



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](#)



Author's Declaration

I, Elaheh Saadar, declare that this thesis entitled, 'mplementation of a pipeline for collecting, ingesting and transforming data into standard formats for the LAME 1 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:

"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. Their advice has been invaluable both for this thesis and my growth as a researcher.

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

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 in research practice.

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	Introduction	9
1.1	Background	9
1.2	Problem Statement	9
1.3	Objective	10
1.4	Thesis Structure	10
2	FAIR Principles and Data Management	12
2.1	Overview of FAIR Principles	12
2.1.1	Findable	12
2.1.2	Accessible	12
2.1.3	Interoperable	13
2.1.4	Reusable	13
2.2	Relevance in Experimental Laboratories	13
2.3	Ontologies and Metadata Standards	14
2.3.1	Ontologies	14
2.3.2	Metadata Standards	14
2.3.3	Implementation in Research Data Management	14
3	LAME Lab and Workflow	15
3.1	Instrumentation and Software	15
3.2	Data Types and Metadata Categories	16
3.3	Challenges in Current Practices	16
3.4	Planned Data-Transfer and Publication Workflow	17
4	System Architecture	18
4.1	Django Interface Design	18
4.1.1	Web API for Data Upload	19
4.1.2	Data Flow	19
4.2	Backend	20
4.2.1	Database Structure	20

4.2.2	Integration with MinIO and NOMAD Oasis on ORFEO . . .	21
4.3	User Roles and Workflow	22
4.3.1	User Roles	22
4.3.2	Workflow	22
5	Metadata Schema Development	24
5.1	Pre-measurement Metadata	24
5.1.1	Core Elements of Pre-measurement Metadata	24
5.2	Ontology Mapping (SEM and FIB-SEM)	25
5.2.1	FAIRmat and NeXus Ontology Integration	25
5.2.2	Required NeXus Classes for Electron Microscopy	26
5.2.3	Metadata to Ontology Mapping for TESCAN AMBER X	26
5.3	Validation and Standardization	27
5.3.1	Validation Procedures	27
5.3.2	Using NOMAD Cloud for Validation	28
5.4	Summary	28
6	NeXus File Generation	29
6.1	Overview	29
6.2	NeXus Format Overview	29
6.3	Case Example: TESCAN Amber X	30
6.3.1	Discussion of the Case Example	32
6.4	Wrapping Up	32
7	Data Pipeline and Sharing	33
7.1	Cleaning and Storage	33
7.1.1	Initial Data Transfer to ORFEO	33
7.1.2	Data Cleaning and Selection	33
7.1.3	Storage in MinIO on ORFEO	34
7.2	Data Lake Integration	34
7.2.1	Naming Conventions and Standardization	34
7.2.2	Integration with NFFA-Europe Standards	34
7.3	FAIR Repository Publishing	35
7.4	Conclusion	36
8	Evaluation and Discussion	37
8.1	Overview	37
8.2	Current Challenges	37
8.3	Future Enhancements	38
8.4	Scientific Impact	38
8.5	Summary	38

9 Conclusion	39
9.1 Summary of Contributions	39
9.2 Challenges and Approaches	39
9.3 Scientific Relevance and Future Directions	40
9.4 Final Remarks	40
References	41

Chapter 1

Introduction

1.1 Background

Scientific research has entered a new era, one marked by an explosion in data generation, largely thanks to cutting-edge laboratory instruments and powerful computational tools. A prime example of this shift is the Laboratory for Advanced Microscopy and Electron Microscopy (LAME), located in the prestigious Area Science Park in Trieste, Italy. This facility is at the forefront of technology, equipped with advanced instruments like the JEOL F200 TEM/STEM microscope with a cold Field Emission Gun (FEG), the JEOL Grand Arm 300 KV TEM/STEM microscope, and the Microscope Plasma FIB-SEM Tescan Amber X. These instruments produce highly detailed datasets that call for well-structured and efficient data management strategies.

At the same time, there's growing momentum within the global scientific community to embrace the FAIR principles—ensuring that data is Findable, Accessible, Interoperable, and Reusable. With that in mind, effective research data management (RDM) has become more critical than ever. It involves using structured metadata, standardized formats, and interoperability protocols that support robust data sharing and reproducibility.

1.2 Problem Statement

Even with the major advancements in electron microscopy, laboratories such as LAME still face real challenges when it comes to managing their data. Typically, electron microscopy data and metadata are captured using proprietary software, such as TESCAN's Essence and Oxford Instruments' Aztec, for instance, in the case of the Microscope Plasma FIB-SEM Tescan Amber X. Unfortunately, this often leads to data silos and a lack of interoperability between systems.

On top of that, documentation practices are frequently manual and inconsistent. This can result in incomplete metadata and diminished data quality, making it harder to fully apply the FAIR principles. In turn, that affects everything from reproducibility to data sharing, ultimately scientific progress.

1.3 Objective

The central objective of this thesis is to design and implement a FAIR-by-design data management platform specifically tailored for electron microscopy research at LAME. This platform is intended to tackle the current limitations by:

- Developing a comprehensive, ontology-based metadata schema to promote consistent documentation and improved interoperability.
- Creating a user-friendly, Django-based web API to enable standardized metadata and data collection directly from experimentalists.
- Automating data integration and cleaning workflows to ensure structured and reliable datasets.
- Implementing automatic generation of standardized NeXus files to support streamlined data sharing and long-term preservation in FAIR-compliant repositories and data lakes.

Through these efforts, the platform aims to significantly enhance data quality, consistency, and accessibility, aligning electron microscopy research at LAME with FAIR standards.

