PAUL SCHERRER INSTITUT

PSI

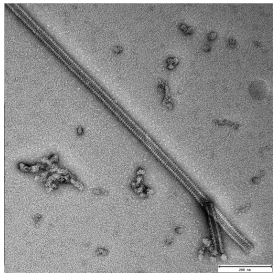
Spencer Bliven :: High Performance Computing & Emerging Technologies :: Paul Scherrer Institute

Open EM Data Network (OpEM)

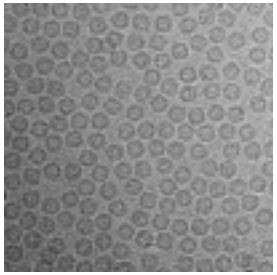
2023-05-02 PSI ORD

Electron Microscopy (EM) Data

Raw Data

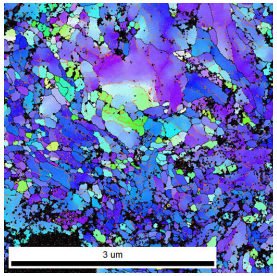


TEM
(Sejwal, PSI)

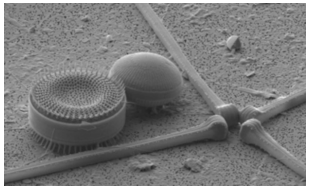


Single particle cryoEM
(EMPIAR-11016, Harder, EPFL)

Typical size:
1-10 TB/dataset
3-4 PB/year for major facilities



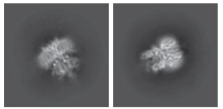
Orientation Maps
(Kunze and Sologubenko, ETHZ)



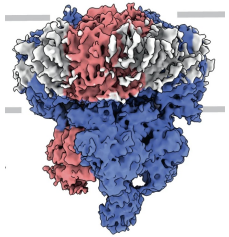
SEM
(Müller, PSI)

Also: Spectrograms, electron tomography, ptychography, 4D STEM, …

Derived Data

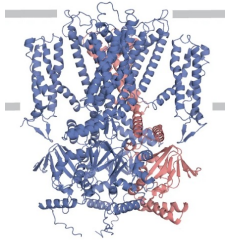


Particles & classes
(Barret, PSI)



Electron Maps (EMDB)

(EMD-12718, Barret, PSI)



Molecular Models (Protein Databank)

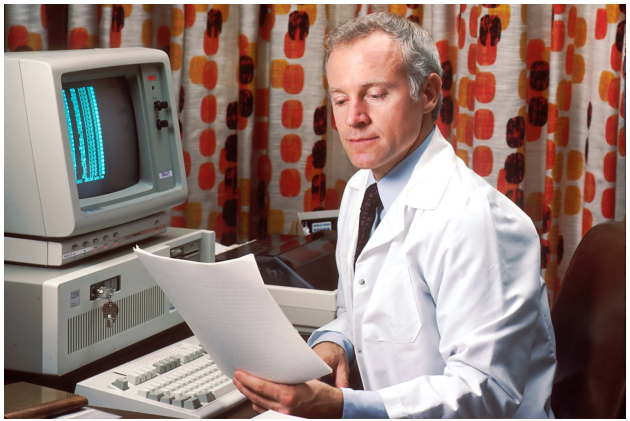
(7o4h, Barret, PSI)

Also: Tomographic reconstructions, segmented models, …

Users

- Data Producers
  - Microscope users
  - EM Facility Scientists
  - Microscope & detector researchers

- Data consumers
  - Data scientists, structural biologists, informaticians
  - Students
  - Training AI/machine learning



Benefits of open data for electron microscopy

- Publishing raw data is essential
  - **Reproducibility.** Check results & apply new methods to old data
  - **Verifiability.** Detect processing mistakes and protect against scientific fraud
  - **Education**. Learn processing techniques by reproducing cutting-edge papers
  - **Method development**. Datasets provide training data for future AI methods
  - **Interdisciplinary**. Images may be relevant for questions beyond the original scope
- Examples from crystallography:
  - PDB redo
  - Phasing old data with molecular replacement (or AlphaFold models)

Open EM Data Network (OpEM)

- How can we improve ORD practices in the electron microscopy (EM) community? → Open EM Data Network
  - ETH ORD M1 Establish application by Henning Stahlberg: 1.5 MCHF
  - swissuniversities by Robbie Loewith: 0.92 MCHF
  - 6.5 new positions
  - ETH: June 2023–Dec 2025 (30 months)
  - Swissuniversities: Jan 2023–Dec 2024 (24 months)

