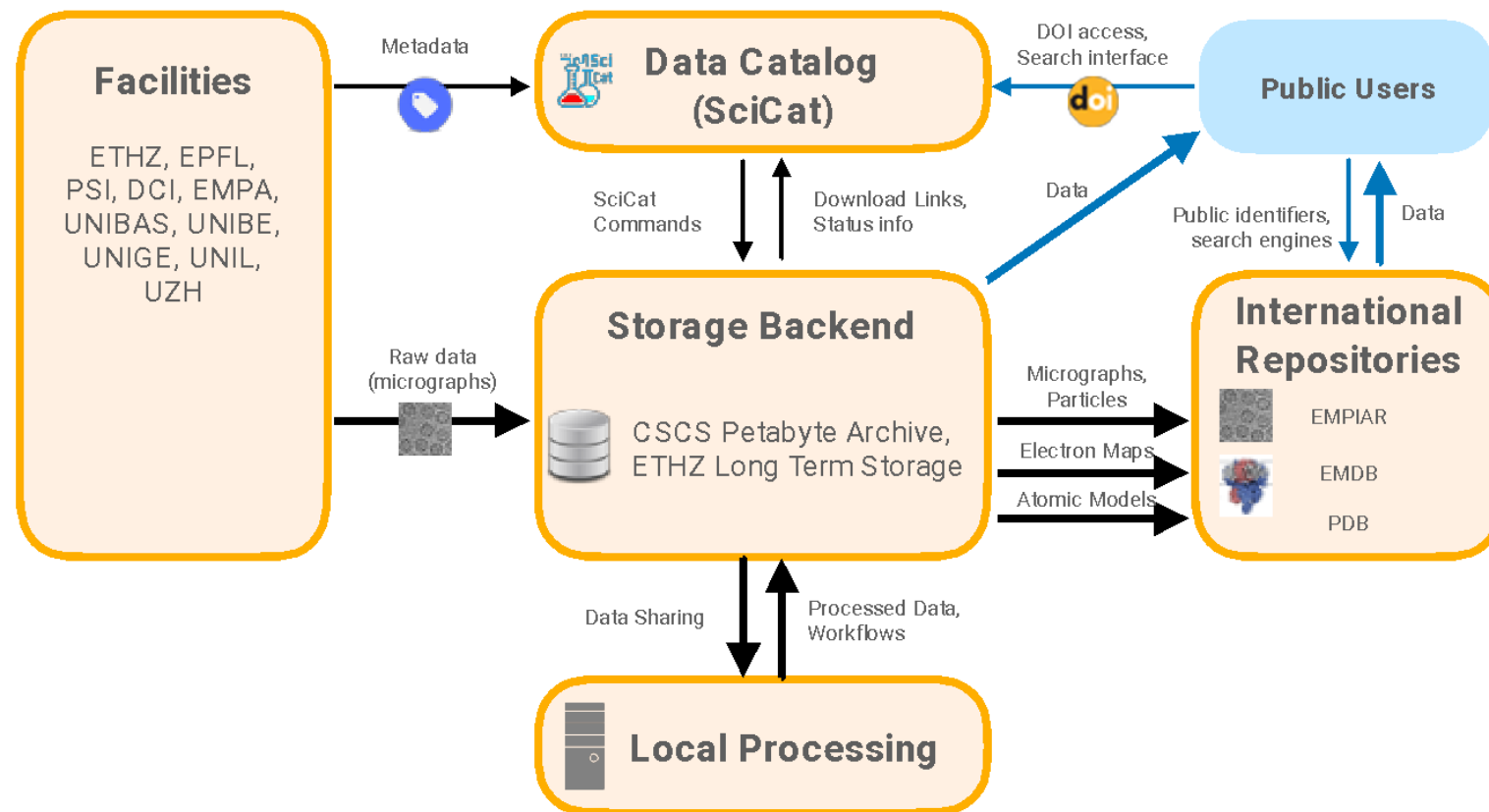
UNIVERSITÄT BERN

Goals

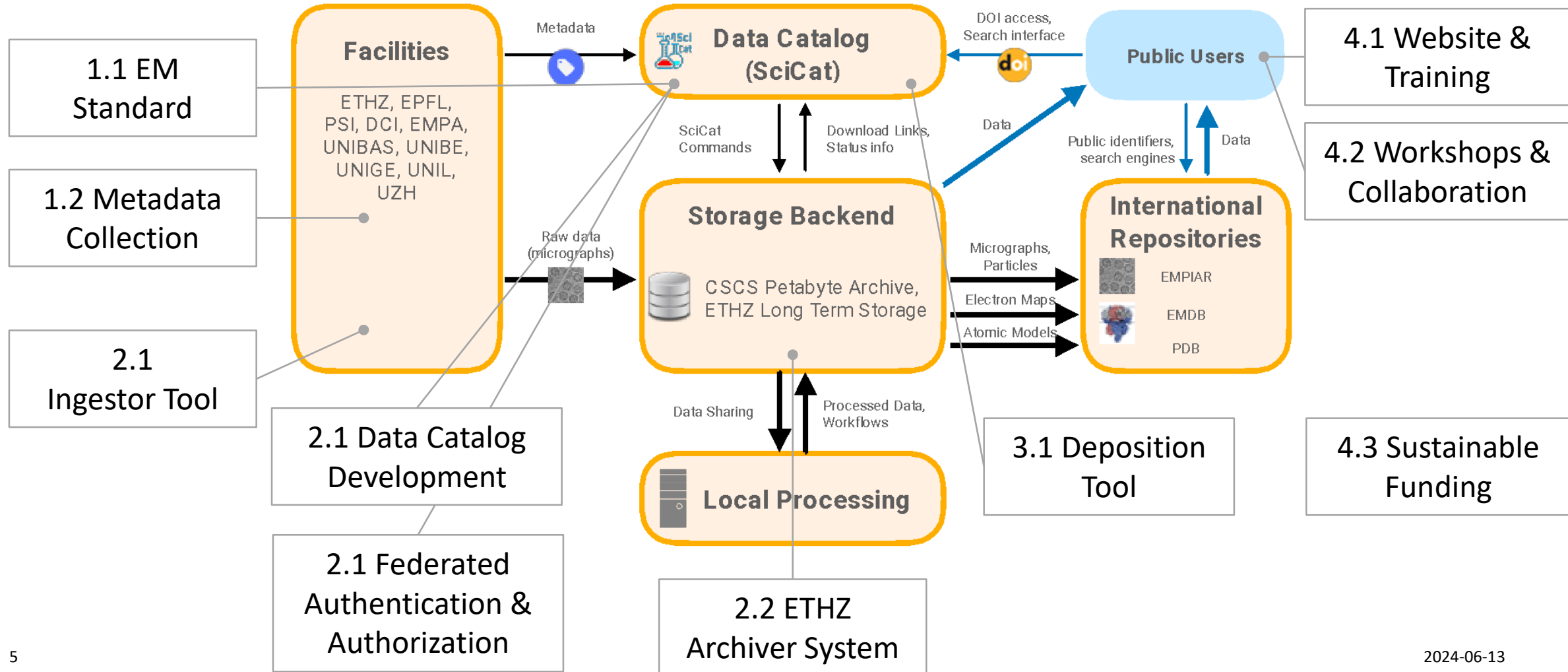
- Electron microscopy (EM) data should be FAIR and Open by default
- Standardized data management at all Swiss cryoEM facilities
- Automatic metadata collection during acquisition
- Streamlined deposition in international community databases (eg EMDB)
- Central data repository providing access to researchers & the public
 - Authenticated access during the embargo period
 - Open access after publication
 - Indexed by search engines or accessible by DOI