1.4 Thesis Structure

To effectively present the development and evaluation of the FAIR-by-design data management platform, this thesis is organized into the following chapters:

- **Chapter 2: FAIR Principles and Research Data Management**
Examines the relevance and application of FAIR principles, particularly within electron microscopy and related research domains, providing the theoretical foundation for the project.
- **Chapter 3: LAME Lab and Workflow**
Offers a detailed overview of the LAME facility, including its electron microscopy instruments, typical workflows, and existing challenges in data management. It also situates the lab within the broader context of the Area Science Park in Trieste.

- **Chapter 4: System Architecture**
Describes the architecture of the developed platform, covering the Django web API, database structure, user management, and the integration and automation of NeXus file creation.
- **Chapter 5: Metadata Schema Development**
Focuses on the design and implementation of an ontology-driven metadata schema, explaining the rationale behind the chosen ontologies, schema structure, and its practical use in capturing experimental metadata.
- **Chapter 6: NeXus File Generation**
Details the methods and tools used to automatically generate NeXus files from collected data and metadata, ensuring compliance with standardized data packaging formats.
- **Chapter 7: Data Pipeline and Sharing**
Describes the integrated data pipeline, including data collection, metadata cleaning and validation, and the procedures for sharing final datasets through data lakes and established repositories.
- **Chapter 8: Evaluation and Discussion**
Evaluates the platform's performance, user feedback, and effectiveness in supporting FAIR compliance. This chapter also discusses encountered challenges and proposes future improvements.
- **Chapter 9: Conclusion**
Summarizes the key contributions and impact of this thesis, emphasizing its role in improving the efficiency, accessibility, and interoperability of research data management in electron microscopy at LAME.

In summary, this introductory chapter has outlined the background, defined the core research problem, set clear objectives, and provided a roadmap for the chapters ahead, laying the groundwork for the detailed exploration that follows.

Chapter 2

FAIR Principles and Data Management

2.1 Overview of FAIR Principles

In today’s rapidly evolving scientific landscape, effective data management and stewardship have become more important than ever. The FAIR principles—Findable, Accessible, Interoperable, and Reusable—were established to enhance the usability and value of research data. These guidelines aim to ensure that data can be efficiently shared and reused across various disciplines and platforms, thereby fostering greater collaboration and innovation in the scientific community [1].

2.1.1 Findable

For data to be truly useful, it first needs to be easy to locate. Assigning globally unique and persistent identifiers, together with rich metadata descriptions, ensures that datasets can be easily discovered by both humans and machines. Storing data in searchable repositories further boosts their findability [2].

2.1.2 Accessible

Once data has been found, clear protocols should define how it can be accessed. Using standardized communication protocols makes data retrieval straightforward and reliable. Importantly, even when access to the data itself is restricted, the metadata should always remain openly available to let potential users know about the dataset’s existence and relevance [2].

2.1.3 Interoperable

To enable integration with other datasets and tools, data should follow recognized formats and make use of standardized vocabularies. This helps ensure that information from different sources can be combined and analyzed smoothly, fostering more collaborative and interdisciplinary research [2].

2.1.4 Reusable

To fully unlock the potential of research data, it is essential to clearly document its origin, the methods used to generate it, and any applicable licensing. Providing comprehensive metadata and following community standards ensures that datasets can be confidently reused in future studies [2].

2.2 Relevance in Experimental Laboratories

Experimental laboratories generate large amounts of data every day. Applying the FAIR principles in these environments brings a range of important benefits:

- **Enhanced Collaboration:** Using standardized data formats and rich metadata makes it much easier for researchers to share and combine their data, helping to support collaborative projects across different teams and disciplines.
- **Improved Data Integrity:** Clear documentation and consistent protocols help reduce mistakes and inconsistencies, making experiments easier to replicate and results more reliable.
- **Efficient Resource Utilization:** Well-organized and reusable data cuts down the need to repeat experiments unnecessarily, saving both time and valuable resources.
- **Compliance and Funding:** With more funding agencies and journals requiring research data to follow FAIR principles, meeting these standards is increasingly important for securing grants and getting published.

By integrating FAIR practices into everyday workflows, experimental labs can significantly boost the quality, visibility, and overall impact of their research outputs [3].

2.3 Ontologies and Metadata Standards

A key element of the FAIR principles is the use of standardized ontologies and metadata schema. These tools offer a structured way to describe data, helping to maintain consistency and ensure interoperability across different systems and studies.

2.3.1 Ontologies

Ontologies provide a shared vocabulary to describe entities and their relationships within a specific domain that researchers can use. By adopting common ontologies, data can be annotated in a way that is universally understood, making it easier to integrate information across different studies and fields [4, 5].

2.3.2 Metadata Standards

Metadata captures essential information about data, such as its origin, format, and the conditions under which it was collected. Using discipline-specific metadata standards ensures that datasets are consistently described, which improves their discoverability and usability. For example, in materials science, platforms like NOMAD provide comprehensive metadata schemas and guidelines for organizing and sharing experimental and computational materials data [6].

2.3.3 Implementation in Research Data Management

Successfully incorporating ontologies and metadata standards into research data management involves:

- **Selection of Appropriate Standards:** Choosing standards that best fit the specific needs and practices of the research domain.
- **Training and Awareness:** Educating researchers about the importance of these standards and how to apply them effectively.
- **Tool Development:** Using or creating tools that make it easier to annotate and manage data according to the selected standards.
- **Continuous Evaluation:** Regularly reviewing and updating data practices to stay aligned with evolving standards and best practices.

By embedding ontologies and metadata standards into everyday research workflows, institutions can ensure their data not only meets FAIR principles but is also ready to support future innovations and collaborations [7].

