



MDMC

Master in Data Management
and Curation

MASTER IN DATA MANAGEMENT AND CURATION

Implementation of a pipeline for collecting, ingesting and transforming data into standard formats for the LAME FIB-SEM

Supervisor(s):
Federica BAZZOCCHI

Candidate:
Elaheh SAADAT

2024–2025





Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA

This Pilot training activity has been funded by the European Union – NextGenerationEU within the project PNRR "PRP@CERIC" IR0000028 and «NFFA-DI" IR0000015 - Missione 4, "Istruzione e Ricerca" – Componente 2, "Dalla ricerca all'impresa" – Linea di investimento 3.1, "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" – Azione 3.1.1, "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti".

The supporting projects:

- [NFFA-DI : Nano Foundries Fine Analysis - Digital Infrastructure](#)
- [PRP@CERIC : Pathogen Readiness Platform for Ceric - Eric Upgrade](#)



nffa-di



prp

Author's Declaration

I, Elahesh Saadar, declare that this thesis entitled, 'mplementation of a pipeline for collecting, ingesting and transforming data into standard formats for the LAME FIB-SEM' and the work presented therein are my own.

I certify that:

- This work was performed wholly or principally during MDMC internship in THE LABORATORY.
- If any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have relied on the published work of others, this is always clearly attributed.
- Where the thesis is based on work I have done in collaboration with others, I have made clear exactly what was done by others and what I contributed.
- With respect to AI technologies (CHOOSE YOUR OPTION):
 - I acknowledge the use of OpenAI's ChatGPT (<https://chat.openai.com>) to provide information for background research and to assist in the drafting writing process with the creation of an outline structure for this essay.
 - I acknowledge the use of OpenAI's ChatGPT (<https://chat.openai.com>) to identify improvements in the writing style.
 - I acknowledge the use of OpenAI's ChatGPT (<https://chat.openai.com>) as a source of information to create materials that will be used in my own words.
- Where I have quoted from the work of others, the source is always cited. Except for such quotations, this thesis is entirely my own work.
- I have acknowledged all major sources of help.

Signed:



Date:

20/06/2025

"In theory, there is no difference between theory and practice. But, in practice, there is."

Jan L. A. van de Snepscheut

Acknowledgments

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Federica Bazzocchi, and our advisor Dr. Stefano Cozzini, for their guidance, support, and patience throughout this project.

I warmly thank the coordinator, Dr. Mariarita de Luca, and all the teachers of the MDMC for the knowledge and support they provided.

I am grateful to the LADE (Laboratory for Advanced Data Exploration) team for providing a supportive environment and for giving me the opportunity to work on this project. I also acknowledge the collaboration with the Laboratory of Electron Microscopy (LAME) at Area Science Park in particular Dr. Jummi Laishram, responsible for the Tescan Amber X, and Dr. Regina Ciancio, the laboratory coordinator, despite the challenges faced during the project, which offered valuable lessons.

Finally, I am deeply grateful to my family, friends, and colleagues for their constant encouragement and support throughout this journey.

Abstract

This thesis presents the development of a FAIR-by-design data management platform for the Laboratory of Electron Microscopy (LAME) at Area Science Park in Trieste, Italy. LAME supports high-resolution materials research using techniques such as Scanning Electron Microscopy (SEM), Focused Ion Beam-SEM (FIB-SEM), and Scanning Transmission Electron Microscopy (STEM), generating large volumes of complex data that require structured, interoperable management.

The platform streamlines data acquisition, metadata capture, validation, and publication in line with the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Built using Django, it enables researchers to upload structured NeXus-format data through an intuitive web interface.

All components are deployed within ORFEO, the centralized data center at Area Science Park, which hosts MinIO for object storage, Authentik for secure authentication, and NOMAD Oasis for FAIR-compliant internal publishing under the NFFA-DI (Nano Foundries and Fine Analysis – Digital Infrastructure) framework. Datasets are also integrated into OFED (Overarching FAIR Ecosystem for Data), ensuring standardized, traceable, cross-institutional access.

The platform is aligned with NFFA-Europe standards via tools like MetaRepo and will soon be publicly hosted on ORFEO. The underlying codebase and documentation will be made openly available on GitHub to support reuse and adoption by other scientific facilities.