Architecture



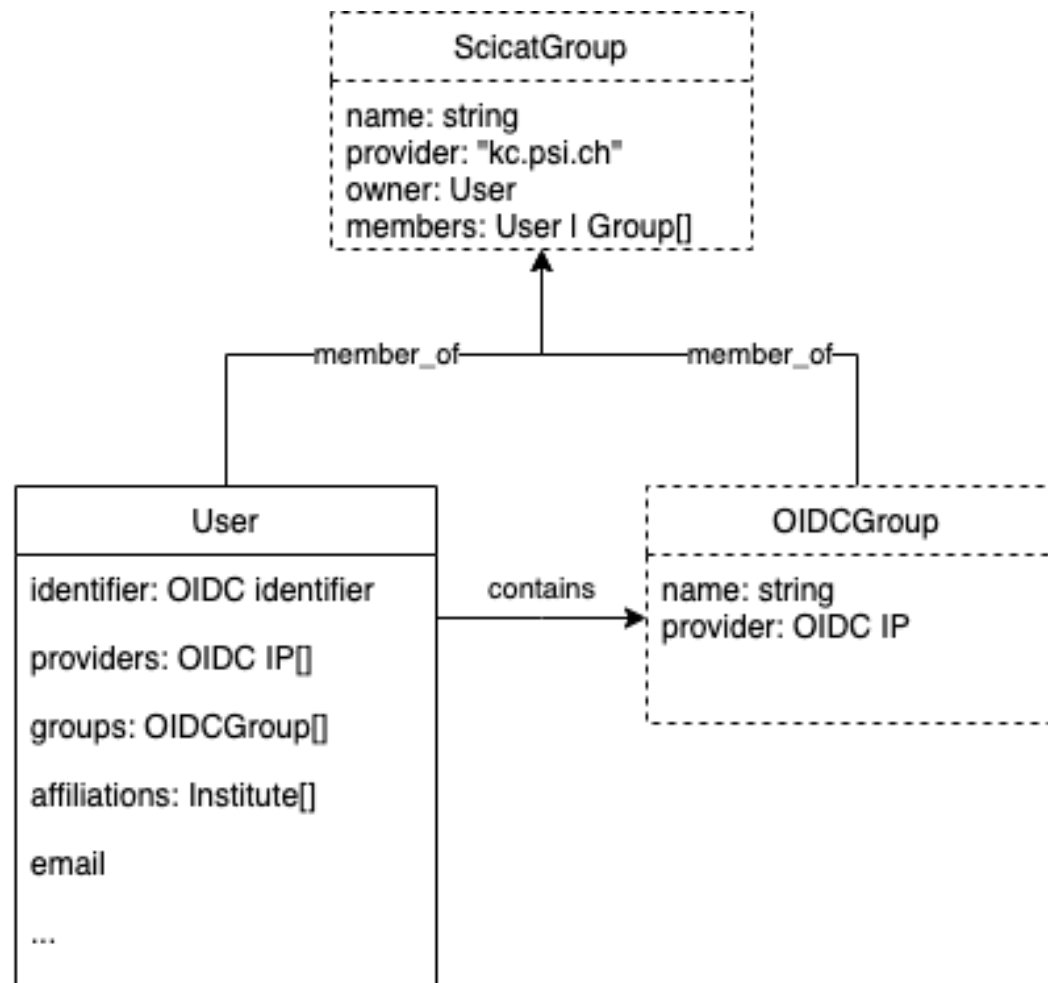
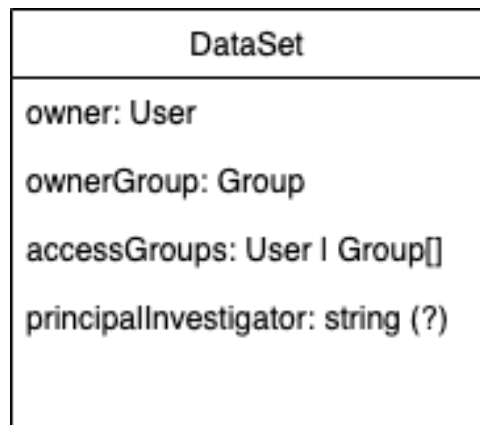
Architecture