Chapter 3

LAME Lab and Workflow

3.1 Instrumentation and Software

Established in 2022, the Laboratory of Electron Microscopy (LAME) at Area Science Park in Trieste, Italy, is dedicated to the advanced characterization of materials, with a particular focus on nanoscience and nanotechnology [8]. The laboratory hosts a suite of state-of-the-art instruments, including:

- **JEOL F200 TEM/STEM Microscope:** Equipped with 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. The system is also optimized for tomographic analysis and in situ experiments.
- **JEOL Grand ARM 300 KV TEM/STEM Microscope:** Featuring both image and probe aberration correction, this microscope achieves sub-angstrom resolutions. It is equipped with dual EDS detectors, a CEOS CEFID EELS spectrometer, a DECTRIS ELA hybrid pixelated camera for energy-filtered 4D STEM, and an electron dose modulator, enabling highly precise and sensitive imaging.
- **Plasma FIB-SEM Tescan Amber X Microscope:** Designed for in situ TEM lamella preparation and advanced materials characterization under low vacuum conditions. It includes an EDS detector and is optimized for "slice and view" applications, as well as for controlled atmosphere sample transfer, supporting complex sample preparation workflows.

These instruments are complemented by specialized software packages that facilitate advanced data acquisition, processing, and analysis, enabling researchers to conduct comprehensive studies of the structural, electronic, and chemical properties of materials.

3.2 Data Types and Metadata Categories

Research activities at LAME produce a wide variety of data types, each requiring careful documentation and management:

- **High-Resolution Imaging Data:** Obtained through TEM, STEM, and SEM techniques, these images reveal atomic- and molecular-scale structures.
- **Spectroscopic Data:** Using techniques such as EDS and EELS, researchers gather detailed information about elemental compositions and electronic structures for understanding material behavior.
- **Tomographic Data:** Three-dimensional reconstructions created via FIB-SEM offer comprehensive insights into complex material geometries.
- **In Situ Experimental Data:** Data collected under varying environmental conditions (e.g., temperature, pressure) helps study how materials behave in real-world applications.

To make sure the data remains usable and reproducible, LAME is supposed to follow standardized metadata practices, which include:

- **Instrument Settings:** Thorough records of microscope parameters, such as voltage settings, magnification, and detector types.
- **Sample Information:** Descriptions of sample sources, preparation methods, and physical characteristics.
- **Data Processing Details:** Clear reporting of the software used for data analysis to ensure transparency and reproducibility.

3.3 Challenges in Current Practices

Despite its advanced capabilities, LAME faces some data management and workflow challenges:

- **Data Volume and Complexity:** The high-resolution and multidimensional data produced require significant storage resources and efficient processing pipelines.
- **Metadata Consistency:** Maintaining consistent metadata documentation across varied experiments remains crucial for data interoperability and reuse.

- **Integration with External Collaborators:** Seamless data sharing with national and international partners demands common standards and robust data exchange mechanisms.
- **Compliance with FAIR Principles:** Aligning data management with FAIR (Findable, Accessible, Interoperable, Reusable) principles requires continuous evaluation and adaptation of workflows.

3.4 Planned Data-Transfer and Publication Workflow

Ongoing efforts at LAME aim to tackle these challenges by implementing data management system and standardizing protocols, ensuring that the laboratory continues to lead in the field of electron microscopy research [8].

To ensure that the rich datasets produced at LAME travel seamlessly along a FAIR-compliant pipeline, we are developing a web application that automatically moves raw and processed data from the laboratory's acquisition workstations to a MinIO object-storage instance hosted on the ORFEO computing infrastructure at Area Science Park. A dedicated micro-service layer will then harvest validated datasets and their metadata from MinIO and publish them to the NOMAD Repository and Archive, making the results openly available to the wider materials-science community.

The architecture, implementation details, and governance of this end-to-end workflow are described in the next chapter.

Chapter 4

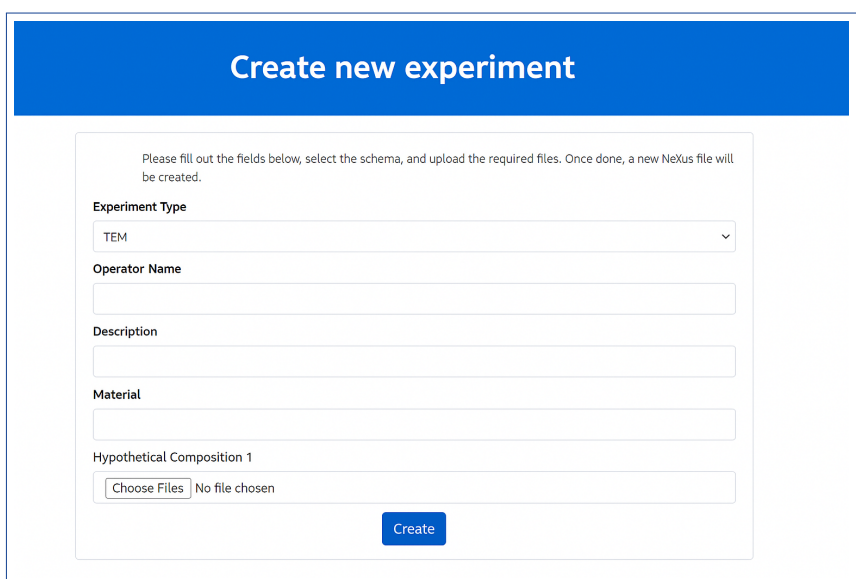
System Architecture

The system architecture developed for the FAIR-by-design data management platform at LAME is built to simplify the entire data handling process, from acquisition all the way to storage. The main goal is to make it easy for experimentalists to upload their data and metadata through an intuitive web API, ensuring everything is properly structured and securely stored in a dedicated data lake infrastructure. To achieve this, the platform takes advantage of modern web technologies, with Django at its core, providing a robust, scalable, and efficient foundation for managing research data.