Contents

1	FAIR Principles and Data Management	11
1.1	Overview of FAIR Principles	11
1.2	Relevance in Experimental Laboratories	12
1.3	Ontologies and Metadata Standards	12
1.3.1	Ontologies	12
1.3.2	Metadata Standards	13
1.3.3	Implementation in Research Data Management	13
2	LAME Lab and Workflow	14
2.1	Instrumentation and Software	14
2.2	Data Types and Metadata Categories	15
2.3	Challenges in Current Practices	15
2.4	Planned Data-Transfer and Publication Workflow	16
3	System Architecture	17
3.1	Django Interface Design	17
3.1.1	Web API for Data Upload	17
3.1.2	Data Flow	18
3.2	Backend Architecture	20
3.2.1	Database Schema	20
3.2.2	Integration with MinIO and NOMAD Oasis	20
3.3	User Roles and Workflow	20
3.3.1	Defined Roles	20
3.3.2	Workflow Summary	21
4	Metadata Modeling and NeXus File Generation	22
4.1	Metadata Schema for SEM and FIB-SEM	22
4.1.1	Pre-measurement Metadata Categories	22
4.2	Ontology Alignment and NeXus Mapping	23
4.2.1	Key NeXus Classes	23

4.2.2	TESCAN AMBER X Metadata Mapping	24
4.3	Validation Procedures	24
4.3.1	Validation via NOMAD Cloud	24
4.4	NeXus File Generation	25
4.4.1	Remarks	25
5	Data Pipeline, Sharing, and Evaluation	27
5.1	Overview	27
5.2	Data Handling and Storage Workflow	27
5.2.1	Data Transfer and Cleaning	27
5.2.2	Structured Storage via MinIO	27
5.3	Standardization and Interoperability	28
5.4	Repository Publishing	28
5.5	System Evaluation	28
5.5.1	Current Limitations	28
5.5.2	Planned Improvements	29
5.6	Impact and Outlook	29
5.7	Conclusion	29
	References	31

Introduction

Background

Scientific research is entering a phase where the amount of data being produced is growing rapidly. This is largely due to the use of advanced laboratory instruments and improved computing tools. A good example is the Laboratory for Advanced Microscopy and Electron Microscopy (LAME), located at the Area Science Park in Trieste, Italy. The lab is equipped with instruments such as the JEOL F200 TEM/STEM with a cold Field Emission Gun (FEG), the JEOL Grand Arm 300 kV TEM/STEM, and the Plasma FIB-SEM Tescan Amber X. These instruments generate large and complex datasets that require organized and sustainable data management.

At the same time, the global research community is increasingly adopting the FAIR principles, which aim to make data Findable, Accessible, Interoperable, and Reusable [1]. In this context, research data management (RDM) is becoming more important. Effective RDM involves using standardized metadata, consistent data formats, and systems that support data sharing and reuse, helping to improve transparency and reproducibility in research.

Problem Statement

Even with advances in electron microscopy, laboratories like LAME continue to face challenges in managing their data. Electron microscopy data and metadata are often produced using proprietary software—such as TESCAN’s Essence and Oxford Instruments’ Aztec in the case of the Plasma FIB-SEM Tescan Amber X—which can lead to limited interoperability and isolated data storage.

In addition, documentation practices are often manual and inconsistent. This can result in incomplete metadata and lower data quality, which makes it more difficult to apply the FAIR principles in practice. These issues can reduce reproducibility, complicate data sharing, and limit the broader use of scientific data.

Objective

The main goal of this thesis is to design and implement a data management platform that follows FAIR principles from the outset, with a focus on the specific needs of electron microscopy research at LAME. The platform is intended to address current challenges by:

- Defining a clear, ontology-based metadata schema to support consistent documentation.
- Developing a user-friendly web API using Django to allow researchers to input metadata and data in a standardized way after the experiment.
- Automating data integration and cleaning processes to produce well-structured, reliable datasets.
- Enabling the automatic creation of standardized NeXus files to support long-term preservation and data sharing through FAIR-compliant repositories.

This work aims to improve the overall quality, consistency, and accessibility of microscopy data at LAME, contributing to better alignment with FAIR data practices.

The focus of this thesis was primarily on the development of the frontend interface and the generation of NeXus files for data produced by the Tescan Amber X microscope. Work related to data from the JEOL instruments, as well as data storage infrastructure and publication workflows, was carried out by a collaborating colleague as part of the their project.