Authentication and Authorization

- Open authentication globally using eduGAIN federation
 - Uses SATOSA to proxy multiple identity providers as a single keycloak SAML provider
 - Need to register SATOSA with a local identity federation (SWITCH AAI in Switzerland)
 - Allows all users to authenticate using existing accounts
- Requires managing roles via SciCat
 - Replaces unix/AD users and groups (but needs to be backwards compatible)
 - Need access management tool for groups, roles, and billing info. Any suggestions before we build our own?

Group Concept



scientificMetadata validation



- *Open Standards Community for EM* (<https://github.com/osc-em>)
 - Workshop 22-23 Feb 2024 with participants from facilities, software, and repositories
 - Draft schema available for EM metadata. The goal is to include metadata required for future processing and deposition. (https://github.com/osc-em/OSCEM_Schemas)
 - Currently JSON Schema, but migration to LinkML in progress
 - Schema terms are defined by existing ontologies where available: [CryoEM ontology](#), [PDBx/mmCIF dictionary](#), [Helmholz EM Glossary](#), [NeXus-FAIRmat NXem format](#)
 - Metadata extraction tools for life sciences (https://github.com/SwissOpenEM/LS_Metadata_reader) and material science (<https://github.com/SwissOpenEM/metadata-extractor>)

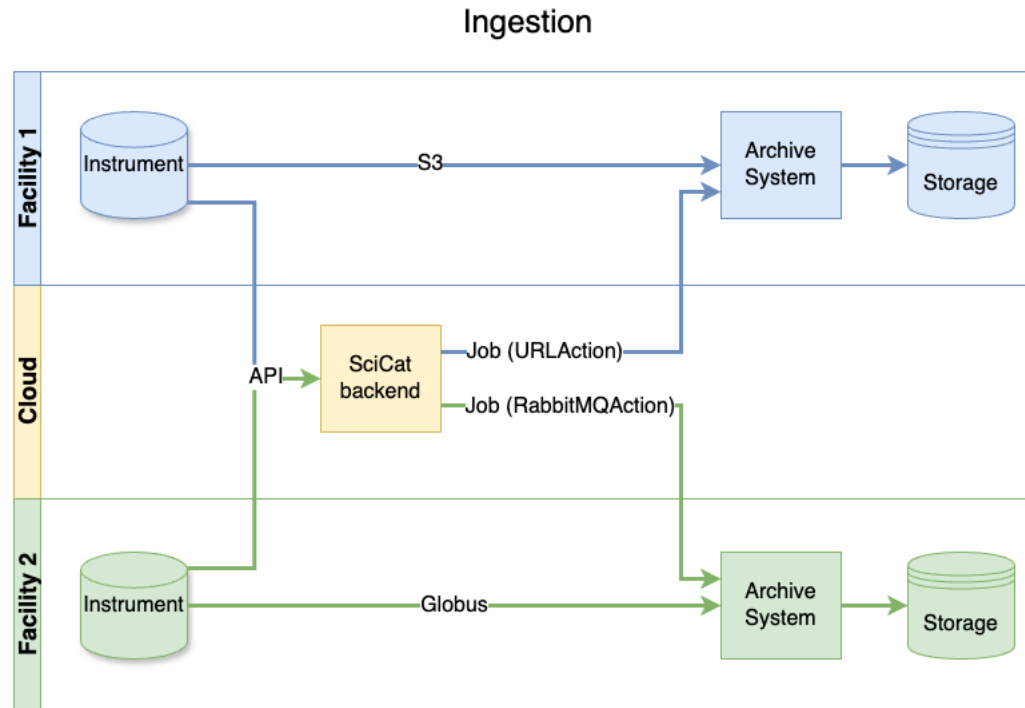


Validation of scientificMetadata #966

- Specify the schema for scientificMetadata
 - `"@context": "https://w3id.org/oscem/sp-cryo-em/1.0/context"`
`"scientificMetadata": { ... }`
- Default schema would be empty/unstructured
- Backend should validate metadata against the schema if specified
- Could be used to selectively enable features, eg:
 - Frontend could change scientificMetadata visualization for some values
 - Augmented search, eg with unit conversion based on semantic units rather than conventions
 - Auto-generated forms for adding and editing metadata (see [jdorn/JSON-Editor](#))