It is important to note that initially, the plan was to integrate an existing solution like eLab-FTW , a free, open-source electronic lab notebook designed for research data management, into the workflow. However, after testing, it became clear that for many experimentalists, the system was too complex and not intuitive enough for daily use. As a result, we decided to develop a custom web application, specifically designed to better fit the laboratory's workflows and simplify the user experience. One can see the picture of the web interface in Figure 4.1.

4.1 Django Interface Design

Django, a high-level Python web framework, was chosen because of its flexibility, strong security features, and its ability to easily connect with databases and RESTful APIs. The web interface built with Django acts as the main bridge between the experimentalists and the data management system, providing a simple and reliable way for them to interact with their data.



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 4.1: Web interface for creating a new experiment. Users can input essential metadata such as experiment type, operator name, description, material details, and upload the corresponding data files.

4.1.1 Web API for Data Upload

The Django-based web API is built to make uploading both data and metadata simple and efficient. It supports standard HTTP methods like POST, GET, and DELETE, letting users easily add, retrieve, or remove datasets.

- **POST Requests:** Used to upload data and metadata files through a straightforward form or API endpoint. The API accepts different file formats to provide maximum flexibility for experimentalists.
- **GET Requests:** Allow users to retrieve uploaded datasets by searching with unique identifiers or specific metadata keywords.
- **DELETE Requests:** Enable users to remove outdated files or entries, helping keep the database clean and up to date.

4.1.2 Data Flow

Researchers access the system through Authentik, an open-source identity provider for authentication and authorization, which is deployed within the ORFEO infrastructure at Area Science Park. Once authenticated via Authentik’s single sign-on

(SSO), Django takes over, assigning the appropriate roles and permissions in the background.

Data uploaded through the API is immediately validated for completeness and consistency. The raw files, typically NeXus metadata and microscopy images, are first stored in MinIO, an S3-compatible object storage system hosted on ORFEO. In parallel, structured metadata is recorded in a relational database to support indexing and retrieval.

From MinIO, datasets are automatically registered within ORFEO's data center, which manages long-term storage, versioning, and secure sharing. After a final review step, the data is transferred to NOMAD Oasis, a FAIR-compliant internal repository also hosted on ORFEO and used within the NFFA-DI framework.

To maximize impact and ensure open access, datasets can then be published on two external FAIR repositories. The first is OFED (Overarching FAIR Ecosystem for Data), the central data-sharing platform developed by NFFA-DI to support standardized, interoperable access across partner institutions. The second is the public NOMAD Repository, a widely recognized platform in the materials science community offering tools for data analysis and reuse.

The complete data workflow is illustrated in Figure 4.2.

4.2 Backend

The web application converts every acquisition into a single NeXus file, assembled on-the-fly from the instrument's .hdr (or other text) header, the raw image stack, and any additional run notes. Once the NeXus package passes validation, the app streams it directly to MinIO object storage on the ORFEO cluster.

4.2.1 Database Structure

Because all binary data live in MinIO, the only database we need is the PostgreSQL instance managed by Django. Its schema is deliberately lightweight and centres on two tables:

- **Experiment** – high-level metadata for each run (instrument settings, sample details, operator, time-stamp) in the form of Nexus files.
- **User** – Django's built-in user table augmented with roles that control upload and publication rights.

During upload, the application performs strict consistency checks (e.g. schema conformity, checksum verification). Regular automated backups of the PostgreSQL database and MinIO buckets ensure data safety and long-term traceability.



Figure 4.2: Architecture of the LAME data management system: from user authentication to data storage in MinIO, registration in ORFEO, and publication to NOMAD Oasis. The system components are represented as interlocking LEGO blocks to emphasize modularity and integration.

4.2.2 Integration with MinIO and NOMAD Oasis on ORFEO

When researchers upload data, it is first stored in MinIO—an S3-compatible object storage system deployed within ORFEO, the central data center at Area Science Park. MinIO handles raw files and NeXus metadata immediately after acquisition, allowing fast access and internal validation.

Once the data passes initial checks, it is registered within the ORFEO infrastructure for long-term storage, structured organization, and secure access. ORFEO serves as the backbone for versioning, collaboration, and compliance with FAIR data principles.

Validated datasets are then published to NOMAD Oasis, a FAIR-compliant repository also hosted within ORFEO, providing access to the broader NFFA-DI research community. In parallel, the data and its metadata become part of OFED (Overarching FAIR Ecosystem for Data), a core digital ecosystem within ORFEO that supports standardized data access, traceability, and interoperability across NFFA-DI infrastructures.

- **MinIO:** Provides high-speed, intermediate storage for uploaded data, enabling efficient internal validation and processing.
- **ORFEO:** Acts as the central data center, supporting structured storage, long-term preservation, and secure sharing.
- **NOMAD Oasis:** Hosts approved datasets for FAIR-compliant access within the NFFA-DI community.
- **OFED:** A digital ecosystem within ORFEO that enables standardized, FAIR-aligned data handling and integration across institutional platforms.

4.3 User Roles and Workflow

The platform uses Authentik to handle single sign-on, making the login process simple and secure. When users log in, Authentik sends a session cookie back to Django, which then matches the user to their internal profile and automatically applies the correct permissions based on their role.