Thesis Structure

This thesis is structured into the following chapters:

- **Chapter 2: FAIR Principles and Research Data Management**
Reviews the FAIR principles and their relevance to research data management, with a focus on electron microscopy. This chapter provides the conceptual foundation for the project. A formal reference to the FAIR principles is suggested [1].
- **Chapter 3: LAME Lab and Workflow**
Describes the LAME facility, including its instrumentation, workflows, and current data management practices.

- **Chapter 4: System Architecture**

Presents the overall architecture of the platform, including the Django-based web API, database structure, and user interaction components.

- **Chapter 5: Metadata and NeXus File Management**

Combines the development of an ontology-based metadata schema with the methods used for automatic NeXus file generation. The chapter explains how metadata is captured, structured, and transformed into a standardized file format.

- **Chapter 6: Data Flow, Evaluation, and Conclusions**

Summarizes the full data pipeline, including metadata validation, dataset publication, and repository integration. It also presents identifies limitations and outlines possible directions for future work.

Chapter 1

FAIR Principles and Data Management

1.1 Overview of FAIR Principles

Effective data management is a requirement in modern scientific research. The FAIR principles—Findable, Accessible, Interoperable, and Reusable—were proposed to improve the usability and long-term value of research data [1]. These principles may support data sharing across institutions.

Findable

To be discoverable, data should be assigned globally unique and persistent identifiers and accompanied by rich metadata. This enables both humans and machines to locate and reference the data. Storing data in searchable repositories further supports this goal [1, 2].

Accessible

Access to data should be governed by clear protocols. Standardized communication methods (e.g., HTTP, FTP) facilitate access. Even if the data itself is restricted, the metadata should remain openly accessible to support transparency and potential reuse [1].

Interoperable

Interoperability requires the use of standardized formats and controlled vocabularies. This allows data to be integrated with other datasets and tools, enabling broader analysis and reuse [1, 2].

Reusable

For data to be reusable, its origin, collection methods, and licensing conditions must be clearly described. Community standards and detailed metadata support reproducibility and further application in new contexts [1].

1.2 Relevance in Experimental Laboratories

Experimental laboratories produce high volumes of heterogeneous data. Applying the FAIR principles in this context can improve data management and scientific workflows in several ways:

- **Data Standardization:** Use of common formats and metadata supports consistency and facilitates collaboration.
- **Reproducibility:** Structured documentation reduces errors and enables the replication of experiments.
- **Efficiency:** Reusing well-documented data helps minimize duplication of experimental efforts.
- **Compliance:** Adherence to FAIR principles is increasingly required by funding agencies and publishers [3].

Integrating FAIR-aligned practices into laboratory workflows supports research transparency and improves long-term data usability.

1.3 Ontologies and Metadata Standards

Standardized metadata and domain ontologies are essential to making data interoperable and reusable.

1.3.1 Ontologies

Ontologies define a controlled vocabulary and structure for describing data and their relationships within a specific research domain. Their use allows for consistent annotation and improves the ability to integrate and compare data across studies [4, 5].

1.3.2 Metadata Standards

Metadata documents contextual information about data, such as experimental conditions, instrumentation, and processing steps. Discipline-specific metadata standards, such as those used by platforms like NOMAD in materials science, help ensure data is described in a consistent and interpretable manner [6].

1.3.3 Implementation in Research Data Management

The practical adoption of ontologies and metadata standards requires:

- **Standard Selection:** Identifying ontologies and metadata schemas that are relevant to the research domain.
- **Training:** Ensuring that research staff are familiar with the standards and their application.
- **Tool Support:** Implementing software that facilitates metadata annotation and validation.
- **Ongoing Review:** Regularly updating standards and workflows in response to new requirements or community practices.

Integrating these elements into research workflows supports compliance with FAIR principles and improves the long-term accessibility and value of research data [7].