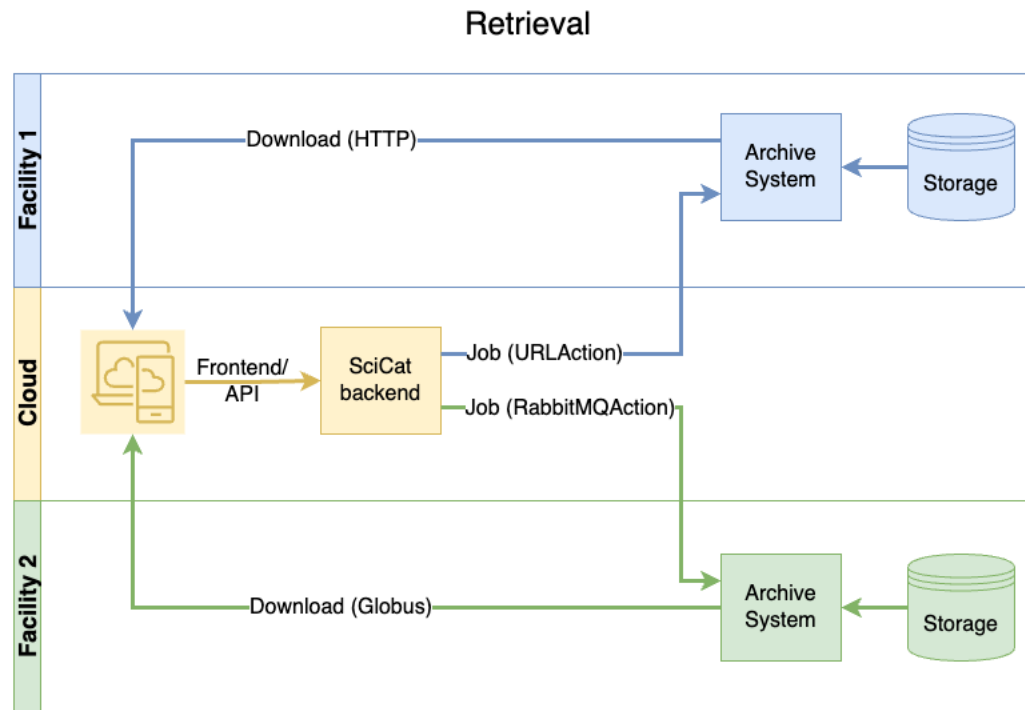
Federated Storage

- Single SciCat database; multiple archive systems (PSI & ETHZ)
- Storage location is determined by the ingestion site. Not envisioned for georedundancy.
- Job configuration dispatches jobs to the correct archive system
- Some sites may require additional authentication (eg with a local LDAP user)