4.3.1 User Roles

- **Experimentalists** — responsible for uploading raw data and metadata through an easy-to-use interface.
- **Data Managers** — oversee automated validation processes, correct any metadata issues, and curate or remove problematic uploads to maintain data quality.
- **Supervisors** — manage users and roles, and oversee the overall system configuration, including storage and API settings.

4.3.2 Workflow

1. **Authentication** — Users sign in via Authentik, which handles verification and passes a secure cookie to Django. Django then assigns permissions according to the user's role.

2. **Data Upload** — Experimentalists upload NeXus files (or raw images with associated metadata) through the web app. Files are first stored in MinIO.
3. **Validation** — Automated checks verify metadata consistency and file integrity. Data Managers step in to manually review and correct any flagged issues.
4. **Archival to ORFEO** — Once validated, datasets are transferred from MinIO to OFED, Area Science Park’s data-lake platform, for structured long-term storage for data related to NFFA-DI project.
5. **Publication** — Approved datasets are published to NOMAD repository, making them accessible to the wider scientific community and ensuring FAIR compliance.
6. **Access and Retrieval** — Depending on their permissions, users can browse, search, and download datasets either through the web interface, ORFEO, or NOMAD.

The data pipeline developed at LAME represents a comprehensive strategy for managing research data—starting from secure acquisition to long-term sharing. Raw data are directly transferred to ORFEO, the centralized data center at Area Science Park, ensuring efficient and reliable handling of large microscopy datasets. The data are then cleaned and standardized in accordance with NFFA-DI and NFFA-Europe guidelines. Final datasets are published on platforms such as OFED and NOMAD to support transparency, reuse, and international collaboration.

The full pipeline will soon be made available on ORFEO and published on GitHub, reflecting LAME’s commitment to open science and robust, FAIR-aligned data practices.

Chapter 5

Metadata Schema Development

Metadata plays an important role in scientific research by providing the context and background needed to make data understandable, reusable, and reproducible. In electron microscopy, and particularly in SEM and FIB-SEM experiments, well-structured metadata not only helps organize datasets efficiently but also ensures they remain accessible and meaningful over time.

This chapter introduces a standardized metadata schema specifically designed for SEM and FIB-SEM workflows, combining the FAIRmat recommendations with the structure offered by NeXus application definitions [9, 10].

5.1 Pre-measurement Metadata

Pre-measurement metadata records all the important details about the sample, the instrument settings, and the experimental conditions before any data is collected. Capturing this information up front is essential to make sure the data can be accurately understood, trusted, and reused later on.

5.1.1 Core Elements of Pre-measurement Metadata

The pre-measurement metadata schema for SEM and FIB-SEM experiments includes the following key elements:

- **Sample Description:**
 - Sample ID: Unique identifier for the sample.
 - Sample Composition: Chemical and physical description.
 - Sample Preparation: Techniques and treatments applied before measurement.

- **Instrument Configuration:**

- Instrument Type: SEM or FIB-SEM.
- Manufacturer: Tescan.
- Model: Specific instrument model used.
- Detector Configuration: Types of detectors activated (SED, BSED, EDS, ...).
- Additional settings from the instrument's software.

- **Measurement Conditions:**

- Imaging Mode: Mode used (e.g., TEM, STEM, SEM).
- Experiment Date: Date and time of data acquisition
- Environment Conditions, if relevant.

- **User Information:**

- Operator ID: Unique ID of the experimentalist.
- Affiliation: Laboratory or institution..
- Other information.

5.2 Ontology Mapping (SEM and FIB-SEM)

To ensure interoperability and data reuse, metadata from SEM and FIB-SEM experiments must be aligned with established ontologies. In this chapter, we focus specifically on how experimental terms are mapped to community-endorsed ontologies, including those recommended by FAIRmat and supported by the NeXus standard. For readers interested in how this metadata is actually structured into NeXus files, a detailed explanation is provided in the next chapter.

5.2.1 FAIRmat and NeXus Ontology Integration

According to the NeXus FAIRmat proposal (https://github.com/FAIRmat-Experimental/nexus_definitions/blob/fairmat/contributed_definitions/NXem.nxd1.xml), the following base classes are essential:

5.2.2 Required NeXus Classes for Electron Microscopy

In the NeXus data format, all information related to an experiment is organized under a single top-level container, `NXentry`. This acts as the root of the data hierarchy and groups together metadata, raw data, and relevant context for a given session.

The `NXem` application definition specifies a few core classes as required, while others are optional but recommended for richer documentation. Each class can include nested fields or groups (children) to provide additional detail.

- **NXinstrument:** Describes the microscope and its configuration, including the model, vendor, and specific components used during the session.
- **NXsample:** Provides details of the sample, such as composition, thickness, and preparation history.
- **NXuser:** Identifies the researchers involved in the session, along with their contact information and roles.
- **NXprogram:** Records the software used for acquisition or data processing, including versioning.

While not mandatory, additional classes such as `NXdetector`, `NXbeam`, and `NXprocess` are highly encouraged to describe detector setups, beam properties, and data transformations, especially when aiming for FAIR-compliant and reproducible datasets.

A minimal NeXus tree structure looks like this: A minimal NeXus tree structure looks like this:

```
NXentry
|-- NXsample
|-- NXinstrument
|-- NXuser
|-- NXprogram
```

5.2.3 Metadata to Ontology Mapping for TESCAN AMBER X

For FIB-SEM experiments using the TESCAN AMBER X, metadata from the instrument software are translated into NeXus fields to ensure interoperability and adherence to FAIR principles. This mapping, developed by the Scientific

Computing Center (SCC) at Karlsruhe Institute of Technology (KIT) [11], enables a standardized, machine-readable representation of essential experimental details.