Chapter 2

LAME Lab and Workflow

2.1 Instrumentation and Software

Established in 2022, the Laboratory of Electron Microscopy (LAME) at Area Science Park in Trieste, Italy, supports advanced materials characterization, with a particular emphasis on nanoscience and nanotechnology [8]. The laboratory is equipped with several advanced electron microscopy systems, including:

- **JEOL F200 TEM/STEM Microscope:** This instrument includes dual energy-dispersive X-ray spectroscopy (EDS) detectors, a CEOS CEFID EELS spectrometer, and a DECTRIS ELA hybrid pixelated camera for energy-filtered 4D STEM imaging. It also supports tomographic analysis and in situ experimental configurations.
- **JEOL Grand ARM 300 kV TEM/STEM Microscope:** Featuring aberration correction for both image and probe modes, this system enables sub-angstrom resolution. It is equipped with dual EDS detectors, a CEOS CEFID EELS spectrometer, a DECTRIS ELA camera for 4D STEM, and an electron dose modulator, making it suitable for sensitive material analysis.
- **Plasma FIB-SEM Tescan Amber X Microscope:** This system is designed for TEM lamella preparation and materials analysis under low-vacuum conditions. It supports in situ sample processing and controlled atmosphere transfer, and includes an integrated EDS detector for elemental analysis.

These instruments are supported by proprietary software such as TESCAN Essence and Oxford Instruments Aztec, used for data acquisition, processing, and spectral analysis. These tools allow detailed characterization of materials at multiple scales and modalities.

2.2 Data Types and Metadata Categories

Research conducted at LAME produces a variety of data types, each associated with specific metadata requirements:

- **High-Resolution Imaging Data:** Generated through TEM, STEM, and SEM methods to study atomic- and nanoscale structures.
- **Spectroscopic Data:** EDS and EELS techniques provide quantitative and spatially resolved elemental and electronic structure information.
- **Tomographic Data:** Acquired through FIB-SEM slice-and-view protocols to reconstruct three-dimensional sample geometries.
- **In Situ Experimental Data:** Collected under variable temperature, pressure, or gas environments to study material behavior in operational conditions.

To support reproducibility and data reuse, metadata is recorded in the following categories:

- **Instrument Settings:** Includes accelerating voltage, magnification, camera length, and detector specifications.
- **Sample Information:** Covers source, preparation method, and physical characteristics.
- **Data Processing Details:** Describes the software and workflows used for image reconstruction, spectral analysis, and data visualization.

The use of standardized metadata categories is essential to support interoperability and alignment with FAIR principles [1].

2.3 Challenges in Current Practices

Despite its capabilities, LAME faces several challenges related to data management:

- **Data Volume and Complexity:** High-throughput instruments generate large datasets, which place demands on storage, processing, and archiving infrastructure.
- **Metadata Consistency:** Heterogeneous experiments require standardized protocols to ensure metadata completeness and consistency.

- **Collaborative Workflows:** Sharing data with national and international partners requires interoperable formats and clear documentation.
- **FAIR Compliance:** Ensuring that data meets FAIR principles requires continuous adaptation of workflows and tools [2].

Addressing these challenges is essential for enhancing data reuse, transparency, and long-term preservation.

2.4 Planned Data-Transfer and Publication Workflow

To address current limitations, LAME is developing an integrated data management workflow designed to improve data handling, documentation, and publication [8].

The workflow begins with the automatic transfer of raw and processed data from acquisition workstations to a MinIO object storage system, hosted on the ORFEO infrastructure at Area Science Park. A micro-service layer will then validate and harvest metadata and datasets, preparing them for publication to the NOMAD Repository and Archive [6]. This system is intended to ensure compliance with FAIR principles while making data openly accessible to the broader materials science community.

Architecture, implementation, and governance details of this workflow are presented in the next chapter.

Chapter 3

System Architecture

The system architecture developed for the FAIR-by-design data management platform at LAME is designed to support the complete research data lifecycle, from data acquisition to long-term preservation. Its primary purpose is to enable experimentalists to upload structured data and metadata through a web-based interface, while ensuring that the data are securely stored and made available in compliance with FAIR principles [1].

The platform is built using modern web technologies, with the Django web framework at its core. Django provides a scalable backend, secure user management, and integration with both object storage and relational databases. The goal is to provide an infrastructure that is accessible to researchers while maintaining institutional standards for data quality and interoperability.