Open EM Data Network (OpEM)

4 ETH Institutes

ETH-RAT

5 Universities

swissuniversities

PAUL SCHERRER INSTITUT

Alun Ashton, Spencer Bliven, Gregor Cicchetti, Stephan Egli, Volodymyr Korkhov, Carlo Minotti, Elisabeth Müller, Gebhard Schertler

University of Basel

Mohamed Chami, Timm Maier

Empa

Materials Science and Technology

Rolf Erni

UNIVERSITÄT BERN

Benoît Zuber

ETH zürich

Nicolas Blanc, Daniel Böhringer, Christophe Briand, Christophe Copéret, Miroslav Peterek, Bilal Qureshi, Andrzej J. Rzepiela

UNIVERSITÉ DE GENÈVE

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

EPFL DCI Lausanne

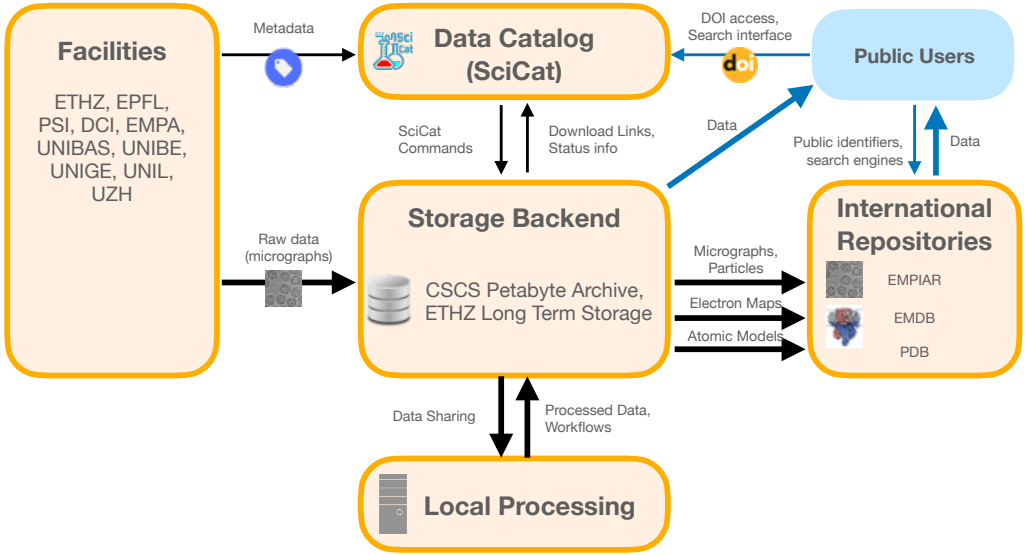
Marco Cantoni, Alexandra Radenovic, **Henning Stahlberg**

UNIL | Université de Lausanne

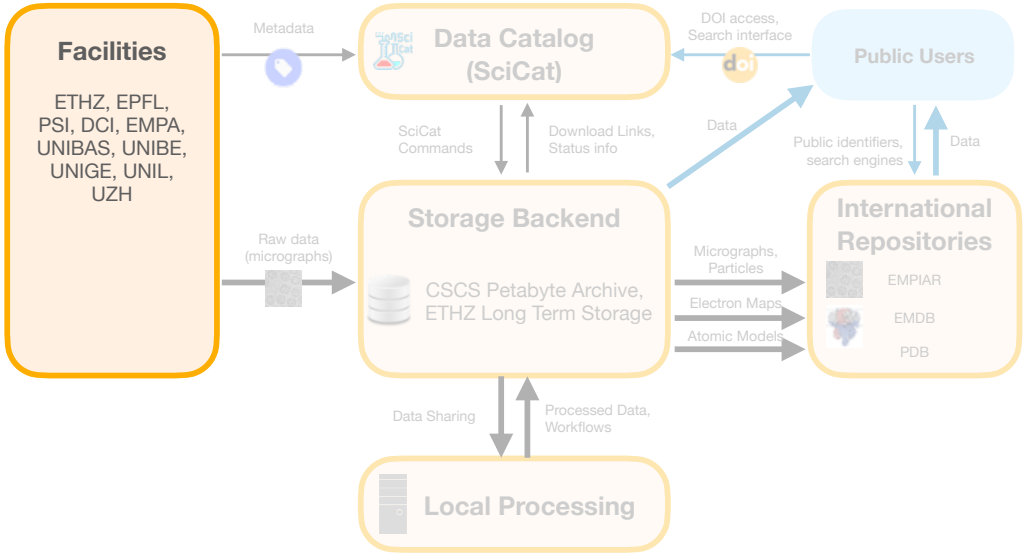
Christel Genoud

Universität Zürich[UZH]

