DATA SCIENCE &
SCIENTIFIC COMPUTING

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Designing and deploying a FAIR-by-design data pipeline and platform for electron microscopy laboratories

Research thesis in: Data Management

Supervisor
Dott. Federica Bazzocchi

Candidate
Nicola Perin

University of Trieste

19 settembre 2025

Agenda

Motivation & Problem

- Electron microscopy (EM) labs generate very large, heterogeneous datasets (images, diffraction, spectra).
- Vendor-specific formats plus sparse metadata hinder interoperability & reuse.
- Informal practices (file names, local notes) do not scale across collaborations.

FAIR Principles (Goal)

- Make data **Findable, Accessible, Interoperable, Reusable**.
- Emphasize rich metadata, standardized formats, and machine actionability.
- Align with funder & journal expectations; enable open science.

Standards & Formats

- **HDF5**: performant hierarchical container for large arrays and attributes.
- **NeXus**: community conventions (e.g., `NXinstrument`, `NXsample`).
- **NXem**: EM application definition (images, diffraction, EDS/EELS, 4D-STEM) + metadata.

Institutional Context

- Area Science Park → RIT institute with labs: LADE (data engineering), LAGE (genomics), LAME (electron microscopy).
- ORFEO data center provides HPC, Ceph storage, and identity services.
- Project targets LAME workflows; scalable across NFFA–DI.

Infrastructure: ORFEO

- **Ceph** distributed storage with tiers for speed/capacity; replication and erasure coding.
- **RADOS Gateway (RGW)**: S3-compatible interface for object storage.
- **Identity and SSO**: FreeIPA (directory & CA) + Authentik (OIDC provider).
- Integrated into **ORFEO HPC cluster**: compute + storage + identity under one umbrella.

Application Platform

- Web app streamlines **upload** + **annotation** in one flow.
- Built on **Django (MVT)** + PostgreSQL; REST API for scripted ingest.
- Containerized services deployed with **Helm charts**.
- Separation of interactive deposit (at the lab) from heavy processing (at data center).

Deployment & Validation

- **VirtualOrfeo**: digital twin of ORFEO HPC, safe for experimentation.
- Built from VMs (KVM/QEMU) managed via Vagrant + Ansible.
- Replicates: directory services, Kubernetes, Ceph cluster.
- Enables fast iteration without impacting production.

K3s Cluster Topology (VirtualOrfeo)

- Lightweight **K3s** Kubernetes with one control-plane/worker VM.
- Namespaces: `authentik`, `lame-fair`, monitoring.
- Ingress via **NGINX** + load balancing with **MetalLB**.
- Certificates issued by FreeIPA CA, automated by **cert-manager**.

Identity & Access Management

- FreeIPA provides directory, groups, CA.
- Authentik acts as OIDC provider integrated with FreeIPA.
- Django app registered as OIDC client (`lame-fair`).
- Secure login: Authentik manages tokens; app never handles raw passwords.

Storage Integration in VirtualOrfeo

- **Ceph RGW**: S3-compatible object store for raw + processed data.
- Optional MinIO for lightweight tests.
- Bucket layout mirrors Project → Proposal → Sample → Experiment.
- NeXus files generated alongside raw TIFFs for FAIR compliance.

Application Deployment on K3s

- Packaged as a **Helm chart** (pods, services, secrets, ingress).
- Gunicorn + NGINX sidecar for Django web service.
- Background worker with Redis queue.
- PostgreSQL database + migrations as Helm hooks.

Domain Model & Data Flow

- Project → Proposal → Sample → Experiment → Measurement hierarchy.
- Metadata captured early; mapped into NeXus/NXem.
- Curated outputs stored in centralized data lake (S3/Ceph).

Metadata & NeXus Construction

- Automated mapping of instrument settings (beam energy, detectors, stage coordinates).
- Human-readable README + machine-readable NeXus containers.
- Libraries (e.g., `pynxtools-em`) lower the adoption barrier.

Storage Gateway & Background Tasks

- Upload pipeline writes to object storage; lifecycle managed server-side.
- Background workers handle conversion, validation, indexing.
- Auditability & provenance preserved end-to-end.

Security, Performance, Scalability

- OIDC tokens with group claims → role-based access in-app.
- Ceph scales horizontally; PostgreSQL tuned for concurrent users.
- Asynchronous jobs decouple ingestion from heavy compute.

Contributions

- FAIR-by-design workflow from acquisition to curated NeXus/NXem.
- Unified upload + annotation web app; API surface for automation.
- Deployment blueprint validated in VirtualOrfeo; ready for ORFEO/NFFA–DI.

Conclusions & Next Steps

- Reproducible, interoperable EM data pipeline proved feasible.
- Scale to additional instruments and labs; expand validators & viewers.
- Prepare for open access portals and cross-lab discovery.