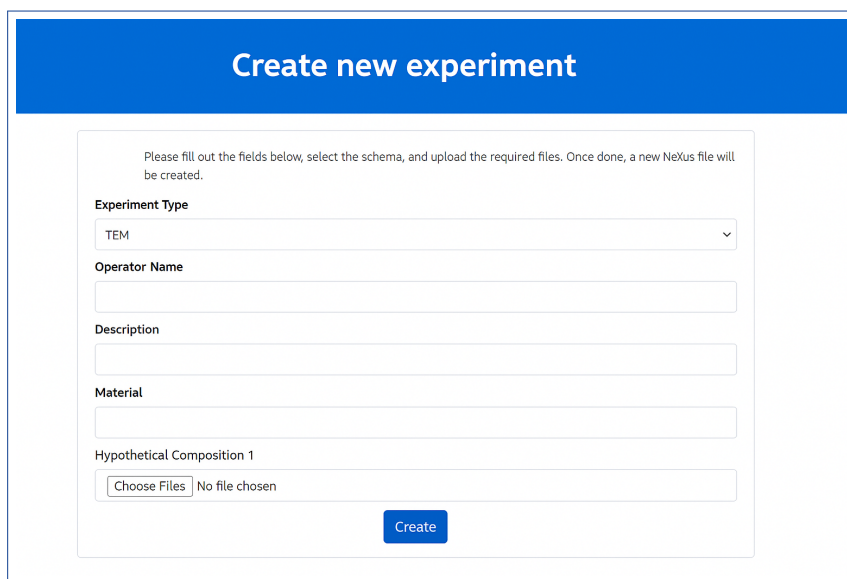
Initially, integration with an existing electronic lab notebook system, such as eLabFTW [9], was considered. However, after testing, it was determined that the interface was not sufficiently intuitive for daily use by experimentalists. As a result, a custom application was developed to better align with existing workflows and usability needs (see Figure 3.1).

3.1 Django Interface Design

Django was selected due to its built-in security features, extensibility, and compatibility with relational databases and RESTful APIs. The web application acts as the primary interface between researchers and the data infrastructure.

3.1.1 Web API for Data Upload

The platform exposes a REST API that supports standard HTTP methods:



Create new experiment

Please fill out the fields below, select the schema, and upload the required files. Once done, a new NeXus file will be created.

Experiment Type
 TEM

Operator Name

Description

Material

Hypothetical Composition 1
 No file chosen

Figure 3.1: Web interface for creating a new experiment. Users can input key metadata such as experiment type, operator name, sample description, and upload associated data files.

- **POST:** Enables users to upload structured metadata and data files via form submission or API endpoint. Multiple formats are accepted.
- **GET:** Allows retrieval of datasets based on unique identifiers or metadata attributes.
- **DELETE:** Permits removal of deprecated or erroneous entries to maintain data integrity.

3.1.2 Data Flow

Authentication is managed through Authentik, an open-source identity provider deployed within the ORFEO infrastructure at Area Science Park. After single sign-on, Django assigns user permissions and roles.

Uploaded data and metadata are validated upon submission. Raw files (including NeXus and microscopy images) are stored in MinIO—an S3-compatible object storage system hosted on ORFEO. Structured metadata is recorded in a PostgreSQL database.

Validated datasets are then registered in ORFEO’s data center for versioned archival and secure access. Approved datasets are transferred to NOMAD Oasis, an internal FAIR repository hosted at ORFEO and aligned with the NFFA-DI framework [6, 10].

Where appropriate, data may also be published to external repositories such as:

- **OFED (Overarching FAIR Ecosystem for Data):** A cross-institutional data-sharing platform developed by NFFA-DI [10].
- **NOMAD Repository:** A public repository supporting FAIR-compliant publication and reuse in materials science [11].

Figure 3.2 provides an overview of the system.



Figure 3.2: System architecture overview: User access and authentication via Authentik; data transfer to MinIO; metadata indexing in PostgreSQL; registration in ORFEO; publication to NOMAD Oasis and external repositories.

3.2 Backend Architecture

The backend processes each uploaded dataset by assembling a NeXus file from instrument headers (e.g., .hdr files), raw data, and user-provided metadata. After validation, the NeXus package is streamed to MinIO storage.

3.2.1 Database Schema

Metadata is stored in a PostgreSQL database managed by Django. The schema is lightweight and includes:

- **Experiment:** High-level metadata per acquisition, including instrument settings, sample details, timestamps, and operator information.
- **User:** Based on Django’s default model, with added fields for roles and access permissions.

During upload, schema validation and checksum verification are performed to ensure data quality. Routine backups of the database and MinIO buckets are scheduled to ensure data resilience and traceability.

