



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

Dr. Federica Bazzocchi

Candidate

Nicola Perin

University of Trieste

19/9/2025

Outline

- ① Data management challenges in electron microscopy
- ② Scientific foundations: FAIR data and standards
- ③ Our infrastructure: LAME and ORFEO
- ④ Pipeline and platform design

Electron microscopy and its data challenges

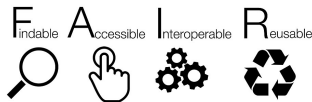
- ▶ **Electron microscopy (EM)**: probe matter at the nanometer scale.
- ▶ Techniques: TEM (internal), SEM (surface), STEM (combo + spectroscopy).
- ▶ Produces huge datasets: images, diffraction patterns, spectra.
- ▶ Issues:
 - Terabytes per session, proprietary formats, poor metadata.
 - Manual handling → lost context.
 - Hard to share and reuse.



Question: how to keep EM data usable and shareable in the long run?

Scientific solution: FAIR and standards

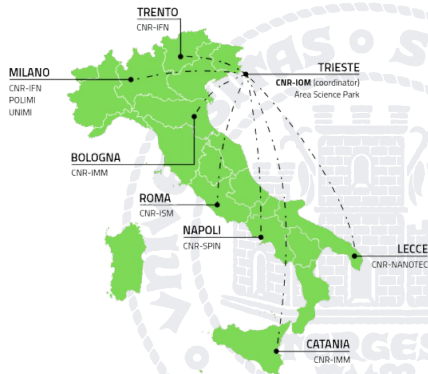
- ▶ **FAIR** = Findable, Accessible, Interoperable, Reusable. Goal: data remain useful beyond the lab and the project.
- ▶ **NeXus**: international standard on top of HDF5 for structured scientific data (e.g. NXinstrument, NXsample).
- ▶ **NXem**: new NeXus application definition for electron microscopy. Ensures images, diffraction, spectra and metadata are stored consistently.



NeXus

Scientific solution: NFFA-DI

- ▶ **NFFA-DI** = Nano Foundries and Fine Analysis – Digital Infrastructure.
- ▶ National initiative linking nanoscience labs across Italy.
- ▶ Mission: **FAIR data practices**, open access to advanced instruments, shared compute.
- ▶ My work contributes to this broader infrastructure effort.



Source: <https://nffa-di.it/en/>

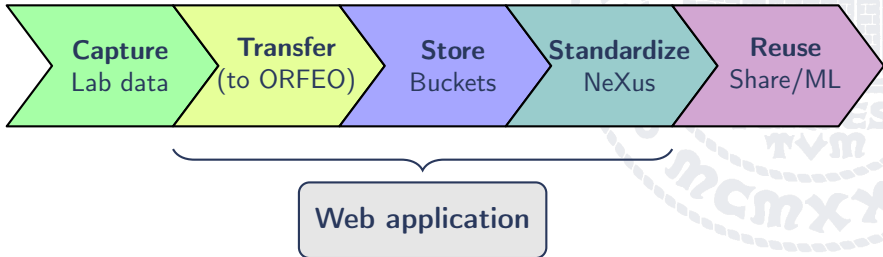
Our infrastructure: LAME and ORFEO

- ▶ **LAME:** advanced EM lab (opened 2022), with TEM/STEM and SEM; affiliated with NFFA-DI.
- ▶ **ORFEO:** datacenter providing storage, HPC, identity services. Core of the NFFA-DI digital infrastructure.
- ▶ **Current gap:** local storage, manual transfers, no smooth link to ORFEO.



Practical solution: a FAIR-by-design pipeline

- ▶ Bridges **LAME** lab practices with **ORFEO** infrastructure.
- ▶ Ensures data move **smoothly** from capture to reuse, without manual gaps.
- ▶ FAIRification happens at the **standardization** step.
- ▶ A **web application** orchestrates transfer, standardization, and storage.



Designing the web application

The infrastructure

- ▶ **Authentik**: centralized single sign-on (SSO).
- ▶ **Ceph storage**: a distributed object store. Data are split into **objects**, replicated across many servers → scalable and fault-tolerant.

The application

- ▶ Built with **Django**, modeling research workflow as: **Project / Proposal / Sample / Experiment / Measure**.
- ▶ Manages user identities through Authentik.
- ▶ Interacts with Ceph via the **Amazon S3 API**.
- ▶ Runs **background tasks** (NeXus conversion).

Using the web application

- 1 Log in with credentials.
- 2 Create a project, add samples and experiments.
- 3 Upload raw data files.

New sample for NFFA_DI / 123

Sample name:

sample001

Sample identifier:

smpl1

Preparation date:

09 / 02 / 2025

Atom types (comma-separated):

Fe,Ag

Physical form:

powder

Create

Cancel

Proposals / samples / experiments

Proposals +

NFFA_DI / 123

Info

Samples for NFFA_DI / 123 +

sample001

Info

Experiments for sample001 +

exp001 — some description

Info

Add measurement to exp001 (sample sample001)

Detector

TEM_JEOL_F200 - TVIPS_camera

Choose file(s)

Choose folder

Upload & register

Testing & deployment: *VirtualOrfeo*

VirtualOrfeo is a lightweight digital twin of the ORFEO datacenter. It consists of multiple **virtual machines** and configuration files, simulating:

- ▶ storage
- ▶ identity
- ▶ compute nodes

The Django web application is packaged as a **container** and deployed inside the **Kubernetes (K3s) cluster**, integrated with storage and identity services as in production.

Conclusions

- ▶ **Pipeline:** from lab capture to FAIR data in ORFEO.
- ▶ **Webapp:** practical tool for projects, uploads, and NeXus conversion.
- ▶ **Validation:** tested end-to-end on VirtualOrfeo.
- ▶ **Impact:** reusable design for NFFA-DI and other labs.
- ▶ **Modularity:** the app can interact with **external services** (e.g. machine learning analysis, data management plan tools) through **APIs**, all testable within VirtualOrfeo.

Thank you!

Questions welcome.