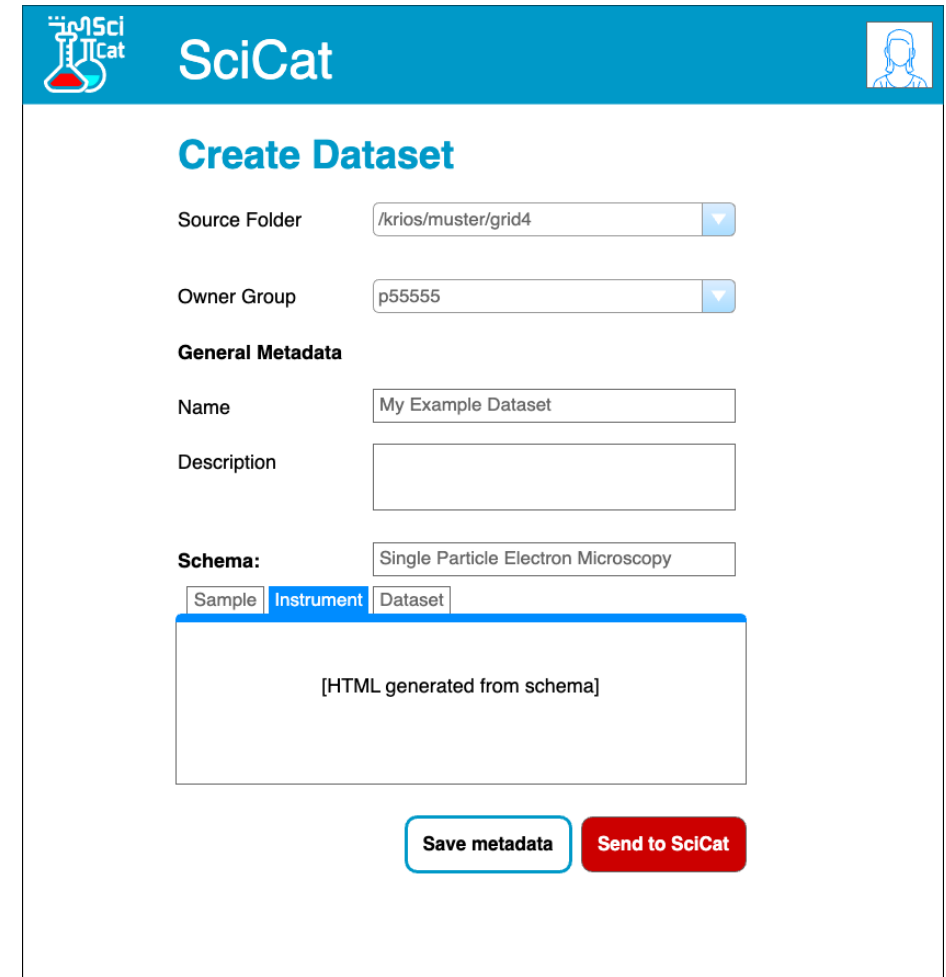


Federated Storage

- Single SciCat database; multiple archive systems (PSI & ETHZ)
- Storage location is determined by the ingestion site. Not envisioned for georedundancy.
- Job configuration dispatches jobs to the correct archive system
- Some sites may require additional authentication (eg with a local LDAP user)



- Use paulscherrerinstitute/scicat-cli for ingesting datasets, retrieving from storage caches, and maintenance tasks
- Golang, Linux/Windows/MacOS, CI/CD
- Qt-based GUI was popular with users but hard to maintain and deploy
- Plan to re-write ingestor GUI using web technologies
 - scientificMetadata editable by users after extraction from dataset files
 - Data transfer via Globus or S3 to archive system
 - Support both facilities and individual users



The image shows a web interface for SciCat titled "Create Dataset". The header includes the SciCat logo and a user profile icon. The form contains the following fields and controls:

- Source Folder:** A dropdown menu with the value "/krios/muster/grid4".
- Owner Group:** A dropdown menu with the value "p55555".
- General Metadata:**
 - Name:** A text input field containing "My Example Dataset".
 - Description:** An empty text area.
 - Schema:** A dropdown menu with the value "Single Particle Electron Microscopy".
- Schema Tabs:** Three tabs labeled "Sample", "Instrument", and "Dataset". The "Instrument" tab is currently selected.
- Content Area:** A large text area containing the placeholder text "[HTML generated from schema]".
- Buttons:** Two buttons at the bottom right: "Save metadata" (white with blue border) and "Send to SciCat" (red).