A few examples from the mapping schema include:

- **Measurement Date and Time** → entry/endTime/Date, entry/endTime/Time
- **User Name** → entry/user/userName
- **Instrument Model and Vendor** → entry/instrument/instrumentName, entry/instrument/instrumentManufacturer/modelName
- **Accelerating Voltage** → entry/instrument/eBeamSource/accelerationVoltage/value
- **Stage Position (XYZ)** → entry/instrument/stage/coordinates/xValue, yValue, zValue
- **Working Distance** → entry/instrument/stage/eBeamWorkingDistance/value
- **Pixel Size** → entry/instrument/imaging/pixelSize/xPixelSize/value, yPixelSize/value
- **Detector Used** → entry/instrument/detectors/detector1/detectorName

This standardized mapping ensures that data exported from the TESCAN AMBER X is fully compatible with the NXem schema, paving the way for integrated storage, searchability, and sharing in FAIR-aligned infrastructures.

5.3 Validation and Standardization

To keep metadata reliable and reusable, the system performs several validation steps—right from when the data is entered, to the final structure of the generated files.

5.3.1 Validation Procedures

- **Schema Validation:** Automated checks make sure the metadata follows the expected structure, as defined by the NeXus application definition.
- **Ontology Consistency:** Metadata terms are cross-checked against established ontologies to catch any mismatches or inconsistencies.
- **User Input Verification:** When users fill out forms, required fields are checked by application, and entries are validated to match standard formats before anything is submitted.

5.3.2 Using NOMAD Cloud for Validation

As part of the FAIRmat infrastructure, the NOMAD Cloud (also known as the NOMAD Repository) provides a public environment to test and validate metadata. Researchers can upload their structured data to check that it complies with semantic and structural expectations.

- **Metadata Compliance Check:** NOMAD automatically reviews the uploaded files to confirm completeness and alignment with domain ontologies.
- **Visualization:** It also provides a preview of the metadata structure to help users verify that everything has been captured correctly.
- **Error Reporting:** If anything's missing or inconsistent, NOMAD provides detailed feedback so it can be fixed before final submission.

5.4 Summary

By following the FAIRmat NeXus guidelines and tailoring them for SEM and FIB-SEM experiments, we've built a metadata schema that promotes transparency, structure, and reuse. Mapping experiment-specific details to standard ontologies makes the data both understandable for humans and usable by machines. And with validation through NOMAD Cloud, we make sure every dataset meets high-quality standards—ready to be shared, reused, and built upon.

Chapter 6

NeXus File Generation

6.1 Overview

Storing research data in a structured and interoperable way is key to making sure it remains useful and accessible over time. In the field of electron microscopy—especially for techniques like Scanning Electron Microscopy (SEM) and Focused Ion Beam-SEM (FIB-SEM)—the NeXus format has become a widely adopted standard. Its flexibility and rich structure make it a strong match for the FAIR data principles. In this chapter, we’ll walk through what the NeXus format is, explain how data can be exported into it, and share a practical example using data from the TESCAN Amber X FIB-SEM system.

6.2 NeXus Format Overview

NeXus is an open, hierarchical data format built on HDF5, designed to store both experimental data and the metadata needed to fully describe it. It has become a widely adopted standard in many scientific disciplines, including electron microscopy, thanks to its balance of structure and flexibility.

The NeXus format relies on well-defined classes that represent different components of an experiment. The required core classes—`NXentry`, `NXinstrument`, `NXsample`, `NXuser`, and `NXprogram`—were introduced in the previous chapter as essential building blocks for SEM and FIB-SEM metadata organization.

The basic structure of a NeXus file follows a clear hierarchy:

- **NXentry:** The root group that contains the overall experiment description and metadata.
- **NXinstrument:** Information about the measurement device, including its configuration and components.

- **NXsample:** Details about the sample under investigation, such as composition and preparation.
- **NXuser:** Information about the person or team conducting the experiment.
- **NXprogram:** Metadata on the software used during acquisition or processing.

To create a NeXus file, metadata—often originating from instrument software—is mapped into these structured groups and fields. This can be done programmatically using scripts or dedicated mapping tools. The result is a standardized HDF5 file that ensures the data remains well-documented, interoperable, and ready for validation, archiving, or reuse.

6.3 Case Example: TESCAN Amber X

Here we present a practical example using hypothetical data and metadata from the TESCAN Amber X FIB-SEM instrument. The following illustrates the hierarchical structure (tree) of the resulting NeXus file:

NeXus File Structure Example

```
entry:NXentry
|   experiment_identifer = "SEM_Exp_001"
|   start_time = "2024-03-15T10:00:00"
|
+-- instrument:NXinstrument
|   +-- name = "TESCAN Amber X FIB-SEM"
|   +-- manufacturer = "TESCAN"
|   +-- model = "Amber X"
|   +-- electron_source:NXsource
|       +-- type = "Field Emission Gun (FEG)"
|       +-- acceleration_voltage = 30.0 (kV)
|
+-- sample:NXsample
|   +-- sample_name = "Alloy_123"
|   +-- preparation_method = "Polishing and coating"
|   +-- composition = "Fe-Ni-Cr Alloy"
|   +-- temperature = 22.0 (°C)
|
+-- detector_SED:NXdetector
|   +-- type = "Secondary Electron Detector"
|   +-- gain = 15
|
+-- detector_BSED:NXdetector
|   +-- type = "Backscatter Electron Detector"
|   +-- gain = 20
|
+-- data:NXdata
|   +-- image_data (2048x2048 array)
|   +-- exposure_time = 30.0 (s)
|   +-- magnification = 5000
|
+-- user:NXuser
|   +-- name = "Dr. ... "
|   +-- affiliation = "LAME, Area Science Park"
|   +-- ...
```

6.3.1 Discussion of the Case Example

In this structured example:

- The **NXinstrument** section clearly identifies the instrument type, model, and key operational parameters, which are critical for interpreting the results.
- The **NXsample** metadata captures comprehensive information about the sample's composition and preparation, helping ensure the experiment can be understood or replicated.
- Each **NXdetector** is described separately to clarify how the different signals were collected and by which components.
- The **NXdata** group organizes the actual datasets, including relevant acquisition metadata such as magnification and exposure time.
- ...