3.2.2 Integration with MinIO and NOMAD Oasis

Once uploaded, raw data and associated metadata are stored in MinIO. After validation, datasets are registered in ORFEO, the data center supporting structured storage, access control, and internal data workflows.

Validated datasets are then published to NOMAD Oasis, a FAIR repository developed within the NOMAD and FAIRmat projects [6]. Simultaneously, data is included in OFED, the internal data ecosystem supporting institutional compliance with FAIR data standards across NFFA-DI facilities.

3.3 User Roles and Workflow

Authentication and user-role management are handled via Authentik. On successful login, Django assigns the corresponding permissions.

3.3.1 Defined Roles

- **Experimentalists:** Upload data and metadata via the web interface.
- **Data Managers:** Review and validate metadata, correct inconsistencies, and curate datasets.

- **Supervisors:** Oversee system configuration, user access, and publication approvals.

3.3.2 Workflow Summary

1. **Authentication:** Users authenticate through Authentik SSO, with roles managed in Django.
2. **Upload:** Experimentalists upload raw data and metadata; NeXus files are generated.
3. **Validation:** The system runs schema checks and metadata consistency validation; Data Managers may perform manual corrections.
4. **Archival:** Validated data is registered in ORFEO and prepared for long-term storage in OFED.
5. **Publication:** Datasets are published to NOMAD Repository and made accessible through FAIR-compliant interfaces.
6. **Access:** Authorized users may retrieve and analyze data through the web interface or external repositories.

This architecture supports secure, scalable, and reproducible data management aligned with institutional and European research infrastructure policies [1, 6].

A public release of the pipeline is planned through ORFEO and GitHub, in alignment with open science practices and the ongoing digitalization efforts at Area Science Park.

Chapter 4

Metadata Modeling and NeXus File Generation

Structured metadata is essential for ensuring that experimental data can be understood, reused, and preserved over time. In electron microscopy, particularly in SEM and FIB-SEM workflows, metadata must capture details about the sample, instrument configuration, and measurement conditions in a consistent and machine-readable format. This chapter outlines the metadata schema developed for the LAME platform, its mapping to NeXus standards, and the procedure for generating NeXus files from instrument output.

4.1 Metadata Schema for SEM and FIB-SEM

The metadata schema is designed to capture key information before and during measurement. It combines guidelines from the FAIRmat initiative [12] with the NeXus standard for scientific data [13]. The goal is to enable interoperability and long-term accessibility in line with FAIR data principles [1].

4.1.1 Pre-measurement Metadata Categories

Pre-measurement metadata includes information required to contextualize the experiment:

- **Sample Information:**
 - Sample ID, composition, and preparation method.
- **Instrument Configuration:**
 - Instrument type, manufacturer, model, and detector setup.

- **Measurement Conditions:**

- Imaging mode, date/time, and relevant environmental parameters.

- **User Information:**

- Operator ID, institutional affiliation, and role.

4.2 Ontology Alignment and NeXus Mapping

To ensure semantic consistency, metadata terms are aligned with ontologies endorsed by the FAIRmat project and the NeXus standard. The NeXus format organizes experimental data into hierarchical groups using defined base classes, with NXentry as the root.

4.2.1 Key NeXus Classes

The following NeXus groups are used:

- **NXentry:** Top-level container for metadata and data.
- **NXinstrument:** Instrument model, manufacturer, configuration.
- **NXsample:** Sample composition, dimensions, and preparation.
- **NXuser:** Responsible researcher or operator.
- **NXprogram:** Software used during data acquisition.

Optional but recommended classes include:

- **NXdetector:** Specific detector details.
- **NXbeam:** Beam settings (e.g., accelerating voltage).
- **NXprocess:** Data processing steps.

4.2.2 TESCAN AMBER X Metadata Mapping

Metadata exported from the TESCAN AMBER X FIB-SEM is translated into the NeXus structure using a mapping developed by the Scientific Computing Center at KIT [14]. Selected mappings include:

- `entry/endTime` \leftarrow Measurement date and time
- `entry/user/userName` \leftarrow Operator name
- `entry/instrument/modelName` \leftarrow Instrument model
- `entry/instrument/eBeamSource/accelerationVoltage` \leftarrow Beam voltage
- `entry/instrument/stage/coordinates` \leftarrow Stage position
- `entry/instrument/detectors/detectorName` \leftarrow Active detectors

This standardized approach supports automated file generation and compatibility with FAIR-compliant infrastructures.