Architecture

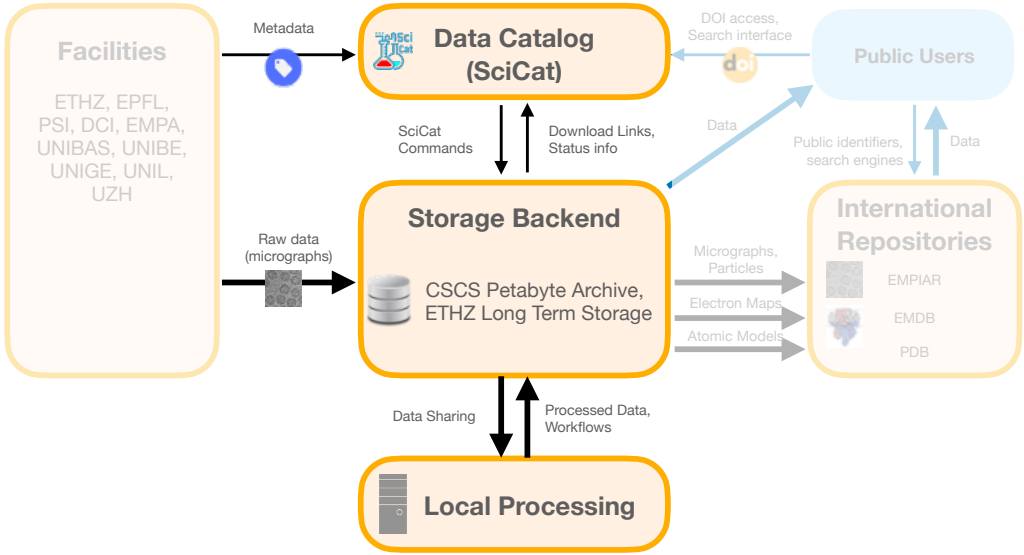


Project Plan



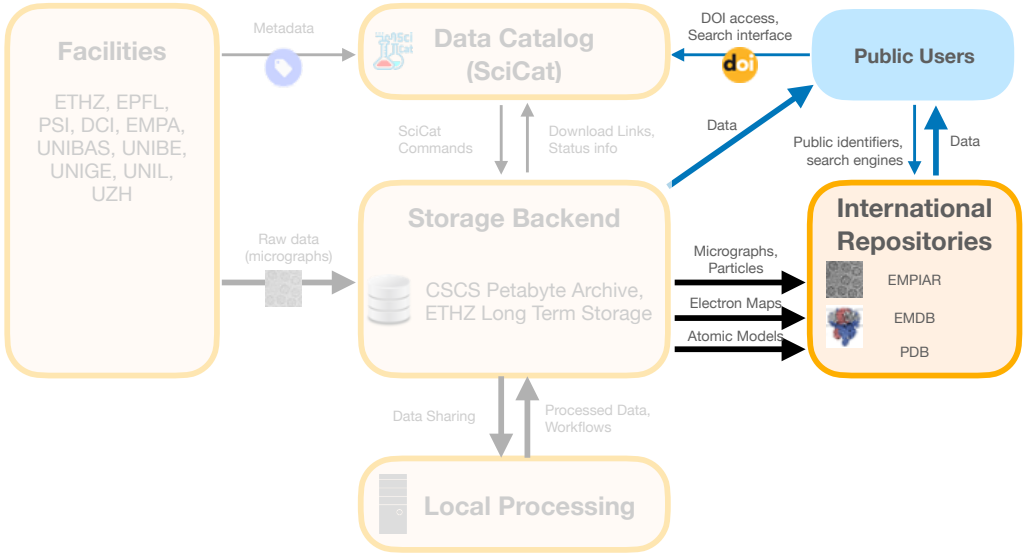
1. Automated collection of metadata during acquisition.
   - Initial focus on single particle cryoEM, but with support for additional domains like material science
   - Coordinate with EU & international standards and initiatives, eg. INSTRUCT-ARIA, 3DEM