It's important to note that this is not a complete NeXus file for SEM or FIB-SEM experiments may contain additional groups such as **NXprocess**, **NXuser**, or ... depending on the complexity of the measurement and level of metadata detail. The structure is flexible by design, allowing users to tailor the schema to their instrument and workflow while maintaining compliance with NeXus standards and FAIR principles.

6.4 Wrapping Up

Creating NeXus files for SEM and FIB-SEM data isn't just about following a standard, it is about making sure your work remains accessible, interpretable, and useful to others over time. By using the FAIRmat-supported NeXus application definitions, researchers can structure their data in a way that supports collaboration, reproducibility, and long-term reuse. It's a practical step toward making scientific data more open and meaningful.

Chapter 7

Data Pipeline and Sharing

Managing experimental data effectively is a key part of doing good science today; not just keeping it safe, but making sure it is easy to find, understand, and reuse. At the Laboratory for Advanced Microscopy and Electron Microscopy (LAME) in Area Science Park, Trieste (Italy), we have built a full data pipeline that takes care of everything from transferring and cleaning data to storing and sharing it. The whole system is designed with the FAIR principles in mind, so that the data stays useful, not just now, but also for the future.

7.1 Cleaning and Storage

7.1.1 Initial Data Transfer to ORFEO

Once acquired, raw data from LAME’s advanced microscopy instruments are automatically transferred to ORFEO—the centralized data center infrastructure at Area Science Park. This direct integration allows large datasets to be securely transmitted without relying on temporary local storage, minimizing the risk of data loss or corruption during the transfer process.

7.1.2 Data Cleaning and Selection

Once housed within ORFEO, experimentalists at LAME undertake the critical task of data cleaning and selection. Given the substantial volume of data generated, this process involves:

- **Quality Assessment:** Evaluating data for completeness and accuracy.
- **Relevance Filtering:** Identifying datasets that are pertinent to ongoing research objectives.

- **Anonymization:** Removing any sensitive information to comply with ethical standards.

This meticulous curation ensures that only high-quality, relevant data proceed further in the pipeline.

7.1.3 Storage in MinIO on ORFEO

Post-cleaning, the selected datasets are uploaded to MinIO, an S3-compatible object storage system operating within ORFEO. MinIO offers scalable and high-performance storage solutions, facilitating efficient data retrieval and management. The integration of MinIO within ORFEO ensures that data remain within the secure infrastructure of Area Science Park, benefiting from its robust data protection measures.

7.2 Data Lake Integration

7.2.1 Naming Conventions and Standardization

To ensure consistency and smooth collaboration across the wider research network, LAME, through its involvement in the NFFA-DI (Nano Foundries and Fine Analysis – Digital Infrastructure) project, follows the naming conventions developed by NFFA-DI. These standards help unify how data is labeled and managed across different research facilities, making it easier to integrate and retrieve information. For example, dataset names follow a structured format that includes key metadata like project codes, instrument identifiers, and experiment dates. This approach not only supports internal organization but also aligns with OFED (Overarching FAIR Ecosystem for Data), the digital backbone of NFFA-DI that ensures research data remains Findable, Accessible, Interoperable, and Reusable.

7.2.2 Integration with NFFA-Europe Standards

By aligning with NFFA-Europe’s data management standards, LAME ensures that its datasets are not only well-structured but also interoperable with data from other NFFA-Europe facilities. This standardization promotes easier collaboration, data sharing, and comparative research across the network. These efforts are supported by MetaRepo, NFFA-Europe’s central platform for managing and accessing research data, which plays a similar role to OFED in NFFA-DI—helping to uphold FAIR (Findable, Accessible, Interoperable, and Reusable) data principles across the ecosystem.

7.3 FAIR Repository Publishing

To make research outputs from LAME as accessible and impactful as possible, two main publication routes are available. This flexible approach ensures data is both well-managed within the NFFA-DI infrastructure and shared with the wider scientific community.

Option 1: Publishing through OFED

The first path begins with preparing and cleaning the data, which is then stored in MinIO. From there, it can be shared through the NOMAD Oasis instance available in ORFEO. Once everything is finalized, the dataset is officially published on the Overarching FAIR Ecosystem for Data (OFED).

OFED is NFFA-DI's central platform for publishing datasets in line with FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Here's what the process involves:

- **Metadata Enrichment:** Adding detailed information to help others find and understand the data.
- **Persistent Identifiers:** Assigning DOIs so the data can be reliably cited and accessed over time.
- **Access Controls:** Setting the right balance between open sharing and protecting sensitive information.

Publishing through OFED not only increases the visibility of LAME's work but also supports collaboration and data reuse across the research community.

Option 2: Uploading to the NOMAD Repository

Another route is to publish directly to the public NOMAD Repository, a well-known platform for materials science data that also follows FAIR guidelines. This is a great option for reaching researchers beyond the NFFA-DI network.

NOMAD offers powerful tools for browsing, analyzing, and visualizing data, making it easier for others to explore and build upon LAME's results.