4.3 Validation Procedures

To ensure metadata quality, several validation steps are implemented:

- **Schema Validation:** Confirms structure against NeXus definitions.
- **Ontology Matching:** Ensures metadata terms are semantically consistent.
- **User Input Checks:** Mandatory fields and format constraints are verified at upload.

4.3.1 Validation via NOMAD Cloud

The NOMAD Cloud environment allows researchers to test and validate NeXus files [6]. Uploaded files undergo:

- **Metadata Compliance Checks:** Against domain ontologies.
- **Visualization:** To inspect metadata hierarchy.
- **Error Reporting:** Feedback for resolving issues before publication.

4.4 NeXus File Generation

The final NeXus files are generated by combining raw image data, user inputs, and pre-collected metadata into a single HDF5 file. A simplified tree structure is shown below.

Example NeXus File Structure

```
entry:NXentry
|   experiment_identifier = "SEM_Exp_001"
|   start_time = "2024-03-15T10:00:00"
|
+-- instrument:NXinstrument
|   +-- name = "TESCAN Amber X"
|   +-- manufacturer = "TESCAN"
|   +-- acceleration_voltage = 30.0 (kV)
|
+-- sample:NXsample
|   +-- sample_name = "Alloy_123"
|   +-- composition = "Fe-Ni-Cr Alloy"
|
+-- detector_SED:NXdetector
|   +-- type = "Secondary Electron Detector"
|
+-- data:NXdata
|   +-- image_data (2048x2048 array)
|   +-- magnification = 5000
|
+-- user:NXuser
|   +-- name = "Operator"
|   +-- affiliation = "LAME, Area Science Park"
```

4.4.1 Remarks

The NeXus format supports the structured organization of datasets, including optional extensions for detectors, stages, beam settings, and processing pipelines. The hierarchical model ensures compatibility with FAIRmat standards and facilitates automated ingestion into platforms like NOMAD and OFED.

The metadata schema and NeXus export pipeline developed for SEM and FIB-SEM data at LAME provide a structured, standards-based approach to research

data management. By integrating FAIRmat ontology recommendations and validating outputs through the NOMAD Cloud, the system supports reliable documentation, interoperability, and long-term accessibility of microscopy datasets.

Chapter 5

Data Pipeline, Sharing, and Evaluation

5.1 Overview

A structured data pipeline is essential for ensuring that scientific data is securely stored, validated, and made accessible for reuse. At the Laboratory for Advanced Microscopy and Electron Microscopy (LAME), based at Area Science Park in Trieste, a data management system has been developed to support the full research data lifecycle—aligned with the FAIR principles [1]. This chapter outlines the implementation of that pipeline, evaluates its current limitations, and identifies areas for future improvement.

5.2 Data Handling and Storage Workflow

5.2.1 Data Transfer and Cleaning

Raw data generated by electron microscopy instruments at LAME is automatically transferred to the ORFEO data center infrastructure. This avoids reliance on local storage and reduces the risk of data loss during acquisition. Upon arrival at ORFEO, experimentalists conduct an initial review to remove incomplete or irrelevant datasets and ensure accuracy. Cleaning also includes optional anonymization where needed.

5.2.2 Structured Storage via MinIO

Validated datasets are stored in MinIO, an S3-compatible object storage system deployed within ORFEO. Metadata are recorded separately in a PostgreSQL

database managed by the web application. This architecture ensures separation of content and metadata while preserving access performance.

5.3 Standardization and Interoperability

To ensure interoperability, LAME follows naming conventions and data structure standards set by the NFFA-DI project. Dataset identifiers include project codes, instrument labels, and timestamps, facilitating traceability and integration with external systems such as OFED (Overarching FAIR Ecosystem for Data). This approach supports consistency and aligns with broader NFFA-Europe guidelines [10].

5.4 Repository Publishing

Validated datasets can be published through two FAIR-compliant routes:

- **OFED:** Internal datasets are enriched with metadata, assigned persistent identifiers (e.g., DOIs), and shared within the NFFA-DI infrastructure. Access permissions are managed according to institutional policies.
- **NOMAD Repository:** Datasets intended for open publication are uploaded to the NOMAD Repository, which provides visualization and analysis tools and promotes reuse across the wider materials science community [6].