Project Plan



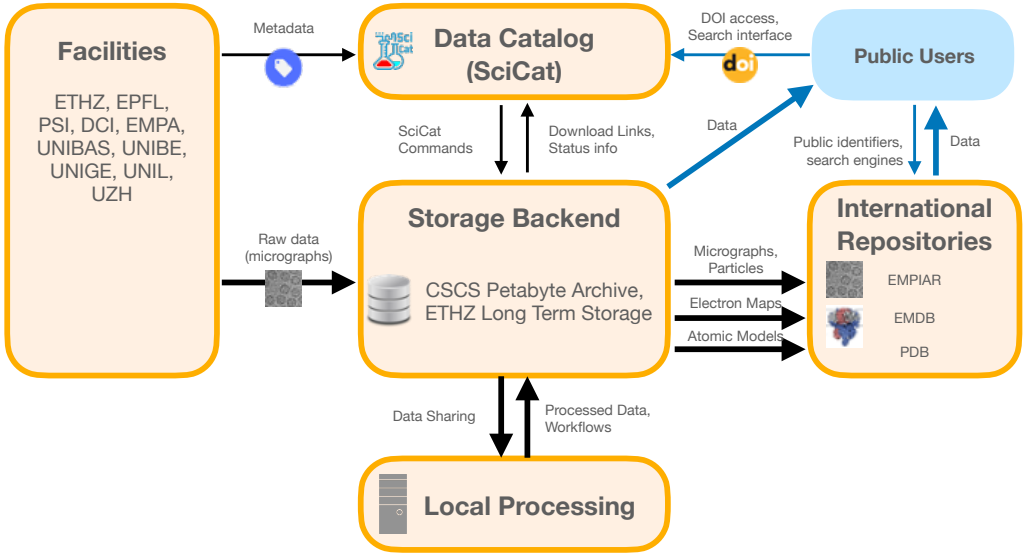
2. Ingestion of metadata from all sites in the Data Catalog
   - PSI will provide data storage for some sites; others adapt existing site-specific storage (e.g. ETHZ LTS)
   - Users deposit derived datasets after processing

Project Plan



3. Seamless deposition in international repositories (EMPIAR/EMDB/PDB)
   - Pre-populate fields from metadata

Project Plan

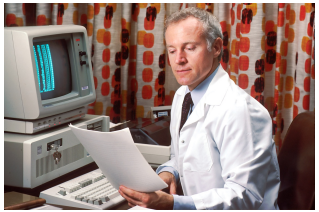


4. Training & outreach
   - Provide user training for researchers, facility managers, and data consumers
   - Find sustainable funding model

For data producers

- Standardized data management at all facilities
- Raw data & metadata are automatically entered in the catalog
- Swiss-wide data transfer with unified authorization (SWITCH AAI)
- Complies with institutional and federal data policies
- Integrates with popular processing software, eg relion, CCPEM-pipeliner, cryoSPARC
- DOI assigned for publication
- Prepares single-particle cryoEM datasets for deposition at EMPIAR+EMDB+PDB

For data consumers



- Findable via OpenAIRE, European Open Science Cloud, Google Dataset search, etc.
- Linked to publications and international repositories
- Searchable via metadata (http://discovery.psi.ch)
- Data retrievable from storage via asynchronous systems
- Scriptable via REST API

Major Tasks

- Now hiring!
- Data Catalog Tasks
  - Federated login outside PSI
  - Permissions model independent of PSI
  - Ingestion from remote sites
  - Support external data storage
- Microscope Facility Tasks
  - Standardize EM scientific metadata
  - Ingestion service for each facility
  - Integrate with institutional storage & services
- Data Publication
  - Collaboration with EBI for EMPIAR/EMDB/PDB deposition
- Train staff & scientists
- Find sustainable funding model

Thanks!

- Henning Stahlberg (EPFL), Robbie Loewith (UNIGE), Gebhard Schertler, Gregor Cicchetti, and other OpEM members
- Stephan Egli, Carlo Minotti, Max Novelli (ESS), Laura Shemilt (RFI) & other SciCat developers
- Derek Feichtinger & the HPCE Group
- Alun Ashton, Leo Sala, & AWI colleagues
- EM Facility staff