Whether using OFED or NOMAD, the goal is the same: to ensure that high-quality data is available, useful, and ready to inspire new discoveries.

7.4 Conclusion

The data pipeline developed at LAME represents a comprehensive strategy for managing research data, starting from secure acquisition all the way to long-term sharing. Raw data are directly transferred to ORFEO, the centralized data center infrastructure at Area Science Park, ensuring efficient and reliable handling of large microscopy datasets. From there, the data undergo cleaning and standardization in alignment with NFFA-DI and NFFA-Europe guidelines. Final datasets are made available through platforms such as OFED and NOMAD, promoting transparency, reuse, and international collaboration. Through this structured approach, LAME fosters open, high-quality science while ensuring that its data remain findable, accessible, interoperable, and reusable.

Chapter 8

Evaluation and Discussion

8.1 Overview

The FAIR-by-design data management platform developed at LAME will enhance the way electron microscopy data is collected, organized, and shared. By integrating standardized formats, centralized storage (via ORFEO), and user-friendly interfaces, it addresses challenges such as data fragmentation and inconsistent metadata. This chapter presents a concise assessment of its limitations, future directions, and scientific impact.

8.2 Current Challenges

While the platform marks clear progress, several challenges persist:

Data Integration and Format Diversity

Integrating data from multiple electron microscopy systems, particularly SEM and FIB-SEM, remains difficult due to the variety of raw data formats and vendor-specific software. Although NeXus serves as a common output format, pre-processing steps still require attention.

Metadata Consistency and Manual Workload

Ensuring complete and accurate metadata remains a challenge. Despite automated checks, inconsistencies arise due to time constraints on experimentalists. Data cleaning and selection, especially for ORFEO uploads, also require significant manual input.

External Dependencies

Publishing to repositories such as OFED and NOMAD increases visibility and reuse, but also introduces dependencies on external infrastructures, which may evolve in ways that impact accessibility or standards.

8.3 Future Enhancements

To address these challenges, several improvements are planned:

- **Automation:** Incorporating AI for metadata extraction and data filtering to reduce manual effort.
- **User Support:** Improving the web interface with structured entry forms and guided data input to support accurate documentation.
- **Interoperability:** Enabling broader repository integration and developing tools to convert proprietary formats into NeXus.
- **Data Preservation:** Expanding long-term archiving strategies within the ORFEO infrastructure to safeguard datasets.

8.4 Scientific Impact

The platform supports more reproducible and collaborative research by improving data quality, metadata clarity, and repository integration. Its alignment with FAIR principles and compatibility with infrastructures such as OFED and NOMAD strengthens both institutional and international research efforts. Additionally, it contributes to open science by making well-documented datasets available for reuse and educational purposes.

8.5 Summary

LAME's data management platform provides a robust foundation for standardized, FAIR-aligned electron microscopy research. While further automation and interoperability are needed, the system already plays a role in advancing transparent and collaborative science within the NFFA-DI and NFFA-Europe ecosystems.

Chapter 9

Conclusion

9.1 Summary of Contributions

This thesis presents the design and implementation of a FAIR-by-design data management platform for electron microscopy research at LAME, Area Science Park in Trieste with the focus on SEM and FIB-SEM data. The platform aims to improve how data is collected, organized, and shared by applying modern technologies and international standards.

The main achievements of this work include:

- Creating a Django-based web API to simplify the upload of data and meta-data.
- Integrating with ORFEO and MinIO to ensure secure, scalable data storage.
- Adopting the NeXus format, guided by FAIRmat principles, to standardize datasets.
- Implementing automated data cleaning and metadata validation to improve quality.
- Publishing datasets on FAIR-compliant repositories such as NOMAD and OFED in the future.

9.2 Challenges and Approaches

One of the primary challenges was integrating data from different microscopy instruments, each producing diverse output formats. This was resolved through the adoption of NeXus and automated transformation pipelines. Ensuring metadata

accuracy was equally important; this was addressed by designing structured, user-friendly forms and incorporating validation mechanisms to support researchers during data entry.

9.3 Scientific Relevance and Future Directions

The platform improves research workflows at LAME by promoting consistency, accessibility, and long-term data preservation. Its design not only benefits local users but also aligns with broader FAIR and open science initiatives, encouraging collaboration and transparency.

Looking ahead, the platform could be expanded with AI-driven tools for automated metadata extraction and smarter data filtering. Additional support for new microscopy techniques would further increase its flexibility and long-term utility.

9.4 Final Remarks

This work shows that integrating FAIR principles into electron microscopy data management is both feasible and beneficial. The platform developed here provides a strong foundation for structured, reusable data and can serve as a reference for other research institutions seeking to enhance their data practices.

References

Bibliography

- [1] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR guiding principles for scientific data management and stewardship," *Sci Data* **3** (Mar., 2016) 160018.
- [2] M. de Luca, "Introduction to open science – fair data management and principles," Sept., 2024. <https://doi.org/10.5281/zenodo.13839535>.
- [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] F. Bazzocchi, "Tools for data management and curation - ontology as key factor for interoperability," Oct., 2024. <https://doi.org/10.5281/zenodo.13949392>.
- [6] J. Rudzinski, "Fair data management for computational materials science using nomad," Nov., 2024. <https://doi.org/10.5281/zenodo.14184871>.
- [7] E. Giglia, "Introduction to open science - open science why, what, how," Sept., 2024. <https://doi.org/10.5281/zenodo.13831946>.

- [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] 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>.
- [10] M. Könnicke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, B. Clausen, S. Cottrell, *et al.*, “The nexus data format,” 2015.
<https://www.nexusformat.org>.
- [11] 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.