Both publishing options support FAIR principles and long-term data accessibility.

5.5 System Evaluation

5.5.1 Current Limitations

Despite its functionality, the platform faces several constraints:

- **Format Diversity:** Integration of data from different vendors (e.g., JEOL, Tescan) requires pre-processing and custom mappings to the NeXus standard.
- **Metadata Inconsistencies:** Although form-based inputs and validation reduce errors, metadata entry still relies on manual input, leading to occasional omissions.

- **External Dependencies:** Data publication depends on external infrastructures such as NOMAD and OFED, which may change their protocols over time.

5.5.2 Planned Improvements

To address these challenges, the following developments are under consideration:

- **Automation:** Use of AI tools to assist in metadata extraction and data validation.
- **User Interface Enhancements:** Improved guidance during data entry to reduce incomplete submissions.
- **Repository Integration:** Expanded support for additional FAIR repositories beyond OFED and NOMAD.
- **Preservation Strategy:** Enhancement of archival workflows within ORFEO to ensure long-term data retention.

5.6 Impact and Outlook

The platform strengthens data quality and reproducibility within LAME's workflows by standardizing metadata and supporting FAIR-compliant publication. Its compatibility with national (OFED) and international (NOMAD) infrastructures promotes institutional alignment with open science policies. Looking forward, the platform can be expanded to support additional microscopy modalities and integrated with broader European research data infrastructures.

5.7 Conclusion

This work presents a data management solution tailored for SEM and FIB-SEM workflows at LAME. The main contributions include:

- A Django-based web interface for structured metadata input.
- Automated validation and data cleaning mechanisms.
- Integration with ORFEO and MinIO for secure storage.
- Use of the NeXus format guided by FAIRmat recommendations.

- Repository publishing via OFED and NOMAD for long-term accessibility.

By implementing a FAIR-by-design approach, the platform enhances data organization, sharing, and preservation. Ongoing development will focus on automation and wider interoperability, supporting LAME's contribution to transparent and collaborative research in electron microscopy.

References

Bibliography

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The fair guiding principles for scientific data management and stewardship,” *Scientific data* **3** (2016) 160018.
- [2] B. Mons, C. Neylon, J. Velterop, *et al.*, “Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud,” *Information Services & Use* **37** no. 1, (2017) 49–56.
- [3] GO FAIR, “Fair principles,” Accessed: 2024-04-04.
<https://www.go-fair.org/fair-principles/>.
- [4] The Turing Way Community, “The fair principles,” Accessed: 2024-04-04.
<https://book.the-turing-way.org/reproducible-research/rdm/rdm-fair>.
- [5] G. Bazzocchi *et al.*, “Ontology-driven metadata schema for electron microscopy data,” *Journal name here* (2022) . To be updated with correct journal and DOI if available.
- [6] C. Draxl and M. Scheffler, “Fairmat: Making materials science data fair,” *npj Computational Materials* **8** no. 1, (2022) 178.
- [7] E. Giglia, “Implementing fair data: The role of institutions and researchers,” *JLIS.it* **10** no. 2, (2019) 54–68.
- [8] “Laboratory of electron microscopy (lame) at area science park.”
<https://www.areasciencepark.it/en/infrastructure/laboratory-of-electron-microscopy-lame/>, 2022. Accessed: 2025-04-27.
- [9] N. Comte, “elabftw: Open source electronic lab notebook.”
<https://www.elabftw.net>, 2023.
- [10] NFFA-Europe Initiative, “Nffa-di project overview.”
<https://www.nffa.eu/projects/nffa-di/>, 2023.

- [11] The FAIRmat Team, “Nomad repository.” <https://nomad-lab.eu>, 2024.
- [12] S. Botti, C. Draxl, A. Heuer, J. Houska, P. Rinke, , and the FAIRmat team, “Fairmat metadata management concepts,” 2024.
<https://doi.org/10.5281/zenodo.15005550>.
- [13] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, B. Clausen, S. Cottrell, *et al.*, “The nexus data format,” 2015.
<https://www.nexusformat.org>.
- [14] E. G. G. Vitali, R. E. Joseph, and R. Aversa, “Metadata extraction tool and schema mapper for scanning electron microscopy (sem) images.”
https://github.com/kit-data-manager/tomo_mapper, 2023.