Mockup

EMDB/EMPIAR/PDB deposition

- Encourage deposition in existing databases following :

- Micrographs



- Density Maps, tomograms



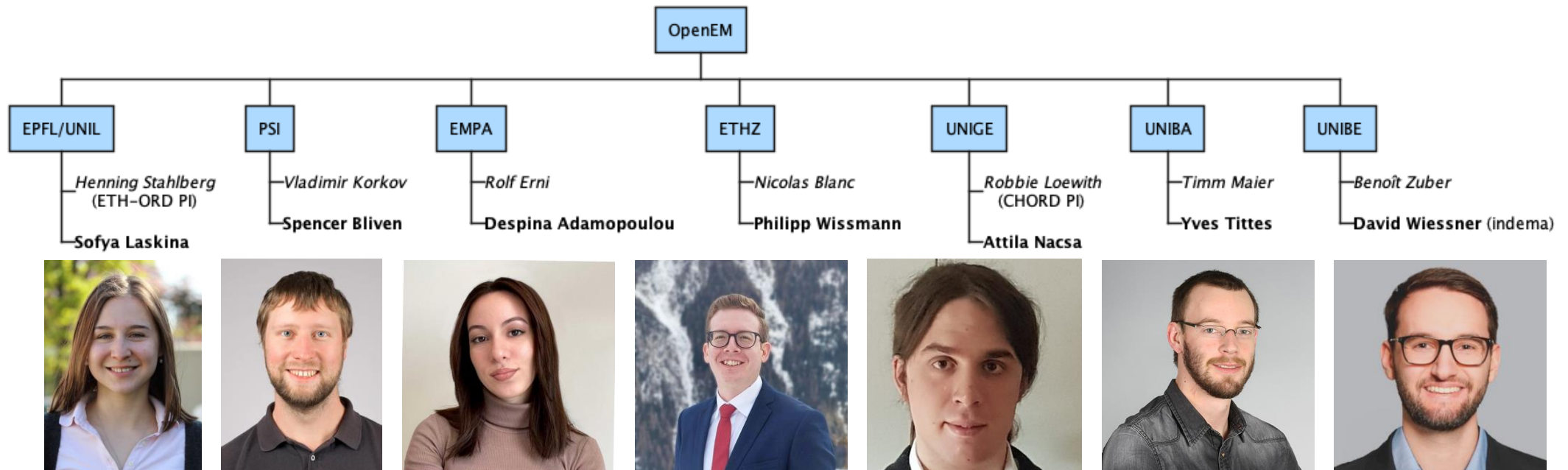
- Molecular Models

- Want to initiate deposition from a SciCat EM dataset, filling forms based on scientificMetadata
- Received early access to the OneDep API, which provides a method for depositing life science datasets to EMDB and PDB. An [empiar-depositor](#) tool is also available
- OSC-EM to mmCIF format converter developed for metadata interoperability: <https://github.com/osc-em/converter-JSON-to-mmCIF>



Thanks!

- OpenEM members
- Carlo Minotti, Ali Rezaee Vahdati, Leonardo Sala
- OpenEM is supported by the Open Research Data Program of the ETH Board.



OpenEM Websites

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

SciCat Data Catalog

- Data repository: <https://discovery.psi.ch>
- Published datasets: <https://doi.psi.ch/>
- SciCat documentation: <https://scicatproject.github.io>

Open Source Software

- SciCat backend: <https://github.com/SciCatProject/scicat-backend-next>
- SciCat CLI <https://github.com/paulscherrerinstitute/scicat-cli>
- ETHZ Archiving Services <https://github.com/SwissOpenEM/ScopeMArchiver>
- Golang Globus transfer library <https://github.com/SwissOpenEM/globus-transfer-request>
- Metadata conversion tools: https://github.com/SwissOpenEM/LS_Metadata_reader and <https://github.com/SwissOpenEM/metadata-extractor>
- OSC-EM format converters <https://github.com/osc-em/converter-JSON-to-mmCIF>
- OSC-EM Schema: https://github.com/osc-em/OSCEM_Schemas